Jak správně měřit kvalitu credit scoringových modelů?

Martin Řezáč

Brno, 11.4.2012

Obsah

Úvod

- Credit scoring je množina prediktivních modelů a jim příslušných statistických technik, které pomáhají finančním institucím při automatickém posuzování žádostí o úvěr.

- Se znalostí pravděpodobnosti selhání žadatele, očekávané míry zamítání, očekávaného zisku/ztráty, popř. dalších business předpokladů, lze efektivně rozhodnout kdo úvěr dostane, v jaké výši a jaké další strategie maximalizují profit plynoucí z daného žadatele/klienta.

Úvod

- Efektivní využití scoringových modelů je nemožné bez znalosti jejich kvality, a to jednak v okamžiku vývoje modelů, ale také po jejich nasazení do praxe.
- Při vývoji je typicky k dispozici několik různých modelů a je třeba vybrat jen jeden. Samozřejmě ten nejlepší (vzhledem k nějakému kritériu).
- Kritériem kvality je většinou nějaký kvantitativní index jako Giniho index nebo KS.
- Jejich hodnota je ovšem silně ovlivněna volbou parametrů v definici dobrého/špatného klienta.
- Navíc je zřejmé, že nejsilnější by měl být scoringový model v oblasti očekávané cutoff hodnoty. Odlišit dobré od superdobrých klientů nebývá typických cílem credit scoringových modelů.

Úvod

- Zatímco historie úvěru sahá 4000 let nazpět (první zaznamenaná zmínka o úvěru pochází ze starověkého Babylonu - 2000 let před n.l.), historie credit scoringu je pouze 50-70 let stará.

- První přístup k řešení problému identifikace skupin v populaci představil ve statistice Fisher (1936). V roce 1941, Durand jako první rozpoznal, že tyto techniky mohou být použity k rozlišování mezi dobrými a špatnými úvěry.

Úvod

- Významným milníkem při posuzování úvěrů byla druhá světová válka.
- Do té doby bylo standardem individuální posuzování žadatele o úvěr. Dále bylo standardem, že ve finanční sféře byli zaměstnáni (téměř) výhradně muži.
- Odchod značné části mužské populace do služeb armády měl za následek potřebu předat zkušenosti dosavadních posuzovatelů žádostí o úvěr novým pracovníkům.
- Díky tomu vznikla jakási rozhodovací pravidla a došlo k „automatizaci“ posuzování žádostí o úvěr.

Úvod

- Příchod kreditních karet ke konci šedesátých let minulého století a růst výpočetního výkonu způsobil obrovský rozvoj a využití credit scoringových technik. Událost, která zajistila plnou akceptaci credit scoringu, bylo přijětí zákonů „Equal Credit Opportunity Acts" (o rovné příležitosti přístupu k úvěrům) a jeho pozdějších znění přijatých v USA v roce 1975 a 1976. Tyto stanovily za nezákonné diskriminace v poskytování úvěru, vyjma situace, pokud tato diskriminace „byla empiricky odvozená a statisticky validní".

Úvod

- V osmdesátých letech minulého století začala být využívána logistická regrese, dodnes v mnoha oblastech považovaná za průmyslový standard, a lineární programování. O něco později se objevily na scéně metody umělé inteligence, např. neuronové sítě. Mezi další používané techniky lze zařadit metody nejbližšího souseda, splajny, waveletové vyhlazování, jádrové vyhlazování, Bayesovské metody, regresní a klasifikační stromy, support vector machines, asociační pravidla, klastrovou analýzu a genetické algoritmy.

Default – definice cílové prom. (good/bad)

- Obvykle je tato definice založena na klientově počtu dnů po splatnosti (Days Past Due, DPD) a částce po splatnosti. S částkou po splatnosti je spojena potřeba stanovení jisté míry tolerance, tedy stanovení co je považováno za významný dluh a co nikoli. Např. nemusí dávat smysl považovat za dluh částky menší než 100 Kč.
- Dále je třeba stanovit časový horizont (performance window), na kterém jsou dva zmíněné parametry sledovány.
- Za dobrého klienta lze např. označit klienta, který:
  - je po splatnosti méně než 60 dnů(s tolerancí 100 Kč) v prvních 6-ti měsících od první splátky,
  - je po splatnosti méně než 90 dnů (s tolerancí 30 Kč) v průběhu celé své platební historie (ever).

Default – definice cílové prom.

- Volba těchto parametrů závisí do značné míry na typu finančního produktu (jistě se bude lišit volba parametrů pro spotřebitelské úvěry pro malé částky se splatností kolem jednoho roku a pro hypotéky, které jsou obvykle spojeny s velmi vysokou finanční částkou a se splatností až několik desítek let) a na dalším využití této definice (řízení rizik, marketing, …).

Default – definice cílové prom.

- Další praktickým problémem definice dobrého klienta je souběh několika smluv jednoho klienta. Například je možné, že zákazník je po lhůtě splatnosti na více smlouvách, ale s rozdílnými dny po splatnosti a s různými částkami. V tomto případě jsou většinou částky klienta dlužné v jednom konkrétním časovém okamžiku sečteny, a ze dnů po splatnosti na jednotlivých smlouvách je brána maximální hodnota. Tento přístup lze uplatnit pouze v některých případech, a to zejména v situaci, kdy jsou k dispozici kompletní účetní data. Situace je podstatně složitější v případě agregovaných údajů, např. na měsíční bázi.

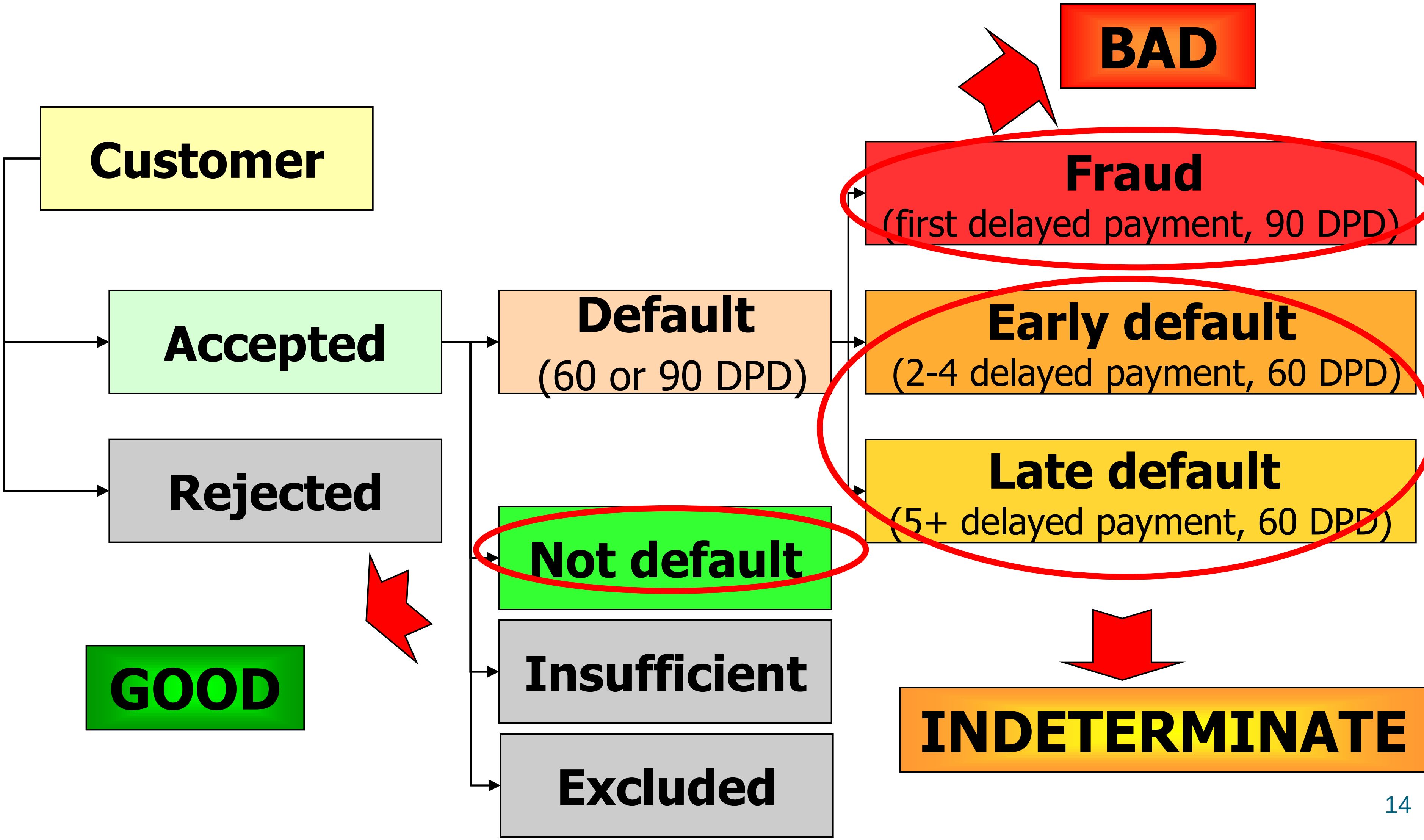
Default – definice cílové prom.

- Obecně uvažujeme následující typy klientů:
  - dobrý (good),
  - špatný (bad),
  - nedefinovaný (indeterminate),
  - s nedostatečnou úvěrovou historií (insufficient),
  - vyřazený (excluded),
  - zamítnutý (rejected).

Default – definice cílové prom.

- První dva typy byly diskutovány. Třetí typ, tj. indeterminate, je na hranici mezi dobrým a špatným klientem a při jeho použití přímo ovlivňuje definici dobrých/špatných klientů. Uvažujeme-li pouze DPD, klienti s vysokými DPD (např. 90 +) jsou typicky označeni za špatné, nedelikventní klienti (jejich DPD je rovno nule) jsou označeni za dobré. Za indeterminate jsou pak označeni delikventní klienti, kteří nepřekročí danou hranici DPD.
- Čtvrtý typ klientů jsou typicky klienti s velmi krátkou platební historií, u kterých je nemožná korektní definice cílové proměnné.
- Vyřazení klienti jsou klienti, jejichž data jsou natolik špatná, že by vedla ke zkreslení modelu(např. fraudy). Další skupinu tvoří klienti, kteří nejsou standardně hodnoceni daným modelem (VIP klienti).
- Poslední typ klientů jsou ti klienti, jejichž žádost o úvěr byla zamítnuta.

Default – definice cílové prom.



Default – definice cílové prom.

- Only good and bad clients are used for further model building. If we do not use the indeterminate category, and if we set up some tolerance level for the amount past due and resolve the issue with simultaneous contracts, there remain two parameters which affect the good/bad definition. They are **DPD and time horizon**.
- Usually it is useful to build up a set of models with varying levels of these parameters. Furthermore, it can be useful **to develop a model with one good/bad definition and measure the model's quality with another**. It should hold that scoring models developed on a harder definition (higher DPD, longer time horizon, or measuring DPD on first payment) perform better than those developed on softer definitions (Witzany, 2009).
- Furthermore, it should hold that a given scoring model performs better if it is measured according to a harder good/bad definition. If not, it usually means that something is wrong.
- Overall, the **development and assessment of credit scoring models on a definition that is as hard as possible, but also reasonable, should lead to the best performance**.

Měření kvality modelu

- Jakmile je k dispozici definice dobrého/špatného klienta a klientovo skóre je možné vyhodnotit kvalitu tohoto skóre. Je-li skóre výstupem nějakého prediktivního modelu (scoringové funkce), posuzujeme kvalitu tohoto modelu. Uvažujeme dvě základní skupiny indexů kvality. První je založena na distribuční funkci. Mezi nejpoužívanější indexy patří
    - Kolmogorovova-Smirnovova statistika (KS)
    - Giniho index (Somersovo D, Kendalovo τ_α , Goodman-Kruskal γ)
    - C-statistika
    - Lift.

  Druhá skupina indexů je založena na pravděpodobnostní hustotě. Mezi nejznámější indexy patří
    - Střední diference (Mahalanobisova vzdálenost)
    - Informační statistika/hodnota (I_{Val}).

Indexy založené na distribuční funkci

$$D_K = \begin{cases} 1, & klient \;\; je\, dobrý \\ 0, & jinak. \end{cases}$$

Počet dobrých klientů: n

Počet špatných klientů: m

Proporce dobrých/špatných klientů: $p_G = \dfrac{n}{n+m}$, $p_B = \dfrac{m}{n+m}$

- Empirické distribuční funkce:

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

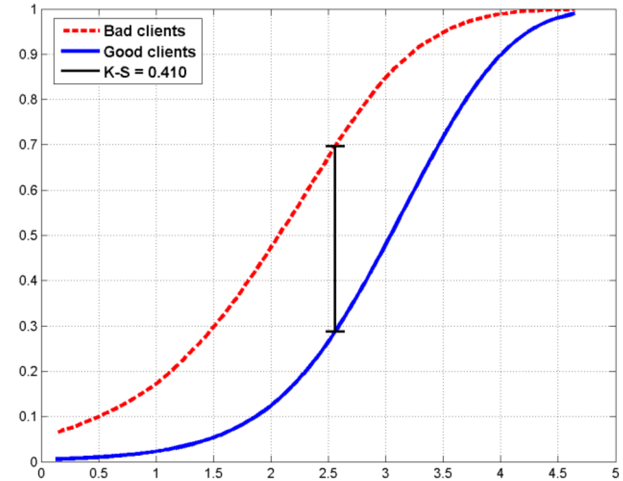
$$F_{m.BAD}(a) = \frac{1}{m}\sum_{i=1}^{m} I(s_i \le a \wedge D_K = 0)$$

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

$$I(A) = \begin{cases} 1 & A \;\; platí \\ 0 & jinak \end{cases}$$

- **Kolmogorovova-Smirnovova statistika (KS)**

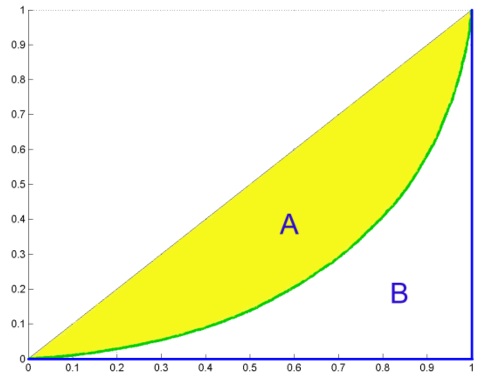
$$\boxed{KS = \max_{a \in [L,H]} \left| F_{m,BAD}(a) - F_{n,GOOD}(a) \right|}$$



Indexy založené na distribuční funkci

- Lorenzova křivka (LC)

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



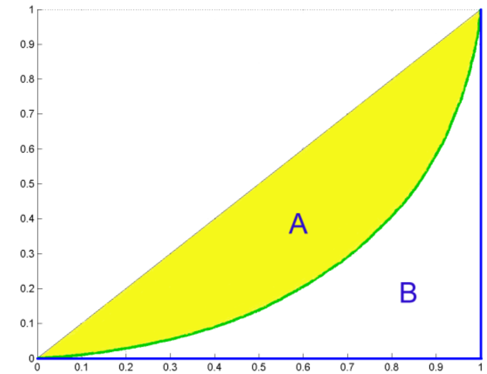
- Tato definice a název (LC) je konzistentní s Müller, M., Rönz, B. (2000). Stejnou definici křivky, ovšem pod názvem ROC lze nalézt v Thomas et al. (2002). Siddiqi (2006) používá název ROC pro křivku s prohozenými osami a LC pro křivku s $F_{m.BAD}(a)$ na svislé ose a $F_{N.ALL}(a)$ na ose horizontální.

Lorenzova křivka, Giniho index

- Lorenzova křivka (LC)

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

- Giniho index



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

$$\boxed{Gini = 1 - \sum_{k=2}^{n+m} \left(F_{m.BAD\,k} - F_{m.BAD\,k-1}\right) \cdot \left(F_{n.GOOD_k} + F_{n.GOOD\,k-1}\right)}$$

kde $F_{m.BAD\,k}$ ($F_{n.GOOD_k}$) je k-tá hodnota vektoru empirické distribuční funkce špatných (dobrých) klientů

Somersovo D, Kendalovo τ_α

- Giniho index je speciální případ Somersova *D* (Somers (1962)), které je pořadovou asociační mírou definovanou jako

$$D_{YX} = \frac{\tau_{XY}}{\tau_{XX}}$$

kde τ_{XY} je Kendallovo τ_a definované jako $\tau_{XY} = E[sign(X_1 - X_2) sign(Y_1 - Y_2)]$

kde $(X_1, Y_1), (X_2, Y_2)$ jsou bivariantní, stochasticky nezávislé, náhodné vektory nad touž datovou populací, a $E[\cdot]$ značí střední hodnotu. V našem případě je *Y=1* jestliže je klient dobrý a *Y=0* jestliže je klient špatný. Proměnná *X* reprezentuje skóre.

Thomas (2009) uvádí, že Somersovo *D* hodnotící výkonnost daného credit scoringového modelu lze vypočítat pomocí

$$D_S = \frac{\sum_i g_i \sum_{j<i} b_j - \sum_i g_i \sum_{j>i} b_j}{n \cdot m}$$

kde g_i (b_j) je počet dobrých (špatných) klientů v i-tém intervalu skóre.

Somersovo D, Mann-Whitney U

- Dále platí, že D_S může být vyjádřeno pomocí Mann-Whitneyho U-statistiky.
  - Seřaď datový vzorek ve vzestupném pořadí podle skóre a sečti pořadí dobrých klientů ve vzniklé posloupnosti. Označme tento součet jako R_G. Potom

$$\boxed{U = R_G - \frac{1}{2}n(n+1)}$$

$$\boxed{D_S = 2\frac{U}{n \cdot m} - 1}$$

Konkordantní, diskordantní páry

- Konkordantní pár (X_1,Y_1), (X_2,Y_2):

$$\boxed{\operatorname{sgn}(X_2 - X_1) = \operatorname{sgn}(Y_2 - Y_1)}$$

- Diskordantní pár:

$$\boxed{\operatorname{sgn}(X_2 - X_1) = -\operatorname{sgn}(Y_2 - Y_1)}$$

- V našem případě X představuje skóre a Y ukazatel dobrého klienta (D_K). Protože dobrý klient má hodnotu Y=1 a špatný Y=0, je zřejmé, že u konkordantního páru má dobrý klient vyšší hodnotu skóre než klient špatný.

Somersovo D, Goodman-Kruskal gamma

- Uvažujme tedy dva náhodně vybrané klienty, přičemž jeden je dobrý ($Y_1=1$) a druhý špatný ($Y_2=0$), skóre prvního označme s_1, druhého s_2. Pak
  - Konkordantní pár (Concordant): $s_1>s_2$
  - Diskordantní pár (Discordant): $s_1<s_2$
  - Vázaný pár (Tied): $s_1=s_2$

➢ Somersovo D:

$$\boxed{D_S = \frac{\#Concordant - \#Discodrant}{\#Concordant + \#Discodrant + \#Tied}}$$

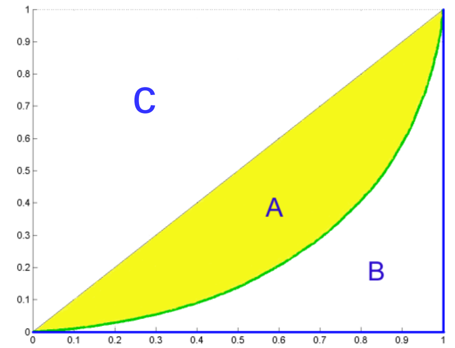
➢ Goodmanovo-Kruskalovo Gamma:

$$\boxed{\gamma = \frac{\#Concordant - \#Discodrant}{\#Concordant + \#Discodrant}}$$

Indexy založené na distribuční funkci

- C-statistika:

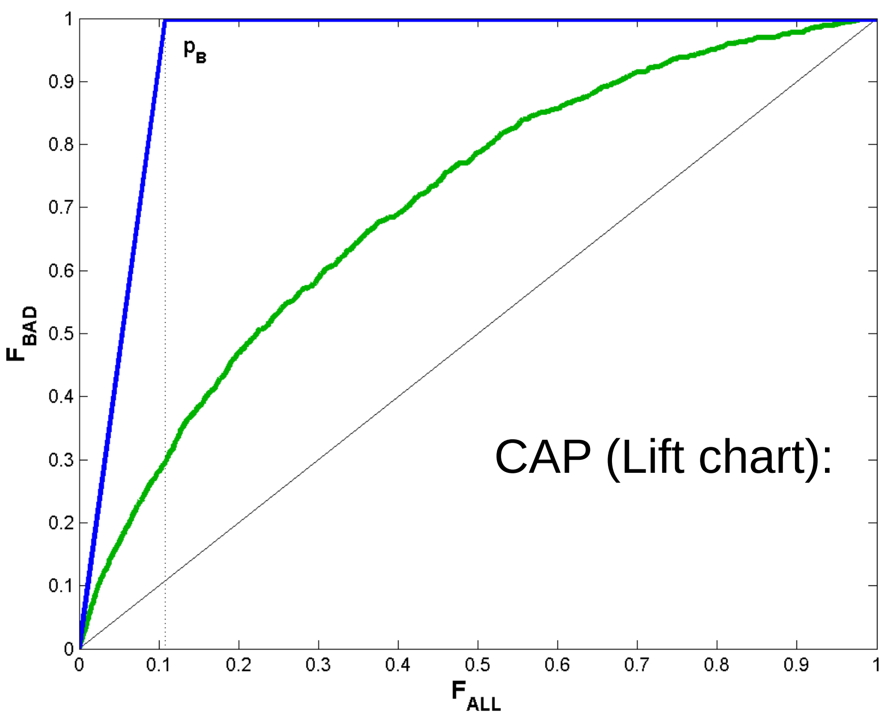
$$\boxed{c - stat = A + C = \frac{1 + Gini}{2}}$$



Tato statistika je rovna pravděpodobnosti, že náhodně vybraný dobrý klient má vyšší skóre než náhodně vybraný špatný klient, tj.

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

CAP – index AR



AR (Accuracy Ratio)

$$AR = \frac{\text{Plocha mezi CAP a diagonálou}}{\text{Plocha mezi CAP ideálního modelu a diagonálou}}$$

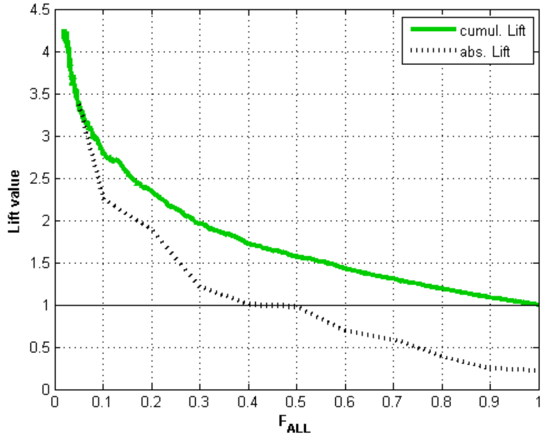
$$= \frac{\text{Plocha mezi CAP a diagonálou}}{0.5(1-p_B)} = Gini$$

V tomto případě máme na x-ové ose proporci všech klientů (F_{ALL}) a na y-vé ose proporci špatných klientů (F_{BAD}). Ideální model je tentokrát reprezentován lomenou čarou z bodu [0, 0] přes [pB, 1] do bodu [1, 1]. Výhoda tohoto obrázku je ta, že je možné odečíst proporci zamítnutých špatných klientů vs. celková proporce zamítnutých klientů. Např. vidíme, že pokud chceme zamítnout 70% špatných klientů, musíme zamítat přibližně 40% všech žadatelů.

Indexy založené na distribuční funkci

- Další možnou mírou kvality scoringového modelu je Lift, který říká kolikrát je daný model, při dané úrovni zamítání, lepší než náhodný model. Přesněji řečeno jde o poměr proporce špatných klientů se skóre menším nebo rovno dané hodnotě skóre a, $a \in [L,H]$, ku proporci špatných klientů v celé populaci. Formálně jej lze zapsat takto:

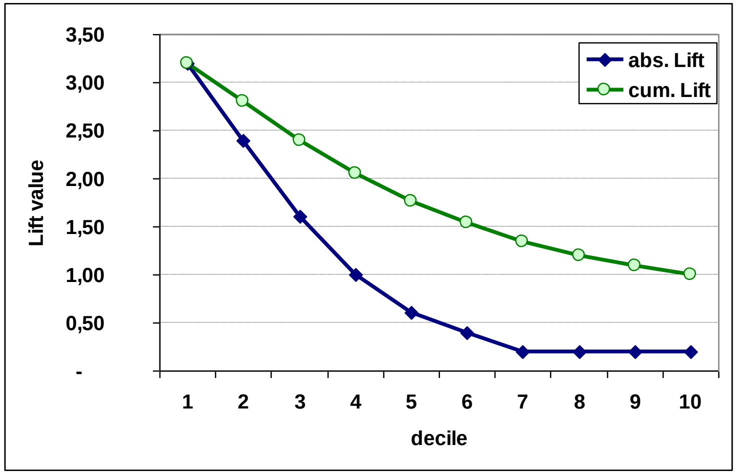
$$Lift(a) = \frac{CumBadRate(a)}{BadRate} = \frac{\dfrac{\sum_{i=1}^{n+m} I(s_i \le a \wedge Y=0)}{\sum_{i=1}^{n+m} I(s_i \le a)}}{\dfrac{\sum_{i=1}^{n+m} I(Y=0)}{\sum_{i=1}^{n+m} I(Y=0 \vee Y=1)}} = \frac{\dfrac{\sum_{i=1}^{n+m} I(s_i \le a \wedge Y=0)}{\sum_{i=1}^{n+m} I(s_i \le a)}}{\dfrac{n}{N}}$$



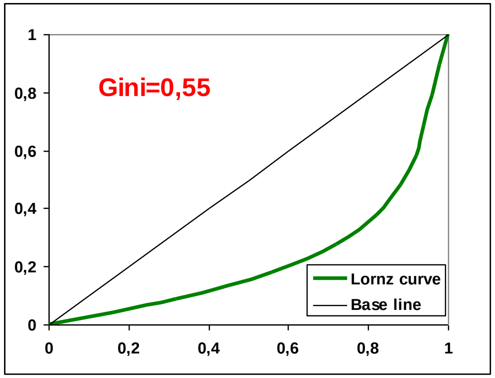
$$absLift(a) = \frac{BadRate(a)}{BadRate}$$

Indexy založené na distribuční funkci

- Pro výpočet lze použít tabulku s počty všech a špatných klientů v daných intervalech skóre (např. decilech).



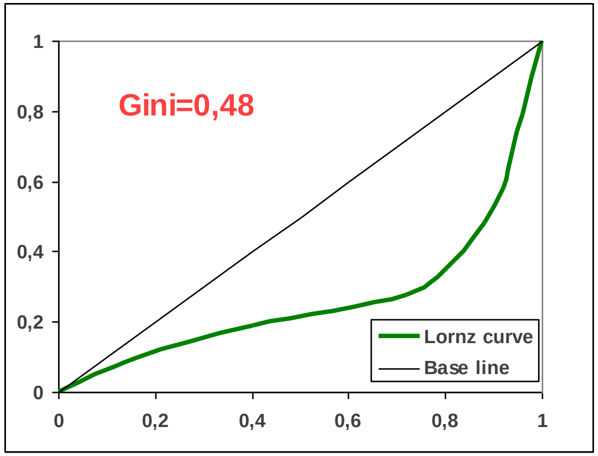
decile	# cleints	absolutely			cumulatively		
		# bad clients	Bad rate	abs. Lift	# bad clients	Bad rate	cum. Lift
1	100	16	16,0%	3,20	16	16,0%	3,20
2	100	12	12,0%	2,40	28	14,0%	2,80
3	100	8	8,0%	1,60	36	12,0%	2,40
4	100	5	5,0%	1,00	41	10,3%	2,05
5	100	3	3,0%	0,60	44	8,8%	1,76
6	100	2	2,0%	0,40	46	7,7%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

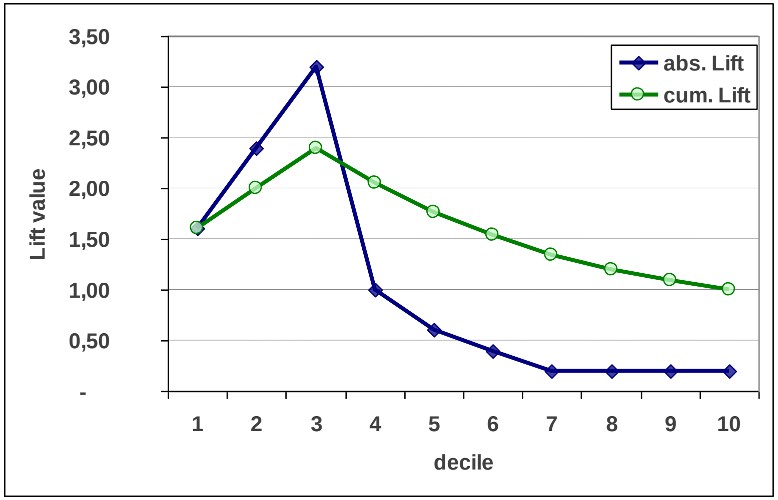


Indexy založené na distribuční funkci

- Pokud bad rate není monotonní:
  - LC vypadá OK
  - Gini se mírně sníží
  - Lift ovšem vypadá podivně

decile	# cleints	absolutely # bad clients	absolutely Bad rate	absolutely abs. Lift	cumulatively # bad clients	cumulatively Bad rate	cumulatively 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%				



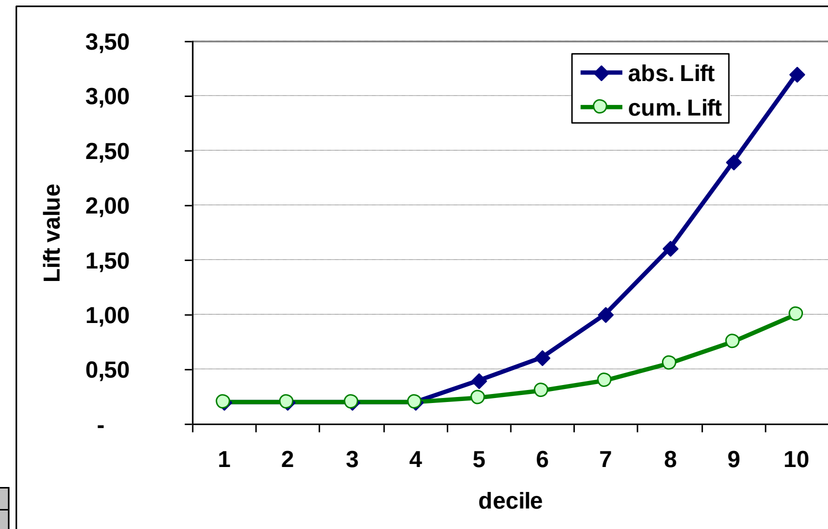


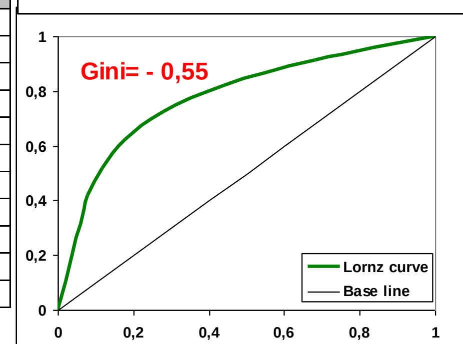
Indexy založené na distribuční funkci

- Pokud má skóre zcela opačný smysl, obdržíme „opačné“ obrázky.

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%				

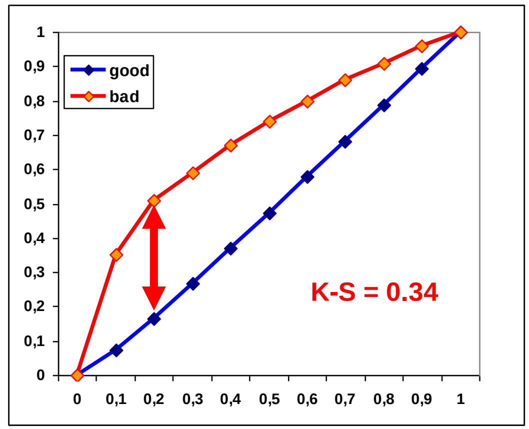


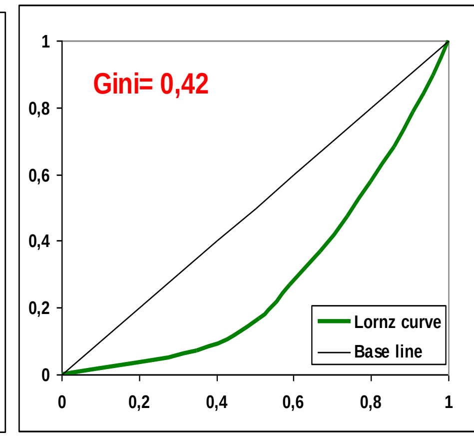


Indexy založené na distribuční funkci

- **SC 1:**

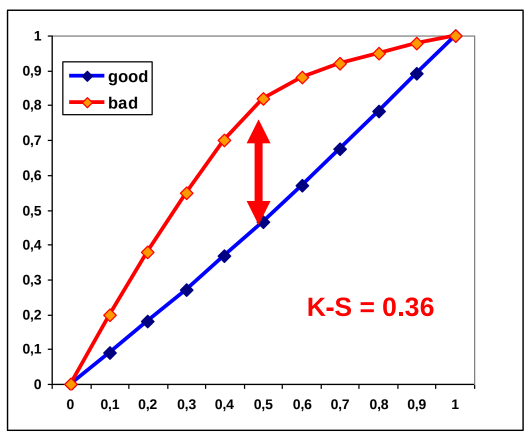
decile	# cleints	# bad clients	Bad rate
1	100	35	35,0%
2	100	16	16,0%
3	100	8	8,0%
4	100	8	8,0%
5	100	7	7,0%
6	100	6	6,0%
7	100	6	6,0%
8	100	5	5,0%
9	100	5	5,0%
10	100	4	4,0%
All	1000	100	10,0%

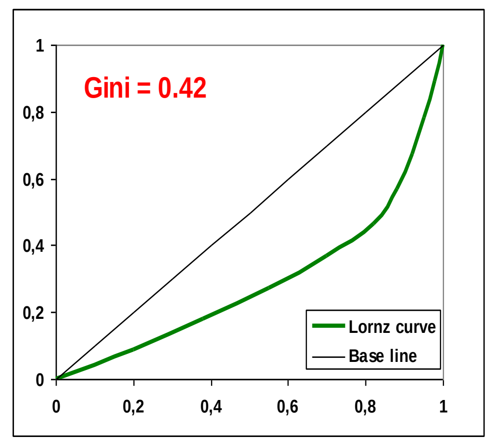




- **SC 2:**

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%

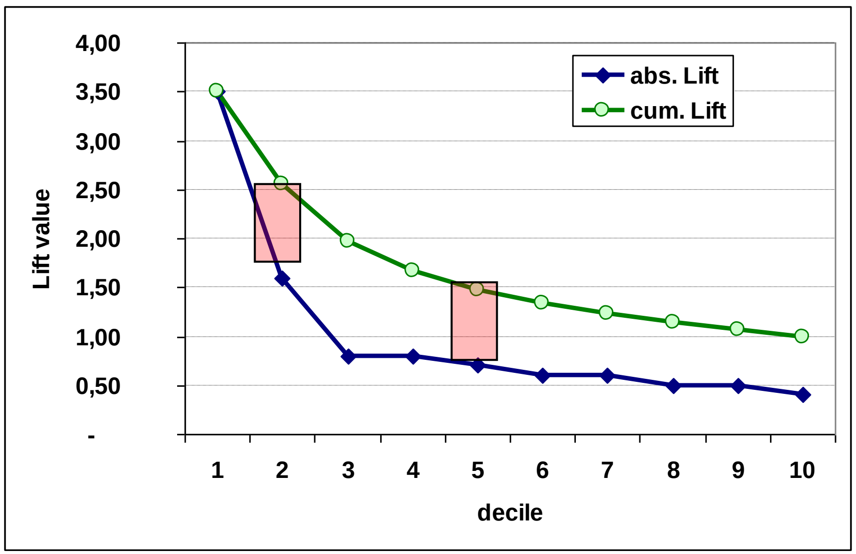




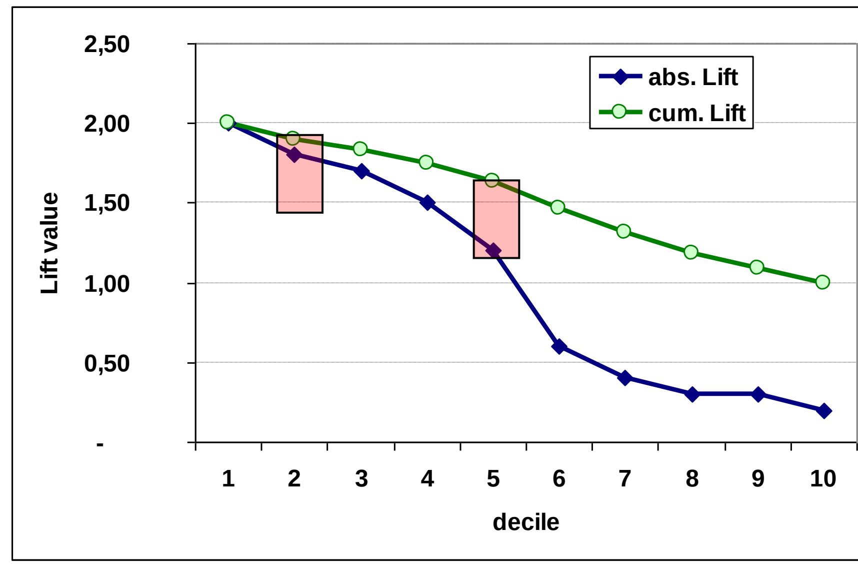
Je evidentní, že pouze Gini a KS nestačí!!!

Indexy založené na distribuční funkci

- **SC 1:**



- **SC 2:**



$Lift_{20\%} = 2.55 \quad > \quad Lift_{20\%} = 1.90$

$Lift_{50\%} = 1.48 \quad < \quad Lift_{50\%} = 1.64$

SC 2 je lepší, pokud je předpokládaná míra zamítaní (reject rate) přibližně 50%.
SC 1 je významně lepší, pokud je předpokládaný reject rate přibližně 20%.

Lift, QLift

- Lift can be expressed and computed by formula:

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

- In practice, Lift is computed corresponding to 10%, 20%, . . . , 100% of clients with the worst score. Hence we define:

$$QLift(q) = \frac{F_{m.BAD}(F_{N.ALL}^{-1}(q))}{F_{N.ALL}(F_{N.ALL}^{-1}(q))} = \frac{1}{q} F_{m.BAD}(F_{N.ALL}^{-1}(q)), \quad q \in (0,1]$$

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

- Typical value of q is 0.1. Then we have

$$QLift_{10\%} = QLift(0.1) = 10 \cdot F_{m.BAD}(F_{N.ALL}^{-1}(0.1))$$

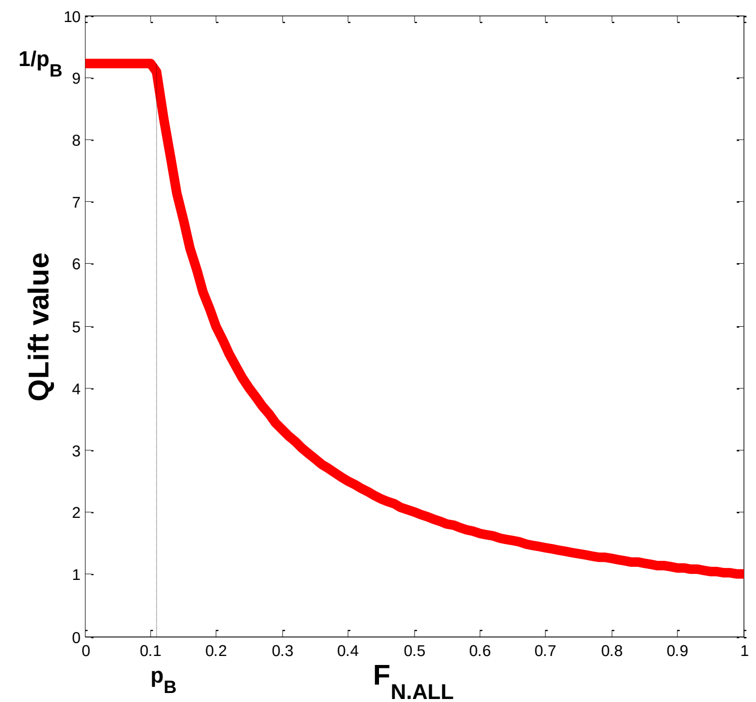
Lift and QLift for ideal model

- It is natural to ask how look Lift and QLift in case of ideal model. Hence we derived following formulas.
- Lift for ideal model:

$$Lift_{ideal}(a) = \begin{cases} \frac{1}{p_B}, & a \leq c \\ \frac{1}{F_{N.ALL}(a)}, & a > c \end{cases}$$

- QLift for ideal model:

$$QLift_{ideal}(q) = \begin{cases} \frac{1}{p_B}, & q \in (0, p_B] \\ \frac{1}{q}, & q \in (p_B, 1] \end{cases}$$

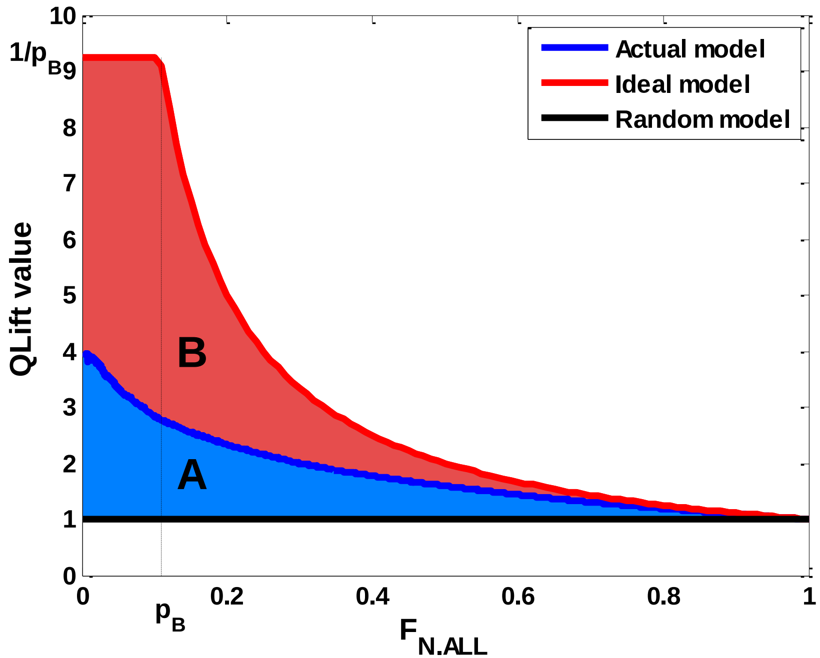


We can see that the upper limit of Lift and QLift is equal to $1/p_B$.

Lift Ratio (LR)

- Once we know form of QLift for ideal model, we can define Lift Ratio as analogy to Gini index.

$$LR = \frac{A}{A+B} = \frac{\int_0^1 QLift(q)\,dq - 1}{\int_0^1 QLift_{ideal}(q)\,dq - 1}$$

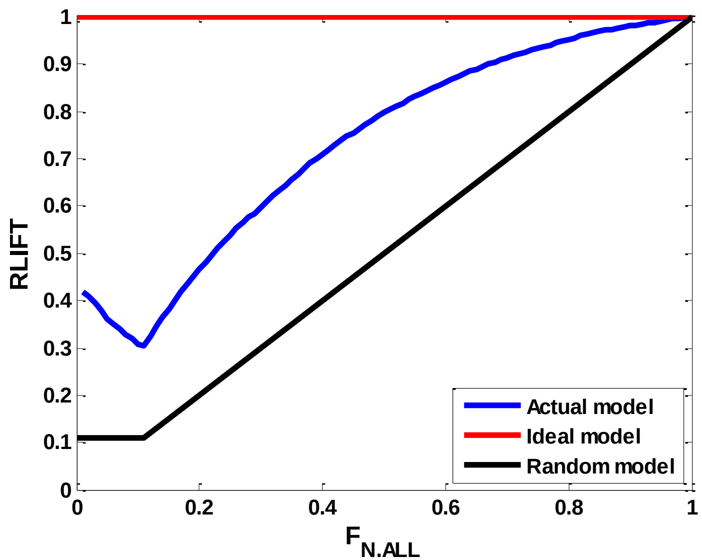


- It is obvious that it is global measure of model's quality and that it takes values from 0 to 1. Value 0 corresponds to random model, value 1 match to ideal model. Meaning of this index is quite simple. The higher, the better. Important feature is that Lift Ratio allows us to fairly compare two models developed on different data samples, which is not possible with Lift.

Rlift, IRL

- Since Lift Ratio compares areas under Lift function for actual and ideal models, next concept is focused on comparison of Lift functions themselves. We define Relative Lift function by

$$RLift(q) = \frac{QLift(q)}{QLift_{ideal}(q)}, \quad q \in (0,1]$$



- In connection to RLift we define Integrated Relative Lift (IRL):

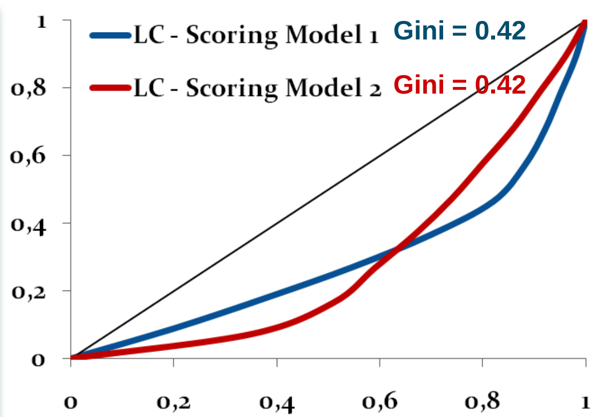
$$IRL = \int_0^1 RLift(q)\, dq$$

- It takes values from $0.5 + \frac{p_B^2}{2}$, for random model, to 1, for ideal model. Following simulation study shows interesting connection to c-statistics.

Příklad

- We consider two scoring models with score distribution given in the table below.
- We consider standard meaning of scores, i.e. higher score band means better clients (the highest probability of default have clients with the lowest scores, i.e. clients in score band 1).
- Gini indexes are equal for both models.
- It is evident from the Lorenz curves, that the first model is stronger for higher score bands and the second one is better for lower score bands.
- The same we can read from values of QLift.

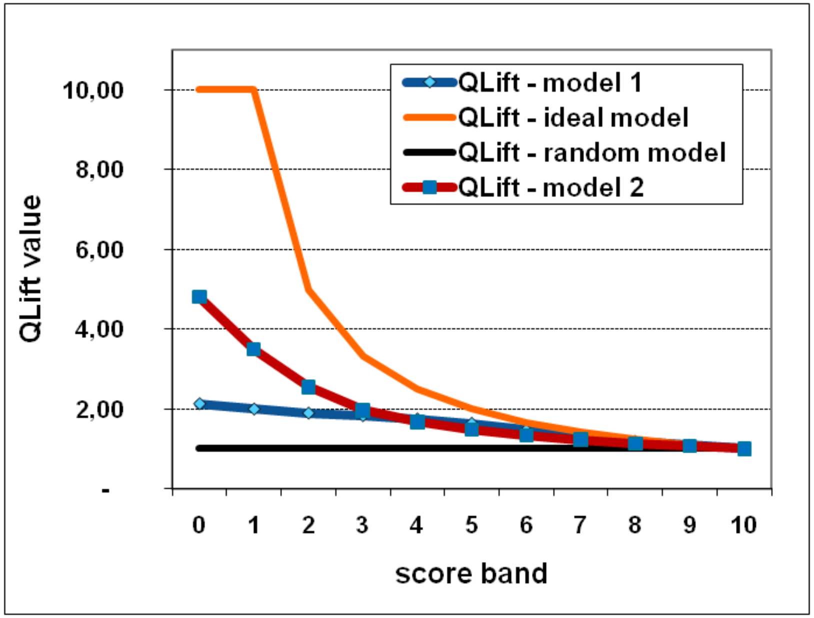
			Scoring Model 1				Scoring Model 2			
score band	# clients	q	# bad clients	# cumul. bad clients	# cumul. bad rate	QLift	# bad clients	# cumul. bad clients	# cumul. bad rate	QLift
1	100	0.1	20	20	20.0%	**2.00**	35	35	35.0%	**3.50**
2	100	0.2	18	38	19.0%	**1.90**	16	51	25.5%	**2.55**
3	100	0.3	17	55	18.3%	**1.83**	8	59	19.7%	**1.97**
4	100	0.4	15	70	17.5%	**1.75**	8	67	16.8%	**1.68**
5	100	0.5	12	82	16.4%	**1.64**	7	74	14.8%	**1.48**
6	100	0.6	6	88	14.7%	**1.47**	6	80	13.3%	**1.33**
7	100	0.7	4	92	13.1%	**1.31**	6	86	12.3%	**1.23**
8	100	0.8	3	95	11.9%	**1.19**	5	91	11.4%	**1.14**
9	100	0.9	3	98	10.9%	**1.09**	5	96	10.7%	**1.07**
10	100	1.0	2	100	10.0%	**1.00**	4	100	10.0%	**1.00**
All	1000		100				100			

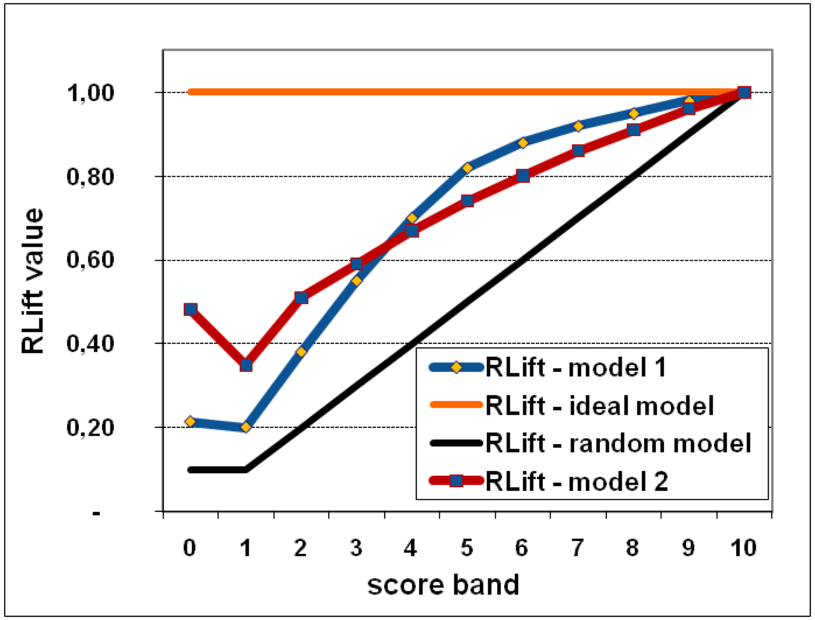


Příklad

- Since Qlift is not defined for *q*=0, we extrapolated the value by

$$QLift(0) = 3 \cdot QLift(0.1) - 3 \cdot QLift(0.2) + QLift(0.3)$$





According to both Qlift and Rlift curves we can state that:

- If expected reject rate is up to 40%, then model 2 is better.
- If expected reject rate is more than 40%, then model 1 is better.

Indexy založené na hustotě

- Střední diference (Mahalanobis distance):

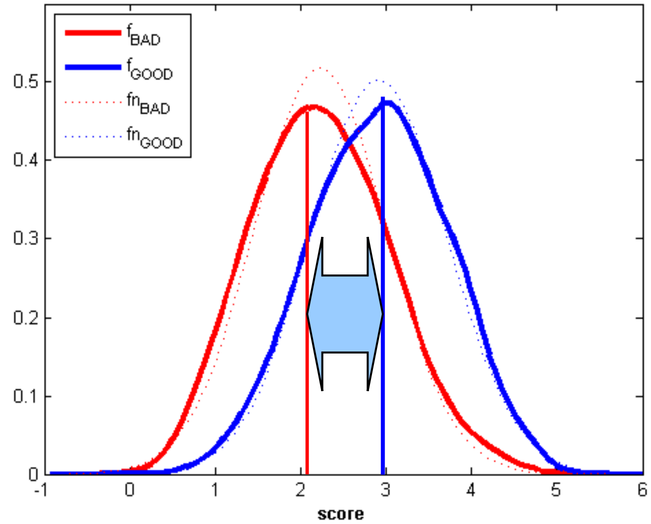
$$D = \frac{M_g - M_b}{S}$$

kde S je společná směrodatná odchylka:

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

M_g ,M_b jsou střední hodnoty dobrých (špatných) klientů

S_g ,S_b jsou příslušné směrodatné odchylky.



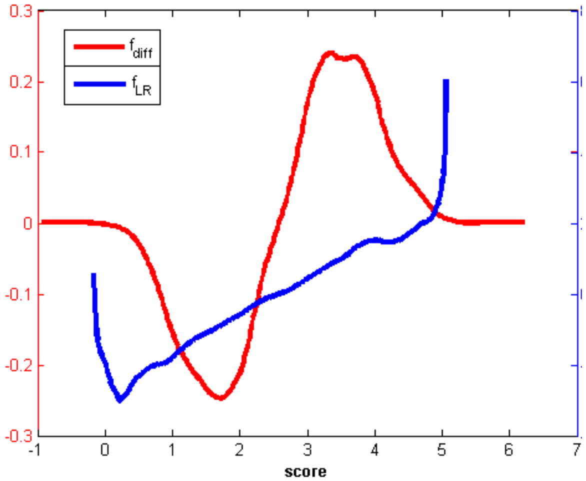
Indexy založené na hustotě

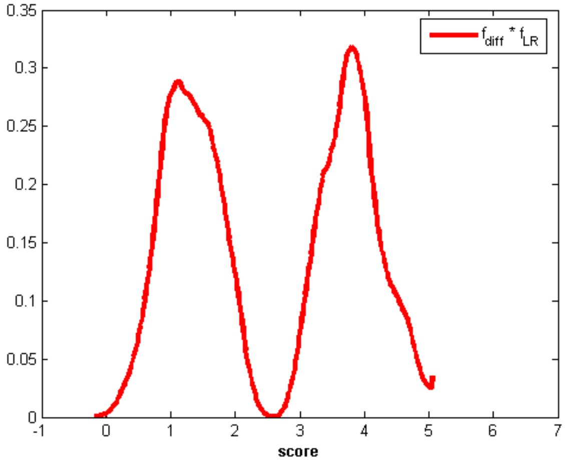
- Informační hodnota (I_{val}) – spojitý případ (Divergence):
  - jde o symetrizovanou Kullback-Leiblerovu divergenci známou také pod názvem J-divergence.

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

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

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





Indexy založené na hustotě

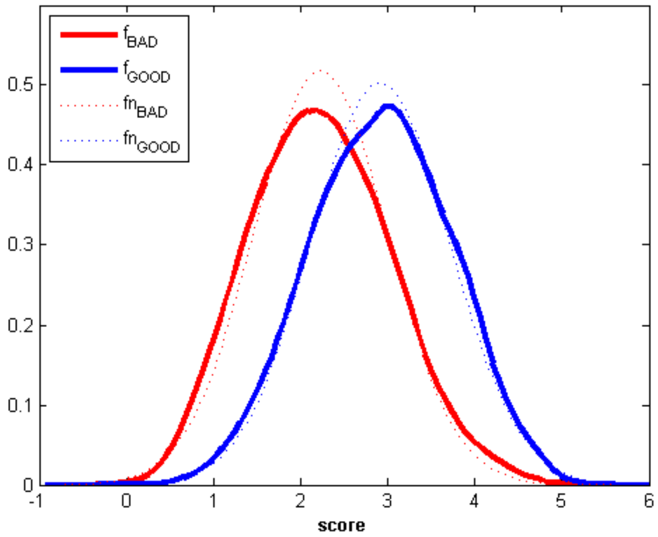
- Pravděpodobnostní hustota: $f_{GOOD}(x) \quad f_{BAD}(x)$

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

- Optimální šířka okna (pomocí principu maximalního vyhlazení):

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

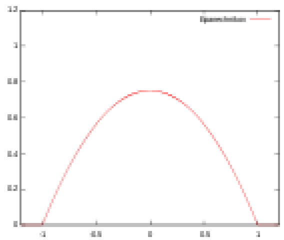
kde: k je řád jádrové funkce (např. 2 pro Epanechnikovo jádro)
n je počet pozorování
$\tilde{\sigma}$ je odhad směrodatné odchylky



Indexy založené na hustotě

- Informační hodnota (I_{val}) – diskretizovaný spojitý případ:

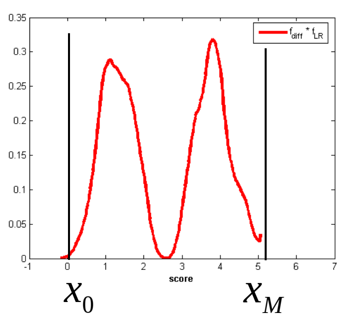
- Nahradíme hustotu jejím jádrovým odhadem a spočteme integrál numericky (např. pomocí složeného lichoběžníkového pravidla).



- S použitím Epanečnikova jádra, $K(x)=\frac{3}{4}(1-x^2)\cdot I(x\in[-1,1])$,
  a optimální šířky vyhlazovacího okna $h_{OS,k}$ dostaneme

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

- Pro daných M+1 bodů $x_0,\ldots,x_M$ dostáváme



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

Indexy založené na hustotě

- Informační statistika/hodnota (I_{val}) – diskrétní případ:

  - Vytvoříme intervaly skóre – typicky decily. Počet dobrých (špatných) klientů v i-tém intervalu označíme g_i (b_i).

  - Musí platit $g_i > 0, b_i > 0 \quad \forall i$

  - Potom dostáváme

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

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

Empirical estimate of I_{val}

The main idea of this approach is to replace unknown densities by their empirical estimates. Let's have n score values, of which n_0 score values s_{0_i}, $i = 1, \ldots, n_0$ for bad clients and n_1 score values s_{1_j}, $j = 1, \ldots, n_1$ 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 $[q_0, q_1], (q_1, q_2], \ldots, (q_{r-1}, q_r]$, where $q_0 = L, q_r = H$. Set

$$n_{0_j} = \sum_{i=1}^{n_0} I\left(s_{0_i} \in (q_{j-1}, q_j]\right)$$

$$n_{1_j} = \sum_{i=1}^{n_1} I\left(s_{1_i} \in (q_{j-1}, q_j]\right), \quad j = 1, \ldots, r$$

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

$$\widehat{I}_{val,DEC} = \sum_{j=1}^{r} \left(\frac{n_{1_j}}{n_1} - \frac{n_{0_j}}{n_0} \right) \ln \left(\frac{n_{1_j} n_0}{n_{0_j} n_1} \right).$$

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.

- 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 (ESIS)

- We want to avoid zero values of n_{0j} or n_{1j} .

- We 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. This is the basic idea of all proposed algorithms.

Empirical estimate with supervised interval selection

- The ESIS:

Set

$$
\begin{aligned}
q_0 &= L - 1 \\
q_i &= \widehat{F_0^{-1}}\left(\frac{k \cdot i}{n_0}\right), i = 1, \ldots, \lfloor \frac{n_0}{k} \rfloor \\
q_{\lfloor \frac{n_0}{k} \rfloor + 1} &= H,
\end{aligned}
$$

where $\widehat{F_0^{-1}}(\cdot)$ 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 n_0 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_{1j} for all actual intervals. If we obtain $n_{1j} < k$ 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 $n_{1j} < k$ 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 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

$$k = \lceil \sqrt{n_0} \rceil$$

- The estimate of the Information value is given by

$$\widehat{I}_{val,ESIS} = \sum_{j=1}^{r} \left(\frac{n_{1_j}}{n_1} - \frac{n_{0_j}}{n_0} \right) \ln \left(\frac{n_{1_j} n_0}{n_{0_j} n_1} \right)$$

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

Simulation results

- Consider n clients, $100p_B\%$ of bad clients with $f_0 : N(\mu_0, \sigma_0)$ and $100(1-p_B)\%$ of good clients with $f_1 : N(\mu_1, \sigma_1)$.
- Because of normality we know $I_{val} = \left(\frac{\mu_1 - \mu_0}{\sigma}\right)^2$.
- Consider following values of parameters:
  - n = 100 000 , n = 1000
  - $\mu_0 = 0$
  - $\sigma_0 = \sigma_1 = 1$
  - $\mu_1 = 0.5, 1, 1.5$
  - $p_B = 0.02, 0.05, 0.1, 0.2$

Simulation results

n=100000, $\mu_1 - \mu_0$ = 0.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,000546	0,000310	0,000224	0,000168
IV_kern	0,000487	0,000232	0,000131	0,000076
IV_esis	0,000910	0,000384	0,000218	0,000127

n=1000, $\mu_1 - \mu_0$ = 0.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,025574	0,040061	0,026536	0,009074
IV_kern	0,038634	0,017547	0,009281	0,004737
IV_esis	0,038331	0,021980	0,016280	0,008028

- worst
- average
- best performance

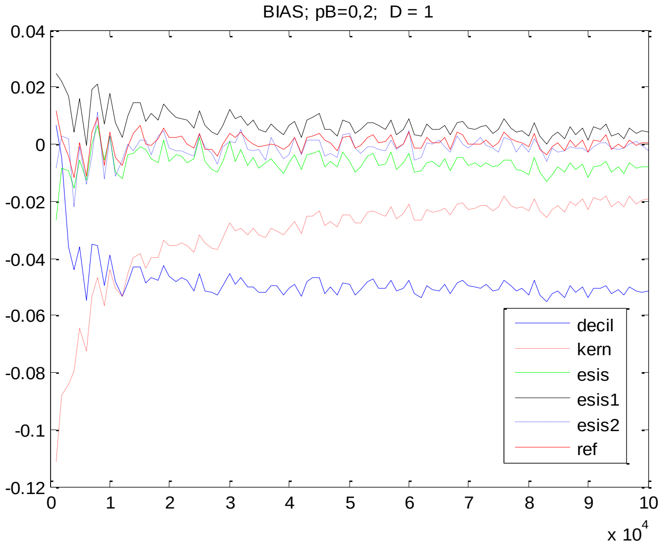
n=100000, $\mu_1 - \mu_0$ = 1.0				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,006286	0,004909	0,004096	0,002832
IV_kern	0,003396	0,001697	0,001064	0,000646
IV_esis	0,002146	0,000973	0,000477	0,000568

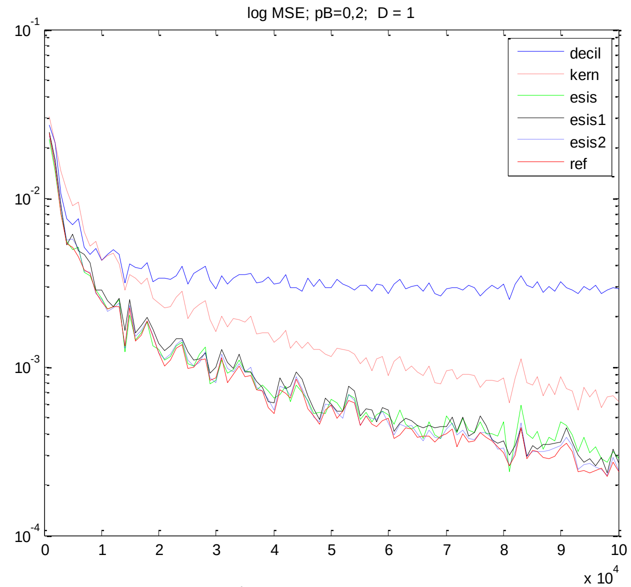
n=1000, $\mu_1 - \mu_0$= 1.0				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,186663	0,084572	0,043097	0,029788
IV_kern	0,117382	0,072381	0,045344	0,032131
IV_esis	0,150881	0,071088	0,036503	0,023609

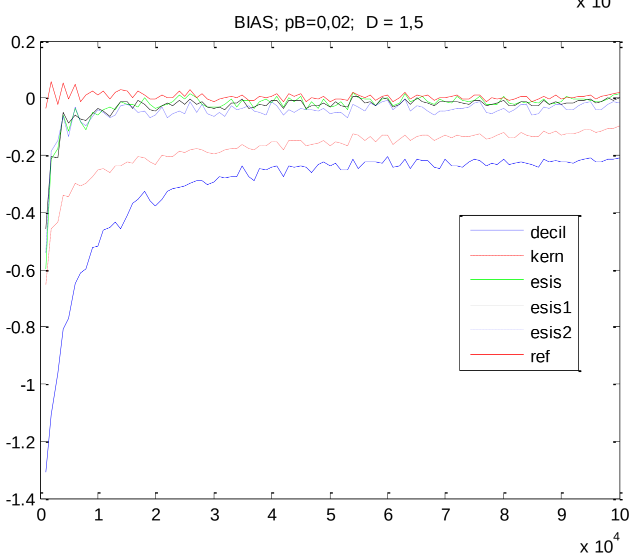
n=100000, $\mu_1 - \mu_0$ = 1.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,056577	0,048415	0,034814	0,020166
IV_kern	0,019561	0,010789	0,006796	0,004862
IV_esis	0,013045	0,008134	0,007565	0,027943

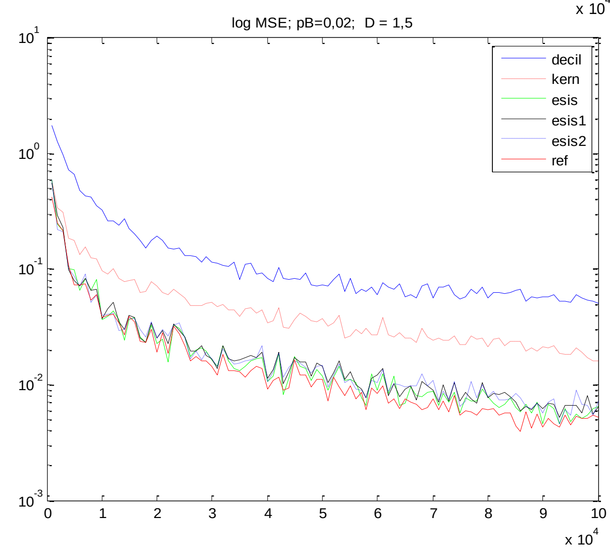
n=1000, $\mu_1 - \mu_0$ = 1.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	1,663859	1,037778	0,535180	0,200792
IV_kern	0,529367	0,349783	0,266912	0,196856
IV_esis	0,609193	0,352151	0,172931	0,194676

Simulation results









„Klasický“ odhad pomocí decilů skóre je značně vychýlený!!!

Normálně rozložené skóre

- Předpokládejme, že skóre dobrých a špatných klientů je normálně rozloženo, tj. jejich pravděpodobnostní hustoty mají tvar

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

- Odhady parametrů μ_b, μ_b, σ_g a σ_b :

  M_g, M_b jsou aritmetické průměty skóre dobrých (špatných) klientů

  S_g , S_b jsou směrodatné odchylky skóre dobrých (špatných) klientů

- Společná směrodatná odchylka: $$S = \left(\frac{nS_g^{\ 2} + mS_b^{\ 2}}{n+m} \right)^{\frac{1}{2}}$$

- Odhady střední hodnoty a směrodatné odchylky skóre všech klientů μ_{ALL}, σ_{ALL}:

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

.

Normálně rozložené skóre

- Předpokládejme, že směrodatné odchylky obou skóre jsou rovny hodnotě σ, pak:

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

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

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

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

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

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

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

Kde $\Phi(\cdot)$ je distribuční funkce standardizovaného normálního rozložení, $\Phi_{\mu,\sigma^2}(\cdot)$ je distribuční funkce s parametry μ, σ^2 a $\Phi^{-1}(\cdot)$ je standardizovaná kvantilová funkce.

Normálně rozložené skóre

- Obecně, tj. bez předpokladu rovnosti směrodatných odchylek skóre:

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

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

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

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

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

Normálně rozložené skóre

- Obecně, tj. bez předpokladu rovnosti směrodatných odchylek skóre:

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

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

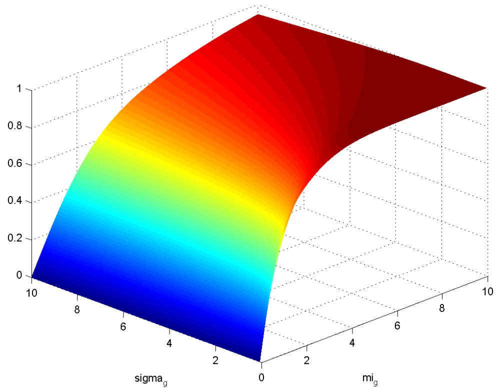
$$\boxed{QLift(q) = \frac{1}{q}\Phi\left(\frac{S_{ALL} \cdot \Phi^{-1}(q) + M - M_b}{S_b}\right)}$$

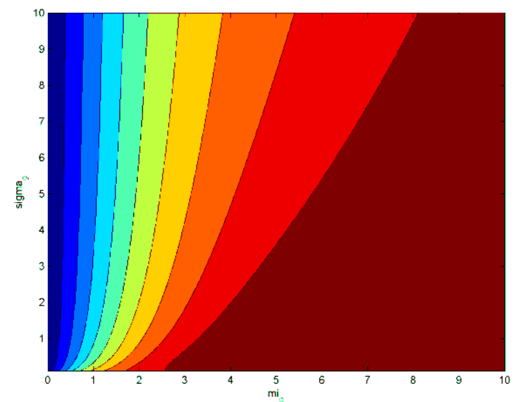
$$I_{val} = (A+1)D^{*2} + A - 1, \quad A = \frac{1}{2}\left(\frac{\sigma_b^2}{\sigma_g^2} + \frac{\sigma_g^2}{\sigma_b^2}\right)$$

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

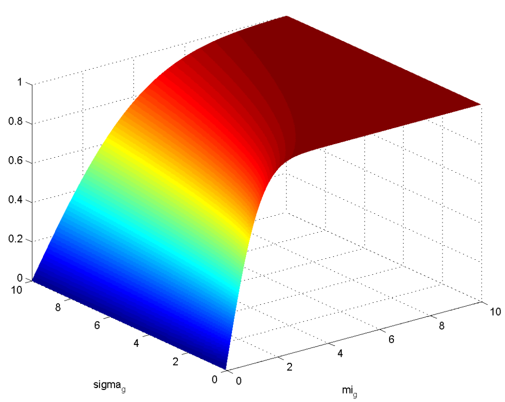
Normálně rozložené skóre

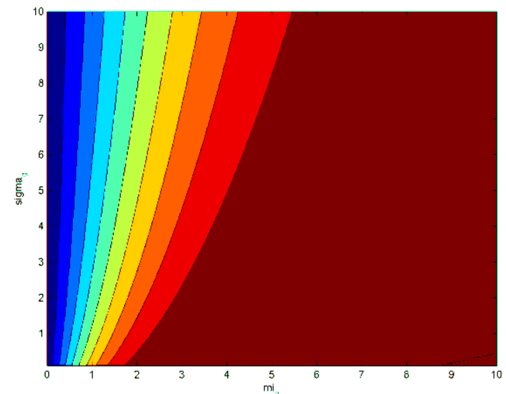
- **KS:** $\mu_b = 0$, $\sigma_b^2 = 1$





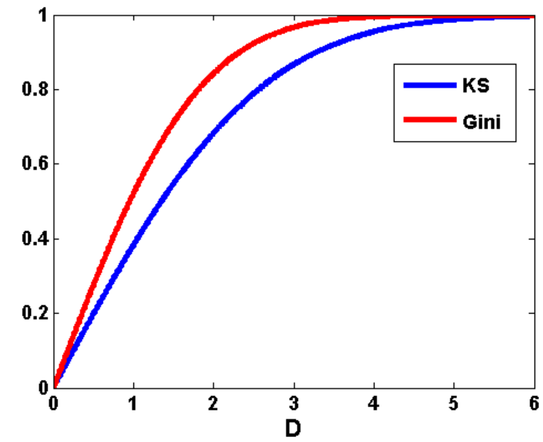
- **Gini** $\mu_b = 0$, $\sigma_b^2 = 1$





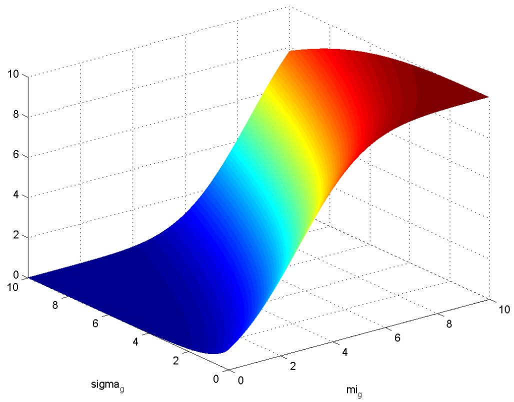
- KS i Gini reagují velmi silně na změnu μ_g , ale zůstávají téměř beze změny ve směru σ_g^2.

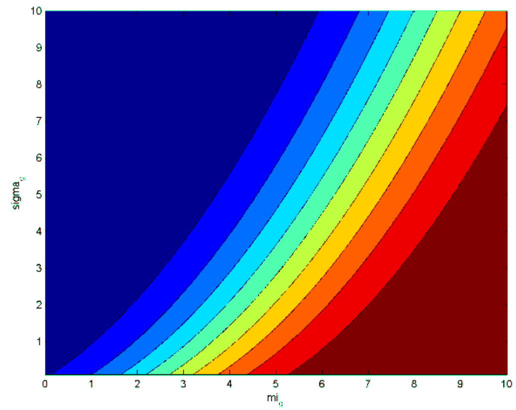
- **Gini > KS**



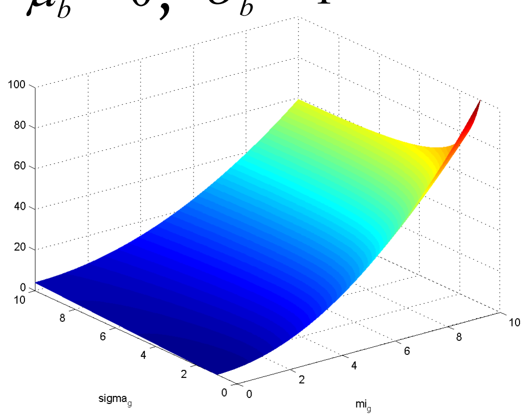
Normálně rozložené skóre

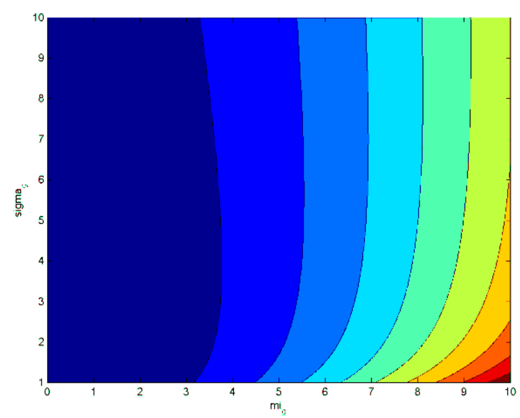
- **Lift$_{10\%}$:** $\mu_b = 0$, $\sigma_b^2 = 1$





- **I$_{val}$:** $\mu_b = 0,\ \sigma_b^2 = 1$





- V případě indexu Lift$_{10\%}$ je evidentní silná závislost na μ_g a významně vyšší závislost na σ_g^2 než v případě KS a Gini.

- Opět velmi silná závislost na μ_g. Navíc hodnota I$_{val}$ míří velmi rychle k nekonečnu pokud se σ_g^2 blíží nule.

Finanční dopad použití scoringových modelů

- We considered the number of credit proposals to be 150,000 per year, the reject rate (RR) to be 40%, and the average default rate (DR) to be 10.5%.
- Furthermore, we considered that the average gain resulting from rejecting a bad client (saved loss) in favor of accepting a good client (earned interest) was €300.
- For further comparison we considered the number of proposals to be 450,000 per year, the reject rate to be 20%, and the gain to be €1,500.

Finanční dopad použití scoringových modelů

- We needed to estimate the number of bad clients who could be rejected by a credit scoring model in addition to rejecting without any model, but with the same reject rate.
- Because $Lift_{RR}$ (i.e., $Lift_q$, *with* $q = RR$, where RR is the reject rate) is defined as the ratio of the proportion of bad clients below a given rejection level (RR) to the proportion of bad clients in the general population, and given our assumptions, we are able to estimate the desired number of bad clients.
- Then, because we know the gain resulting from rejecting a bad client in favor of accepting a good client, we can estimate the profit resulting from using a credit scoring model. The profit is given by

$$\boxed{profit = \#\, proposals \cdot DR \cdot RR \cdot (Lift_{RR} - 1) \cdot gain}$$

Finanční dopad ...

Quality indices / Portfolio parameters								proposals: 150 000/year				prop. : 450 000/year	
								gain: 300 € /credit		gain: 1500 € /credit		gain: 1500 € /credit	
								RR: 40%	RR: 20%	RR: 40%	RR: 20%	RR: 40%	RR: 20%
D	**KS**	**Gini**	**c-stat**	**Lift$_{10\%}$**	**Lift$_{20\%}$**	**Lift$_{40\%}$**	**Ival**	**Profit**	**Profit**	**Profit**	**Profit**	**Profit**	**Profit**
0,2500	**0,0995**	**0,1403**	**0,5702**	**1,4422**	**1,3376**	**1,2197**	**0,0625**	**415 318 €**	**319 019 €**	**2 076 589 €**	**1 595 095 €**	**6 229 766 €**	**4 785 284 €**
0,5000	**0,1974**	**0,2763**	**0,6382**	**1,9794**	**1,7156**	**1,4395**	**0,2500**	**830 718 €**	**676 264 €**	**4 153 588 €**	**3 381 320 €**	**12 460 764 €**	**10 143 959 €**
0,7500	**0,2923**	**0,4041**	**0,7021**	**2,5987**	**2,1187**	**1,6489**	**0,5625**	**1 226 474 €**	**1 057 182 €**	**6 132 369 €**	**5 285 909 €**	**18 397 106 €**	**15 857 726 €**
0,8620	**0,3335**	**0,4578**	**0,7289**	**2,8977**	**2,3028**	**1,7370**	**0,7430**	**1 392 838 €**	**1 231 152 €**	**6 964 189 €**	**6 155 762 €**	**20 892 566 €**	**18 467 285 €**
1,0000	**0,3829**	**0,5205**	**0,7602**	**3,2801**	**2,5294**	**1,8391**	**1,0000**	**1 585 984 €**	**1 445 248 €**	**7 929 919 €**	**7 226 240 €**	**23 789 757 €**	**21 678 719 €**
1,2500	**0,4680**	**0,6232**	**0,8116**	**3,9988**	**2,9304**	**2,0041**	**1,5625**	**1 897 678 €**	**1 824 194 €**	**9 488 388 €**	**9 120 970 €**	**28 465 165 €**	**27 362 911 €**
1,5000	**0,5467**	**0,7112**	**0,8556**	**4,7287**	**3,3068**	**2,1406**	**2,2500**	**2 155 813 €**	**2 179 903 €**	**10 779 067 €**	**10 899 516 €**	**32 337 201 €**	**32 698 548 €**

- Firstly, we can see that a firm with 150,000 credit proposals per year, a 40% reject rate, and a €300 gain per credit can earn approximately €1.4M per year when using the given credit scoring model compared to the case of using no model.
- Secondly, we can see that improving a model, by means of improving the quality indices, leads to a situation where a smaller reject rate results in a higher profit.
- And finally, we can see that a firm with an only three times bigger portfolio and five times higher gain per credit, i.e., 450,000 proposals per year and €1,500 per credit, and with an excellent model can increase its profit by more than €32M per year, which is quite a noticeable amount of money.

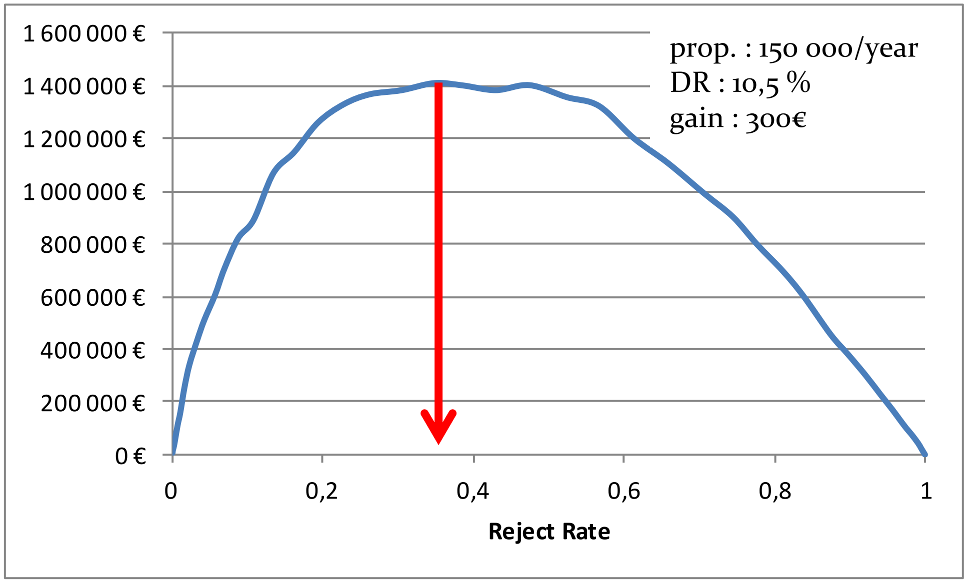
Finanční dopad …

Portfolio parameters / Quality indices								proposals: 150 000/year				prop. : 450 000/year	
								gain: 300 € /credit		gain: 1500 € /credit		gain: 1500 € /credit	
								RR: 40%	RR: 20%	RR: 40%	RR: 20%	RR: 40%	RR: 20%
D	**KS**	**Gini**	**c-stat**	**Lift$_{10\%}$**	**Lift$_{20\%}$**	**Lift$_{40\%}$**	**Ival**	**Profit**	**Profit**	**Profit**	**Profit**	**Profit**	**Profit**
0,2500	**0,0995**	**0,1403**	**0,5702**	**1,4422**	**1,3376**	**1,2197**	**0,0625**	**415 318 €**	**319 019 €**	**2 076 589 €**	**1 595 095 €**	**6 229 766 €**	**4 785 284 €**
0,5000	**0,1974**	**0,2763**	**0,6382**	**1,9794**	**1,7156**	**1,4395**	**0,2500**	**830 718 €**	**676 264 €**	**4 153 588 €**	**3 381 320 €**	**12 460 764 €**	**10 143 959 €**
0,7500	**0,2923**	**0,4041**	**0,7021**	**2,5987**	**2,1187**	**1,6489**	**0,5625**	**1 226 474 €**	**1 057 182 €**	**6 132 369 €**	**5 285 909 €**	**18 397 106 €**	**15 857 726 €**
0,8620	**0,3335**	**0,4578**	**0,7289**	**2,8977**	**2,3028**	**1,7370**	**0,7430**	**1 392 838 €**	**1 231 152 €**	**6 964 189 €**	**6 155 762 €**	**20 892 566 €**	**18 467 285 €**
1,0000	**0,3829**	**0,5205**	**0,7602**	**3,2801**	**2,5294**	**1,8391**	**1,0000**	**1 585 984 €**	**1 445 248 €**	**7 929 919 €**	**7 226 240 €**	**23 789 757 €**	**21 678 719 €**
1,2500	**0,4680**	**0,6232**	**0,8116**	**3,9988**	**2,9304**	**2,0041**	**1,5625**	**1 897 678 €**	**1 824 194 €**	**9 488 388 €**	**9 120 970 €**	**28 465 165 €**	**27 362 911 €**
1,5000	**0,5467**	**0,7112**	**0,8556**	**4,7287**	**3,3068**	**2,1406**	**2,2500**	**2 155 813 €**	**2 179 903 €**	**10 779 067 €**	**10 899 516 €**	**32 337 201 €**	**32 698 548 €**

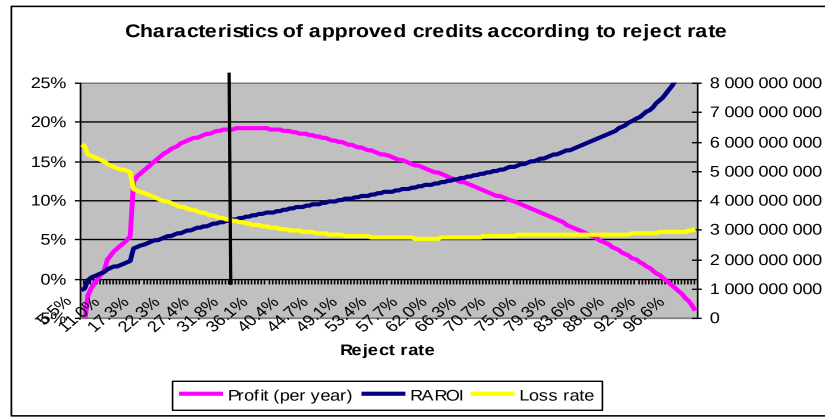
- Furthermore, one can compare the values of the expected profit within the columns as well as within the rows. This means that it is possible to compare the profit of the different portfolios provided by a credit scoring model with a given quality.
- But one can also compare the profit for a given portfolio according to the quality of the credit scoring model. For instance, if a firm with 150,000 credit proposals per year, a €300 gain per credit, and a 40% reject rate enhances the model and its Gini index increases from 0.4578 to 0.5205 (i.e., an increase of 0.06, which is an improvement of approximately 14%), the expected profit is approximately €193K per year (€1,585,984 minus €1,392,838).
- The typical potential for improving the Gini index is between 10% and 20% in the case of scoring models for consumer credit, provided that the redevelopment is carried out once or twice a year, which is usually the optimal time period. If a firm has credit-scoring-model development costs of around €20K, obviously it is profitable to redevelop the model and to maintain its quality, in the sense of the listed indices, at as high a level as possible.

Finanční dopad …

Odhad zisku pomocí uvedeného vzorce:

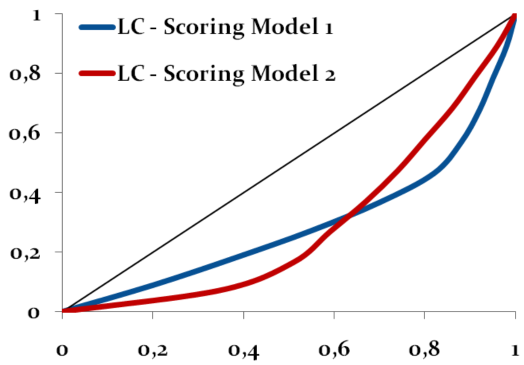


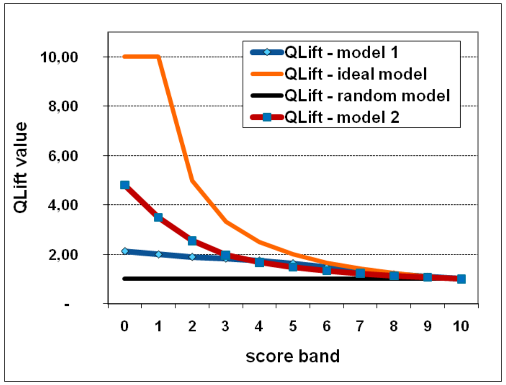
Vyjádření reálného zisku na jiném portfoliu, ale s podobným průběhem křivky Liftu:



- Pomocí uvedeného vzorce pro odhad zisku lze také zkoumat závislost zisku na hladině zamítání (Reject Rate).
- Tímto způsobem lze přibližně určit optimální hladinu zamítání. V našem případě je to 34,85 %.
- Obrázek odpovídá modelu s charakteristikami kvality zvýrazněnými v předchozí tabulce (žlutě zvýrazněný řádek).
- Změna ostatních parametrů (#proposals, DR a gain) vede pouze ke změně hodnot na vertikální ose. Tvar křivky zisku, tedy i bod maxima, zůstává stejný a ovlivňuje ho jen a pouze průběh křivky Liftu.

Finanční dopad …

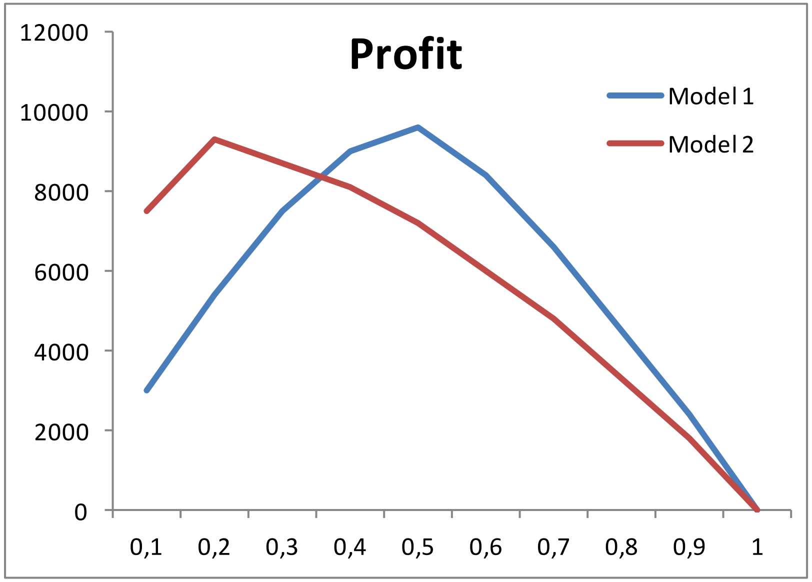




	scoring model 1	scoring model 2
GINI	0.420	0.420
KS	0.356	0.344
QLift(0.1)	2.000	3.500
LR	0.242	0.372
IRL	0.699	0.713

- Uvažujme příklad scoringových modelů ze str. 36
  - totožný Gini,
  - podle KS je lepší model 1
  - podle indexů založených na Liftu je lepší model 2

- Vyjádříme zisk pomocí našeho vzorce
  - absolutní úroveň zisku je srovnatelná
  - max. zisku lze s **modelem 1** dosáhnout při **RR = cca 50%**
  - max. zisku lze s **modelem 2** dosáhnout při **RR = cca 20%**



Závěr

- Efektivní využití scoringových modelů je nemožné bez znalosti jejich kvality.
- Posuzování kvality je silně závislé na definici dobrého/špatného klienta. Zavedení typu „indeterminate“ většinou nepřináší žádné zlepšení.
- Je potřeba posuzovat scoringové modely podle jejich síly v oblasti skóre, kde je očekávána cutoff hodnota.
- Giniho index ani KS nestačí!
- Výsledky týkající se Liftu a informační hodnoty lze použít k výběru nejlepšího scoringového modelu.
- Výsledky pro normálně rozložené skóre lze použít pro výpočet všech uvedených indexů. Navíc mohou pomoci v porozumění, jak se tyto indexy chovají v závislosti na uvedených parametrech.
- Navržený vzorec pro odhad zisku plynoucího z použití scoringového modelu lze použít nejen pro samotný odhad zisku, ale také pro přibližné stanovení optimální (vzhledem k zisku) míry zamítání a pro porovnání daných modelů vzhledem k takto určené optimální míře zamítání, přičemž je zřejmé, že model generující srovnatelný zisk s nižší mírou zamítání je lepší.

Děkuji za pozornost.