Dipartimento di
Economia, Statistica e Diritto
Università di Pavia

Serie Statistica

n. 7/2011

Silvia Figini, Lijun Gao, Paolo Giudici

Model averaged credit risk models

QUADERNI DEL DIPARTIMENTO DI ECONOMIA, STATISTICA E DIRITTO
UNIVERSITÀ DI PAVIA

REDAZIONE

Claudia Banchieri
Dipartimento di Economia, Statistica e Diritto
Università degli Studi di Pavia
Corso Strada Nuova 65
27100 PAVIA
tel. 0039-0382-984406
fax 0039-0382-984402
E-MAIL banchier@unipv.it
I Quaderni sono disponibili sul sito: http://www-5.unipv.it/webdesed/quaderni.php

COMITATO SCIENTIFICO

I **QUADERNI DEL DIPARTIMENTO DI ECONOMIA, STATISTICA E DIRITTO** hanno lo scopo di favorire la tempestiva divulgazione, in forma provvisoria o definitiva, di ricerche scientifiche originali. La pubblicazione di lavori nella collana è soggetta a referaggio e all'approvazione del Comitato Scientifico.
Questa nuova edizione dei **QUADERNI** rappresenta la continuazione di tre serie di pubblicazioni pre-esistenti: Quaderni del Dipartimento di Economia Pubblica e Territoriale, Quaderni di ricerca del Dipartimento di Statistica ed Economia Applicate "L. Lenti" e Osservatorio dei contratti della P.A.

Model averaged credit risk models

Silvia Figini, Lijun Gao and Paolo Giudici*

June 17, 2011

Abstract

In credit risk statistical models are usually chosen according to a model selection procedure that aims at selecting the most performing structure. The chosen model is, once selected, taken as the basis for further actions, such as parameter estimation, default prediction and predictive classification. Relying upon a single model may not be the best strategy. In this paper we investigate whether the usage of more models, in a model averaging perspective, improves the performance of credit risk models. To achieve this aim we propose, starting from the most employed credit risk model - logistic regression - to average results obtained from a collection of models, using either bootstrapping or a Bayesian perspective.
In order to compare the results obtained by single models and model averaged ones, we propose a set of performance measures , some of which are known in the statistical literature, and some others that are newly designed to fit the application needs present in credit risk modelling. Our results show that model averaged models perform better than single models and that the choice between bootstrapped averages and Bayesian model averages depends on the performance measures that are considered.
In order to show how our proposal works, we have used a real data set provided by a credit rating agency composed of small and medium enterprises, that has already been considered in the statistical literature.

1 Introduction

Credit risk models can be included in a more general class of statistical models aimed at explaining a binary target response variable as a function of a set of explanatory variables.
There have been several methods proposed in the literature, to predict a binary

* *silvia.figini@unipv.it*

target as a function of a set of covariates, such as logistic regression, classification trees, neural networks, Bayesian networks (see, for a review of these models, Figini and Giudici, 2009), support vector machines (Crook et al. 2006), multivariate adaptive regression splines (Taylan et al. 2007, Hastie et al. 2001). The main focus of all these methods is to predict a dependent (response) variable from a set of independent variables (predictors).

In the above context, statistical models are usually chosen according to a model selection procedure that aims at selecting the most performing structure. The chosen model is, once selected, taken as the basis for further actions, such as parameter estimation, default prediction and predictive classification. However, relying upon a single model may not be the best strategy, as the uncertainty over model is not adequately taken into account. In our opinion, Bayesian Model Averaging (BMA) provides a coherent way to form a weighted average of a class of possible models using the relative likelihood of each model (see e.g. Hoeting et al. 2001).

Few papers have investigated the comparison between single selected models and model averaging. Among them, we recall the paper of Hayden et al. (2009), which presents a comparison between stepwise selection in logistic regression and BMA for credit risk models. An other reference is Tsai et al. (2010) that show a statistical criterion and a financial market measure to compare the forecasting accuracy of different model selection approaches: Bayesian information criterion, model averaging and model mixing.

The main objective of this paper is to extend the previous contributions, evaluating the relative performance of logistic credit risk models and that of averaged models constructed by means of Bootstrap and Bayesian model averaging. In order to carry out such comparison we propose different kind of measures aimed at measuring the predictive ability, the stability, the discriminant power and the selectivity of each model proposed. Some of these measures are known in the statistical literature (see e.g. Hand et al. 2010, Hand et al. 2008 and Figini and Giudici 2009); others are more known in credit risk practices. The comparison will be based on a real data set that concerns credit repayment behaviour of a sample of German small and medium enterprises.

The paper is structured as follows: Section 2 describes the statistical models most employed in credit risk as well as the model averaging procedures that will be employed. It also contains a description of the performance measures that will be used to compare models of interest to select the best model in a cross validation exercise.

Section 3 reports and compares the empirical evidences achieved on the described real data set.

Finally, conclusions and further ideas of research are reported in Section 4.

2 Credit risk models

Most rating agencies usually analyze each company on site and evaluate the probability of default (PD) on the basis of different quantitative financial criteria considered in a single year, or over a multiple-year time horizon. However, this kind of evaluation does not take into account unstructured and qualitative information for PD estimation. In our opinion, unstructured information - such as business knowledge and expert opinions - is relevant and should be considered for rating measurement.

The problem of data integration can be solved in a two step approach or in a one step approach (see e.g. Figini and Giudici, 2011).

In a two step approach the integration problem could be formalised as follows. For a common set of n statistical units, that will be in our running example, Small and Medium Enterprise (SME for short), let $X_1, ..., X_k$ be a set of k quantitative financial ratios and let $X_1^*, ..., X_p^*$ be a set of p categorical variables derived from opinions and business knowledge. Finally, let Y a target binary variable that describes whether the company has defaulted ($Y = 1$) or not ($Y = 0$).

On the basis of the information collected as above, our aim is to estimate, for each SME, the probability of default that represents a measure of relative risk for the lenders.

For each observation i, $(i = 1, ..., n)$, let Y_i denote the default response variable; let $X_{1i}, ..., X_{ki}$ denote a set of k quantitative candidate predictors, and, finally let $X_{1i}^*, ..., X_{pi}^*$ a set of p qualitative candidate predictors.

In a two step approach the elements of $Y_i = (Y_1, ..., Y_n)'$ are modelled as conditionally independent random variables from an exponential family separately for the quantitative and qualitative explanatory vectors:

$$\pi(y_i|X_{1i}, ..., X_{ki}) \propto exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \qquad (1)$$

$$\pi^*(y_i|X^*_{1i},...,X^*_{pi}) \propto exp\left\{\frac{y_i\theta^*_i - b'(\theta^*_i)}{a'_i(\phi)} + c'(y_i,\phi^*)\right\}, \qquad (2)$$

where θ_i is the canonical parameter functionally related to the linear predictor $\eta_i = X'_i\beta$, with β a $k \times 1$ vector of fixed effects regression coefficients, ϕ a scale parameter and a_i, a'_i b, $b^{'}$, c and $c^{'}$ are known functions. For the sake of interpretational simplicity, we assume that $a_i = a'_i$ $b = b^{'}$, $c = c^{'}$.
According to generalised linear modelling theory (see e.g. Dobson, 2002), we assume $a_i(\phi) = a'_i(\phi) = \frac{\phi}{\omega_i}$, where ω_i is a known positive weight.
The logit link is appealing because it produces a linear model for the log of the odds, $ln\left\{\frac{\pi(y_i=1|X_{1i},...,X_{ki})}{1-\pi(y_i=1|X_{1i},...,X_{ki})}\right\}$ for the quantitative data and $ln\left\{\frac{\pi^*(y_i=1|X^*_{1i},...,X^*_{pi})}{1-\pi^*(y_i=1|X^*_{1i},...,X^*_{pi})}\right\}$ for the qualitative data, implying a multiplicative model for the odds themselves (for more details, see e.g. Dobson, 2002). As a results we derive the corresponding π_i (probability of default estimated following Equation 1) and π^*_i (probability of default estimated following Equation 2) for each SME $i = 1,...,n$. When the models in Equation 1 and 2 are applied to the data at hand, two distinct estimates of the probability of default are obtained according to generalised linear modelling theory (see e.g. Dobson, 2002): π_i which is the probability of default estimated using the quantitative explanatory data set and π^*_i which is probability of default estimated using the qualitative explanatory data set.

In a two step approach, in order to obtain an integrated estimate of the probability of default, one can employ the proposal of Figini and Giudici, 2011, that is essentially a weighted linear combination of the default probabilities obtained separately with the qualitative and quantitative exploratory data sets.

The advantage of the two step approach that we have described so far is the ability to underline separately the contribution of each of the two data components that often lead to different evaluations. However, from a statistical view point, such approach is a less efficient usage of the data, with respect to a one step approach, that puts all explanatory variables together in a single model, as follows:

$$\pi(y_i|X_{1i},...,X_{ki},X^*_{1i},...,X^*_{ki}) \propto exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\}, \qquad (3)$$

where a_i, b, and c are known functions as before.
The aim of this paper is to propose an efficient model that, however, maintains the interpretational advantage of the two step model, namely a model that is

able to enucleate the separate contribution of quantitative and qualitative variables.

Integrating qualitative information such as expert opinions with quantitative information such as financial ratios, in a one step statistical model is a difficult task, especially because the databases are different, and may lead to different conclusions in terms of estimated default probabilities, increasing considerably model uncertainty. A possible solution to this problem is take a collection of subsamples from the data, replicate model selection and estimation for each subsample and, finally, average the result obtained, in a bootstrap approach (see e.g. Efron et al. 1994). The disadvantage of this latter approach is, besides the computational burden, the need to have an adequate number of observations to make the bootstrap subsamples representative. Another possible solution is Bayesian Model Averaging (BMA, for short) (see e.g. Hoeting et al., 2001) that provides a coherent mechanism to account for model uncertainty.

Both the bootstrap and the Bayesian model averaging perspectives provide appealing ways to integrate quantitative and qualitative estimated default probabilities in a one step model that makes explicit the contribution of each component.

We claim that Bayesian Model Averaging provides more efficient estimates of default probabilities with respect to the bootstrap method: this result that shall be shown in the next section derives from the fact that Bayesian Model Averaging considers only the models with the highest posterior probability in an Occam's Razor paradigm (see e.g. Hoeting et al. 2001).

We now formalise the expression of BMA in our context. Suppose that Y_i is a dependent variable and that X_i is a data matrix composed of p and k qualitative and quantitative variables respectively, for $i = 1, ..., n$.

A one step generalised linear model M_1 is defined for both subsets of variables by specifying $P(Y_i|X_i,)$ in such a way that $E(Y_i|X_i) = \mu_i$, $V(Y_i|X_i) = \sigma^2 \nu(\mu_i)$ and $g(\mu_i) = X_i$, where $({}_1, \ldots, {}_p)'$ and $({}_1, \ldots, {}_k)'$ are the parameters estimate for the qualitative and quantitative variables and g is the link function. Raftery (1996) describes a useful parametric form for the prior parameters distributions that will be adopted here.

Concerning the prior over the model space, in our proposal, since there is little prior information about the relative plausibility of the models, the assumption that all models are equally likely a priori is a reasonable choice.

We now briefly describe our proposed bootstrap model averaging algorithm for logistic regression models.
We consider 1000 bootstrap samples each of which is further split into two subsamples. The first subsample contains 70% of the observations and is used to construct logistic regression models for the estimation of the default probabilities. The second subsample contains 30% of the observations and is used for model validation. An important aspect of our bootstrap algorithm is that we assign the observations to the two subsamples such that the fraction of defaults is held constant as in the observed sample.

In order to evaluate the proposed models, in the next section we shall compare them with classical logistic regression on a real data set. The models will be compared using out of sample performance, that is, their performance in predicting defaults in the validation sample. Models will also be compared in terms of discriminant power and stability. More precisely, We now discuss the issue of model performance measure specification. In order to compare different models, the empirical literature typically uses criteria based on statistical tests (see Burnham et al., 1998), criteria based on scoring functions (see Akaike, 1974, Schwarz, 1978, Vapnik 1998), computational criteria (see e.g. Hastie et al., 2001) and criteria based on loss functions (see e.g. Kohavi et al., 1997).
In our opinion, taking credit risk practices into account, a predictive model should be evaluated not only using predictive performance, but also discriminant power and stability.
In order to detect the *predictive ability* of a model, a clear comparison can be derived using the confusion matrix and related measures of interest (see e.g. Kohavi et al., 1997). Let A denote the credit decision variable, and Y be the variable denoting the status of a statistical unit of interest. A is equal to 1 when credit is not granted to the customer; and it is 0 otherwise. Likewise, Y is equal to 1 when credit of the customer goes into default; and it is 0 otherwise. Let $N_{A_i Y_j}$ denote the number of borrowers with the states of $A = i$, $Y = j$. Then, the contingency table in Table 1 shows the number of statistical units depending on their credit decisions and status. The first diagonal in the contingency table reported in Table 1, express the correctly classified borrowers, and the second diagonal shows the wrongly classified borrowers.
On the basis of Table 1, the percentage of correct classifications (PC) of good

Y/A	$A=0$	$A=1$
$Y=0$	$N_{A_0Y_0}$	$N_{A_1Y_0}$
$Y=1$	$N_{A_0Y_1}$	$N_{A_1Y_1}$

Table 1: Theoretical confusion matrix

and bad borrowers is a first index which can be employed. However, considering credit risk, a loss from a bad applicant is expected to be larger than profit obtained from a good borrower. Hence, in our evaluation, we use different weights for $N_{A_0Y_0}$ and $N_{A_1Y_1}$, such as 1:2, differently from other researchers that take them as equal (see e.g. Parnitzke, 2005; Crook and Banasik, 2004). Under this assumption a performance indicator (PI) can be formulated as follows:

$$PI = \frac{N_{A_0Y_0} + 2N_{A_1Y_1}}{N_{Y_0} + 2N_{Y_1}} \tag{4}$$

The above formulation implies that the true positives $N_{A_1Y_1}$, that is, the bad SME classified as bad, weight twice the true negatives $N_{A_0Y_0}$, that is the good SME classified as good.

Note that both PC and PI depend on the chosen cut off threshold. In our application we select the best models using different choices of cut off and compare the results.

We remark that the cut-off point could also be selected optimising performance measures as the one described before or by maximising the statistic *Kappa* (see e.g. Cohen, 1960). Although we shall not pursue this issue in the paper, we refer the reader to what shown in Figini et al. 2009. A third predictive performance measure that we shall consider the Receiver Operating Characteristic curve (ROC) and the area under it (AUC) (see e.g. Hand et al. 2010).

A last measure of predictive performance is represented by the Breir score (Breir, 1950) which combines the quality of the ranking with the accuracy of the estimated probabilities of default. The Breir score (BS) is defined as follows:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(PD_i - I(Default_i))^2, \tag{5}$$

where PD_i is the estimated default probability and $I(Default_i)$ is an indicator variable that takes the value 1 if the SME i defaults and zero otherwise. Thus, the Breir score is the mean squared error of the observed PDs. Note that the

BS is an inverse measure of performance: a smaller score indicates better performance.

So far we have discussed the issue of predictive performance. Another important characteristic that a model should have is discriminant power. The *discriminant power* of a predictive model can be measured by the lift and the captured response. As described in Figini and Giudici 2009, the lift puts the observations in the validation data set into increasing or decreasing order on the basis of their score, which is the probability of the response event (default), as estimated on the basis of the training data set. It subdivides these scores into deciles and then calculates the observed probability of default for each of the deciles classes in the validation data set.
The captured response shows for each decile the percentage of predicted defaults: if the model were perfect, this percentage would be higher in the fist deciles.

Finally descriptive statistical measures of variability for the estimated probabilities of default as well as for some measures of performance give a simple, but effective idea on the *stability* of the results for the considered models.

3 Application

Our empirical analysis is based on annual 1996–2004 data from 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 Y_i and a set of explanatory variables: $X_{1i}, ..., X_{ki}$ that are quantitative financial ratios and $X^*_{1i}, ..., X^*_{pi}$ that are qualitative features. The sample size available is composed of 1000 SMEs. The observed probability of default is equal to 12.5%.

There is a wide range of financial ratios that can be used to learn more about a company. We have chosen the following based on Creditreform experience:

- *Supplier target days*: it is a temporal measure of financial sustainability expressed in days that considers all short and medium term debts as well as other payables.

- *Outside capital structure*: this ratio expresses the capability of a company to receive forms of financing other than banks' loans.
- *Cash ratio*: it indicates the cash flows a company can generate in relation to its size.
- *Capital tied up*: this ratio evaluates the turnover of short term debts with respect to sales.
- *Equity ratio*: it measures the financial leverage of a company calculated by dividing a measure of equity by the total assets.
- *Cash flow to effective debt*: it indicates the cash a company can generate in relation to its debts;
- *Cost income ratio*: is an efficiency measure which is useful to assess how costs are changing compared to income.
- *Trade payable ratio*: this ratio reveals how often the company payables turn over during the year.
- *Liabilities ratio*: it is a measure of financial leverage calculated by dividing a gross measure of long-term debt by the assets of the company.
- *Result ratio*: this is an index of how profitable a company is relative to its total assets.
- *Liquidity ratio*: this ratio measures the extent to which a firm can quickly liquidate assets and cover short-term liabilities.

As already pointed out, we also have qualitative opinions expressed by Creditreform business experts. These are:

- *KdtUrt*: this information is relevant in order to define if the business relationship between an SME and Creditreform is acceptable or is not recommended. This feature is based on past credit decisions and it shows a value 0 if the relationship is acceptable and 1 otherwise.
- *ZwsUrt*: this variable summarises the payment history of each SME. The levels are 0 if the payment is within time and 1 if irregular payments are present.

- *Entw*: this variable describes information on the company in terms of business development level. A level equal to 2 or 1 or 0 means respectively: a positive company development, a stagnating company development and a declining company development.

- *Auft*: it is a categorical variable that reports the order situation for each SME. If the order situation is good, the variable is equal to 2; if the order situation is declining or bad it is equal to 0 or 1.

- *AnzMta*: this feature is a grouped variable derived from a quantitative information. We have computed three groups composed of different number of employees in relation to special company structures.

We now move to data modelling.

In a one step perspective, we have used the whole data set, putting together quantitative and qualitative variables. Classical logistic regression analysis (CLR) allows to select the best predictive model and to obtain parameter estimates conditionally on such model. In order to assess predictive ability, we have implemented a cross-validation procedure: we have used 70% of the observations as a training set and 30% as a validation set on which to calculate predictive accuracy.

As discussed in the previous section, we have then tried to improve CLR analysis by means of a bootstrap and a BMA approach.

The bootstrap analysis takes as input for each randomly chosen subsample, the results obtained with the most predictive logistic regression model and the corresponding performance measures. Such results and measures can then be summarised to obtain bootstrapped mean estimates as well as corresponding variability measures. In particular, here we report the mean of the performance measures that give an indication of predictive accuracy of the bootstrapped CLR method, the standard deviation of some of them that gives an indication on the stability of the method and, finally, the relative incidence of each variable calculated as the percentage of subsamples (out of the 1000 considered) in which a variable appears in the final selected model. This latter measure is particularly useful for interpretation purposes because it allows to see how different variables impact on the estimated PDs.

The BMA analysis has been conducted assuming that all models are equally likely a priori and implementing the algorithm described in Hoeting et al., 2001. We remark that BMA provides on its own the relative incidence of each variable without recurring to the bootstrap analysis. However, in order to better compare in terms of efficiency BMA with CLR we have carried out a bootstrapped version of BMA following the procedure previously described for CLR.

We now move to model comparison of the four considered models: CLR, BMA, Bootstrapped Classical Logistic Regression (BCLR) and Bootstrapped Bayesian Model Averaging (BBMA). We first check if the different models are different in terms of predictive ability. To reach this objective, on the basis of the percentage of correct classifications (PC) and the performance indicator (PI) described in Section 2, we have compared the models.
Table 2 reports the results obtained using a large set of cut offs.

Table 2 about here

From Table 2, note that, fixing a cut off equal to 0.1 or 0.2 the best model is classical logistic regression; if the cut off is equal to 0.3 or 0.4 the best model becomes the bootstrapped classical logistic regression. If the cut off is greater than 0.4, the classical logistic regression and the bootstrapped classical logistic regression are similar in terms of performance. Note also that bootstrapped BMA logistic regression increases predictive performance in correspondence to higher cut offs: for example if the cut off is greater than 0.5, this model becomes the best one.
We have then computed for each model the percentage of correct classifications (PC) as shown in Table 3.

Table 3 about here

From Table 3 we note that fixing a cut off greater than 0.4 the BBMA is the best model in terms of correct classifications. We have then computed the AUC with the relative confidence interval on the basis of a bootstrap percentile method (see e.g. Hosmer et al, 2000). Table 4 shows for all models the AUC and the correct classification rate.

Table 4 about here

From Table 4 the BBMA model shows the best performance. In addition we have obtained that the AUC test based on the confidence intervals confirms that the

AUCs computed are significantly different from 0.7. In business practice, this means that the discrimination made by the corresponding model is acceptable.

For the sake of comparison, we have also compared our approach based on one step integration with the results described in Figini and Giudici 2011 which are based on a two step model. In the merged model proposed in Figini and Giudici 2011, the obtained AUC is equal to 0.909 and the percentage of correct classification using a cut off equal to 0.8 is equal to 0.915.
Therefore, our proposal leads to better results in terms of model performance.

The final measure of predictive performance that we consider is the Breir Score. Table 5 shows for all the considered models the Breir score.

Table 5 about here

From Table 5, we remark that the best model is based on the bootstrapped BMA logistic regression.

We now compare the discriminant power of each model proposed. Table 6 shows the lift and the cumulated response (CCR) calculated in the original data on the validation data set.

Table 6 about here

As we can observe from Table 6, the BMA model performs better than classical logistic regression in terms of lift and cumulated captured response. It is interesting to note that considering the first decile BMA captures 51.22% of the defaults, while logistic regression only 39.66%.
In order to confirm this empirical evidence, we have computed the lift and the CCR on 1000 bootstrapped data sets using 70% of observations and training set and the rest as validation. Table 7 shows the results on the validation set.

Table 7 about here

As we can observe in Table 7 Bayesian approach yields models that tend to outperform models selected through heuristics approaches. Considering the first decile, BMA captures the 61.52% of the defaults, while logistic regression captures only 51.69%.

Finally, in order to compare the stability of the proposed models we have derived statistical measures of location and variability for the PD estimated under the different models. The results are in Table 8.

Table 8 about here

In Table 8 we report for each model the mean, the standard deviation, the minimum and the maximum values for the PDs estimated. In terms of stability of the results, we note that classical and Bayesian models give similar results.

To conclude our analysis we compare in terms of variable inclusion, the one step BBMA model with the two step model of Figini and Giudici, 2011. Our result confirm that the significant variables in classical logistic regression model are: Result ratio, Liabilities ratio and ZwsUrt. Table 9 reports for each model the relative importance of each variable that describes how many of the considered models contain that variable.

Table 9 about here

In terms of model complexity expressed by number of variables included in each model, we note that classical logistic regression is the best one because it includes only 3 variables. The second best is the bootstrapped classical logistic regression and finally the Bayesian models.
On the basis of the results achieved we think that Bayesian models lead better results in terms of predictive performance, while classical approaches provide an efficient and parsimonious method to select the most important variables.

4 Concluding remarks

In this paper we have presented a comparison between classical and model averaged models for credit risk estimation. Our results suggest that model averaged models could be considered as an alternative to classical logistic regression. In particular, we have found that, in comparison with classical logistic regression models, Bayesian Model Averaging provides risk models with superior discriminatory power and comparable predictive performance. This conclusion has been confirmed by the application of the proposed methods to other databases: indeed, the superiority in performance is higher the more rare the default event is.

To conclude, if the aim of the research is to find a model able to discriminate bad from good companies, BMA is the right answer; on the other hand, if the aim is to find a simple and effective tool in terms of predictive accuracy, classical logistic regression and its bootstrapped version provides a solution that is hardly beaten by Bayesian model averaged models.
We believe that more research should be conducted in this area, for example considering other risk management problems and developing performance indicators suited for credit risk model comparison.

5 Tables

cut off	**CLR**	**BMA**	**BCLR**	**BBMA**
0.1	0.8155	0.0672	0.0667	0.076
0.2	0.881	0.3707	0.8586	0.456
0.3	0.9126	0.473	0.942	0.876
0.4	0.9259	0.5575	0.9414	0.906
0.5	0.9316	0.6293	0.9333	0.934
0.6	0.942	0.7069	0.9333	0.936
0.7	0.9379	0.7851	0.9333	0.937
0.8	0.9316	0.8563	0.9333	0.937
0.9	0.9322	0.9006	0.9333	0.954
1	0.9322	0.9333	0.9333	0.957

Table 2: Performance indicator for different cut offs

cut off	**CLR**	**BMA**	**BCLR**	**BBMA**
0.1	0.8168	0.8125	0.125	0.8127
0.2	0.8728	0.8772	0.8545	0.857
0.3	0.8987	0.8944	0.9084	0.899
0.4	0.9073	0.9019	0.8922	0.912
0.5	0.9106	0.9084	0.875	0.912
0.6	0.9084	0.8998	0.875	0.912
0.7	0.8901	0.8847	0.875	0.9234
0.8	0.875	0.8761	0.875	0.9238
0.9	0.8739	0.8739	0.875	0.938
1	0.875	0.875	0.875	0.938

Table 3: Correct classification for different cut offs

Model	AUC
Classical Logistic Regression	0.837
BMA Logistic Regression	0.8916
Boostrapped Classical Logistic Regression	0.913
Boostrapped BMA Logistic Regression	0.945

Table 4: Model assessment based on AUC

Model	Breir Score
Classical Logistic Regression	0.0676
BMA Logistic Regression	0.234
Boostrapped Classical Logistic Regression	0.086
Boostrapped BMA Logistic Regression	0.045

Table 5: Model assessment based on Breir Score

Decile	Lift BMA	Lift CLR	CCR BMA	CCR CLR
1	5.12	3.97	51.22	39.66
2	2.07	2.33	71.94	62.93
3	1.13	1.55	83.20	78.45
4	0.81	1.03	91.30	88.79
5	0.32	0.60	94.46	94.83
6	0.25	0.09	96.97	95.69
7	0.25	0.26	99.45	98.28
8	0.06	0.17	100.00	100.00
9	0.00	0.00	100.00	100.00
10	0.00	0.00	100.00	100.00

Table 6: Lift and cumulated captured response for the original data

Decile	Lift BMA	Lift CLR	CCR BMA	CCR CLR
1	6.15	5.17	61.52	51.69
2	1.71	2.19	78.57	73.61
3	0.29	0.38	81.45	77.44
4	0.29	0.38	84.33	81.27
5	0.29	0.38	87.20	85.11
6	0.29	0.38	90.08	88.94
7	0.29	0.38	92.96	92.77
8	0.29	0.38	95.84	96.61
9	0.29	0.34	98.72	100.00
10	0.13	0.00	100.00	100.00

Table 7: Lift and cumulated captured response for the bootstrap data sets

	CLR	**BMA**	**Boot CLR**	**Boot BMA**
Mean	0.125	0.123	0.166	0.185
Standard Deviation	0.042	0.04	0.006	0.005
Max	0.9709	0.9707	0.4609	0.876
Min	0	0	0.1049	0.082

Table 8: Lift and cumulated captured response for the bootstrap data sets

Variable	**BMA**	**BCLR**	**BBMA**
Supplier target days	12.5%	0.2%	15.7%
Outside capital structure	1.2%	5.0%	4.5%
Cash ratio	0.8%	0%	0%
Capital tied up	9.9%	11.7%	12.7%
Equity ratio	40.4%	7.3%	10.7%
Cash flow to effective debt	0%	2.6%	1.3%
Cost income ratio	7.8%	0%	10.3%
Trade payable ratio	4.2%	0%	8.5%
Liabilities ratio	64.8%	7%	80%
Result ratio	100%	30.4%	100%
Liquidity ratio	0.2%	0%	1.3%
KdtUrt	1%	2.4%	7.9%
ZwsUrt	100%	38.3%	100%
Entw	1%	0%	4.6%
Auft	0%	0%	0%
AnzMta	1.2%	0%	5.1%

Table 9: Incidence of each variable in the models

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Second Interantional Symposium on Information Theory*, 267-281.

Breir, G.W. (1950) Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, 113-122.

Burnham, K.P. and Anderson D.R. (1998). Model selection and inference: a practical information- theoretic approach, Springer-Verlag, New York.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37-46.

Crook, J., Edelman, D. and Thomas, L. (2006). Recent developments in consumer credit risk assessment. *European Journal of Operational Research.*

Crook, J., Banasik, J. (2004). Does rejection inference really improve the performance of application scoring models? *Journal of Banking and Finance* vol. 28, 857-874.

Dobson, A.J. (2002). Introduction to Generalised Linear Model, Chapman Hall.

Efron, B. and Tibshirani, R. (1994) Introduction to Generalised Linear Model, Chapman Hall.

Figini, S. and Giudici, P. (2009). Applied Data Mining, Wiley, London.

Figini, S. and Giudici, P. (2011). Statistical Merging of rating models, *Journal of the Operational Research Society*, **62**, 1067 -1074.

Hand, D. and Jamain A. (2008). Mining supervised classification performance studies: a meta analytic investigation, *Journal of Classification* **25**, 87-112.

Hand, D. and Zhou F. (2010). Evaluating models for classifying customers in retail banking collections, *Journal of the Operational Research Society* **61**, 1540-1547.

Hayden, E., Stomper, A. and Westerkamp A. (2009). Selection vs Averaging of Logistic Credit Risk models, submitted.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). The elements of statistical learning: data mining, inference, and prediction, Springer.

Hoeting J., Madigan D., Raftery A. and Volinsky C. (2001). Bayesian Model Averaging, *Statistical Science* **14**, 382-401.

Hosmer, D.W. and Lemeshow, S. (2000). Applied logistic regression, Wiley New York.

Kohavi R. and John G. (1997). Wrappers for feature selection. *Artificial intelligence*, **97**, 273-324.

Parnitzke, T. (2005) Credit scoring and the sample selection bias, working paper.

Raftery, A. and and Richardson, S. (1996) Model selection for generalized linear models via GLIB, with application to epidemiology. In *Bayesian Biostatistics* (D.A. Berry and D.K. Stangl, eds.), New York: Marcel Dekker, pp. 321–354.

Schwarz G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**, 461-464.

Tayan, P. and Weber G. (2007). Multivariate adaptive regression spline and continuous optimization for modern applications in science, economy and technology, submitted.

Tsai C.L., Wang H. and Zhu N. (2010). Does a Bayesian approach generate robust forecasts? Evidence from applications in portfolio investment decisions, *Annals of the Institute of Statistical Mathematics*, vol. 62, pp. 109-116.

Vapnik, V. (1998). Statistical Learning Theory, Wiley.

QUADERNI DEL DIPARTIMENTO DI ECONOMIA, STATISTICA E DIRITTO

n. 1/2011 Paolo Giudici, Emanuela Raffinetti, *A Gini concentration quality measure for ordinal variables,* Serie Statistica.
n. 2/2011 Silvia Figini, Lijun Gao, Paolo Giudici, *Bayesian efficient capital at risk estimation,* Serie Statistica.
n. 3/2011 Luigi Bernardi, *Tendenze dei prelievi tributari ed effetti fiscali della crisi finanziaria nell'Unione europea e in Svizzera,* Serie Economia.
n. 4/2011 Silvio Beretta, Renata Targetti Lenti, *India in the Outsourcing/Offshoring Process: A Western Perspective,* Serie Economia.
n. 5/2011 Paolo Giudici, Eleonora Lorenzini, *SMEs, e-commerce and territorial development: the experience of a "web district",* Serie Statistica.
n. 6/2011 Paola Cerchiello, *On the distribution of functionals of discrete ordinal variables,* Serie Statistica.
n. 7/2011 Silvia Figini, Lijun Gao, Paolo Giudici, *Model averaged credit risk models,* Serie Statistica.

COLLANE PRECEDENTI

QUADERNI DEL DIPARTIMENTO DI ECONOMIA PUBBLICA E TERRITORIALE

n. 1/2010 Silvio Beretta, *Variabili finanziarie ed economia globale in tempo di crisi*
n. 2/2010 Silvio Beretta, Renata Targetti Lenti, *L'India nel processo di integrazione internazionale. Dal primo al secondo* unbundling *e la posizione dell'Italia*
n. 3/2010 Margarita Olivera, *Challenges to Regional Integration in Latin America*
n. 4/2010 Italo Magnani, *Un economista liberale guarda alla economia dell'ambiente: impressioni e riflessioni*
n. 5/2010 Italo Magnani, *A cinquant'anni dalla scomparsa di Benvenuto Griziotti: Riflessioni*
n. 6/2010 Luca Mantovan**,** *Class-bias in Technology Adoption: Stagnation and Transformation of Subsistence Agriculture in the Ethiopian Northern Highlands*
n. 7/2010 Marco Missaglia, Giovanni Valensisi, *A trade-focused, post-Keynesian CGE model for Palestine*
n. 8/2010 Giovanni Valensisi, Marco Missaglia, *Reappraising the World Bank CGE model on Palestine: macroeconomic and financial issues*
n. 1/2009 Giorgio Panella, Andrea Zatti, Fiorenza Carraro, *Market Based Instruments for Energy Sustainability*
n. 1/2008 Italo Magnani, *Il pubblico e il privato nella economia della città*
n. 2/2008 Italo Magnani, *Note a margine di una recente opera sull'indirizzo sociologico della scienza delle finanze italiana*
n. 3/2008 Italo Magnani, *La riforma sociale nella formazione di Nitti economista*
n. 4/2008 Marisa Bottiroli Civardi, Renata Targetti Lenti and Rosaria Vega Pansini, *Multiplier Decomposition, Poverty and Inequality in Income Distribution in a SAM Framework: The Vietnamese Case*
n. 5/2008 Luca Mantovan, *A Study on Rural Subsistence in the Ethiopian Northern Highlands*

[per i Quaderni precedenti si rinvia a http://www-5.unipv.it/webdesed/ept/quaderni.php]

QUADERNI DI RICERCA DEL DIPARTIMENTO DI STATISTICA ED ECONOMIA APPLICATE "L. LENTI"

Carla Ge Rondi, *L'après mariage en Italie au début du XXIe siècle* (2005, n. 27)
Carla Ge Rondi, *Casalinga: popolazione attiva senza retribuzione* (2005, n. 25)

Bruno Scarpa, David Dunson, *Bayesian Methods for Searching for Optimal Rules for Timing Intercourse to Achieve Pregnancy* (2005, n. 24)
Bruno Scarpa, *Lo stress in azienda. Modelli di analisi di un'indagine per l'identificazione delle cause di stress* (2004, n. 23)
Bruno Scarpa, *La Customer Satisfaction per un'azienda di servizi informatici. Impostazione e analisi di un'indagine via web* (2004, n. 22)

[per i Quaderni precedenti si rinvia a http://www-5.unipv.it/webdesed/lenti/quaderni.php]

OSSERVATORIO DEI CONTRATTI DELLA P.A.

[Si rinvia a http://www.contratti-appalti.it/]