JEL Classification: C10, C53, D81, G32
Keywords: credit scoring, quality indices, lift, profit, normally distributed scores

How to Measure the Quality of Credit Scoring Models*

Martin ŘEZÁČ – Masaryk University, Brno, Czech Republic (mrezac@math.muni.cz)
corresponding author
František ŘEZÁČ – Masaryk University, Brno, Czech Republic (rezac@econ.muni.cz)

Abstract
Credit scoring models are widely used to predict the probability of client default. To measure the quality of such scoring models it is possible to use quantitative indices such as the Gini index, Kolmogorov-Smirnov statistics (KS), Lift, the Mahalanobis distance, and information statistics. This paper reviews and illustrates the use of these indices in practice.

1. Introduction

Banks and other financial institutions receive thousands of credit applications every day (in the case of consumer credit 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 the credit reliability of individuals who ask for credit. The assessment of the risk associated with the granting of credit is underpinned by one of the most successful applications of statistics and operations research: credit scoring.

Credit scoring is a set of predictive models and their underlying techniques that aid financial institutions in the granting of credit. These techniques decide who will get credit, how much credit they should get, and what further strategies will enhance the profitability of borrowers to lenders. Credit scoring techniques assess the risk in lending to a particular client. They do not identify "good" applications and "bad" applications (where negative behavior, e.g., default, is expected) on an individual basis, but they forecast the probability that an applicant with any given score will be "good" or "bad". These probabilities or scores, along with other business considerations, such as expected approval rates, profit, churn, and losses, are then used as a basis for decision making.

Several modeling methods for credit scoring have been introduced during the last six decades. The best known and most widely used are logistic regression, classification trees, the linear programming approach, and neural networks. See Hand and Henley (1997) or Vojtek and Kočenda (2006) for more details.

It is impossible to use a scoring model effectively without knowing how accurate it is. First, one needs to select the best model according to some measure of quality at the time of development. Second, one needs to monitor the quality of the model after its deployment into real business. The methodology of credit scoring models and some measures of their quality have been discussed in surveys conducted by Hand and Henley (1997), Thomas (2000), and Crook at al. (2007). However, until just ten years ago, the general literature devoted to the issue of credit scoring was not

* We thank Martin Fukač and our colleagues for their valuable comments, and our departments (Dpt. of Mathematics and Statistics, and Dpt. of Finance) for supporting our research.

substantial. Fortunately, the situation has improved in the last decade with the publication of works by Anderson (2007), Crook et al. (2007), Siddiqi (2006), Thomas et al. (2002), and Thomas (2009), all of which address the topic of credit scoring.

Nevertheless, despite the existence of several recent books and various articles in scientific journals, there is no comprehensive work devoted to the assessment of the quality of credit scoring models in all their complexity. Because of this, we decided to summarize and extend the known results in this area. We begin with the definition of good/bad clients, then consider each of the most popular indices and their expressions for normally distributed scores, generally with unequal variances of scores.

The most used indices in practice are the Gini index, which is most widely used in Europe, and the KS, which is most widely used in North America, despite the fact that their use may not be optimal. It is obvious that the best performance of a given scoring model needs to be near the expected cut-off value. Hence, we should judge quality indices from this point of view. The Gini index is a global measure; hence, it is impossible to use it for the assessment of local quality. The same holds for the mean difference D. The KS is ideal if the expected cut-off value is near the point where the KS is realized. Although information statistics are a global measure of a model's quality, we propose using graphs of f_{diff} and f_{LR} and the graph of their product to examine the local properties of a given model. In particular, we can focus on the region of scores where the cut-off is expected. Overall, Lift seems to be the best choice for our purpose. Since we propose to express Lift by means of the cumulative distribution functions of the scores of bad and all clients, it is possible to compute the value of Lift for any level of score.

In this paper, we aim to contribute to current practice by presenting a comprehensive, discursive overview of widely used techniques for assessing the quality of credit scoring models. Firstly, we discuss the definition of good/bad client, which is crucial for further computation. The result of a quality assessment process depends greatly on this definition. In the following section we review widely used quality indices, including their properties and mutual relationships, and bring some new theoretical results connected to them. Especially, it is the expression of Lift by means of the cumulative distribution functions of the scores of bad and all clients, and the expressions of selected indices for normally distributed data, namely, the Gini index and Lift in the case of the common variance of scores, and the mean difference D, the KS, the Gini index, Lift, and information statistics in the general case, i.e., without assuming equality of variances. The normality of scores has to be tested. On the other hand, it holds that by using logistic regression with categorical inputs transformed as the weight of evidence (a very common way of building up a credit scoring model) one obtains scores with distributions very close to the normal. And it is clear that once one can accept the assumption of normality of scores, computations of quality indices are much more accurate than in the case of using empirical estimates. Finally, applications of all the listed quality indices, including appropriate computational issues, are illustrated in a case study based on real financial data.

2. Definition of Good/Bad Client

In fact, the most important step in predictive model building is establishing the correct definition of dependent variable. In the 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 the case of the past due amount. This means that we need to define what is considered as debt and what is not. It may be that the client delays payment innocently (because of technical imperfections in the system). Also, it makes little sense to regard a small amount past due (e.g., less than €3) as debt. In addition, it is necessary to determine the time horizon along which the previous two characteristics are tracked. For example, a client is marked "good" if he has less than 60 DPD (with a tolerance of €3) in the 6 months from the first due date, or if he has less than 90 DPD (with a tolerance of €1) ever. The choice of these parameters depends greatly on the type of financial product (there would certainly be different parameters for small consumer loans with original maturities of around one year, on the one hand, and for mortgages, which are typically connected to very large amounts with maturities of up to several tens of years, on the other) and on the further use of this definition (credit scoring, fraud prevention, marketing,...). Another practical issue with respect to the definition of good client is the accumulation of several agreements. In this case, all amounts past due connected with the client at one particular point in time are usually added together and the maximum value of days past due is taken.

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 have been discussed. The third type of client is on the borderline between good and bad clients, and directly affects their definition. If we consider only DPD, clients with a high DPD (e.g., 90+) are typically identified as bad, while clients who are not delinquent (e.g., their DPDs are less than 30 or equal to zero) are identified as good. Clients are considered to be indeterminate if they are delinquent but have not exceeded the given DPD threshold. When we use this type of client, then we model very good clients against very bad ones. The result is that we obtain a model with amazing predictive power. However, this power dives immediately after assessing the model in the context of the whole population, where indeterminates are considered to be good. Thus, the use of this type of client is highly suspect and usually does not lead to any improvement in a model's quality. The next type of client is typically a client with a very short credit history, which makes correct definition of the dependent variable (good/bad client) all but impossible. The excluded clients are typically clients with significantly misleading data (e.g., fraudsters). They are also marked as "hard bad". The second group of excluded clients consists of applicants who belong to a category that will not be assessed by a model (scorecard), e.g., VIPs. The meaning of "rejected client" is obvious. See Anderson (2007), Thomas et al. (2002) or Thomas (2009) for more details.

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.

3. Measuring Quality

Once the definition of good/bad client and the client's score are 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 can evaluate the quality of this model. We can consider two basic types of quality indices: first, indices based on cumulative distribution functions such as Kolmogorov-Smirnov statistics, the Gini index and Lift; second, indices based on likelihood density functions such as mean difference (Mahalanobis distance) and informational statistics. For further available measures and appropriate remarks see Wilkie (2004), Giudici (2003) or Siddiqi (2006).

3.1. Indices Based on Distribution Function

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

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

The empirical cumulative distribution functions (CDFs) of the scores of good (bad) clients are given by the relationships

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

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

where s_i is the score of the i^{th} client, n is the number of good clients, m is the number of bad clients, and I is the indicator function, where I(true) = 1 and I(false) = 0. L is the minimum value of a given score, H is the maximum value. We denote the proportion of bad clients by $p_B = \frac{m}{n+m}$ and the proportion of good clients by $p_G = \frac{n}{n+m}$. The empirical distribution function of the scores of all clients is given by

$$F_{N.ALL}(a) = \frac{1}{N}\sum_{i=1}^{N} I\left(s_i \le a\right) \qquad a \in \left[L, H\right] \tag{2}$$

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

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

$$KS = \max_{a \in [L,H]} \left| F_{m,BAD}(a) - F_{n,GOOD}(a) \right| \tag{3}$$

Figure 1 Distribution Functions, KS

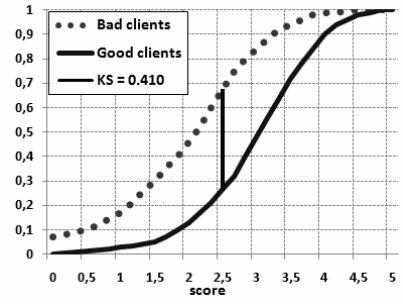


Figure 2 Lorenz Curve, Gini index

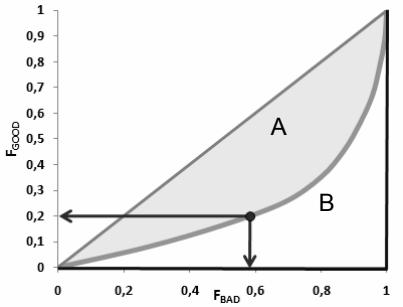


Figure 1 gives an example of the estimation of distribution functions for good and bad clients, including an estimate of the KS statistics. It can be seen, for example, that a score of around 2.5 or smaller has a population including approximately 30% of good clients and 70% of bad clients.

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

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

The definition and name (LC) is consistent with Müller and Rönz (2000). One can find the same definition of the curve, but called the ROC, in Thomas et al. (2002). Siddiqi (2006) used the name ROC for a curve with reversed axes and LC for a curve with the CDF of bad clients on the vertical axis and the CDF of all clients on the horizontal axis. For a short summary of currently used credit scoring methods and the quality testing thereof by using the ROC on real data with interpretations, see Kočenda and Vojtek (2011).

Each point on the curve represents some value of a given score. If we assume this value to be the cut-off value, we can read the proportion of rejected bad and good clients. An example of a Lorenz curve is given in *Figure 2*. We can see that by rejecting 20% of good clients, we reject almost 60% of bad clients at the same time.

In connection with the LC, we will now consider the next quality measure, the Gini index. This index describes the global quality of a scoring function. It takes values between -1 and 1. The ideal model, i.e., a scoring function that perfectly separates good and bad clients, has a Gini index equal to 1. On the other hand, a model that assigns a random score to the client has a Gini index equal to 0. Negative values correspond to a model with reversed meanings of scores. Using *Figure 2* the Gini index can be defined as

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

The actual calculation of the Gini index can be made using

$$Gini = 1 - \sum_{k=2}^{n+m} \left[\left(F_{m.BAD_k} - F_{m.BAD_{k-1}} \right) \cdot \left(F_{n.GOOD_k} + F_{n.GOOD_{k-1}} \right) \right] \tag{4}$$

where $F_{m.BAD_k}$ ($F_{n.GOOD_k}$) is the k^{th} vector value of the empirical distribution function of bad (good) clients. For further details see Thomas et al. (2002), Siddiqi (2006) or Xu (2003).

The Gini index is a special case of Somers' D (Somers, 1962), which is an ordinal association measure defined in general as $D_{YX} = \frac{\tau_{XY}}{\tau_{XX}}$, where τ_{XY} is Kendall's τ_a defined as $\tau_{XY} = E\left[sign\left(X_1 - X_2\right) sign\left(Y_1 - Y_2\right) \right]$, where $\left(X_1, Y_1\right)$, $\left(X_2, Y_2\right)$ are bivariate random variables sampled independently from the same population, and $E\left[\cdot\right]$ denotes expectation. In our case, $X = 1$ if the client is good and $X = 0$ if the client is bad. Variable Y represents the scores. It can be found in Thomas (2009) that the Somers' D assessing the performance of a given credit scoring model, denoted as D_S, can be calculated as

$$D_S = \frac{\sum_i g_i \sum_{j<i} b_j - \sum_i g_i \sum_{j>i} b_j}{n \cdot m} \tag{5}$$

where g_i (b_j) is the number of "goods" ("bads") in the i^{th} interval of scores. Furthermore, it holds that D_S can be expressed by the Mann-Whitney U-statistic in the following way. Order the sample in increasing order of score and sum the ranks of goods in the sequence. Let this sum be R_G. D_S is then given by $D_S = 2\frac{U}{n \cdot m} - 1$, where U is given by $U = R_G - \frac{1}{2} n\left(n+1\right)$. Further details can be found in Nelsen (1998).

Another type of quality assessment figure available is the CAP (Cumulative Accuracy Profile). Other names used for this concept are the Lift chart, the Lift curve, the Power curve, and the Dubbed curve. See Sobehart et al. (2000) or Thomas (2009) for more details.

In the case of the CAP 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 advantage of this figure is that one can easily read the proportion of rejected bads vs. the proportion of all rejected. It is called a Gains chart in a marketing context (see Berry and Linoff, 2004). In this case, the horizontal axis represents the proportion of clients who can be addressed by some marketing offer and the vertical axis represents the proportion of clients who will accept the offer.

When we use the CAP instead of the LC, we can define the Accuracy Rate (AR) (see Thomas, 2009). Again, it is defined by the ratio of some areas. We have

$$AR = \frac{\text{Area between CAP curve and diagonal}}{\text{Area between ideal model's CAP and diagonal}} = \\ = \frac{\text{Area between CAP curve and diagonal}}{0.5 \cdot (1 - p_B)}$$

Although the ROC and the CAP are not equivalent, it is true that the Gini index and the AR are equal for any scoring model. The proof for discrete scores is given in Engelmann et al. (2003); that for continuous scores can be found in Thomas (2009).

In connection with the Gini index, the c-statistic (Siddiqi, 2006) is defined as

$$c\_stat = \frac{1 + Gini}{2} \tag{6}$$

It represents the likelihood that a randomly selected good client has a higher score than a randomly selected bad client, i.e., $c\_stat = P\left(s_1 \geq s_2 \;\middle|\; D_{K_1} = 1 \wedge D_{K_2} = 0\right)$. It takes values from 0.5, for the random model, to 1, for the ideal model. Other names, such as Harrell's c (Harrell et al., 1996; Newson, 2006), AUROC (Thomas, 2009) or AUC (Engelmann et al., 2003), can be found in the literature.

Another possible indicator of the quality of a scoring model is *cumulative Lift*, which states how many times, at a given level of rejection, the scoring model is better than random selection (the random model). More precisely, it indicates the ratio of the proportion of bad clients with a score of less than a, $a \in [L,H]$, to the proportion of bad clients in the general population. In practice, the calculation is done for Lift corresponding to 10%, 20%,..., 100% of clients with the worst score (see Coppock, 2002). One of the main contributions of this paper is our proposal to express Lift by means of the cumulative distribution functions of the scores of bad and all clients (expressions (7) and (8)). We define Lift as

$$Lift(a) = \frac{F_{n.BAD}(a)}{F_{N.ALL}(a)} \qquad a \in [L,H] \tag{7}$$

In connection with Coppock's approach, we define

$$Lift_q = \frac{F_{n.BAD}\left(F^{-1}_{N.ALL}(q)\right)}{F_{N.ALL}\left(F^{-1}_{N.ALL}(q)\right)} = \frac{1}{q} F_{n.BAD}\left(F^{-1}_{N.ALL}(q)\right) \tag{8}$$

where q represents the score level of $100q\%$ of the worst scores and $F^{-1}_{N.ALL}(q)$ can be computed as $F^{-1}_{N.ALL}(q) = min\{a \in [L,H],\, F_{N.ALL}(a) \geq q\}$. Since the expected rejection rate is usually between 5% and 20%, q is typically assumed to be equal to 0.1 (10%), i.e., we are interested in the discriminatory power of a scoring model at the 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. Indices Based on Density Function

Let M_g and M_b be the means of the scores of good (bad) clients and S_g and S_b be the standard deviations of good (bad) clients. Let S be the pooled standard deviation of good and bad clients, given by $S = \left(\frac{nS_g^{\,2} + mS_b^{\,2}}{n+m}\right)^{\frac{1}{2}}$. Estimates of the mean and standard deviation of the scores for all clients $(\mu_{ALL}, \sigma_{ALL})$ are given by

$$M = M_{ALL} = \frac{nM_g + mM_b}{n+m}$$

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

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

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

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

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

$$I_{val} = \int_{-\infty}^{\infty} \left(f_{GOOD}(x) - f_{BAD}(x)\right) ln\left(\frac{f_{GOOD}(x)}{f_{BAD}(x)}\right) dx \tag{9}$$

We propose to examine the decomposed form of the right-hand side of the 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 statistic is a global measure of a model's quality, one can use graphs of f_{diff} and f_{LR} and the graph of their product to examine the local properties of a given model (see section 4 for more details).

We have two basic ways of computing the value of this index. The first way is to create bins of scores and compute it empirically from a table with the counts of good and bad clients in these bins. The second way is to estimate unknown densities using kernel smoothing theory. Consequently, we compute the integral by a suitable numerical method.

Let's have m score values $s_{0,i}$, $i = 1,...,m$ for bad clients and n score values $s_{1,j}$, $j = 1,...,n$ for good clients and recall that L denotes the minimum of all values and H the maximum. Let's divide the interval $[L,H]$ into 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

$$n_{0,k} = \sum_{i=1}^{m} I\left(s_{0,i} \in (q_{k-1}, q_k]\right), \qquad n_{1,k} = \sum_{j=1}^{n} I\left(s_{1,j} \in (q_{k-1}, q_k]\right), \quad k = 1,...,r$$

as the observed counts of bad and good in each interval. Then, the empirical information value is calculated by

$$I_{val} = \sum_{k=1}^{r} \left(\frac{n_{1,k}}{n} - \frac{n_{0,k}}{m} \right) ln\left(\frac{n_{1,k} \cdot m}{n_{0,k} \cdot n} \right) \tag{10}$$

Choosing the number of intervals is also very important. In the literature and also in many applications in credit scoring, the value $r = 10$ is preferred. An advanced algorithm for interval selection can be found in Řezáč (2011).

Another way of computing this index is by estimating appropriate densities using kernel estimations (Wand and Jones, 1995). Consider $f_{GOOD}(x)$ and $f_{BAD}(x)$ to be the likelihood density functions of the scores of good and bad clients, respectively. The kernel density estimates are defined by

$$\tilde{f}_{GOOD}(x, h_1) = \sum_{j=1}^{n} \frac{1}{n} K_{h_1}\left(x - s_{1,j}\right)$$

$$\tilde{f}_{BAD}(x, h_0) = \sum_{i=1}^{m} \frac{1}{n} K_{h_0}\left(x - s_{0,i}\right)$$

where $K_{h_i}(x) = \frac{1}{h_i} K\left(\frac{x}{h_i}\right)$, $i = 0,1$, and K is some kernel function, e.g., the Epanechnikov kernel. Bandwidth h_i can be estimated by the maximal smoothing principal (see Terrel, 1990, or Řezáč, 2003) or by cross-validation techniques (see Wand and Jones, 1995).

As the next step, we need to estimate the final integral. We use the composite trapezoidal rule. Set

$$\tilde{f}_{IV}(x) = \left(\tilde{f}_{GOOD}\left(x, h_1\right) - \tilde{f}_{BAD}\left(x, h_0\right)\right) \cdot ln\left(\frac{\tilde{f}_{GOOD}\left(x, h_1\right)}{\tilde{f}_{BAD}\left(x, h_0\right)} \right)$$

Then, for given $M+1$ equidistant points $L = x_0, ..., x_M = H$ we obtain

$$I_{val} = \frac{H - L}{2M} \left(\tilde{f}_{IV}(L) + 2 \sum_{i=1}^{M-1} \tilde{f}_{IV}\left(x_i\right) + \tilde{f}_{IV}(H) \right) \tag{11}$$

The value of M is usually set between 100 and 1,000. As one has to trade off between computational speed and accuracy, we propose using $M = 500$. For further details see Koláček and Řezáč (2010).

3.3. 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 the F-test.

The mean difference D (see Wilkie, 2004) is now defined as $D = \frac{\mu_g - \mu_b}{\sigma}$ and is calculated by

$$D = \frac{M_g - M_b}{S} \tag{12}$$

The maximum difference between the cumulative distributions, denoted KS before, is calculated, as proposed in Wilkie (2004), at the point where the distributions cross, halfway between the means. The KS value is therefore given by

$$KS = \Phi\left(\frac{D}{2}\right) - \Phi\left(\frac{-D}{2}\right) = 2 \cdot \Phi\left(\frac{D}{2}\right) - 1 \tag{13}$$

where $\Phi(\cdot)$ is the standardized normal distribution function. We derived a formula for the Gini index. It can be expressed by

$$G = 2 \cdot \Phi\left(\frac{D}{\sqrt{2}}\right) - 1 \tag{14}$$

The computation for Lift statistics is quite easy. Denoting $\Phi^{-1}(\cdot)$ as the standard normal quantile function we have

$$Lift_q = \frac{1}{q}\Phi\left(\frac{S_{ALL}}{S}\Phi^{-1}(q) + p_G \cdot D\right) \tag{15}$$

This expression (15), as well as (14), is specific to this paper. A couple of further interesting results are given in Wilkie (2004). One of them is that, under our assumptions concerning the normality and equality of standard deviations, it holds that

$$I_{val} = D^2 \tag{16}$$

We derived expressions for all the mentioned indices in the general case, i.e., without assuming equality of variances. This means that the following expressions (17) to (21) are specific to this paper and cannot be found in the literature. The mean difference is now in the form

$$D = \sqrt{2}\, D^* \tag{17}$$

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

The empirical form of KS can be expressed by

$$KS = \Phi\left(\frac{\sqrt{S_b^2 + S_g^2}}{S_b^2 - S_g^2} S_b \cdot D^* - \frac{1}{S_b^2 - S_g^2} S_g \cdot \sqrt{\left(S_b^2 + S_g^2\right) D^{*2} + 2 \cdot \left(S_b^2 - S_g^2\right) ln\left(\frac{S_g}{S_b}\right)}\right) - \Phi\left(\frac{\sqrt{S_b^2 + S_g^2}}{S_b^2 - S_g^2} S_g \cdot D^* - \frac{1}{S_b^2 - S_g^2} S_b \cdot \sqrt{\left(S_b^2 + S_g^2\right) D^{*2} + 2 \cdot \left(S_b^2 - S_g^2\right) \ln\left(\frac{S_g}{S_b}\right)}\right) \tag{18}$$

The Gini coefficient can be expressed as

$$G = 2 \cdot \Phi\left(D^*\right) - 1 \tag{19}$$

Lift is given by the formula

$$Lift_q = \frac{1}{q} \Phi\left(\frac{S_{ALL} \cdot \Phi^{-1}(q) + M - M_b}{S_b}\right) \tag{20}$$

Finally, the information statistic is given by

$$I_{val} = \left(A + 1\right) D^{*2} + A - 1 \tag{21}$$

where $A = \dfrac{1}{2}\left(\dfrac{\sigma_g^2}{\sigma_b^2} + \dfrac{\sigma_b^2}{\sigma_g^2}\right)$; in computational form it is $A = \dfrac{1}{2}\left(\dfrac{S_g^2}{S_b^2} + \dfrac{S_b^2}{S_g^2}\right)$. For this index, one can find a similar formula in Thomas (2009). To explore the behavior of expressions (12)–(21) it is possible to use the tools offered by the Maple system. See Hřebíček and Řezáč (2008) for more details. Further comments on the behavior of the listed indices can be found in Řezáč and Řezáč (2009).

4. Case Study

Applications of all the listed quality indices, including appropriate computational issues, are illustrated in this case study. Based on real financial data, we aim to provide computations with a commentary and to note what might be computational issues and what might be appropriate interpretation of the results obtained. First, we describe our data, including basic statistical characteristics and box plots. Then we test the normality of our data (Q-Q plot, Lilliefors test for subsamples) and the equality of the standard deviation (*F*-test). After that we provide figures and indices based

Figure 3 Box Plot of Scores of Good and Bad Clients

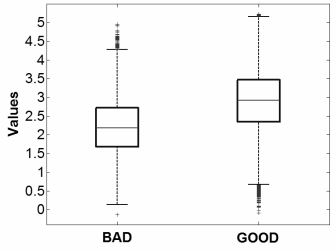


Table 1 Basic Characteristics

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

on the cumulative distribution function, i.e., the CDF, the Lorenz curve, the CAP, the KS, the Gini index and Lift. Subsequently, we estimate the likelihood densities, compute the mean difference and information statistics, and discuss the curves which are used for the computation of the information statistics. Finally, we focus on the profit that a firm can make. We estimate this profit according to the quality indices obtained and according to a set of portfolio parameters.

All numerical calculations in this section are based on scoring data provided by a financial company operating in Central and Eastern Europe[1] providing small- and medium-sized consumer loans. Data were registered between January 2004 and December 2005. To preserve confidentiality, the data were selected in such a way as to provide heavy distortion in the parameters describing the true solvency situation of the financial company. The examined data set consisted of 176,878 cases with two columns. The first one showed a score (the outcome of the application of the credit scoring model based on logistic regression) representing a transformed estimate of the probability of being a good client, and the second showed an indicator of whether a client was "good" or "bad" (defaulted 90 DPD ever). The number of bad clients was 18,658, which means a 10.5% bad rate. *Table 1* and *Figure 3* give some basic characteristics.

In 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 wanted to use the results for normally distributed scores, we needed to test the hypothesis that data come from a distribution in the normal family. Q-Q plots (see *Figure 4*) show that the distributions of the given scores were quite similar to the normal one. With a very large sample size (which was our case), a normality test may detect statistically significant but unimportant deviations from normality. Unless the Q-Q plot indicates a source for the nonnormality, the normality test result may not be useful in this case. For that reason, we decided to take the Q-Q plots as proof of normality. Furthermore, when we took 10,000 random subsamples of length 100 for each score group and saved the results of the Lilliefors test (Lil-

[1] The financial institution does not wish to be explicitly identified and we respect this wish, as according to the contract to obtain the data.

Figure 4 Q-Q Plots for Scores of Good and Bad Clients

Figure 4a

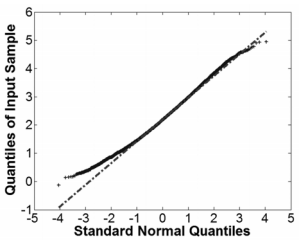


Figure 4b

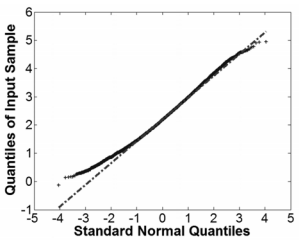
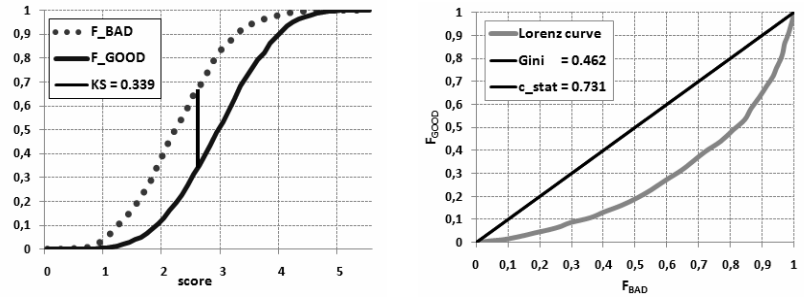


Figure 5

(a) Cumulative Distribution Functions **(b) Lorenz curve**



liefors, 1967) at the 5% significance level, in around 94% of cases the test confirmed normality (in the case of bad client scores and good client scores as well).

Furthermore we needed to test that the standard deviations σ_g, σ_b, are equal. Using the *F*-test at the 5% significance level, this hypothesis was not rejected. More precisely, the *p*-value was equal to 0.186, the value of the test statistic was equal to 1.063, and the 95% confidence interval for the true variance ratio was [0.912, 1.231]. According to this result, we could use the expressions (12) to (16), i.e., the expressions assuming equality of variances. Generally, if the *F*-test rejects the hypothesis, then one should use expressions (17) to (21).

We obtain the first insight into the discriminatory power of the score if we use the graph of the cumulative distribution functions of bad and good clients (see *Figure 5a*). The KS statistic derived from this figure was equal to 0.3394. The value of the score at which it was achieved was equal to 2.556. Using the results for normally distributed data, it turned out that the KS was equal to 0.3335.

Figure 5b shows the Lorenz curve computed from our data set. It can be seen, for example, that by rejecting 20% of good clients, we reject 50% of bad clients at the same time. The Gini index was equal to 0.4623 and the *c*-statistic was equal to 0.7311. Using the expression for normally distributed data, the Gini index was equal to 0.4578. The *c*-statistic was equal to 0.7289 in this case.

The CAP for our data is displayed in *Figure 6*. The ideal model is now represented by the polyline from [0, 0] through [p_B, 1] to [1, 1]. We can easily read the proportion of rejected bads vs. the proportion of all rejected. For example, we can see that if we want to reject 70% of bads, we have to reject about 40% of all applicants.

Figure 6 CAP

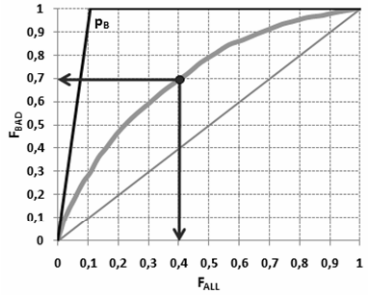


Table 2 Absolute and Cumulative Lift – Computational Scheme

decile	# clients	absolutely			cumulatively		
		# bad clients	Bad rate	abs. Lift	# bad clients	Bad rate	cum. Lift
1	17688	5233	29.6%	2.80	5233	29.6%	2.80
2	17688	3517	19.9%	1.88	8750	24.7%	2.34
3	17641	2239	12.7%	1.20	10989	20.7%	1.96
4	17735	1879	10.6%	1.00	12868	18.2%	1.72
5	17688	1810	10.2%	0.97	14678	16.6%	1.57
6	17688	1315	7.4%	0.70	15993	15.1%	1.43
7	17688	1077	6.1%	0.58	17070	13.8%	1.31
8	17687	723	4.1%	0.39	17793	12.6%	1.19
9	17688	461	2.6%	0.25	18254	11.5%	1.09
10	17687	404	2.3%	0.22	18658	10.5%	1.00
All	176878	18658	10.5%				

Table 3 Absolute and Cumulative Lift

	% rejected (F_{ALL})									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
abs. Lift	2.25	1.88	1.20	1.00	0.97	0.70	0.58	0.39	0.25	0.22
cum. Lift	2.80	2.34	1.96	1.72	1.57	1.43	1.31	1.19	1.09	1.00
cum.Lift_norm	2.90	2.30	1.97	1.74	1.56	1.42	1.29	1.19	1.09	1.00

The last-mentioned indicator of a scoring model's quality based on a distribution function was Lift. Let's demonstrate the procedure for computing this index. We sorted customers according to the score and split them into ten groups using deciles of the score. Then we counted the number of bad clients in each group, in our case around 17,688 clients. This yielded their proportion in the group (the Bad Rate). Absolute Lift in each group was then given by the ratio of the share of bad clients in the group to the proportion of bad clients in total. Cumulative Lift was given by the ratio of the proportion of bad clients in groups up to the given group to the proportion of bad clients in total. The results are presented in *Table 2.*

Table 3 contains values of absolute and cumulative Lift corresponding to selected points on the rejection scale. It is common to focus on the cumulative Lift value at 10% on this scale. In our case it was 2.80, which means that the given

Figure 7 Absolute and Cumulative Lift

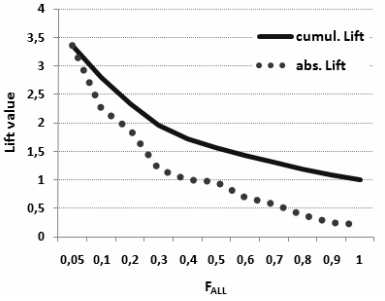
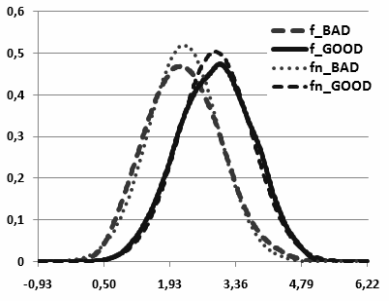


Figure 8 Density Functions



scoring model was 2.8 times better than the random model at this level of rejection. Values in the last row of the table are computed by assuming normality. *Figure 7* shows the Lift values on the whole rejection scale.

Estimates of the densities of bad and good clients are shown in *Figure 8*. The thick lines represent kernel estimations, with the bandwidth based on the maximal smoothing principle. The thin lines are the densities of the normally distributed scores with parameters equal to M_b, S_b, and S_g, M_g,, respectively. It can be seen that, in both cases, the intersection of the densities of bad and good clients was approximately equal to 2.56, which was the value of the score where the KS was achieved.

The mean difference D was equal to 0.8620. *Figures 9* and *10* show the shapes of curves f_{diff}, f_{LR}, and f_{diff}, f_{LR}, which are used for the computation of information statistics. The first one is based on the kernel estimation of density functions using the maximal smoothing principle. *Figure 10* is based on the parametrical estimation of densities assuming normality of scores.

We can see that curves f_{diff}, f_{LR} have three points of intersection. The middle point is the point of intersection of our densities, i.e., where $f_{GOOD} = f_{BAD}$ holds, and, further, it is the point where f_{diff}, f_{LR} is equal to zero. As we can see, the curve f_{diff}, f_{LR} has two peaks. Generally, when variances differ enough, we can have two "middle" points and three peaks, but this is not our case. Since the information statistic is the integral of f_{diff}, f_{LR}, we can easily examine the local properties of a model. It is obvious that the higher left peak of f_{diff}, f_{LR} means a stronger model for the area of low scores, i.e., the area of bad clients, and vice versa. In our case, we can see in *Figures 9* and *10* that our model is very slightly better for higher scores. However,

Figure 9 Kernel Based f_{diff}, f_{LR}, and f_{diff}, f_{LR}

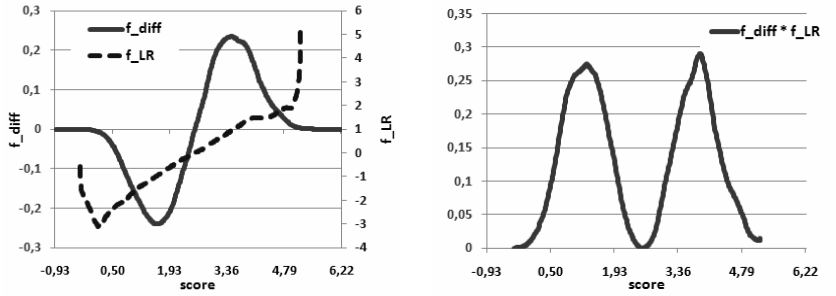


Figure 10 f_{diff}, f_{LR}, and f_{diff}, f_{LR} Based on Normally Distributed Densities

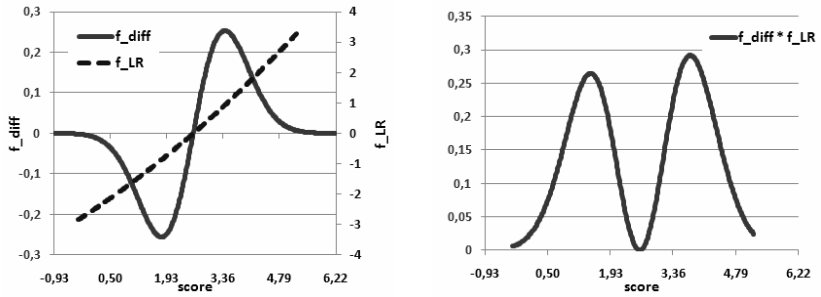


Table 4 Informational Statistics

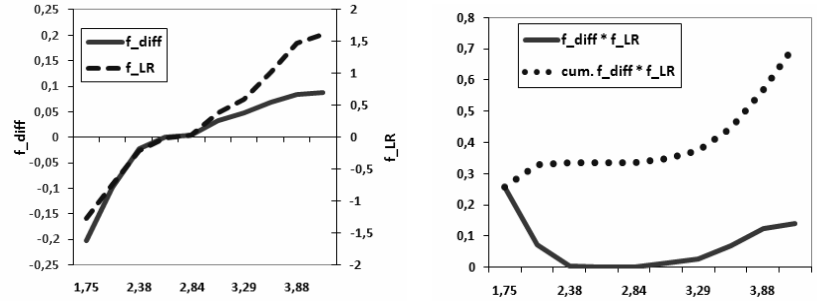
decile	# cleints	# bad clients	#good	% bad [1]	% good [2]	[3] = = [2] - [1]	[4] = = [2] / [1]	[5] = = ln[4]	[6] = = [3] * [5]
1	17688	5233	12455	28.0%	7.9%	-0.20	0.28	-1.27	0.26
2	17688	3517	14171	18.8%	9.0%	-0.10	0.48	-0.74	0.07
3	17641	2239	15402	12.0%	9.7%	-0.02	0.81	-0.21	0.00
4	17735	1879	15856	10.1%	10.0%	-0.00	1.00	-0.00	0.00
5	17688	1810	15878	9.7%	10.0%	0.00	1.03	0.03	0.00
6	17688	1315	16373	7.0%	10.3%	0.03	1.47	0.38	0.01
7	17688	1077	16611	5.8%	10.5%	0.05	1.82	0.60	0.03
8	17687	723	16964	3.9%	10.7%	0.07	2.77	1.02	0.07
9	17688	461	17227	2.5%	10.9%	0.08	4.41	1.48	0.12
10	17687	404	17283	2.2%	10.9%	0.09	5.04	1.62	0.14
All	176878	18658	158220					Info. Value	***0.7120***

the difference is so small that we can state that the model has practically the same strength for low and high scores.

Now we will demonstrate the computational scheme for informational statistics in the case of discretized data. We sorted customers according to score and split them into ten groups using deciles of the score. Then we counted the number of all, bad, and good clients in each group. The results are in *Table 4*.

The second, third, and fourth columns contain the counts of all, bad, and good clients. The next two columns, [1] and [2], contain the relative frequencies of bad and

Figure 11 f_{diff}, f_{LR}, and f_{diff}, f_{LR} Based on Empirical Computation



good clients in each score interval. The last four columns, [3] to [6], represent mathematical operations employed in (10). Adding the last column [6], we get the information value.

Computed empirically, with the 10 bins corresponding to the divisions using the deciles of the score, the information statistic was equal to 0.7120. It was 0.7431 using the result for normally distributed data, i.e., $I_{val} = D^2$. When the f_{diff} f_{LR} curve is used, i.e., when I_{val} is computed numerically as the area under this curve, then the result was 0.7109 (in the case of the kernel estimate) and 0.7633 (in the case of the parametrical estimate).

Parts a) and b) of *Figure 11* are the last figures we will examine. Similarly to the previous two figures, we are interested in f_{diff}, f_{LR} and f_{diff}, f_{LR}. *Figure 11a* is a graphical representation of the values in column [3] and [5] in *Table 4*. At first sight, the curves seem to be very different from the curves in *Figure 9*. As we can see, both curves are increasing. This is what we expect in the case of f_{LR}. The increase in this quantity means an increase in the natural logarithm of the ratio of the densities of the scores of good and bad clients. Because the natural logarithm is a continuous increasing function, it means an increase in the ratio itself. Indeed, it means that with increasing score we have an increasing proportion of good clients, which is our case. The increase of f_{diff} is once again what we expected. Again, it means that we have higher density values for bad clients for low scores, i.e., for low, but fixed, scores we have a higher probability of a client being bad compared to the probability of a client being good; and we have higher density values for good clients for high scores, i.e., for high, but fixed, scores we have a higher probability of a client being good compared to the probability of a client being bad. The main difference compared to *Figure 9* is the behavior near the boundaries of the scores, i.e., near the minimum and maximum values of the scores. The reason for this behavior is quite simple. Because the number of scores near these boundaries is very small, and because the empirical approach works with deciles, the first and last empirically computed value of f_{diff} is, in fact, an average which smooths the appropriate parts of f_{diff} in *Figure 9*. Finally, we must mention the fact that both curves intersect when the score is approximately equal to 2.56, which is the point at which the densities of the scores also cross.

Figure 11b is a graphical representation of column [6] in *Table 4*. We can see that the curve of f_{diff} f_{LR} touches zero when the score is around 2.5. In contrast to

Figure 9, we can also see that the curve of $f_{diff}\, f_{LR}$ has a *U* shape without falls to zero around the boundaries of the scores. Again, as for *Figure 11a*, it holds that the first and last empirically computed value of $f_{diff}\, f_{LR}$ is, in fact, an average which smooths the appropriate parts of $f_{diff}\, f_{LR}$ in *Figure 9*. Finally, the cumulative values of $f_{diff}\, f_{LR}$ are displayed in *Figure 11b*.

Now, it is natural to ask what is the financial impact of all these computations and indices. It is clear that one should have knowledge (or at least a feeling) about the financial impact to be able to decide what the worthiness of a credit scoring model is. To compute the profit which a firm can make, one would need to know the full financial data, especially the credit amount, the term, the interest rate, the recovery rate, the recovery costs, and the initial/fixed costs per credit. Despite the fact that we did not have these data, we estimated the profit using some simplifying assumptions. 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%. All of these correspond to data examined in the case study. 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.

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., $Lift_q$ given by (8), *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. We propose to compute the profit by

$$profit = \#\, proposals \cdot DR \cdot RR \cdot (Lift_{RR} - 1) \cdot gain \tag{22}$$

Table 5 contains estimates of the profit for the credit scoring model examined in this case study. Furthermore, it contains profits for other considered portfolio parameters. Moreover, a comparison of the values of the selected quality indices (with the assumption of normality and equality of variances) can be found there too. The bold row corresponds to the model from this case study, the three rows above correspond to models with worse quality, and the last three rows correspond to models with higher performance, i.e., higher values of the quality indices and higher profits.

Firstly, we can see in *Table 5* 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. Compared with the costs of developing a new credit scoring model, which are around 200 man days (including

Table 5 Saved Profit According to Quality Indices

Quality indices							
D	***KS***	***Gini***	***c_stat***	***Lift*$_{10\%}$**	***Lift*$_{20\%}$**	***Lift*$_{40\%}$**	***Ival***
0.2500	0.0995	0.1403	0.5702	1.4422	1.3376	1.2197	0.0625
0.5000	0.1974	0.2763	0.6382	1.9794	1.7156	1.4395	0.2500
0.7500	0.2923	0.4041	0.7021	2.5987	2.1187	1.6489	0.5625
0.8620	**0.3335**	**0.4578**	**0.7289**	**2.8977**	**2.3028**	**1.7370**	**0.7430**
1.0000	0.3829	0.5205	0.7602	3.2801	2.5294	1.8391	1.0000
1.2500	0.4680	0.6232	0.8116	3.9988	2.9304	2.0041	1.5625
1.5000	0.5467	0.7112	0.8556	4.7287	3.3068	2.1406	2.2500

Portfolio parameters (in €)					
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%**
Profit (in €)					
415 318	319 019	2 076 589	1 595 095	6 229 766	4 785 284
830 718	676 264	4 153 588	3 381 320	12 460 764	10 143 959
1 226 474	1 057 182	6 132 369	5 285 909	18 397 106	15 857 726
1 392 838	**1 231 152**	**6 964 189**	**6 155 762**	**20 892 566**	**18 467 285**
1 585 984	1 445 248	7 929 919	7 226 240	23 789 757	21 678 719
1 897 678	1 824 194	9 488 388	9 120 970	28 465 165	27 362 911
2 155 813	2 179 903	10 779 067	10 899 516	32 337 201	32 698 548

development, testing, and deployment), this represents about €20K (in Eastern Europe), so it is obvious that these expenses are in fact negligible and the resulting profit is really considerable.

Furthermore, one can compare the values of the expected profit within the columns as well as within the rows in *Table 5*. 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.

5. Conclusions

We considered DPD and time horizon to be the crucial parameters affecting the definition of good/bad client. Issues relating to indeterminate clients were also discussed. However, in the light of this discussion we concluded that this category

should not be used at all. Finally, the dependence of a scoring model's performance on the definition was discussed. On the basis of this discussion, we suggested using a definition which is as hard as possible, but also reasonable, e.g., 90 DPD ever, or 90 DPD on the first payment.

We derived a formula for the Lift curve based on the ratio of the cumulative distribution functions of the scores of bad and all clients. This allows the value of Lift for any given score to be calculated. On the other hand, it is much more useful to know the value of Lift corresponding to some quantile of a score. Because of this, we proposed the quantile form of Lift. Despite the high popularity of the Gini index and the KS, we conclude that Lift and figures of decomposed information statistics are more appropriate for assessing the local quality of a credit scoring model. In particular, it is better to use them in the case of an asymmetric Lorenz curve. Using the Gini index or KS during the development process could lead to the selection of a weaker model in this case.

Known formulas for mean difference, the KS and information statistics were supplemented with formulas for the Gini index and Lift in the case of normally distributed scores with common variance. Afterwards, we did not assume equality of standard deviations and derived expressions for all the mentioned indices in general. The behavior of these indices was illustrated in appropriate figures. We should realize that all the listed indices are estimations of appropriate random statistics, whose exact values are unknown. Despite the fact that the scores of credit scoring models are not usually exactly normally distributed, they are often very close to this distribution. In this case, we suggest using the expressions from this paper, because we obtain more accurate estimates.

The case study demonstrated the use of the discussed indices, highlighted some computational issues, and presented possible interpretations. It is clear that a financial institution has to use quality indices of credit scoring models that are as accurate as possible. In the case of monitoring or reporting usage, this can support more accurate decision making. In the case of selection of the credit scoring model to be deployed, the economic impacts are obvious. With a better model, a company can acquire less risky clients into its portfolio. Better business profitability is the direct consequence. But if one has a wrong estimate of the quality of a credit scoring model, then one can easily choose the wrong model to be deployed. In this case, more risky clients and lower profitability are the direct consequences. It is clear that empirical expressions of the quality indices can be used every time. But, if one can accept the assumption of normality of scores, then using expressions (12) to (16) results in more accurate estimates of the quality indices. Moreover, if the assumption of equality of the variances of scores cannot be proved, and the assumption of normality is proved, then expressions (17) to (21) should lead to the most accurate estimates of the quality indices. There is no general rule that tells us the right value of a given quality index. This value is specific to country, product, and type of scoring (application/behavioral/other). Very roughly, we can expect higher values of quality indices in emerging countries, for more risky products, and for behavioral models. Finally, it was shown how to estimate the financial impact of using a credit scoring model, and that the amount saved by a financial institution may be really significant.

Although it may seem that most issues relating to the measurement of the quality of credit scoring models have been resolved, many questions for further research

still remain. For instance, what is the effect of the type of inputs into a model (continuous, categorized, WOE,…) and the type of model (logistic regression, NN, Trees,…) on the degree of asymmetry of the Lorenz curve? How does the theoretical distribution of scores depend on the type of inputs and type of model? And what is the confidence band of the Lift curve? Last but not least, an important further goal of research in this area will be to generalize the expressions presented in section 3.4 so as to make them applicable to a more general class of distributions, such as generalized gamma or generalized beta distributions.

REFERENCES

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

Berry MJA, Linoff GS (2004): *Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management*. 2nd ed. Indianapolis, Wiley.

Coppock DS (2002): Why Lift? *DM Review Online*. www.dmreview.com/news/5329-1.html. Accessed on December 1, 2009.

Crook JN, Edelman DB, Thomas LC (2007): Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3):1447–1465.

Engelmann B, Hayden E, Tasche D (2003): *Measuring the Discriminatory Power of Rating System*. http://www.bundesbank.de/download/bankenaufsicht/dkp/200301dkp_b.pdf. Accessed on October 4, 2010.

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

Hand DJ, Henley WE (1997): Statistical Classification Methods in Consumer Credit Scoring: a review. *Journal. of the Royal Statistical Society*, Series A, 160(3):523–541.

Harrell FE, Lee KL, Mark DB (1996): Multivariate prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387.

Hřebíček J, Řezáč M (2008): Modelling with Maple and MapleSim. In: *22nd European Conference on Modelling nad Simulation ECMS 2008 Proceedings*, Dudweiler, pp. 60–66.

Kočenda E, Vojtek M (2011): Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data. Forthcoming in: *Emerging Markets Finance and Trade.*

Koláček J, Řezáč M (2010): Assessment of Scoring Models Using Information Value. In: *Keynote, Invited and Contributed Papers. 19th International Conference on Computational Statistics*, Paris, SpringerLink, pp. 1191—1198.

Lilliefors HW (1967): On the Komogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62:399–402.

Müller M, Rönz B (2000): Credit Scoring using Semiparametric Methods. In: Franke J, Härdle W, Stahl G (Eds.): *Measuring Risk in Complex Stochastic Systems*. New York, Springer-Verlag.

Nelsen RB (1998): Concordance and Gini's measure of association. *Journal of Nonparametric Statistics*, 9(3):227–238.

Newson R (2006): Confidence intervals for rank statistics: Somers' D and extensions. The *Stata Journal*, 6(3):309–334.

Řezáč M (2003): Maximal Smoothing. *Journal of Electrical Engineering*, 54:44–46.

Řezáč M (2011): Advanced Empirical Estimate of Information Value for Credit Scoring Models. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis,* LIX(2):267–273.

Řezáč M, Řezáč F (2009): *Measuring the Quality of Credit Scoring Models*. http://www.crc.man.ed.ac.uk/conference/archive/2009/presentations/Paper-65-Paper.pdf. Accessed on October 4, 2010.

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

Sobehart J, Keenan S, Stein R (2000): *Benchmarking Quantitative Default Risk Models: A Validation Methodology, Moody's Investors Service.* http://www.algorithmics.com/EN/media/pdfs/Algo-RA0301-ARQ-DefaultRiskModels.pdf. Accessed on October 4, 2010.

Somers RH (1962): A new asymmetric measure of association for ordinal variables. *American Sociological Revie*w, 27:799–811.

Terrell GR (1990): The Maximal Smoothing Principle in Density Estimation. *Journal of the American Statistical Association*, 85:470–477.

Thomas LC (2000): A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2):149–172 .

Thomas LC (2009): *Consumer Credit Models: Pricing, Profit, and Portfolio*. Oxford, Oxford University Press.

Thomas LC, Edelman DB, Crook JN (2002): Credit Scoring and Its Applications. Philadelphia, *SIAM Monographs on Mathematical Modeling and Computation*.

Vojtek M, Kočenda E (2006): Credit Scoring Methods. *Finance a úvěr-Czech Journal of Economics and Finance*, 56(3-4):152–167.

Wand MP, Jones MC (1995): *Kernel Smoothing*. London, Chapman and Hall.

Wilkie AD (2004): Measures for comparing scoring systems. In: Thomas LC, Edelman DB, Crook JN (Eds.): *Readings in Credit Scoring*. Oxford, Oxford University Press.

Witzany J (2009): *Definition of Default and Quality of Scoring Functions*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1467718. Accessed on September 1, 2010.

Xu K (2003): *How has the literature on Gini's index evolved in past 80 years?* http://economics.dal.ca/RePEc/dal/wparch/howgini.pdf. Accessed on December 1, 2009.