MASARYK UNIVERSITY
Czech Republic

On Aspects of Quality Indexes for Scoring Models

Martin Řezáč, Jan Koláček
Dept. of Mathematics and Statistics, Faculty of Science,
Masaryk University

COMPSTAT' 2010, Paris

Content

Introduction

- Credit scoring is the set of predictive models and their underlying techniques that aid financial institutions in the granting of credits.

- While it does not identify "good" or "bad" applications on an individual basis, it provides statistical odds, or probability, that an applicant with a given score turns to be "good" or "bad".

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 (according to some criteria).

- Before measuring the quality of models one should know (among other things):
  - expected reject rate (expected cutoff)

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 will consider following widely used quality indexes:
  - Kolmogorov-Smirnov statistics (KS)
  - Gini index
  - C-statistics
  - Lift.

Measuring the quality

- We consider following markings:

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

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 cumulative distribution functions (CDF):

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

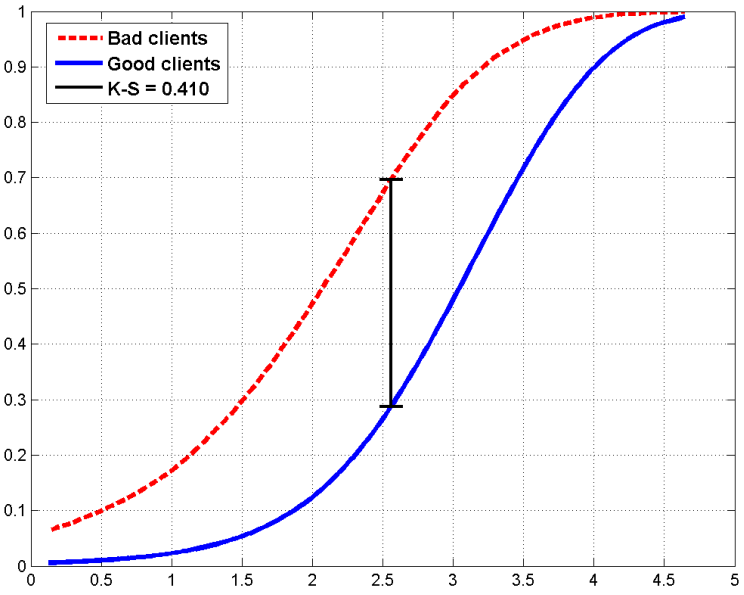
$$F_{m.BAD}(a) = \frac{1}{m}\sum_{i=1}^{m} I(s_i \le a \wedge D_K = 0) \qquad I(A) = \begin{cases} 1 & A\ is\ true \\ 0 & A\ is\ false \end{cases}$$

KS statistics

- KS is defined as maximal absolute difference between CDFs of good and bad clients:

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

- It takes values from 0 to 1. Value 0 corresponds to random model, value 1 corresponds to ideal model.



Gini index

- Lorenz curve is defined paramertrically:

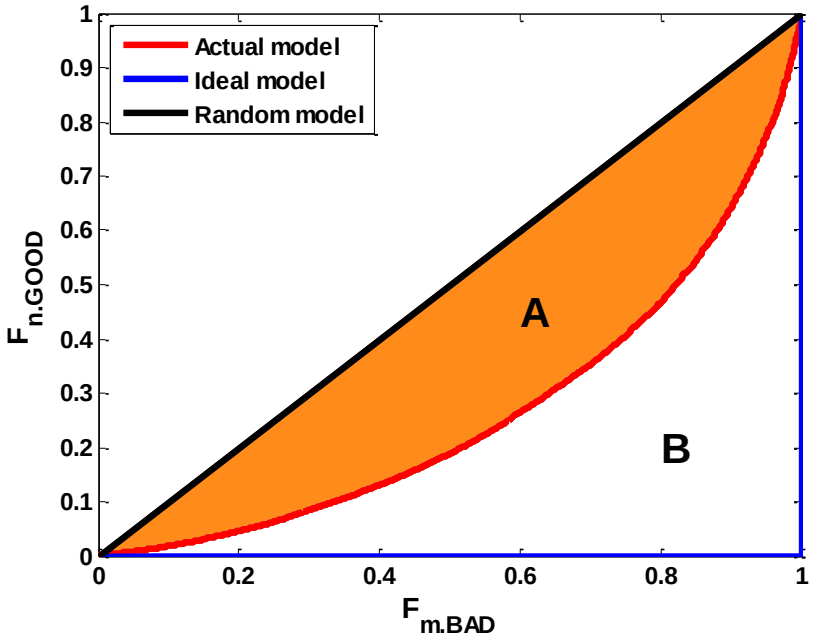
$$x = F_{m.BAD}(a)$$

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

- Gini index is defined as

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

- It takes values from 0 to 1. Value 0 corresponds to random model, value 1 corresponds to ideal model.



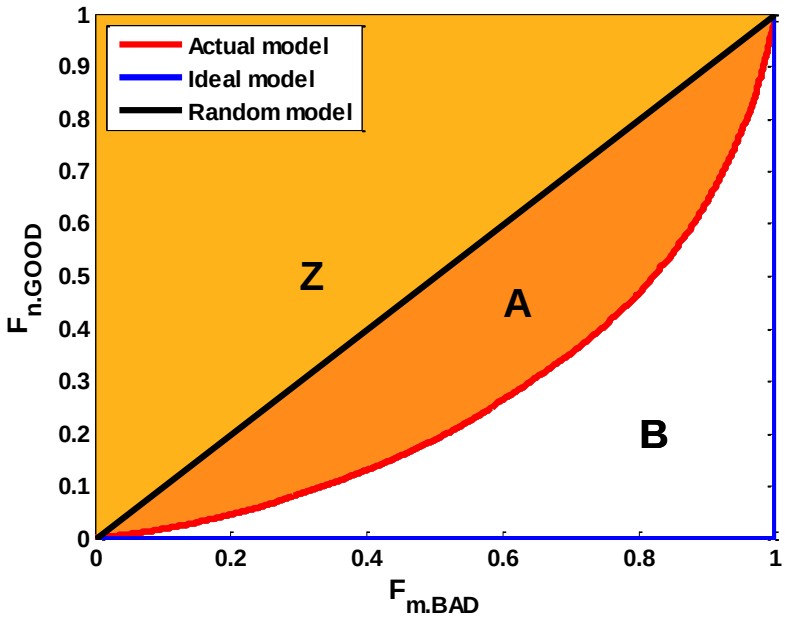
$$Gini = 1 - \sum_{k=2}^{n+m} (F_{m.BAD\,k} - F_{m.BAD\,k-1}) \cdot (F_{n.GOOD\,k} + F_{n.GOOD\,k-1})$$

where $F_{m.BAD\,k}$ ($F_{n.GOOD_k}$) is kth vector value of empirical distribution function of bad (good) clients

C-statistics

- C-statistics is defined as area over Lorenz curve:

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



- It takes values from 0.5 to 1. Value 0.5 corresponds to random model, value 1 corresponds to ideal model.

- Using ROC methodology it is equal to AUROC (AUC).

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

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

Lift

- Another possible indicator of the quality of scoring model is *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 smaller score than a score *a*, $a \in [L, H]$, to the proportion of bad clients in the whole population. Formally, it can be expressed by:

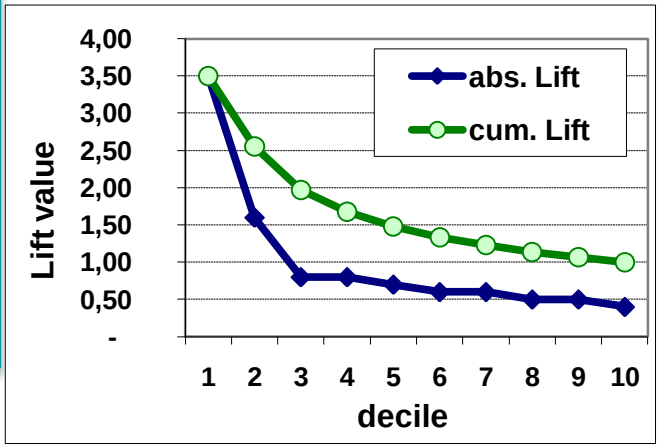
$$Lift(a) = \frac{CumBadRate(a)}{BadRate} = \frac{\frac{\sum_{i=1}^{n+m} I(s_i \le a \wedge Y = 0)}{\sum_{i=1}^{n+m} I(s_i \le a)}}{\frac{\sum_{i=1}^{n+m} I(Y = 0)}{\sum_{i=1}^{n+m} I(Y = 0 \vee Y = 1)}} = \frac{\frac{\sum_{i=1}^{n+m} I(s_i \le a \wedge Y = 0)}{\sum_{i=1}^{n+m} I(s_i \le a)}}{\frac{n}{N}}$$

- It is possible to consider also absolute Lift $absLift(a) = \frac{BadRate(a)}{BadRate}$, but we will focus on the cumulative form.

Lift

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

decile	# cleints	absolutely			cumulatively		
		# bad clients	Bad rate	abs. Lift	# bad clients	Bad rate	cum. Lift
1	100	35	35.0%	3.50	35	35.0%	3.50
2	100	16	16.0%	1.60	51	25.5%	2.55
3	100	8	8.0%	0.80	59	19.7%	1.97
4	100	8	8.0%	0.80	67	16.8%	1.68
5	100	7	7.0%	0.70	74	14.8%	1.48
6	100	6	6.0%	0.60	80	13.3%	1.33
7	100	6	6.0%	0.60	86	12.3%	1.23
8	100	5	5.0%	0.50	91	11.4%	1.14
9	100	5	5.0%	0.50	96	10.7%	1.07
10	100	4	4.0%	0.40	100	10.0%	1.00
All	1000	100	10.0%				



- It takes positive values. Cumulative form ends in value 1.
- Upper limit of Lift depends on p_B .

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:

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

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

- Typical value of q is 0.1. Then we have

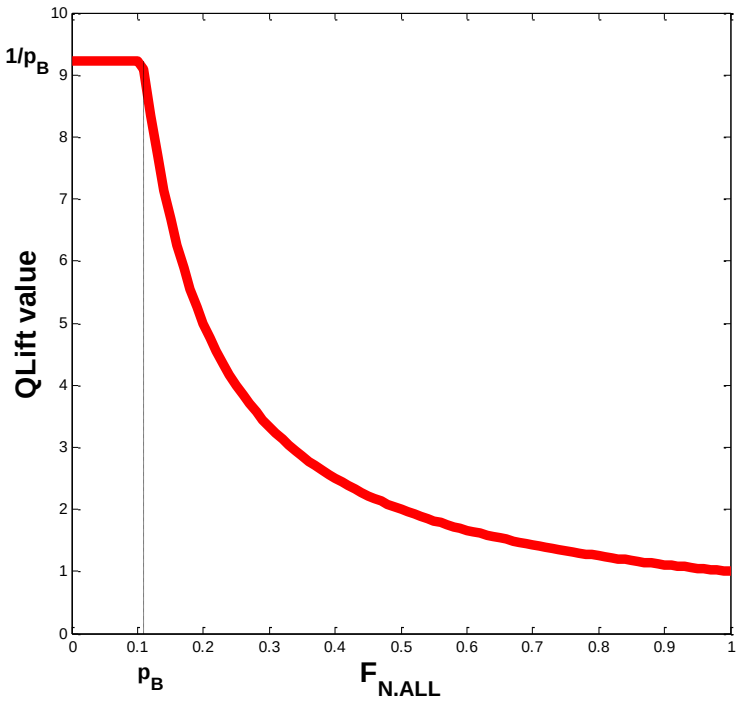
$$QLift_{10\%} = QLift(0.1) = 10 \cdot F_{m.BAD}(F^{-1}_{N.ALL}(0.1))$$

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:

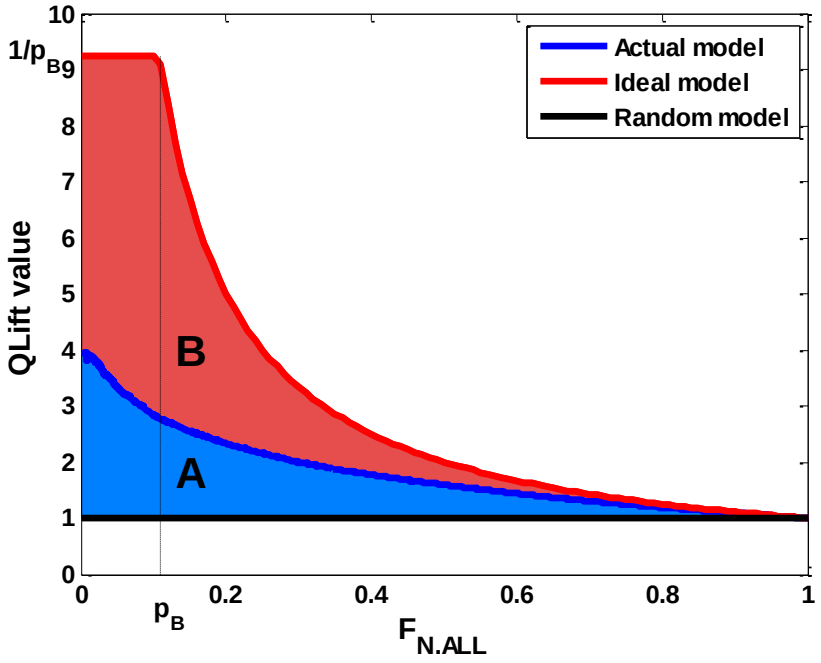
$$QLift_{ideal}(q) = \begin{cases} \frac{1}{p_B}, & q \in (0, p_B] \\ \frac{1}{q}, & q \in (p_B, 1] \end{cases}$$

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.

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

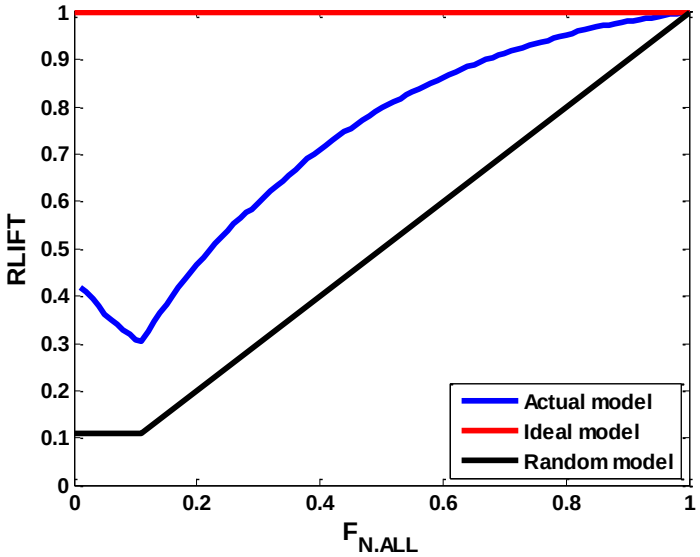


- It is obvious that it is global measure of model's quality and that it takes values from 0 to 1. Value 0 corresponds to random model, value 1 match to ideal model. Meaning of this index is quite simple. The higher, the better. Important feature is that Lift Ratio allows us to fairly compare two models 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):

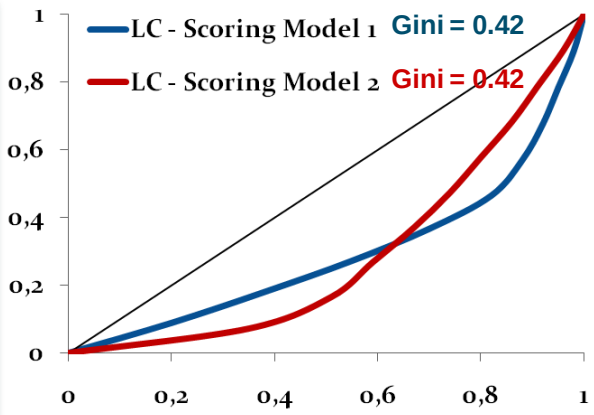
$$IRL = \int_0^1 RLift(q)\, dq$$

- 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.

Example

- 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.
- From the Lorenz curves is evident, 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.

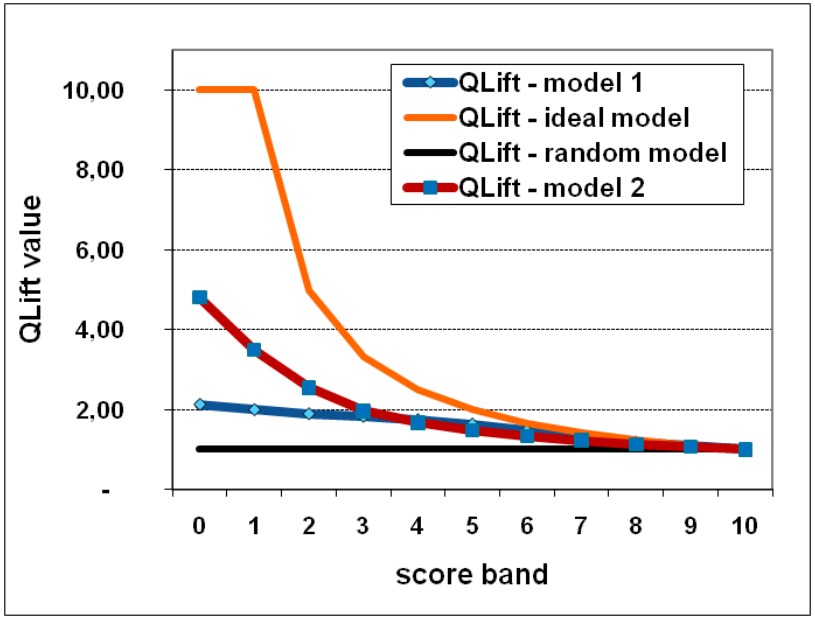
			Scoring Model 1				Scoring Model 2			
score band	# clients	q	# bad clients	# cumul. bad clients	# cumul. bad rate	QLift	# bad clients	# cumul. bad clients	# cumul. bad rate	QLift
1	100	0.1	20	20	20.0%	**2.00**	35	35	35.0%	**3.50**
2	100	0.2	18	38	19.0%	**1.90**	16	51	25.5%	**2.55**
3	100	0.3	17	55	18.3%	**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			

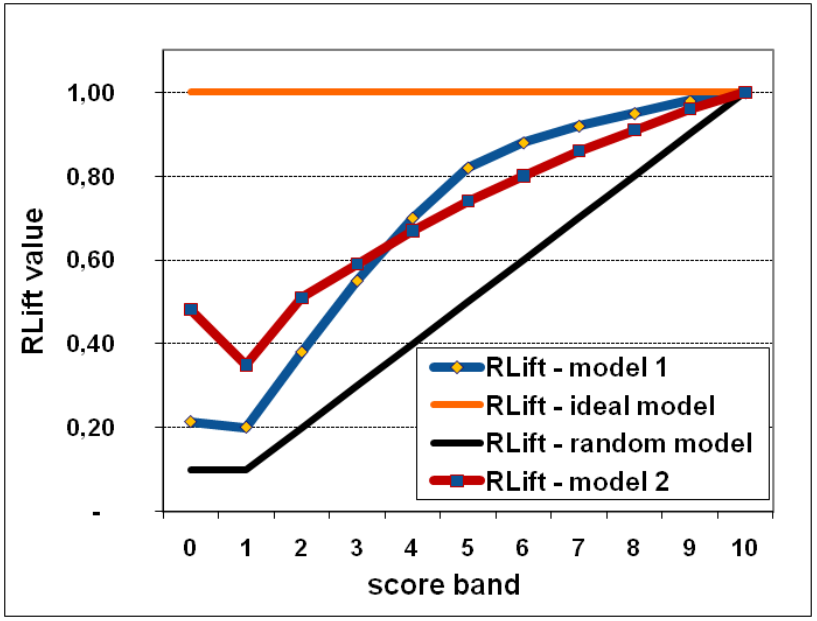


Example

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

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



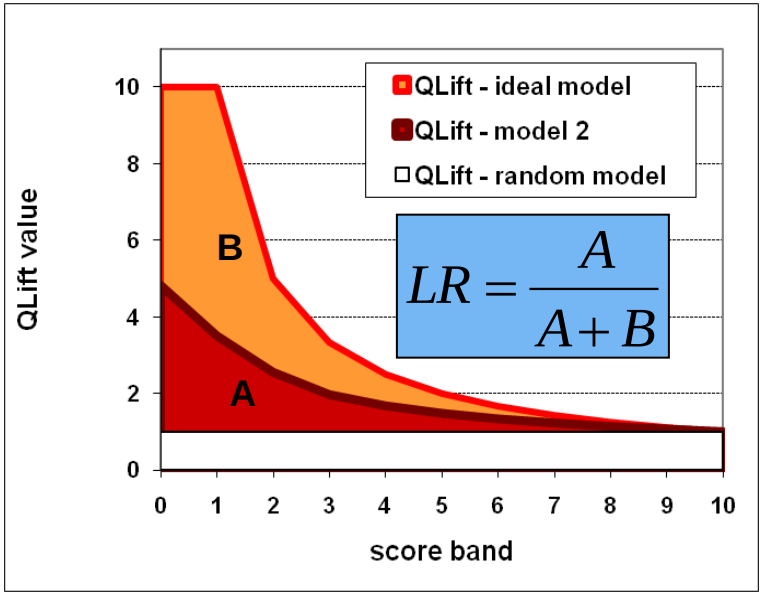


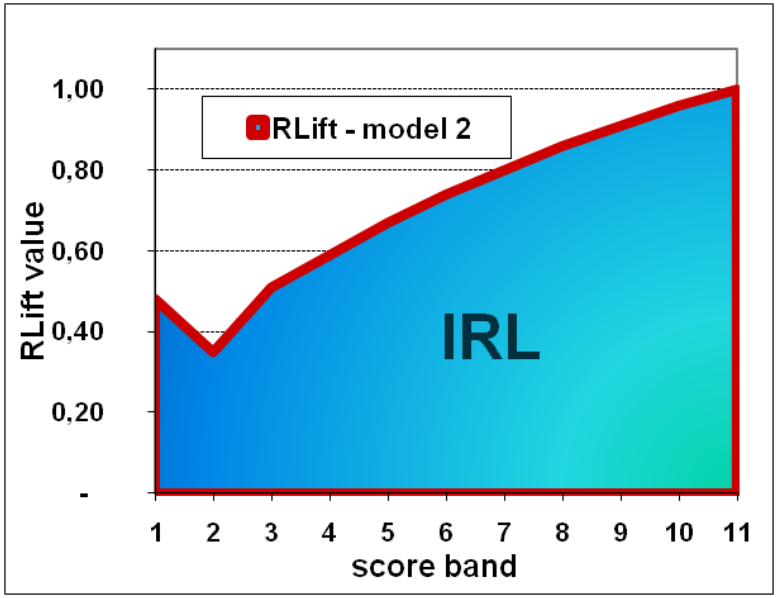
According to both Qlift and Rlift curves we can state that:

- 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.

Example

- Now, we consider indexes LR and IRL:



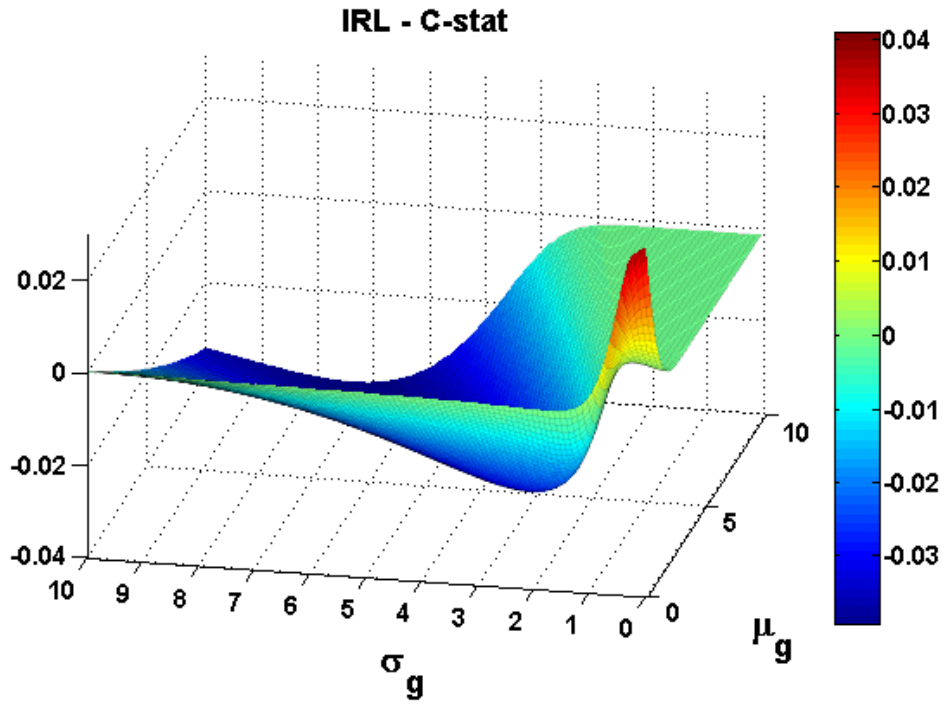


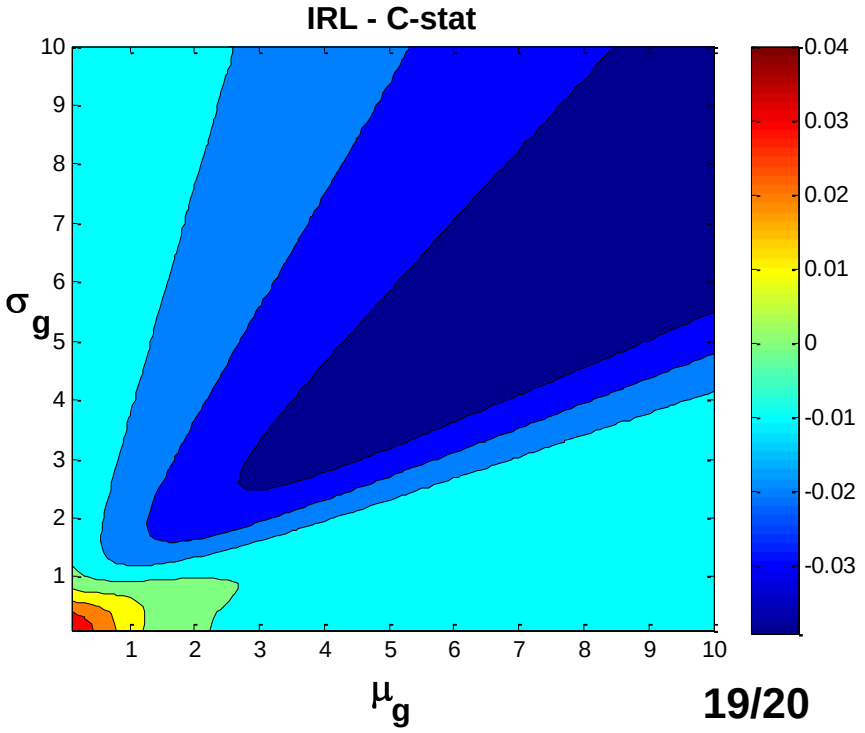
	scoring model 1	scoring model 2
GINI	**0.420**	**0.420**
QLift(0.1)	**2.000**	**3.500**
LR	**0.242**	**0.372**
IRL	**0.699**	**0.713**

Using LR and IRL we can state that model 2 is better than model 1 although their Gini coefficients are equal.

Simulation study

We made a simulation with scores generated from normal distribution. Scores of bad clients had mean equal to 0 and variance equal to 1. Scores of good clients had mean and variance from 0.1 to 10 with step equal 0.1. Number of samples and sample size was 1000, p_B was equal to 0.1. IRL and c-statistics were computed for each sample and each value of mean and variance of good client's scores. Finally, means of IRL and c-statistics were computed.





Conclusions

- It is necessary to judge scoring models according to their strength in score range where cutoff is expected.
- The Gini and KS are not enough!
- Results concerning Lift can be used to obtain the best available scoring model.
- Formula for Lift (QLift) for ideal model was derived. This allowed to propose new advanced indexes – Lift Ratio and Integrated Relative Lift.
- The simulation shows that IRL and c-statistics are approximately equal in case that variances of good and bad clients are equal. Furthermore it shows that they significantly differ in another cases.