Measuring the Quality of Credit Scoring Models

Martin Řezáč[1]
František Řezáč[2]
Masaryk University, Czech Republic

Abstract
In the current strong competitive environment it is quite fundamental good care of the quality of client portfolio. Credit scoring models are widely used to achieve this business aim. For a measurement of quality of the scoring models it is possible to use quantitative indexes such as Gini index, K-S statistics, Lift, Mahalanobis distance and Information statistics. They can be used for comparison of several developed models at the moment of development. It is possible to use them for monitoring of quality of models after the deployment into real business as well. Figures like ROC curve (Lorenz curve), Lift chart (Gains chart) can be used as well.
This paper deals with definition of good/bad client, which is crucial for further computations. Parameters affecting this definition are discussed. The main part is devoted to quality indexes based on distribution functions (Gini, K-S and Lift) and on density functions (Mahalanobis distance, Information statistics). It brings some interesting results connected to the Lift, cumulative and absolute Lift as well. In case of density based indexes it is mentioned kernel density estimate and bandwidth estimation based on maximal smoothing approach. Some results for normally distributed scores are included.

Key Words: Predictive modeling, Credit scoring, Quality indexes, Normally distributed scores

1. Introduction

Banks and other financial institutions receive thousands of credit applications every day (in case of consumer credits it can be tens or hundreds of thousands every day). Since it is impossible to process them manually, automatic systems are widely used by these institutions for evaluating credit reliability of individuals who ask for credit. The assessment of the risk associated with granting of credits has been underpinned by one of the most successful applications of statistics and operations research: credit scoring.

Credit scoring is the set of predictive models and their underlying techniques that aid financial institutions in the granting of credits. These techniques decide who will get credit, how much credit they should get, and what further strategies will enhance the profitability of the borrowers to the lenders. Credit scoring techniques assess the risk in lending to a particular client. It does not identify "good" or "bad" (negative behaviour is expected, e.g. default) applications on an individual basis, but it provides probability, that an applicant with any given score will be "good" or "bad". These probabilities or scores, along with other business

[1] Dept. of Mathematics and Statistics, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic, mrezac@math.muni.cz

[2] Dept. of Finance, Faculty of Economics and Administration, Masaryk University, Lipová 41a, 602 00 Brno, Czech Republic, rezac@econ.muni.cz

considerations such as expected approval rates, profit, churn, and losses, are then used as a basis for decision making.

Several methods connected to credit scoring were introduced during last six decades. The most known and widely used are logistic regression, classification trees and neural networks. These and many other methods like nearest-neighbour, discriminant analysis and linear programming approach are discussed in Anderson (2007), Crook at al. (2007), Hand and Henley (1997), Siddiqi (2006) or Thomas et al. (2002). Further remarks connected to credit scoring issues can be found there as well.

2. Definition of good/bad client

In fact, the most important step in predictive model building is the correct definition of dependent variable. In case of credit scoring it is necessary to precisely define good and bad client. Usually this definition is based on the client's number of days after the due date (days past due, DPD) and the amount past due. We need to set some tolerance level in case of the past due amount. It means what it is considered as the debt and what is not. It may be that the client gets into payment delay innocently (because of technical imperfections of the system). It does not make sense to regard as debt small amount (e.g. less than 3€) past due as well. Furthermore, it is necessary to determine the time horizon in which the previous two characteristics are traced. For example, as a good is marked client who:

- Has less than 60 DPD (with tolerance 3€) in 6 months from the first due date
- Has less than 90 DPD (with tolerance 1€) ever

Choice of these parameters depends greatly on the type of financial product (certainly will be different parameters for consumer loans for small amounts with original maturities around one year and for mortgages, which are typically connected to very large amounts with maturities up to several tens of years) and on further usage of this definition (credit scoring, fraud prevention, marketing, ...). Another practical issue of the definition of good client is the accumulation of several agreements. For example, it may be that the customer is overdue on more contracts, but with different days past due and with different amounts. In this case, all amounts past due connected to the client in one particular point in time are usually added together and it is taken the maximum value from days past due. This approach can be applied only in some cases and especially in a situation where there is a complete accounting data. The situation is considerably more complex in case of aggregated data.

In connection with the definition of good client we can generally talk about the following types of clients:

- Good
- Bad
- Indeterminate
- Insufficient
- Excluded
- Rejected.

The first two types were discussed. The third type of client is on the border between good and bad clients, and directly affects their definition. If we are considering only DPD, clients with a high DPD (e.g. 90 +) are typically identified as bad, clients who are not delinquent (e.g. their DPD are less than 30 or equal to zero) are identified as good. As indeterminate are then

considered delinquent customers who have not exceeded given threshold of DPD. The next type is typically case of the clients with the very short history, which makes impossible the correct definition of dependent variable (good / bad client). The excluded clients are for example clients who have exited the system, or they are so close to the point of no return that their classification is indisputable. They are also marked as "hard bad". The meaning of rejected client is obvious. See Anderson (2007) or Thomas et al. (2002) for more details.

3. 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, Gini index and Lift. The second, indexes based on likelihood density function like Mean difference (Mahalanobis distance) and Informational statistics. For further available measures and appropriate remarks see Wilkie (2004), Giudici (2003) or Siddiqi (2006).

3.1. Indexes based on distribution function

Assume that score s is available for each client and put the following markings.

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

Distribution functions, respectively their empirical forms, of scores of good (bad) clients are given by relationship

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

$$F_{m.BAD}(a) = \frac{1}{m}\sum_{i=1}^{m} I\left(s_i \le a \wedge D_K = 0\right),\ a \in \left[L, H\right],$$

where s_i is score of i-th client, n is number of good, m is number of bad clients. L is the minimum value of given score, H is the maximum value. The proportion of bad clients we denote by

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

proportion of good clients by

$$p_G = \frac{n}{n+m}.$$

Furthermore, empirical distribution function of scores of all clients is given by

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

where $N = n + m$ is number of all clients.

An often-used characteristic in describing the quality of the model (scoring function) is Kolmogorov-Smirnov statistics (K-S or KS). It is defined as

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

Figure 1 gives an example of estimation of distribution functions of good and bad clients, including an estimate of KS statistics. It can be seen, for example, that the score around 2.5 and smaller has population of approximately 30% of good clients and 70% of bad clients.

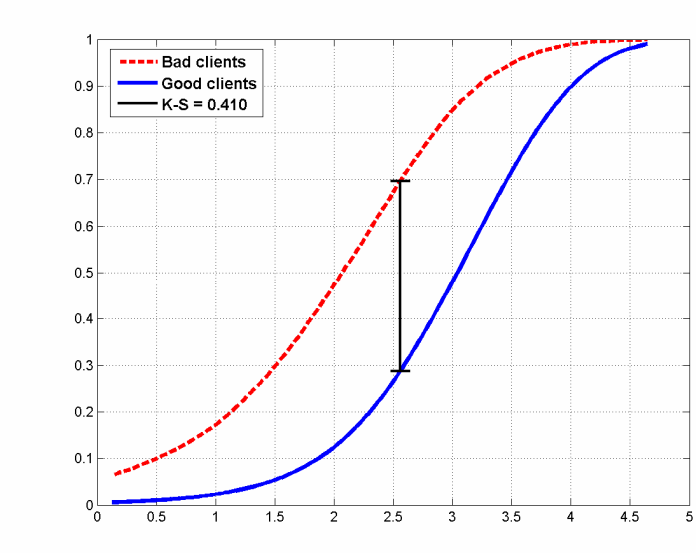

Figure 1: Distribution Functions, KS.

The Lorenz curve (LC), sometimes called ROC curve (Receiver Operating Characteristic curve), can also be successfully used to show the discriminatory power of scoring function, i.e. the ability to identify good and bad clients. The curve is given parametrically by

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

In connection to LC we assume next quality measure, Gini index. This index describes a global quality of scoring function. It takes values between 0 and 1. The ideal model, i.e. scoring function that perfectly separate good and bad clients, has the Gini index equal to 1. On the other hand, model that assigns a random score to the client has this index equal to 0. Using Figure 2 it can be defined

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

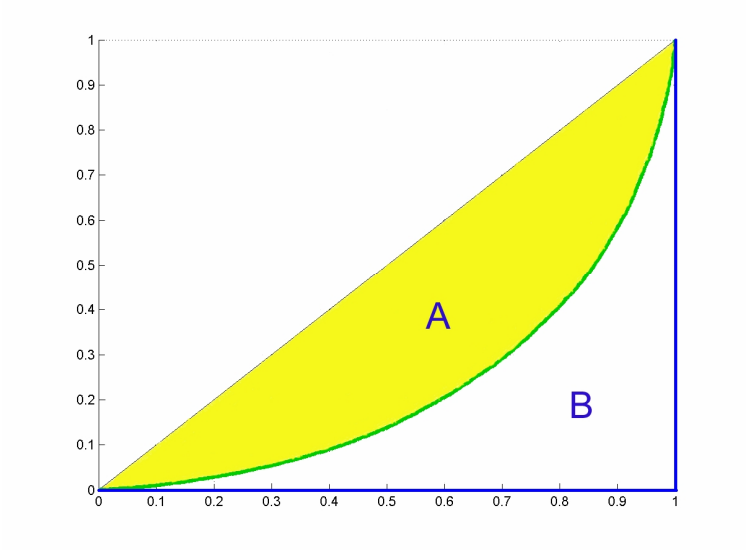

Figure 2: Lorenz Curve, Gini Index.

The actual calculation of Gini index can be, given the previous markings, made using

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

where $F_{m.BAD\,k}$ ($F_{n.GOOD_k}$) is k-th vector value of empirical distribution function of bad (good) clients. For further details see Anderson (2007) or Xu (2003). In connection to the Gini index, c-statistics (see Siddiqi 2006) is defined as

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

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

$$c - stat = P\left(s_1 \geq s_2 \;\middle|\; D_{K_1} = 1 \wedge D_{K_2} = 0\right).$$

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:

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

It can be easily verified that the Lift can be equivalently expressed as

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

In practice, this calculation is done for Lift corresponding to 10%, 20%, ..., 100% of clients with the worst score. Let's demonstrate this procedure by the following example, taken from Coppock (2002). Assume that we have a score of 1,000 clients, of which 50 are bad. The proportion of bad clients is 5%. Sort customers according to score and split into ten groups, i.e., divide it by deciles of score. In each group, in our case around 100 clients, then count bad clients. This will get their share in the group (Bad Rate). Absolute Lift in each group is then given by the ratio of the share of bad clients in the group to the proportion of bad clients in total. Cumulative Lift is given by the ratio of the share of bad clients in groups up to the given group to the proportion of bad clients in total. See Table 1.

Table 1: Absolute and Cumulative Lift

		absolutely			cumulatively		
decile	# cleints	# 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%				

In connection to the previous example we define

$$Lift_q = \frac{F_{n.BAD}(F^{-1}_{N.ALL}(q))}{F_{N.ALL}(F^{-1}_{N.ALL}(q))} = \frac{1}{q} F_{n.BAD}\left(F^{-1}_{N.ALL}(q)\right),$$

where q represents the score level of $100q\%$ of the worst scores and $F^{-1}_{N.ALL}(q)$ can be computed as

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

Usually, q is assumed to be equal 0.1 (10%), i.e. we are interested in discriminatory power of scoring model in point of 10% of the worst scores. In this case we have

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

3.2. Indexes based on density function

Consider $f_{GOOD}(x)$ and $f_{BAD}(x)$ to be likelihood density functions of good or bad clients respectively. The empirical estimates we can get using kernel density estimates

$$\widetilde{f}_{GOOD}(x,h) = \sum_{\substack{i=1,\\ D_K=1}}^{n} \frac{1}{n} K_h(x - s_i),$$

$$\widetilde{f}_{BAD}(x,h) = \sum_{\substack{i=1\\ D_k=0}}^{m} \frac{1}{n} K_h(x - s_i),$$

where $K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$, K is some kernel function, e.g. Epanechnikov kernel. For further details see Wand and Jones (1995). The estimation of bandwidth h can be given by maximal smoothing principal approach, i.e.

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

where k is the order of kernel K, $\widetilde{\sigma}$ is an appropriate estimation of standard deviation and n is the number of observations. For further details see Řezáč (2002).

Let M_g and M_b be means of scores of good (bad), clients and S_g and S_b be standard deviations of good (bad) clients. Let S be the pooled standard deviation of the good and bad clients given by

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

Estimates of mean and standard deviation of scores for all clients (μ_{ALL}, σ_{ALL}) are given by

$$M = M_{ALL} = \frac{nM_g + mM_b}{n+m}, \; S_{ALL} = \left(\frac{n^2 S_g{}^2 + m^2 S_b{}^2}{(n+m)^2} \right)^{\frac{1}{2}}.$$

The first quality index based on density function is the standardized difference between the means of two groups of scores, i.e. scores of bad and good clients. Denote by D this mean difference, calculated as

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

Generally, good clients are supposed to get high scores and bad clients low scores, so that we would expect that $M_g > M_b$, so that D is positive. Another name for this concept is Mahalanobis distance; see Thomas et al. (2002).

The second index based on densities is the information statistics (value) I_{val} , defined in Hand and Henley (1997) as

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

It is useful to examine decomposed form of right-hand side expression. For this purpose we mark

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

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

Although the information statistics is global measure of model's quality, one can use graphs of f_{diff}, f_{LR} and graph of their product to examine local properties of given model.

The following Table 2 gives an example of computational scheme for informational statistics in case of discretized data.

Table 2: Informational Statistics

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

3.3 Figures for quality assessment

The concept of Lorenz curve was already mentioned. Each point of curve represents some value of given score. If we assume this value as cutoff value, we can read the proportion of rejected bad and good clients. An example of Lorenz curve is given in Figure 3. We can see that by rejection of 20% of good clients we reject 50% of bad clients at the same moment.

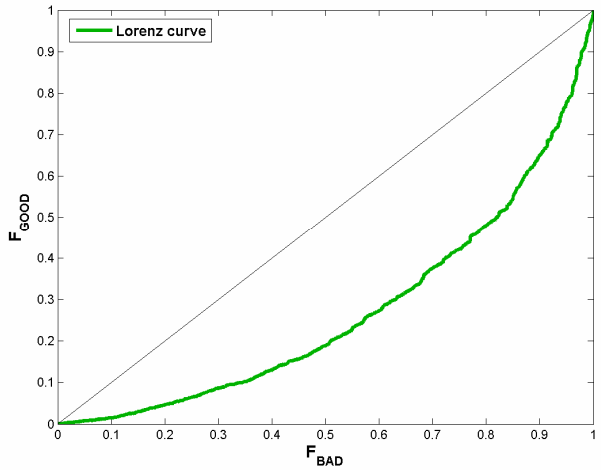

Figure 3: Lorenz Curve (ROC).

Another available type of quality assessment figure is 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. An example of Lift chart is displayed in Figure 4. 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 in case of Figure 4 we can see that if we want to reject 70% of bads, we have to reject about 40% of all applicants.

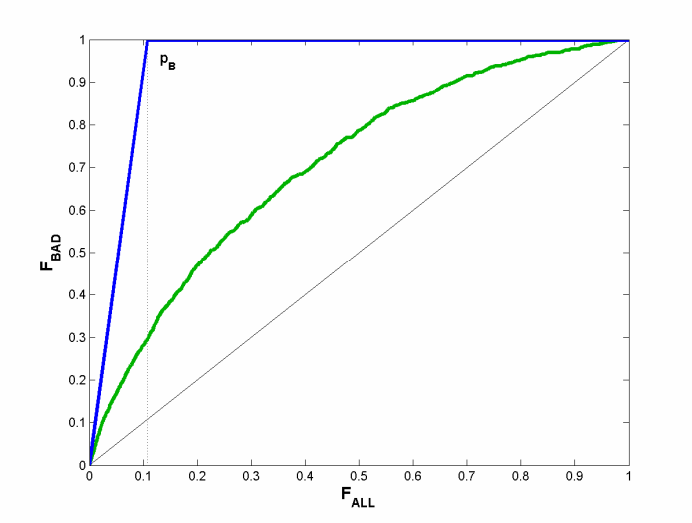

Figure 4: Lift chart.

It is called Gains chart in case of marketing usage. In this case, the horizontal axis represents proportion of clients who can be addressed by some marketing offer and the vertical axis represents proportion of clients who will accept the offer.

3.4 Some results for normally distributed scores

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

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

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

The values of M_g, M_b and S_g, S_b can be taken as estimates of μ_g, μ_b and σ_g, σ_b. Finally we assume that standard deviations are equal to a common value σ. In practice, this assumption should be tested by F-test.

The mean difference D is now defined as

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

and is calculated by

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

The maximum difference between the cumulative distributions, denoted KS before, is calculated at the point where the distributions cross, halfway between the means. The value KS is therefore given by

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

where $\Phi(\cdot)$ is the standardized normal distribution function.

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

For the Lift statistics is computation quite easy. Denoting $\Phi^{-1}(\cdot)$ the standard normal quantile function and $\Phi_{\mu,\sigma^2}(\cdot)$ the normal distribution function with expected value μ and variance σ^2, we have

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

Computational form is then

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

A couple of further interesting results are given in Wilkie (2004). One of them is that, under our assumptions on normality and equality of standard deviations, it holds

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

It is possible to find expressions for all mentioned indexes in general case, i.e. without assumption of equality of variances. The mean difference is now in form

$$D = \sqrt{2}\, D^*, \text{ where } D^* = \frac{\mu_g - \mu_b}{\sqrt{\sigma_g^2 + \sigma_b^2}}, \text{ for } \quad \sigma_g \neq \sigma_b.$$

The KS is given by

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

where $a = \sqrt{\sigma_b^2 + \sigma_g^2}$, $b = \sigma_b^2 - \sigma_g^2$, $c = \ln\left(\frac{\sigma_g}{\sigma_b}\right)$. Empirical form can be expressed by

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

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

Gini coefficient can be expressed as

$$G = 2 \cdot \Phi\left(D^*\right) - 1.$$

Lift is given by

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

When we replace theoretical means and standard deviations by their estimates we obtain

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

Finally, information statistics is given by

$$I_{val} = (A+1)D^{*2} + A - 1,$$

where $A = \frac{1}{2}\left(\frac{\sigma_g^2}{\sigma_b^2}+\frac{\sigma_b^2}{\sigma_g^2}\right)$, in computation form it is $A = \frac{1}{2}\left(\frac{S_g^2}{S_b^2}+\frac{S_b^2}{S_g^2}\right)$.

Some of these results are graphically expressed in relation to μ_b, μ_g and σ_b^2, σ_g^2 in the following figures. In case of Figures 5 to 8 it was selected $\mu_b = 0$, $\sigma_b^2 = 1$. There is displayed dependence of examined characteristics on μ_g a σ_g^2. In case of Figure 9 it was set $\mu_b = 0$, $\mu_g = 1$ and displayed value of I_{val} depending on the σ_b^2, σ_g^2. Right-hand side of all these figures is contour graph of appropriate graph at left-hand side.

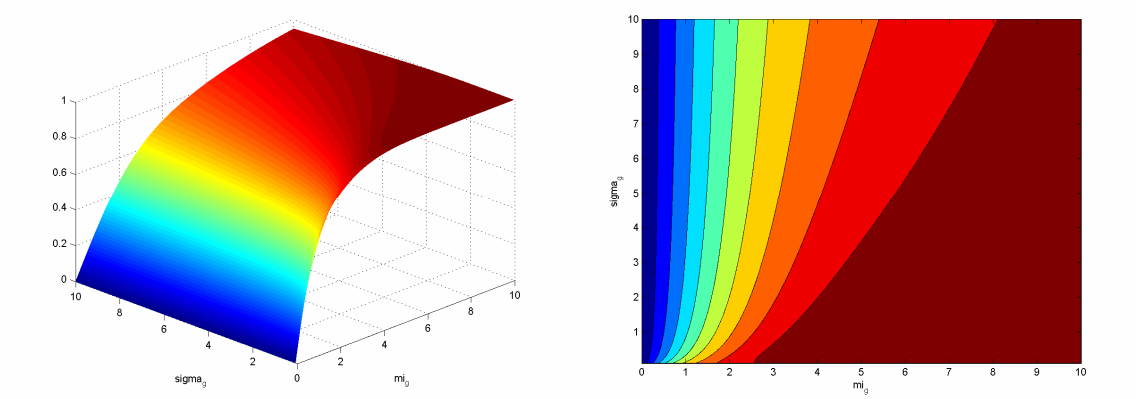

Figure 5: *KS*, $\mu_b = 0$, $\sigma_b^2 = 1$

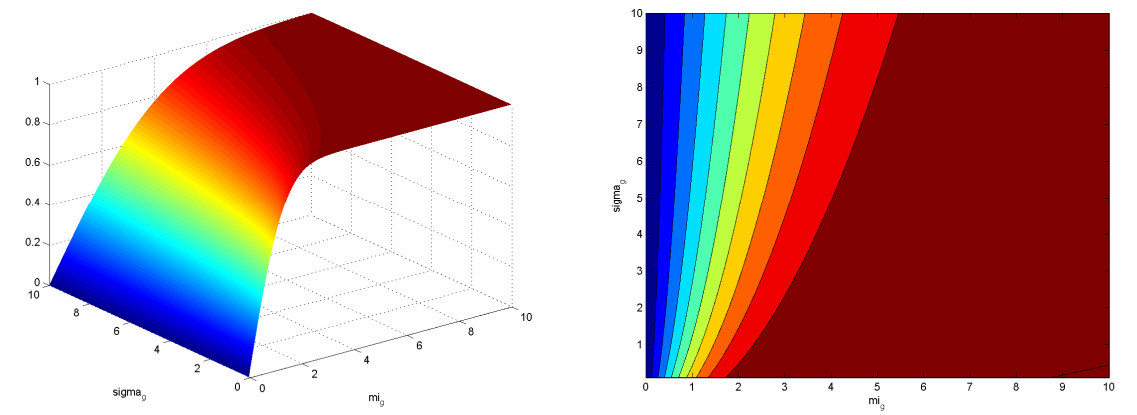

Figure 6: Gini coefficient, $\mu_b = 0$, $\sigma_b^2 = 1$

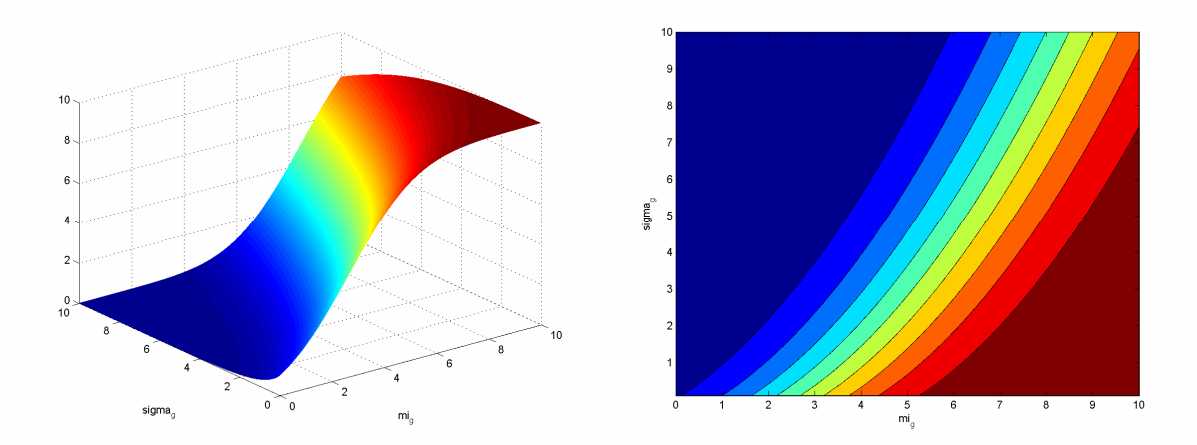

Figure 7: Lift$_{10\%}$, $\mu_b = 0$, $\sigma_b^2 = 1$

It is evident from the figures that KS statistics and the Gini react much more to change of μ_g and are almost unchanged in the direction of σ_g^2. Its theoretical maximum, i.e. 1, is approximately reached in value $\mu_g = 4$, which is the value where relevant probability densities of good and bad clients almost do not overlap and hence the perfect separation of these two groups is reached. In case of Lift$_{10\%}$, see Figure 7, it is evident strong dependence on μ_g. This time, however, the value of this index significantly affects the σ_g^2.

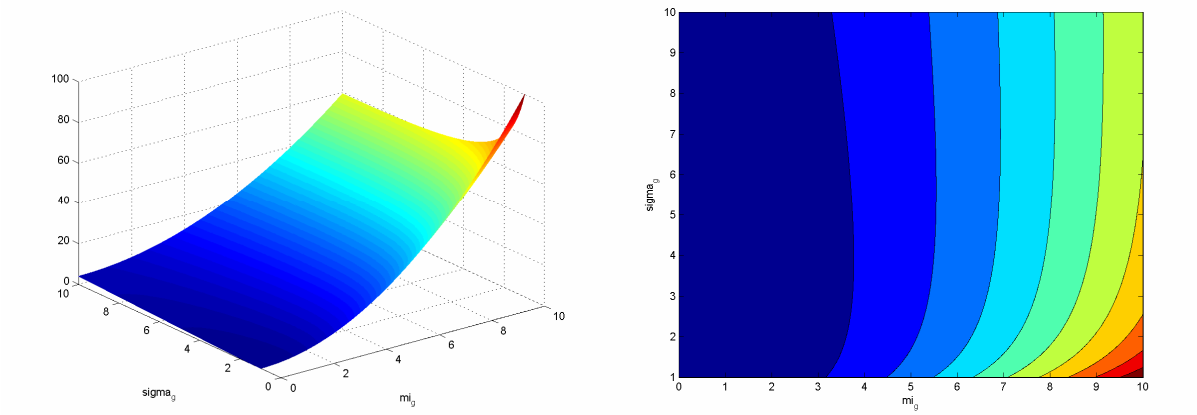

Figure 8: I_{val} , $\mu_b = 0$, $\sigma_b^2 = 1$

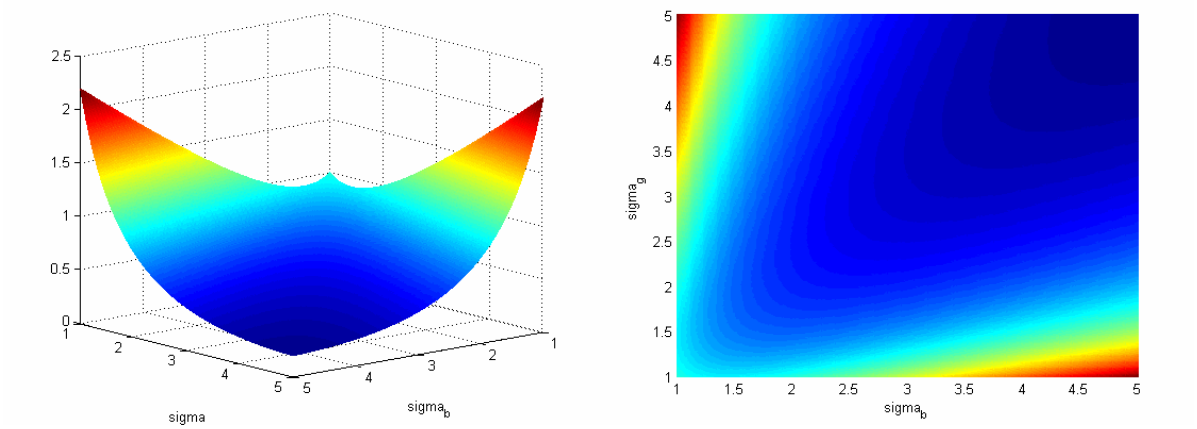

Figure 9: I_{val} $\mu_b = 0$, $\mu_g = 1$

Information statistics is again much more responsive to change of μ_g than to change of σ_g^2, see Figure 8. But there is one significant difference. If it is $\sigma_g^2 < \sigma_b^2$, i.e. $\sigma_g^2 < 1$ in our case, the value of information statistics is increasing very quickly. It is caused by the fact that overlap of the relevant probabilistic densities is significantly smaller and hence we can see faster growth towards its theoretical maximum, which lies in infinity. The dependence of information statistics on variances σ_b^2, σ_g^2 is captured in Figure 9. It takes high values when both variances approximately equal to 1, it grows to infinity if ratio of the variances tends to infinity or is near zero.

4. Case study

All numerical calculations in this chapter are based on the score distributions corresponding to an unnamed financial company in Europe. The examined data set consisted of two columns. The first one was score, representing a transformed estimate of probability of being good client, and the second was indicator of being good. Following Table 3 and Figure 10 give some basic characteristics.

Table 3: Basic Characteristics

Mg	Mb	M	Sg	Sb	S
2.9124	2.2309	2.8385	0.7931	0.7692	0.7906

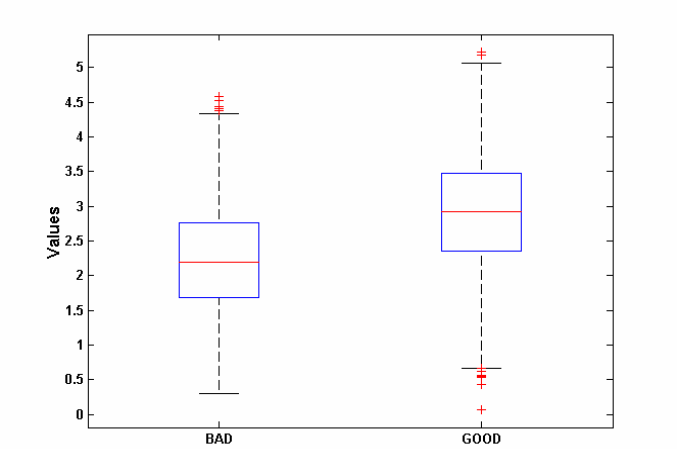

Figure 10: Box plot of Scores of Good and Bad Clients

On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

Because we want to use the results for normally distributed scores, we need to test hypothesis that data comes from a distribution in the normal family. Using Lilliefors test (Lilliefors 1967) at the 5% significance level, the hypothesis of normality of both scores is rejected. On the other hand, when we took all possible random subsamples of length 100 for each score group and saved the test result, in around 94% cases the test confirmed normality (in case of bad clients score and good clients score as well). Hence we will assume that given scores are normally distributed. Furthermore we need to test that standard deviations σ_g, σ_b are equal.

Using F-test at the 5% significance level, this hypothesis is not rejected. More precisely, the p-value is equal to 0.186, value of test statistics equals to 1.063 and the 95% confidence interval for the true variance ratio is [0.912, 1.231].
The first insight into discriminatory power of the score we obtain using the graph of cumulative distribution functions of bad and good clients, see Figure 11. KS statistics, derived from this figure, is equal to 0.3394. The value of score, in which it is achieved, equals to 2.556. Using result for normally distributed data we have, that KS equals to 0.3335.

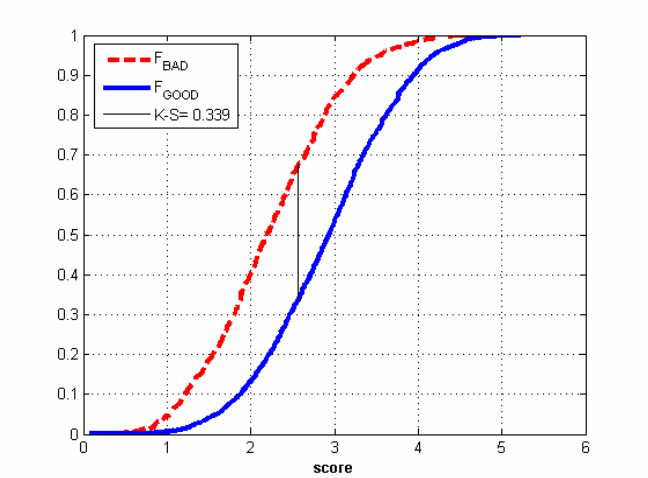

Figure 11: Cumulative Distribution Functions

The next Figure 12 shows Lorenz curve computed from our data set. It can be seen, that for example by rejection of 20% of good clients we reject 50% of bad clients at the same moment. The Gini index is equal to 0.4623 and c-statistics equals to 0.7311.

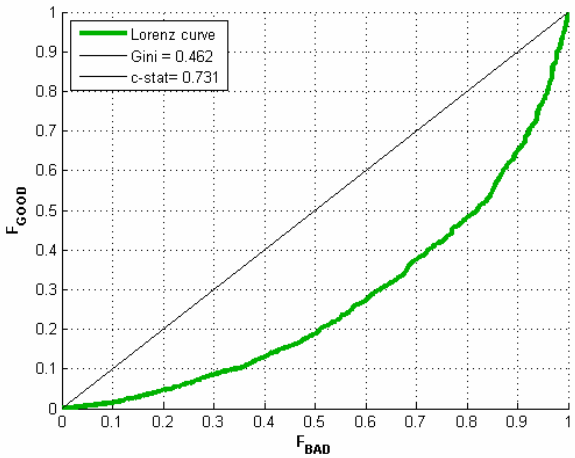

Figure 12: Lorenz Curve

The last mentioned indicator of scoring model quality based on distribution function is the Lift. The following Table 4 contains values of absolute and cumulative Lift corresponding to selected points on rejection scale. It is common to focus on the cumulative Lift value at 10% on this scale. In our case it is 2.80, which means that the given scoring model is 2.8 times better than random model at this level of rejection. Values in the last row of the table are computed by assumption of normality. The following Figure 13 shows Lift values on the whole rejection scale.

Table 4: Absolute and Cumulative Lift

	% rejected (F_{ALL})										
	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Abs. Lift	3,36	2,25	1,88	1,20	1,00	0,97	0,70	0,58	0,39	0,25	0,22
Cum.Lift	3,36	2,80	2,34	1,96	1,72	1,57	1,43	1,31	1,19	1,09	1,00
Cum.Lift_norm	3,67	2,95	2,30	1,95	1,72	1,54	1,40	1,28	1,18	1,09	1,00

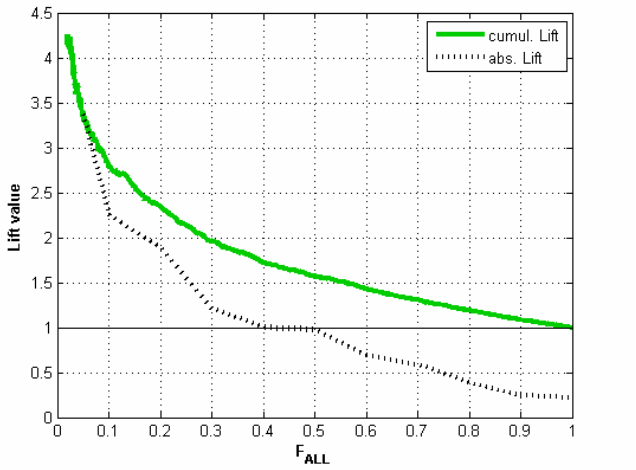

Figure 13: Absolute and Cumulative Lift

Estimations of densities of bad and good clients are shown in next Figure 14. Thick lines represent kernel estimations with bandwidth based on maximal smoothing principle. Thin lines are densities of normally distributed scores with parameters equals to M_b, S_b and S_g, M_g respectively. It can be seen that, in both cases, the cross point of densities of bad and good clients is approximately equal to 2.56, which is the value of score where the KS is achieved.

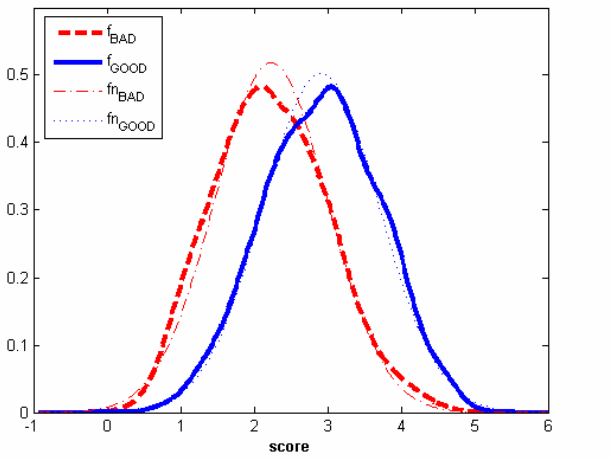

Figure 14: Density functions

The mean difference *D* is equal to 0.8620. Next Figures 15 and 16 show shapes of curves f_{diff}, f_{LR} and $f_{diff} \cdot f_{LR}$, which are used for computation of information statistics. First one is based on kernel estimation of density functions using maximal smoothing principle. Figure 16 is based on normal estimation of densities. Empirically computed information statistics is equal

to 0.7436. It is 0.7431 using result for normally distributed data, i.e. $I_{val} = D^2$. When curves $f_{diff} \cdot f_{LR}$ are used, i.e. I_{val} is computed numerically as the area under this curves, then the results are 0.7109 (in case of kernel estimate) and 0.7633 (in case of normal estimate).

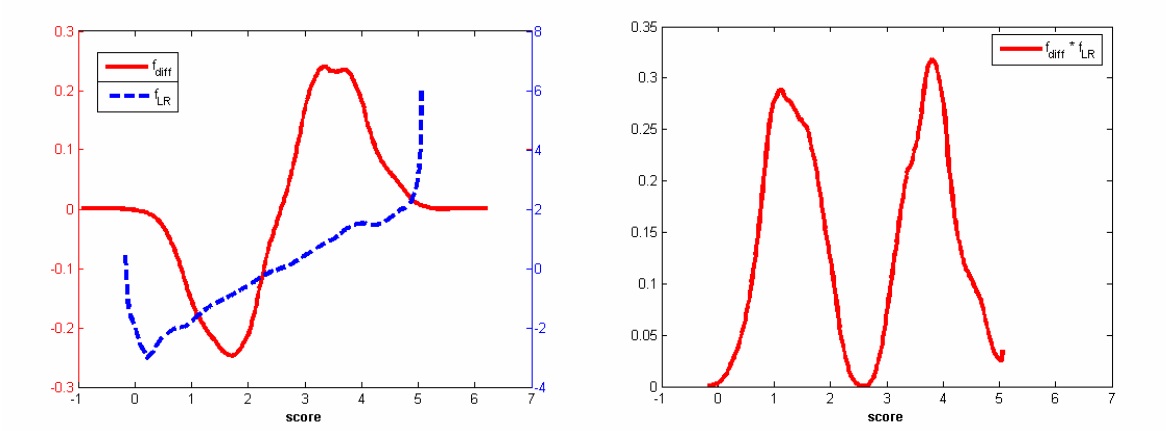

Figure 15: Kernel based f_{diff}, f_{LR} and $f_{diff} \cdot f_{LR}$

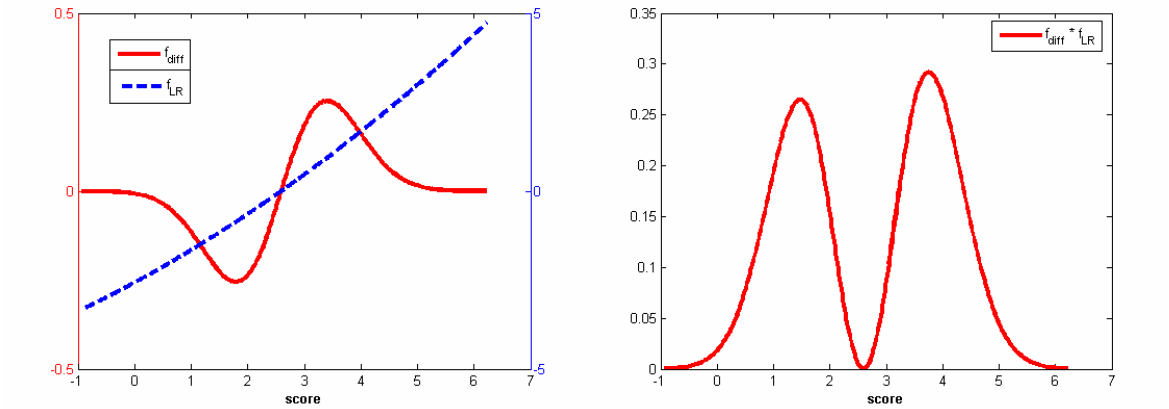

Figure 16: f_{diff}, f_{LR} and $f_{diff} \cdot f_{LR}$ based on normally distributed densities

5. Conclusions

All referred quantitative indicators (indexes) can be successfully used to measure the quality of different models (credit scoring functions) or their individual parts (predictors entering the models). They can be used as benchmarks for comparison of several proposed models at the time of model development. However, it is necessary to correctly interpret the received values of these indicators. We probably receive very different results if we use good / indeterminate / bad definition for development of a model and then we measure the quality of model only for good/bad clients or we measure it on the whole sample (where indeterminates are marked as good).

It should also be remembered that although the developed model can achieve excellent results, its actual quality is shown through time, i.e. after its deployment in practice. For this reason, it is therefore necessary to regularly monitor performance of the model (process). Once the performance of the model falls below given threshold the model has to be redeveloped.

It is obvious that we need to have the best performance of given scoring model near by expected cutoff value. Hence we should judge quality indexes from this point of view. Gini index is global measure, hence it is impossible to use it for assessment of local quality. The same holds for mean difference *D*. The KS is ideal if the expected cutoff value is near that point where KS is realized. Although the information statistics is global measure of model's quality, it is possible to use graphs of f_{diff}, f_{LR} and graph of their product to examine local properties of given model. Especially we can focus on region of scores where the cutoff is expected. The Lift seems to be the best choice for our purpose.

The application study shows that indexes based on assumptions of normality and empirical indexes take quite similar values. However, in general, one should be careful and a confirmation of all assumptions is needed. Anyway, presented results can help to understand how mentioned indexes behave.

References

[1] Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press , Oxford.

[2] Crook, J.N., Edelman, D.B., Thomas, L.C. (2007). 'Recent developments in consumer credit risk assessment', *European Journal of Operational Research* 183 (3), 1447-1465.

[3] Coppock, D.S. (2002). 'Why Lift?', *DM Review Online* [online], [cit. 2009-01-15] <http://www.dmreview.com/news/5329-1.html>.

[4] Giudici, P. (2003). Applied Data Mining: statistical methods for business and industry, Wiley, Chichester.

[5] Hand, D.J. and Henley, W.E. (1997). 'Statistical Classification Methods in Consumer Credit Scoring: a review'. *Journal. of the Royal Statistical Society*, Series A, 160,No.3:523-541.

[6] Lilliefors, H.W. (1967). 'On the Komogorov-Smirnov test for normality with mean and variance unknown'*, Journal of the American Statistical Association*, 62:399-402.

[7] Řezáč, M. (2003). 'Maximal Smoothing', *Journal of Electrical Engineering*, 54:44-46.

[8] Siddiqi, N. (2006). *Credit Risk Scorecards: developing and implementing intelligent credit scoring*, Wiley, New Jersey.

[9] Thomas, L.C., Edelman, D.B., Crook, J.N. (2002). *Credit Scoring and Its Applications*, SIAM Monographs on Mathematical Modeling and Computation, Philadelphia.

[10] Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*, Chapman and Hall, London.

[11] Wilkie, A.D. (2004). 'Measures for comparing scoring systems', In *Readings in Credit Scoring* 2004, Thomas, L.C., Edelman, D.B., Crook, J.N. (Eds.). Oxford University Press, Oxford, 51-62.

[12] Xu, K. (2003). '*How has the literature on Gini's index evolved in past 80 years?'* [online], [cit. 2009-01-15]. <http://economics.dal.ca/RePEc/dal/wparch/howgini.pdf>.