ISSN: 2281-1346

Department of Economics and Management

DEM Working Paper Series

Performance of credit risk prediction models via proper loss functions

Silvia Figini
(Università di Pavia)

Mario Maggi
(Università di Pavia)

 # 64 (01-14)

Via San Felice, 5
I-27100 Pavia
http://epmq.unipv.eu/site/home.html

January 2014

Performance of credit risk prediction models via proper loss functions *

Silvia Figini, Mario Maggi †

January 2014

Abstract

The performance of predictions models can be assessed using a variety of methods and metrics. Several new measures have recently been proposed that can be seen as refinements of discrimination measures, including variants of the AUC (Area Under the ROC curve), such as the H index. It is widely recognized that AUC suffers from lack of coherency especially when ROC curves cross. On the other hand, the H index requires subjective choices. In our opinion the problem of model comparison should be more adequately handled using a different approach. The main contribution of this paper is to evaluate the performance of prediction models using proper loss function. In order to compare how our approach works with respect to classical measures employed in model comparison, we propose a simulation studies, as well as a real application on credit risk data.
Keywords: Model Comparison, AUC, H index, Loss Function, Proper Scoring Rules, Credit Risk

1 Introduction and motivation

The performance of prediction models can be assessed using a variety of methods and metrics (see e.g., Hand, 2012). A large number of measures have been proposed to evaluate the performance of a classification rules. Some of these have been developed to meet the practical requirements of specific applications, but many others which here we call *classification accuracy criteria* represent different ways of balancing the different kinds of misclassification which may be made.

In this paper we focus on the assessment of prediction models for a dichotomous outcome. Following Hand (2012), the performance measures employed in this framework are classified into threshold-dependent criteria (sensitivity, specificity, positive predictive value, negative predictive value, probability of correct classification, error rate, kappa statistic, Youden Index, F-measure), threshold-independent criteria such as the Kolmogorov Smirnov test, and methods depending on all the possible thresholds in $[0, 1]$, such as the Area Under the *Receiver Operating Characteristic* (ROC) Curve (AUC), the Gini Index and the H measure (Hand, 2009).

The ROC curve describes the performance of a classification or diagnostic rule. This curve is generated by plotting the fraction of true positives out of the positives (true positive rate) versus the fraction of false positives out of the negatives (false positive rate), at all threshold in $[0, 1]$. However, comparing curves directly has never been easy, especially when those curves cross each other. Hence, summaries, such as the whole and the partial areas under the ROC curve have been proposed (see e.g. Hand, 2009).

The AUC is one of the most common measures to evaluate the discriminative power of a predictive model. It is defined as the integrated true positive rate over all false positive rate ranges. One of the most favorable properties of AUC is to be threshold-independent. In practice, there are usually many different classifiers that have different ROC curves and perform very

*This work has been financed by the research project PRIN MISURA (cod. 2010RHAHPL).

†Department of Economics and Management, University of Pavia – Italy: maggma@eco.unipv.it; Department of Political and Social Sciences, University of Pavia – Italy: silvia.figini@unipv.it

differently at all reasonable thresholds but they may have similar AUC values. For example, the AUC indices of any two ROC curves that are symmetric around the negative diagonal of the unit square (hence ROC curves cross each other) are the same. This property becomes a shortcoming of AUC being used as a classification performance measure.

In fact, AUC has another well-understood weakness when comparing ROC curves which cross. As shown in Gigliarano et al. (2012), when two ROC curves cross, then one curve has larger specificity for some choices of sensitivity, and the other has larger specificity for the other choices of sensitivity. Aggregating over all choices, as the AUC does, it could clearly lead to conclusions which misrepresent the performance of the model. In particular, as stated in Krzanowski and Hand (2009, p. 108), "one can easily conjure up examples in which the AUC for classifier 1 is larger than the AUC for classifier 2, even though classifier 2 is superior to classifier 1 for almost all choices of the classification threshold".

Partly in an attempt to overcome this problem, and partly in recognition of the fact that it is likely that not all values of cut-offs will be regarded as relevant, various researchers have proposed to compare crossing ROC curves by restricting the performance evaluation to proper subregions of scores and to use a measure of a partial area under the ROC curve (see e.g., Krzanowski and Hand, 2009; Hand and Till, 2001), defined as an integration of the ROC curve over a confined range of false positive rate instead of over the whole. However, a clear shortfall of this measure is to be dependent on a range of false positive rate values that has to be specified; therefore, it is no longer threshold independent. Various alternative measures to ROC curves and AUC have been proposed, to take into account misclassification costs (see e.g., Hand, 2009) with particular focus on model selection for predictive models for binary outcome. To overcome the shortcomings of AUC, especially when ROC curves cross, Gigliarano et al. (2012) propose a novel family of indexes for model selection coherent with the stochastic dominance approach.

In this framework, more recently the H index has been proposed for model comparison and assessment by Hand (2009) to overcome the main drawbacks of the AUC. The H measure is the mean loss from a classifier when the distribution of relative costs is distributed according to a Beta with specific settings for the parameters. We underline that the H index is subjective because it requires an appropriate elicitation for the parameters in the Beta distribution, thus it produces different results in terms of model selection based on the fixed settings.

Our approach proposed in this paper is rather different. Instead of using a specific class of performance indicators, our main objective is to put model comparison in a decision theoretic framework developing a characterization of the proper scoring rules (see e.g. Gneiting and Raftery, 2007). As a result we obtain for the set of models under comparison a clear ordering measured with the average loss function. We underline that our approach is rather general and it can be useful for a wide range of applications.

We make a comparison between classification accuracy criteria proposed in the literature and our methodology on simulated and real data sets.

In this paper we compare parametric models based on logistic regression (LR) and non parametric models based on classification trees (CT) which are two common single competitive models comparable in terms of predictive performance when the dependent variable is binary. Secondly, to take into account model uncertainty, we have considered the corresponding averaged models, such as Bayesian model averaging (BMA) and Random Forest (RF) (Breiman, 2001).

We are able to overcome the cut-off selection and the weakness previously underlined on the classification accuracy criteria present in the literature.

On the basis of the evidence achieved on simulated and real data we highlight that in terms of model selection the AUC and loss functions often do not agree. We observed also that theoretical properties relevant for the BMA do not hold in terms of empirical evidence achieved on simulated and real data (Raftery and Zheng, 2003). We note also that single models perform better (or at least not worse) than the corresponding averaged models in terms of AUC measure; however a more clear comparison can be obtained using our approach. Simulations show that loss functions deliver more stable model ranking than AUC, in particular for what concerns the comparison between single and averaged models.

The paper is organized as follows: Section 2 describes the theoretical approach for model

comparison; Section 3 reports a simulation study and the empirical evidences achieved on a real credit risk data; Section 4 ends the paper and underlines further ideas of research.

2 Proper loss functions for model comparison

In order to assess the forecasting power of a classification model, the forecasting errors have to be measured. The loss functions allow to take into account not only the misclassification (as the criteria described in Section 1), but also the magnitude of the errors. In terms of results, for each model and for a given loss function, the corresponding average prediction loss can be easily computed. The most commonly used loss function is the mean square error (MSE); for binary outcomes, the MSE is defined as:

$$\frac{1}{n}\sum_{i=1}^{n}\left(P_i - Y_i\right)^2, \tag{1}$$

where $P_i = P\left[Y_i = 1 \mid X_1, \ldots, X_k\right]$ is the predicted probability, Y_i is the observed target binary variable ($Y_i = 1$ event of interest), $\{X_1, \ldots, X_k\}$ is the set of covariates employed to explain Y, n is the validation sample size. We recall that loss functions are related to *scoring rules* (or functions).

Scoring rules which have been introduced in probabilistic forecast to estimate the most accurate model (see e.g., Gneiting and Raftery, 2007; Selten, 1998), are functions which assign a reward to good predictions. This family of functions have been used as an element of the estimation procedure in cost-weighted class probability estimation. A relevant class of scoring functions is the set of *proper scoring function* (see e.g., Shuford et al., 1966; Savage, 1973; Schervish, 1989). We recall that a scoring function $S(x)$ is proper if the expected score $E\left[S\left(P \mid Y\right)\right]$ is minimized for $P = Y$. Therefore, proper scoring functions ensure that the prediction is as "honest" as possible (see e.g., Buja et al., 2005; Gneiting and Raftery, 2007; Selten, 1998, for applications to dichotomous prediction models).

The loss functions are linked to scoring function in a straightforward way: the functional form of a loss function can be obtained by changing the sign of a scoring function and by applying a suitable affine transformation to have zero losses for correct forecasts. There are many kinds of scoring rules and thus of loss functions, see, e.g., Buja et al. (2005); Gneiting and Raftery (2007); Selten (1998); Winkler (1994). Therefore, loss functions related to proper scoring rules can be defined *proper loss functions*. Beside, scoring rules can also be used to compare the prediction performances of different parameter values of a given model and/or different models.

In this paper we apply proper loss functions to assess the prediction power of different models, evaluated on validation sample of size n. For a given loss function L, the corresponding average value can be computed as $\frac{1}{n}\sum_{i=1}^{n} L(P_i)$. On the basis of different loss functions, we evaluate the predicting performances of the models under comparison. Following Hansen et al. (2011), for a given loss function and a fixed confidence level, we compute *Model Confidence Sets* (MCS) which contains the best significant models. Uninformative data yield a MCS with many models, whereas informative data yield a MCS with only a few models. MCS are estimated following the procedure set up by Hansen et al. (2011)[1], adopting χ^2 test for the model elimination rule.[2]

[1]The results are obtained running the Ox package Mulcom (`http://mit.econ.au.dk/vip_htm/alunde/mulcom/mulcom.htm`). This package can be run with the Ox console which is free for academic research, study and teaching purposes (`http://www.doornik.com/download.html`).

[2]Other test can be applied, for instance, as the authors suggest, F statistics or statistics built on t-statistics without the need of the model covariance matrix. In our applications F statistics delivers similar results to χ^2 and the t-statistic, which is much slower in terms of computing time, is convenient when the number of models is large relative to the sample case, which is not our case.

2.1 Applied loss functions

In our analysis we will consider the following loss functions related to the corresponding scoring rule.

$$\begin{array}{cc} \text{Square loss} & \text{Brier score} \\ L(P_i) = (P_i - Y_i)^2 & S(P_i) = -(P_i - Y_i)^2 \end{array} \tag{2}$$

$$\begin{array}{cc} \text{Spherical loss} & \text{Spherical score} \\ L(P_i) = \begin{cases} 1 - \dfrac{P_i}{\sqrt{P_i^2 + (1-P_i)^2}}, & \text{if } Y = 1 \\ 1 - \dfrac{1-P_i}{\sqrt{P_i^2 + (1-P_i)^2}}, & \text{if } Y = 0 \end{cases} & S(P_i) = \begin{cases} \dfrac{P_i}{\sqrt{P_i^2 + (1-P_i)^2}}, & \text{if } Y = 1 \\ \dfrac{1-P_i}{\sqrt{P_i^2 + (1-P_i)^2}}, & \text{if } Y = 0 \end{cases} \end{array} \tag{3}$$

$$\begin{array}{cc} \text{Logarithmic loss} & \text{Logarithmic score} \\ L(P_i) = \begin{cases} -\log(P_i), & \text{if } Y = 1 \\ -\log(1-P_i), & \text{if } Y = 0 \end{cases} & S(P_i) = \begin{cases} \log(P_i), & \text{if } Y = 1 \\ \log(1-P_i), & \text{if } Y = 0 \end{cases} \end{array} \tag{4}$$

We remark that functions (2), (3) and (4) are symmetric around $\frac{1}{2}$; this means that the magnitude of the corresponding value of the loss function does not depend on the sign of the error:

$$L(P_i \mid Y_i = 0) = L(1 - P_i \mid Y_i = 1) \quad \text{and} \quad S(P_i \mid Y_i = 0) = S(1 - P_i \mid Y_i = 1).$$

Following Winkler (1994), it is possible to consider a class of loss functions able to generalize symmetric loss functions, assigning different weights to the different kinds of errors. Starting from a scoring rule S, an increasing function S_1, a decreasing function S_2 and a level $c \in (0,1)$, the scoring rule is[3]

$$S^*(P_i) = \frac{S(P_i) - S(c)}{T(c)} \tag{5}$$

where

$$T(c) = \begin{cases} S_1(1) - S_1(c), & P_i \geq c \\ S_2(1) - S_2(c), & P_i < c \end{cases} \tag{6}$$

As a special case, the quadratic asymmetric scoring function can be obtained with $S_1(P_i) = -(1-P_i)^2$ and $S_1(P_i) = -P_i^2$, so that (5) and (6) lead to

$$S(P_i) = \begin{cases} \frac{(1-c)^2 - (1-P_i)^2}{T(c)}, & Y_i = 1 \\ \frac{c^2 - P_i^2}{T(c)}, & Y_i = 0 \end{cases}$$

where

$$T(c) = \begin{cases} (1-c)^2, & P_i \geq c \\ c^2, & P_i < c \end{cases}$$

The corresponding asymmetric quadratic loss function is easily obtained as follows

$$L(P_i) = \begin{cases} k\left(1 - \frac{(1-c)^2 - (1-P_i)^2}{T(c)}\right), & Y_i = 1 \\ k\left(1 - \frac{c^2 - P_i^2}{T(c)}\right), & Y_i = 0 \end{cases} \tag{7}$$

[3]Often, the nature of the problem may suggest the need to assign asymmetric penalties to different kinds of errors. The level c was originally introduced as the *least skillful forecast*, that is, the least skillful forecast is the "(yet not unreasonable) [...] forecasts that could be made without any real expertise and with access only to readily available data" (see Winkler, 1994, p. 1398). In either interpretation, the choice of c is strictly related to the nature of the empirical problem faced by the researcher.

where

$$T(c) = \begin{cases} (1-c)^2, & P_i \geq c \\ c^2, & P_i < c \end{cases} \tag{8}$$

$$k = \begin{cases} \frac{c}{2}, & c \leq \frac{1}{2} \\ \frac{1-c}{2}, & c > \frac{1}{2} \end{cases}$$

It is easy to show that if $c = \frac{1}{2}$, the function (7) corresponds to (2), i.e. it is symmetric. All these functions, except necessarily the log function (4), take values in the interval $[0, 1]$.

Note that all the functions (2), (3), (4) and (7) are associated to scoring functions which are proper in the sense of Gneiting and Raftery (2007).

3 Application

This section reports the empirical evidence achieved comparing classification accuracy criteria summarized in Section 1 and our proposed methodology based on loss functions described in Section 2. Our empirical analysis is based on credit risk data provided by Creditreform, one of the major rating agencies for SMEs in Germany. When handling bankruptcy data it is natural to label one of the categories as success (healthy) or failure (default) and to assign them the values 0 and 1 respectively. Therefore, our data set consists of a binary response variable (default) values and a set of 10 financial ratios as covariates (see e.g. Figini and Fantazzini, 2009). The sample size available is composed of about 1000 observations and the a priori probability of default is equal to 13%.

In order to explain the target variable as a function of the covariates at hand, we compare single models based on LR and CT with averaged models such as BMA and RF. Furthermore, in order to underline how different a priori probabilities affect the results in terms of model selection using classical measures based on classification accuracy criteria and loss functions, we introduce a simulation study described in more detail below.

Since we are working on credit risk data, we remark that a default probability equal to 0.5 is a sound signal of risk. In this situation, it is easy to figure out that the corresponding type I and type II errors (see e.g. Hand and Till, 2001) should be weighted differently. Also, the economic consequences of a wrong prediction are more serious in the case of default.

In order to take into account of this aspect a possible solution is the application of asymmetric loss functions with the parameter $c < 0.5$. Afrer a careful sensitivity analysis made on real and simulated data, we have fixed a value of $c = 0.15$ in the loss function (7).

In terms of results, we report the AUC and the average loss function, together with the corresponding significance measures (DeLong and MCS, respectively). Boldfaces indicates the best models and a star indicate the models not distinguishable from the best one (using the corresponding tests)[4].

3.1 Simulated data

In this section we design different scenarios for the a priori probability of default simulating the dependent variable from a Binomial distribution with $n = 1000$ and different levels of a priori probability. The covariates are fixed and correspond to the real data introduced above. Tables 1 to 5 report the results on the simulated data. The tables are structured as follows: the first column shows the simulated a priori probability, the rest of the columns reports the AUC index (Tables 1) or the average loss (Tables 2 to 5) for the models under comparison.

From Table 1 we notice that for the large part of the scenarios the LR performs better than the other competing models. We underline also that the AUC values are rather constant across the simulated a priori probability levels. On the basis of the DeLong test we underline that the best models are often indistinguishable, thus the model selection appears difficult and further classification accuracy criteria should be applied (i.e. threshold dependent criteria) to select

[4]The corresponding results in terms of H index and the KS index are available upon request to the authors.

the best model. However, as pointed out in Section 1, threshold dependent criteria require a subjective choice for the cut-off, thus the resulting model selection is affected.

The results concerning the application of loss functions to the simulated data are reported in Tables 2 to 5. The best models are always single models (LR or CT). The RF model is never included into the MCS.

In general, loss functions give a more clear picture to select the best model, without resorting to subjective choices. In fact, the number of models indistinguishable from the best one is on average smaller than using the AUC as criterion.

The empirical evidence at hand shows that using the logarithmic (Table 4) or the asymmetric quadratic loss functions (Table 5), the number of elements in MCS is minimized. It is interesting to notice that the introduction of the asymmetry in the quadratic loss function clearly decrease the number of elements of the MCS; this shows that the asymmetry can take into account a relevant feature of the data.

In our opinion, our simulation exercise show that loss functions allow to better identify the best model, extracting more information from the data.

3.2 Real data

In this section we report in Table 6 the results obtained on the real credit risk data set described above. From Table 6, the AUC underline as best model is CT which is not statistically different from LR (DeLong test). In terms of loss function, the best model is CT as well. We remark that in this application the classical approach based on AUC agree with the model selection obtained using loss fucntions in all cases. Square, spherical and logarithmic loss functions identify CT as distinguishable best model.

4 Conclusion

The performance of predictions models can be assessed using a variety of methods and metrics. In this paper we proposed a methodological approach to evaluate the performance of prediction models using proper loss functions. The proposed approach is threshold independent, thus it overcomes the cut-off selection problem.

On the basis of the evidence achieved on simulated and real data we underline that in terms of model selection the AUC and loss functions often do not agree. We observed also that theoretical properties relevant for the BMA do not hold in terms of practical application. In fact BMA does not (always) minimize square errors. Furthermore, the empirical evidence at hand revealed that out of sample predictive performance of BMA are not superior with respect to the other models.

We note also that single models perform better (or at least not worse) than the corresponding averaged models in terms of AUC measure, with a more clear comparison using loss functions.

The simulated data show that the introduction of the asymmetry into the quadratic loss function decreases the size of the MCS reaching a value close to the logarithmic loss function. With respect to the logarithmic, the asymmetric quadratic is limited, so remains usable even in the case of an observation with an error of size 1.

Further idea of research will consider loss function and related measures of assessment to improve the averaging among the models.

5 Tables

p	LR	BMA	CT	RF
0.01	**0.6573**	0.5	0.5627	0.5
0.02	**0.6918**	0.6085	0.5	0.5225
0.03	**0.6338**	0.6144 *	0.5923 *	0.5515 *
0.04	**0.7328**	0.6505	0.6067	0.5037
0.05	**0.5974**	0.5	0.5	0.5885 *
0.06	**0.6479**	0.6004 *	0.5	0.5647 *
0.07	**0.5939**	0.5732 *	0.5634 *	0.5235 *
0.08	**0.5884**	0.5	0.5	0.5452 *
0.09	0.6376 *	0.6102 *	**0.6673**	0.5100
0.10	**0.5905**	0.5644 *	0.5	0.5269
0.15	0.5761 *	0.5484	**0.6024**	0.5246
0.20	0.5717 *	0.5635 *	**0.5778**	0.5061
0.25	**0.5754**	0.5486 *	0.5	0.5609 *
0.30	**0.5515**	0.5498 *	0.5	0.5183
0.35	**0.5466**	0.5330 *	0.5	0.5276 *
0.40	0.5397*	**0.5530**	0.5457 *	0.5110 *
0.45	**0.5637**	0.5464 *	0.5	0.5025
0.50	**0.5737**	0.5581*	0.5412	0.5282
0.60	**0.5219**	0.5038 *	0.5*	0.5106 *
0.70	**0.5451**	0.5373 *	0.5	0.5006*
0.80	0.5683	0.5	**0.6143**	0.5824 *
0.90	**0.5930**	0.5662*	0.5321	0.5170 *

Table 1: AUC: Bold figures indicate the best model, * indicate AUC not distinguishable from the best one at 0.05 level.

p	LR	BMA	CT	RF
0.01	0.0106*	0.0107*	**0.0106**	0.0116
0.02	**0.0259**	0.0262*	0.0262	0.0278
0.03	0.0366*	0.0371*	**0.0345**	0.0413
0.04	0.0260*	0.0261*	**0.0238**	0.0294
0.05	**0.0337**	0.0338	0.0338	0.0372
0.06	**0.0455**	0.0461	0.0462	0.0505
0.07	0.0652	0.0661	**0.0621**	0.0723
0.08	**0.0861**	0.0865*	0.0865*	0.0930
0.09	0.0697	0.0706	**0.0650**	0.0789
0.10	**0.0860**	0.0865*	0.0865*	0.0934
0.15	0.1371*	0.1381	**0.1349**	0.1489
0.20	0.1494*	0.1514	**0.1486**	0.1663
0.25	**0.2017**	0.2047	0.2052	0.2273
0.30	**0.2074**	0.2084*	0.2086*	0.2247
0.35	**0.2198**	0.2225	0.2230	0.2422
0.40	**0.2367**	0.2383*	0.2360*	0.2562
0.45	**0.2423**	0.2475	0.2480	0.2653
0.50	0.2461*	0.2488	**0.2441**	0.2619
0.60	**0.2367**	0.2381*	0.2381*	0.2550
0.70	**0.2145**	0.2157*	0.2159*	0.2299
0.80	0.1604	0.1621	**0.1560**	0.1653
0.90	0.0832*	0.0843*	**0.0816**	0.0904

Table 2: Average square loss: Bold figures indicate the best model, * indicate the other elements of the MCS.

p	LR	BMA	CT	RF
0.01	0.0107*	0.0107*	**0.0107**	0.0112
0.02	**0.0265**	0.0266*	0.0266	0.0278
0.03	0.0377*	0.0380*	**0.0355**	0.0417
0.04	0.0265*	0.0265*	**0.0244**	0.0293
0.05	**0.0343**	0.0344	0.0344	0.0369
0.06	**0.0469**	0.0472	0.0473	0.0506
0.07	0.0678	0.0686	**0.0644**	0.0737
0.08	**0.0904**	0.0906*	0.0906*	0.0970
0.09	0.0730	0.0739	**0.0679**	0.0812
0.10	**0.0902**	0.0906*	0.0906	0.0964
0.15	0.1486	0.1493	**0.1466**	0.1596
0.20	0.1631*	0.1653	**0.1625**	0.1814
0.25	**0.2284**	0.2316	0.2322	0.2574
0.30	**0.2352**	0.2363*	0.2366*	0.2557
0.35	**0.2516**	0.2551	0.2556	0.2796
0.40	0.2747*	0.2765*	**0.2734**	0.2990
0.45	**0.2825**	0.2893	0.2901	0.3118
0.50	0.2873*	0.2912	**0.2852**	0.3073
0.60	**0.2745**	0.2762*	0.2762*	0.2968
0.70	**0.2446**	0.2460*	0.2462*	0.2629
0.80	0.1763	0.1779	**0.1726**	0.1819
0.90	0.0874*	0.0882*	**0.0854**	0.0938

Table 3: Average spherical loss: Bold figures indicate the best model, * indicate the other elements of the MCS.

p	LR	BMA	CT	RF
0.01	**0.0583**	0.0596*	0.0586*	0.2124
0.02	**0.1169**	0.1236	0.1240	0.1770
0.03	0.1582	0.1606	**0.1511**	0.2205
0.04	0.1166*	0.1214	**0.1111**	0.2160
0.05	**0.1486**	0.1518	0.1518	0.1862
0.06	**0.1858**	0.1932	0.1941	0.2288
0.07	0.2510	0.2560	**0.2422**	0.2904
0.08	**0.3119**	0.3155	0.3155	0.3374
0.09	0.2624	0.2667	**0.2471**	0.3314
0.10	**0.3117**	0.3153*	0.3155*	0.3520
0.15	0.4446	0.4485	**0.4370**	0.4899
0.20	0.4744*	0.4796	**0.4718**	0.5239
0.25	**0.5913**	0.5996	0.6007	0.6571
0.30	**0.6051**	0.6074*	0.6079*	0.6488
0.35	**0.6312**	0.6371	0.6381	0.6823
0.40	0.6654*	0.6695	**0.6647**	0.7100
0.45	**0.6769**	0.6881	0.6892	0.7277
0.50	0.6855*	0.6907	**0.6803**	0.7207
0.60	**0.6661**	0.6691*	0.6691*	0.7086
0.70	**0.6199**	0.6228*	0.6233*	0.6580
0.80	0.5003	0.5052	**0.4838**	0.5157
0.90	0.3032*	0.3091	**0.3012**	0.3374

Table 4: Average logarithmic loss: Bold figures indicate the best model, * indicate the other elements of the MCS.

p	LR	BMA	CT	RF
0.01	0.0585*	0.0589*	**0.0584**	0.0757
0.02	**0.1293**	0.1386	0.1390*	0.1577
0.03	0.1772	0.1833	**0.1674**	0.2189
0.04	0.1336	0.1362	**0.1180**	0.1656
0.05	**0.1729**	0.1754	0.1754	0.2174
0.06	**0.2188**	0.2289	0.2305	0.2850
0.07	0.2949	0.3050	**0.2848**	0.3618
0.08	**0.3577**	0.3693*	0.3693*	0.3876
0.09	0.3008	0.3066	**0.2818**	0.3766
0.10	**0.3608**	0.3687*	0.3693*	0.4255
0.15	0.4174*	0.4248	**0.4068**	0.4864
0.20	0.4207*	0.4237	**0.4220**	0.4703
0.25	**0.4093**	0.4134	0.4137	0.4470
0.30	**0.4117**	0.4123*	0.4124	0.4411
0.35	**0.4029**	0.4045	0.4047	0.4227
0.40	**0.3884**	0.3904	0.3891*	0.4047
0.45	**0.3663**	0.3698	0.3701	0.3793
0.50	0.3451*	0.3467	**0.3439**	0.3542
0.60	**0.3002**	0.3010*	0.3010*	0.3109
0.70	**0.2560**	0.2567*	0.2569*	0.2651
0.80	0.1781	0.1791	**0.1755**	0.1810
0.90	0.0872*	0.0879	**0.0863**	0.0914

Table 5: Average asymmetric quadratic loss: $c = 0.15$. Bold figures indicate the best model, * indicate the other elements of the MCS.

AUC – loss function	LR	BMA	CT	RF
AUC	0.8527*	0.8472	**0.8854**	0.8195
square loss	0.0893	0.0896	**0.0728**	0.0951
shperical	0.0983	0.0985	**0.0784**	0.1032
logarithmic	0.2926	0.2953	**0.2570**	0.3485
asymmetric quadratic	0.2490*	0.2549*	**0.2452**	0.3060

Table 6: AUC and average loss function values for the actual data. Bold figures indicate the best model, * indicate AUC not statistically distinguishable from the best one, and the other elements of the MCS.

References

Adams, N.M. and Hand, D.J. (1999). Mining for unusual patterns in data. Working paper, Dept. Mathematics, Imperial College, London. Agrawal, R., Stolorz, P. and Piatetsky-Shapiro, G. (eds.)

Breiman, R. (2001). Random Forests. Machine Learning 45, Issue 1, 5–32.

Buja, S., Stuetzle, W., and Shen, Y. (2005). Loss Functions for Binary Class Probability Estimation: Structure and Applications, Wharton University Working Paper (http://stat.wharton.upenn.edu/ buja/PAPERS/paper-proper-scoring.pdf).

Crook, J. Edelman D.B., Lyn C.T. (2007) Recent developments in consumer credit risk assessment , European Journal of Operational Research, Volume 183, Issue 3, 1447–1465.

DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.

Figini, S. and Fantazzini, D. (2009) Random Survival Forests Models for SME Credit Risk Measurement, Methodology and Computing in Applied Probability, Volume 11, Issue 1, 29–45.

Gigliarano, C., Figini, S. and Muliere P. (2012). Making classifier performance comparisons when Receiver Operating Characteristic curves intersect. Quaderni di Statistica. vol. 14.

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

Hand, D.J. and Till, R.J. (2001). A simple generalization of the area under the ROC curve to multiple class classification problems. Machine Learning, 45, 171–186

Hand, D.J. and Anagnostopoulos C. (2013) A better Beta for the H measure of classification performance, Preprint, http://arxiv.org/abs/1202.2564

Hand, D. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning 77, 103–123.

Hand, D.J. (2012). Assessing the performance of classification methods. International Statistical Review 80, 400–414.

Hansen, P.R., Lunde, A., Nason, J.M. (2011). The model confidence set. Econometrica 79(2), 453-497.

Krzanowski, W.J. and Hand, D.J. (2009) ROC curves for continuous data. CRC/Chapman and Hall.

Raftery, A. and Zheng Y. (2003). Discussion: Performance of bayesian model averaging. Journal of the American Statistical Association 98, 931–938.

Savage, L.J. (1973). Elicitation of Personal Probabilities and Expectations, J. of the American Statistical Association 66, No. 336, 783–801.

Schervish, M.J. (1989). A General Method for Comparing Probability Assessors, The Annals of Statistics 17, No. 4, 1856–1879.

Selten, R. (1998). Aximatic characterization of the quadratic scoring rule. Experimental Economics 1, 43–62.

Shuford, E.H., Albert, A., Massengill, H.E. (1966). Admissible Probability Measurement Procedures, Psychometrika 31, 125–145.

Winkler, R.L. (1994). Evaluating Probabilities: Asymmetric Scoring Rules. Management Science 40(11), 1395–1405.