UCD GEARY INSTITUTE
DISCUSSION PAPER SERIES

Improving Classifier Performance Assessment of Credit Scoring Models

Raffaella Calabrese
Dynamics Lab, Geary Institute
University College Dublin

Geary WP2012/04
February 2012

Improving classifier performance assessment of credit scoring models

Raffaella Calabrese
Dynamics Lab, Geary Institute
University College Dublin
raffaella.calabrese@ucd.ie

Abstract

In evaluating credit scoring predictive power it is common to use the Receiver Operating Characteristics (ROC) curve, the Area Under the Curve (AUC) and the minimum probability-weighted loss. The main weakness of the first two assessments is not to take the costs of misclassification errors into account and the last one depends on the number of defaults in the credit portfolio. The main purposes of this paper are to provide a curve, called curve of Misclassification Error Loss (MEL), and a classifier performance measure that overcome the above-mentioned drawbacks. We prove that the ROC dominance is equivalent to the MEL dominance. Furthermore, we derive the probability distribution of the proposed predictive power measure and we analyse its performance by Monte Carlo simulations. Finally, we apply the suggested methodologies to empirical data on Italian Small and Medium Enterprisers.

1 Introduction

In this paper, the authors address the problem of assessing the predictive power of credit scoring models. We assume we have a database which contains the characteristics of borrowers. This information is used to construct a scoring model (Crook

et al., 2007; Thomas et al., 2002) that permits banks to discriminate between those borrowers that will pay on time and those borrowers that will pay late or default. A commonly used decision rule is an optimal cut-off, where borrowers with scores greater than or equal to the optimal cut-off are classified as non-defaulters; others with scores below this optimal cut-off are classified as potential defaulters.

A significant innovation of the revised Framework on International Convergence of Capital Measurement and Capital Standards (Basel Committee on Banking Supervision (BCBS), 2004) is the greater use of assessments of risk provided by banks' internal systems as inputs to capital calculations. When following the "Internal Ratings-Based" (IRB) approach to the revised Framework, banking institutions are allowed to use their own internal measures as input for their minimum regulatory capital calculations, subject to certain conditions and to explicit supervisory approval. This is forcing banks and supervisors to develop methodologies to evaluate the accuracy of internal rating models. In this context, validation comprises a range of approaches and tools used to assess the soundness of IRB systems. Therefore, the field of model validation is one of the major challenges for financial institutions and supervisors.

Performance assessments are used by banks to choose between alternative scoring models (Stein and Jordao, 2003) and to monitor rating models over time to decide when the discriminatory power has deteriorated to the extent that the scoring model needs replacing by a new one. For this decision process it is pivotal to understand the classifier performances of scoring models across credit portfolios with different characteristics.

The Basel Committee on Banking Supervision (BCBS, 2005) summarizes a number of statistical methodologies for assessing discriminatory power described in the literature. Credit scoring models are usually evaluated using power curve such as the Cumulative Accuracy Profile (CAP) and the Receiver Operating Characteristic (ROC) curves (Kraznowski and Hand, 2009). Unlike the ROC curve, the CAP curve depends on the composition of the portfolio (BCBS, 2005). Hence, the CAP curve cannot be used for monitoring scoring models over time when the composition of the portfolio changes and for comparing classifier performances of rating models across

different portfolios (Sobehart and Keenan, 2001). Therefore, in this paper we focus only on the ROC curve and its summary index known as the Area Under the Curve (AUC).

Both the ROC curve and the AUC do not depend on the proportion of defaulters in the credit portfolio. Therefore, they could be used to monitor the performance of credit models over time. The main drawback of the ROC curve and the AUC is the assumption of equal misclassification error costs. There are usually large costs associated with extending credit to defaulting obligors and usually smaller costs associated with not granting credit (or granting credit with overly restrictive terms) to subsequently non-defaulting obligors. For this reason, many authors (Beling et al., 2005; Crook at al., 2007; Oliver and Thomas, 2009; Oliver and Wells, 2001; Stein, 2004) take the costs of misclassification errors into account. Most of these authors (Beling et al., 2005; Crook at al., 2007; Oliver and Wells, 2001; Oliver and Thomas, 2009) compute the optimal cut-off by maximizing the expected profit, which is equivalent to minimizing the Probability-Weighted (PW) loss function. For this reason we consider the minimum of the PW loss function as a classifier performance measure in this paper. Analogously to the Bayesian error rate (BCBS, 2005), in the PW loss function the misclassification errors are weighted by the the proportion of defaulters. Hence, the minimum of the PW loss function should be estimated on portfolios with *representative* default probability and cannot be used by banks for monitoring scoring models across different portfolios (Hand and Henley, 1997; Hand and Vinciotti, 2003).

Within this research field, in order to overcome the drawbacks of the above-mentioned methodologies, the main aim of this work is to propose both a curve and a performance measure that take the costs of misclassification errors into account and are robust for different numbers of defaulters in the portfolio. Based on our knowledge, this is the first paper with this aim. In particular, we propose the curve of *Misclassification Error Loss* (MEL) which represents graphically the discriminatory power when the cut-off changes and its shape depends on the ratio of the misclassification error costs. Coherently with the MEL curve, we propose considering the minimum of the MEL curve as a performance measure that depends on the

ratio of the misclassification error costs, but not on the number of defaults in the portfolio. We prefer to consider the ratio of misclassification error costs since it is usually known, unlike the misclassification error costs (Adams and Hand, 1999).

Some important theoretical results are obtained in this work: the ROC dominance is equivalent to the MEL dominance, the normalized area under the MEL curve is equal to the Gini index, the slope of the MEL curve is obtained and the probability density function of the minimum of the MEL curve is derived. Moreover, the minimum of the MEL curve is compared with the minimum of the PW loss function using both simulations and real data. The most innovative aspect of this work is that we incorporate the main characteristics of credit model validation in our simulations. Based on our knowledge, this is the first work that performs Monte Carlo simulations on the classifier performance by drawing from skewed score distributions and by considering low proportions of defaulters in credit portfolios. The simulation results show that our proposal exhibits a definitely better performance than the minimum of the PW loss function.

Another innovative aspect of this paper is the application of the methodological proposals to Italian Small and Medium Enterprises (SMEs). Basel II (BCBS, 2004) establishes that banks should develop credit risk models specifically addressed to SMEs. To the authors' knowledge, no empirical studies are mainly focused on the validation of scoring models for SMEs, only a few studies hint at this topic (Altman and Sabato, 2006; Fantazzini and Figini, 2008). In particular we consider 34,290 Italian SMEs over the years 2005-2009. The main result is that our methodology, unlike the minimum of the PW loss function, allows to classify correctly two scoring models according to their classifier performances.

The present paper is organized as follows. Section 2 analyses the ROC curve, the AUC index and the minimum of the PW loss function. In section 3 the MEL curve and its minimum are suggested. In the following section we compare the properties of our proposal to those of the minimum of the PW loss function by simulations. Successively, in Section 5 we compare our proposals with the ROC curve and the minimum of the PW loss function on a database of Italian SMEs. Finally, the last section is devoted to conclusions.

2 The validation of credit scoring models

Let S be the score on a continuous scale that is assigned to a borrower and which is intended to forecast the borrower's creditworthiness. The borrower's future state at the end of a fixed time period could be default or non-default. The conditional distribution functions of S given the borrower's future state default or non-default are denoted respectively by $F_d(\cdot)$ and $F_n(\cdot)$. Analogously, the conditional probability density functions of S given the future state default or non-default are indicated by $f_d(\cdot)$ and $f_n(\cdot)$.

The institution's intention with the score variable S is to forecast the borrower's future state by relying on the information on the borrower's creditworthiness that is summarized in S. A commonly used decision rule is a cut-off s^* where each debtor with a score lower than s^* is classified as a potential defaulter and each debtor with a score higher than s^* as a non-defaulter (see Thomas et al., 2002). For a given cut-off, the errors of the scoring model are given by $1 - F_d(s^*)$ and $F_n(s^*)$ which represent respectively the Type I and the Type II errors by choosing that the borrower is a future defaulter as null hypothesis.

The research field of this work is to evaluate how well credit models can discriminate between the future defaults and non-defaults. The most basic approach to assessing the performance of a default prediction model is to consider the number of predicted defaults (or non-defaults) and compare this with the actual number of defaults (or non-defaults) experienced. A common means of representing this is a contingency table or confusion matrix, as in Table 1.

In particular, True Default (TD) and True Non-default (TN) are respectively the number of defaults and non-defaults that are predicted correctly. Conversely, False Default (FD) indicates the number of predicted defaults that do not occur and False Non-default (FN) is the number of predicted non-defaults that actually default.

The total number of defaults in the credit portfolio is indicated by D and the total number of non-defaults by ND. For a given cut-off s^*, the false positive rate is defined as

$$\hat{F}_n(s^*) = \frac{FD}{ND}$$

	Actual default	Actual non-default
Default forecast (score below s^*)	TD	FD
Non-default forecast (score above s^*)	FN	TN
	D	ND

Table 1: Contingency table or confusion matrix.

and the true positive rate is

$$\hat{F}_d(s^*) = \frac{TD}{D}.$$

The alarm rate is given by

$$\hat{F}(s^*) = \frac{TD + FD}{D + ND}$$

and it represents the proportion of defaulters. For different cut-off values, any model would exhibit different performances; thus, contingency tables could be used as a means of assessing competing models only for a given cut-off value s^*. In order to represent the model performance for all possible cut-off values, the most popular graphic representation is the Receiver Operating Characteristic (ROC) curve (BCBS, 2005).

2.1 ROC curve

The ROC curve is defined as the plot of the non-diagonal elements combination of a contingency table for all possible cut-off points. This means that the ROC curve is represented by the plot of the true positive rate on the vertical axis, versus the false positive rate on the horizontal axis, for all possible cut-off points

$$ROC(u) = F_d[F_n^{-1}(u)], \qquad u \in (0, 1).$$

In Figure 1, the ROC curve is plotted. A perfect model would correctly predict the full number of defaults and it is represented by the horizontal line at the unit true positive rate. On the other side, a model with zero predictive power is represented by

the straight line 45. Finally, any other case of some predictive power is represented by a concave curve positioned between the two extreme cases.

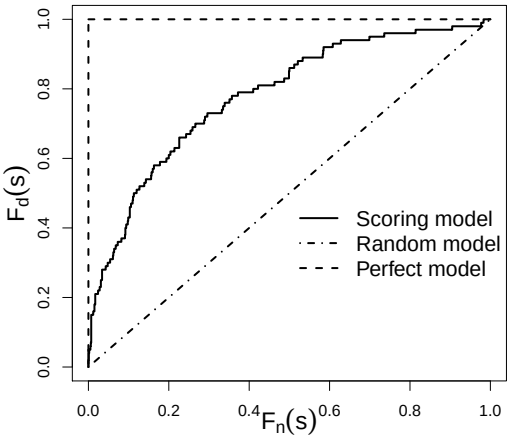


Figure 1: The Receiver Operating Characteristic (ROC) curve.

In the case that the ROC curve of a particular model lies uniformly above the ROC curve of a competing model, the former exhibits superior discriminatory power for all possible cut-off points. In analytic terms, this relationship is defined as follows.

Definition 2.1. The credit scoring model S^1 *ROC dominates* the scoring model S^2 whenever $ROC^1(s) > ROC^2(s)\ \forall s \in \mathbb{R}$.

In the case that the two curves intersect, it is not clear which model has the higher discriminatory power. The slope of the ROC at each point on the curve is the ratio of the probability density functions $f_d(s)$ and $f_n(s)$ for a given score s (Tasche, 2002).

A drawback of the ROC curve is that it assumes equal misclassification error costs for Type I and II Errors (Provost and Fawcett, 2001). This assumption could be very risky for banks. The reason is that it is much more costly to classify a borrower as non-defaulter when he is a defaulter than to classify a borrower as defaulter when he is a non-defaulter (Stein, 2005). In particular, when a defaulted borrower is classified as non-defaulter by scoring models, banks give him a loan. When the borrower becomes defaulter, the bank may lose the whole or part of the

credit exposure, which represents the costs corresponding to Type I error for False Negative. On the contrary, when a non defaulter is classified as defaulter, the bank loses only the interest on loans.

The aim of this work is to propose a curve that incorporates misclassification error costs.

2.2 Classifier performance measures

In order to validate a credit model, how the discriminatory power can be measured is a trivial question. From Figure 1, the stronger the slope of the ROC curve for $F_n(\cdot)$ is close to 0, implying the default probability estimate being close to 1 for low scores (Tasche, 2006), and the weaker the slope of the respective curve for $F_n(\cdot)$ is close to 1, implying the default probability estimate being close to 0 for high scores, the distribution functions of S $F_d(\cdot)$ and $F_n(\cdot)$ differ more and the discriminatory power of the underlying score variable S is better. For the assessment of credit model performance, a synthetic index of the discriminatory power for all possible cut-offs is known as Area Under the Curve (AUC) (see Kraznowski and Hand, 2009).

From Figure 1, it is intuitively clear that the area between the axis of abscissa and the ROC curve can be considered a measure of discriminatory power

$$AUC = \int_0^1 ROC(u)du.$$

It takes values in the [0.5,1] interval where the two bounds correspond to models with zero and full discriminatory power, respectively. By normalizing the AUC index $\dfrac{AUC - 0.5}{0.5} = G$ the Gini index G is obtained.

Since the AUC does not incorporate the costs of misclassification errors, the minimum of the Probability-Weighted (PW) loss function is used (Hand and Henley, 1997; Hand and Vinciotti, 2003)

$$\min_s \{C(FN)p[1 - F_d(s)] + (1 - p)C(FD)F_n(s)\} \quad s \in \mathbb{R} \tag{2.1}$$

where p is the default probability and C(FN) and C(FD) are the costs corresponding to Type I error for FN $1 - F_d(s)$ and Type II error for FD $F_n(s)$, respectively. High values of the minimum of the PW loss function correspond to lower classifier

performance. The optimal cut-off s^* that minimizes the PW loss function coincides with the one that maximizes the expected profit (Beling et al, 2005; Crook et al., 2007; Oliver and Wells, 2001; Oliver and Thomas, 2009).

Since the minimum of the PW loss function is dependent on the sample default probability, banks and regulators cannot apply this assessment for monitoring credit scoring over time and across credit portfolios when the default probability changes. Moreover, since default is a rare event (Calabrese and Osmetti, 2011), the important costs C(FN) for banks are multiplied by a too small value p, this could imply the underestimation of the losses for FN.

The classifier performance measure proposed in the next section aims at overcoming this disadvantage.

3 New classifier performance assessments

3.1 The curve of Misclassification Error Loss (MEL) and the area under the MEL curve

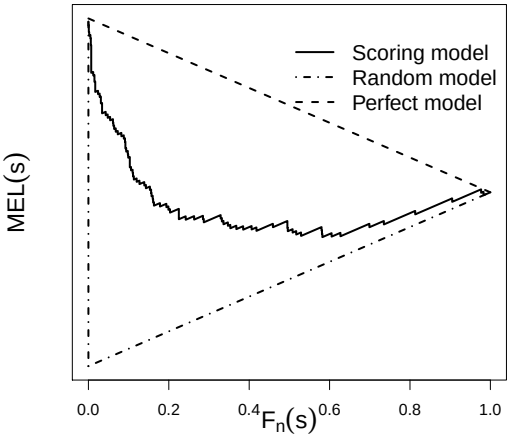


Figure 2: The Misclassification Error Loss (MEL) Curve.

The first aim of this section is to propose a curve that does not depend on the default probability (a sample characteristic) but depends on the misclassification

error costs. In terms of the conditional cumulative distributions of scores, the curve of Misclassification Error Loss (MEL) is proposed, defined as

$$MEL(u) = \frac{C(FN)}{C(FD)}[1 - F_d(s)] + F_n(s) = k[1 - F_d(s)] + F_n(s) \quad s \in \mathbb{R} \qquad (3.1)$$

where k is the ratio $C(FN)/C(FD)$ of the costs of misclassification errors. We point out that the costs $C(FN)$ are often much higher than $C(FD)$ since the first depends on the loss given default and the workout fees on default, by contrast the latter depends on the interest spread. This means that the costs ratio $k = C(FN)/C(FD)$ is usually higher than 1.

Unlike the PW loss function (2.1), Type I and II errors are not weighted by the probabilities p and $1 - p$, since it would imply the underestimation of the loss for Type I error when p is too small. As above-mentioned, the measurements of the discriminatory power of scoring models should be independent from the characteristic of the credit portfolio, such as the proportion of defaulters.

From Figure 2, we highlight that all the MEL curves pass through the points $(0, k)$ and $(1, 1)$. In particular, the MEL curve of the random model, with zero discriminatory power, is given by the dotted line that joins the points $(0, k)$ and $(1, 1)$. By contrast, the MEL curve of the perfect credit model is given by two dotted lines, the first joining the points $(0, k)$ and (0,0), the second joining the points (0,0) and $(1, k)$. Any other model with some predictive power is given by a curve positioned between the two extreme cases.

In the case that the MEL curve of a particular model lies uniformly above the MEL curve of a competing model, the latter exhibits superior discriminatory power for all possible cut-off points. In analytic terms, this relationship is defined as follows.

Definition 3.1. The credit scoring model S^1 MEL dominates the scoring model S^2 whenever $MEL^1(s) > MEL^2(s)\ \forall s \in \mathbb{R}$.

In the case that the two curves intersect, it is not clear which model shows the higher discriminatory power.

Proposition 3.1. *The ROC dominance is equivalent to the MEL dominance.*

Proof. At first we prove that if the scoring model S^1 ROC dominates the scoring model S^2 then S^1 MEL dominates S^2. From the definition 2.1 we deduce that

$$F_d^1(s) > F_d^2(s) \quad \forall\, s \in \mathbb{R}. \tag{3.2}$$

The condition (3.2) can be written also as

$$1 - F_d^1(s) < 1 - F_d^2(s) \quad \forall\, s \in \mathbb{R}. \tag{3.3}$$

By multiplying both sides of the inequality (3.3) for the costs ratio k and by summing up the probability $F_n(s)$ which takes the same value for both scoring models S^1 and S^2, it is obtained

$$k\{1 - F_d^1[F_n^{-1}(u)]\} + F_n^1(s) < k\{1 - F_d^2[F_n^{-1}(u)]\} + F_n^2(s) \quad \forall\, s \in \mathbb{R}. \tag{3.4}$$

Analogously, we can prove that if the scoring model S^1 MEL dominates the scoring model S^2. that the scoring model S^1 MEL dominates the scoring model S^2. □

The previous result is coherent with the one obtained by Beling et al. (2005) that the ROC dominance is equivalent to the expected-profit dominance.

Proposition 3.2. *The slope of the MEL curve is*

$$\frac{\partial CC[F_n(s)]}{\partial F_n(s)} = -k\frac{f_d(s)}{f_n(s)} + 1. \tag{3.5}$$

Proof. The following results are useful in order to compute the slope of the MEL curve. From the equation

$$1 = \frac{\partial F_n(s)}{\partial F_n(s)} = \frac{\partial F_n(s)}{\partial s}\frac{\partial s}{\partial F_n(s)} = f_n(s)\frac{\partial s}{\partial F_n(s)}$$

we get the result

$$\frac{\partial s}{\partial F_n(s)} = \frac{1}{f_n(s)}. \tag{3.6}$$

By applying the equation (3.6), we derive

$$\frac{\partial F_d(s)}{\partial F_n(s)} = \frac{\partial F_d(s)}{\partial s}\frac{\partial s}{\partial F_n(s)} = f_d(s)\frac{\partial s}{\partial F_n(s)} = \frac{f_d(s)}{f_n(s)}. \tag{3.7}$$

By considering the equation (3.7), we obtain

$$\frac{\partial CC[F_n(s)]}{\partial F_n(s)} = -k\frac{\partial F_d(s)}{\partial F_n(s)} + 1 = -k\frac{f_d(s)}{f_n(s)} + 1.$$

□

By setting the slope (3.5) of the MEL curve equal to zero, the score s at which the MEL curve reaches its minimum satisfies the following equation

$$\frac{f_d(s)}{f_n(s)} = \frac{C(FD)}{C(FN)}.$$

In order to understand the behaviour of the MEL curve, it is useful that Tasche (2006) proves that the score variable S is optimal in a test-theoretic sense if and only if the likelihood ratio $\frac{f_d(s)}{f_n(s)}$ is monotonous. If high scores indicate high credit-worthiness, the score density function for defaulters $f_d(s)$ is small for high scores and large for low scores and the score density function for non-defaulters $f_n(s)$ is large for high scores and small for low scores. This means that the likelihood ratio $\frac{f_d(s)}{f_n(s)}$ is decreasingly monotonous. From this result and the equation (3.5) it is deduced that the MEL curve is decreasing for scores lower than the one at which the MEL curve has the minimum and it is increasing for scores higher, as Figure 3 shows.

Figure 2 shows that the area between the MEL curve and the dotted line of the perfect model that joins the points $(0, k)$ and $(1, 1)$ represents a classifier performance measure. The relationship between the AUC and the area under the MEL curve is shown by the following proposition.

Proposition 3.3. *The normalized area under the MEL curve is equal to the Gini index G.*

Proof. The area under the MEL curve is

$$\int_0^{F_n(s)} MEL[F_n(s)]dF_n(s) = \int_0^1 MEL[F_n(s)]dF_n(s) - \int_{F_n(s)}^1 MEL[F_n(s)]dF_n(s). \tag{3.8}$$

We compute the two integrals on the left side of the equation (3.8)

$$\int_0^1 MEL[F_n(s)]dF_n(s) = \int_0^1 \{k[1-F_d(s)]+F_n(s)+F_n(s)\}]dF_n(s) = k-k\,AUC+0.5 \tag{3.9}$$

$$\int_{F_n(s)}^M EL[F_n(s)]dF_n(s) = 0.5 \tag{3.10}$$

By substituting the results (3.9) and (3.10) in the equation (3.8), the following result is obtained

$$\int_0^{F_n(s)} MEL[F_n(s)]dF_n(s) = k - k\,AUC. \tag{3.11}$$

By normalizing the area under the MEL curve the Gini index G is obtained

$$\frac{k - k\,AUC}{0.5\,k} = 2 - AUC = G$$

□

3.2 The minimum of the MEL curve

A coherent classifier performance measure with the MEL curve that is not dependent on the sample characteristics and considers the misclassification costs of Type I and II errors is the minimum of the MEL curve

$$\min_s \{k[1 - F_d(s)] + F_n(s)\} = \max_s [kF_d(s) - F_n(s)]\,. \tag{3.12}$$

Proposition 3.4. *The probability density function of the minimum of the MEL curve is*

$$f_S(s) = \begin{cases} \frac{m(1+s)}{k}\left[\frac{1}{k}\left(\frac{(s)^2}{2} + s + \frac{1}{2}\right)\right]^{m-1} & -1 \le s < 0; \\ \frac{m}{k}\left[\frac{1+2s}{2k}\right]^{m-1} & 0 \le s < k-1; \\ \frac{m(-s+k)}{k}\left[\frac{1}{k}\left(-\frac{(s)^2}{2} + ks + \frac{-k^2+2k}{2}\right)\right]^{m-1} & k-1 \le s < k; \\ 0 & otherwise \end{cases} \tag{3.13}$$

where m is the number of the points at which the differences $kF_d(s) - F_n(s)$ are calculated.

Proof. Let $F_d(S) = U$ and $F_n(S) = V$. Therefore, U and V are two continuous uniform random variables with support [0,1]. We consider the following transformation

$$\begin{cases} T = V, \\ Z = kU - V \end{cases}$$

We compute the joint density function

$$f_{TZ}(t,z) = |J| f_{VU}\left(t, \frac{z+v}{k}\right) = \frac{1}{k} I_{(0,1)}(t) I_{(0,1)}(\frac{z+v}{k}) \tag{3.14}$$

where J is the Jacobian of the transformation and U and V are independent random variables.

To find the marginal density function of Z we integrate out t

$$f_Z(z) = \int_{-\infty}^{\infty} f_{TZ}(t,z)dt = \begin{cases} \frac{1+z}{k} & -1 \leq z < 0; \\ \frac{1}{k} & 0 \leq z < k-1; \\ \frac{-z+k}{k} & k-1 \leq z < k. \end{cases} \tag{3.15}$$

In the previous result we consider $k \geq 1$ since k is equal to the ratio $\dfrac{C(FD)}{C(FN)}$ of misclassification error costs. From the probability density function (3.15) we compute the cumulative distribution function of Z

$$F_Z(z) = \begin{cases} \frac{1}{k}\left(\frac{z^2}{2} + z + \frac{1}{2}\right) & -1 \leq z < 0; \\ \frac{1+2z}{2k} & 0 \leq z < k-1; \\ \frac{1}{k}\left(-\frac{z^2}{2} + kz + \frac{-k^2+2k}{2}\right) & k-1 \leq z < k; \\ 1 & z \geq k. \end{cases} \tag{3.16}$$

The probability density function of the maximum of Z is (Herbert and Nagaraja, 2003)

$$f_{max\{Z\}}(z) = m[F_Z(z)]^{m-1} f_Z(z) \tag{3.17}$$

where m is the number of the points at which the differences $kF_d(s) - F_n(s)$ are calculated. By substituting the equations (3.15) and (3.16) in the equation (3.17), we obtain the expression (3.13). □

When $k = 1$, the expression (3.12) is the Kolmogorov-Smirnov statistic (Gibbons, 1971) for testing $H_0 : F_d(s) = F_n(s) \quad vs \quad H_1 : F_d(s) > F_n(s)$. Kraznowski and Hand (2009) consider the Kolmogorov-Smirnov statistic for the maximum vertical distance for ROC curve.

4 Simulation results

Based on our knowledge, few simulations (e.g. Satchell and Xia, 2007; Stein and Jordao, 2003; Stein, 2005) are performed in the literature on the accuracy of credit scoring models. We generate 1,000 samples of credit scores of both defaulters and non-defaulters from two random variables and we denote these as samples 1. Then, we generate 1,000 samples of credit scores from the same parametric model but with different parameters in order to change the classifier performance of the scoring model. We denote this second set of samples as samples 2. Both the minima of the PW loss function (2.1) and of the MEL curve (3.12) are computed on the simulated samples 1 and 2. The value of a given measure evaluated on the samples 1 are compared with those of the same measure evaluated on the samples 2. Hence, we compute the proportions

$$\frac{\sharp\{CPM_1 > CPM_2\}}{1,000} \tag{4.1}$$

for each pair of samples, where CPM_1 and CPM_2 are the same Classifier Performance Measure (CPM) evaluated on the sample 1 and 2, respectively. These values are reported in Table 2.

Similarly to Satchell and Xia (2007), we consider two different sample sizes (500 and 1000) and two different default proportions (0.05 and 0.01). The default probability of 0.05 is chosen since it represents the default percentage for Italian SMEs (Cerved Group, 2011) examined in the following section. The ratio of misclassification error costs is considered equal to 2 in the following simulations.

Analogously to Satchell and Xia (2007), the first parametric model for credit scores is given by the normal distribution $N(\mu, \sigma^2)$ with expectation μ and variance σ^2. At first, the score of defaulters and non-defaulters are simulated from the normal distributions $N_D(0, 1)$ and $N_N(1, 1)$. In order to increase the classifier performance of the scoring model, the mean of the distribution of non-defaulters for the sample 2 is increased. Therefore, we simulate the defaulters' scores from the normal distribution $N_D(0, 1)$ and the non-defaulters' scores from $N(1.1, 1)$.

Much empirical evidence shows asymmetric distributions of the scores for defaulters and non-defaulters (e.g. Christodoulakis and Satchell, 2006), even the score

sample sizes	*PD*	$N_D(0,1);N_N(1,1)$ $N_D(0,1);N_N(1.1,1)$	$SN_D(0,2,-0.5);\ SN_N(0,2,0.5)$ $SN_D(0,2,-0.5);\ SN_N(0.1,2,0.5)$
500	.05	1	0.552
500	.01	0.571	0.467
1000	.05	1	0.635
1000	.01	0.747	0.228

Table 2: The results of Monte Carlo simulations on 1,000 samples where $N(\mu, \sigma^2)$ indicates the normal random variable with expectation μ and variance σ^2; $SN(\xi, \omega, \alpha)$ indicates the skewed normal random variable where ξ is the location parameter ($\xi \in \mathbb{R}$), ω is the scale parameter ($\omega \in \mathbb{R}^+$) and α is the shape parameter ($\alpha \in \mathbb{R}^+$).

distributions in the empirical evidence of this paper show these characteristics. For this reason, Christodoulakis and Satchell (2006) analyze the theoretical characteristics of the ROC curve when the credit scores follow a skew normal distribution. Hence, we generate the scores also from two skew normal random variables (Azzalini, 1985) $SN(\xi, \omega, \alpha)$ where ξ is the location parameter ($\xi \in \mathbb{R}$), ω is the scale parameter ($\omega \in \mathbb{R}^+$) and α is the shape parameter ($\alpha \in \mathbb{R}^+$). In particular, we generate the defaulters' scores from the skew normal distribution $SN_D(0, 2, -0.5)$ and the non-defaulters' scores from $SN_N(0, 2, -0.5)$. By increasing the distance between the expectations of the scores S_D and S_{ND} we increase the classifier performance of the scoring model, so we simulated the score samples from skew normal distributions $SN_D(0, 2, -0.5)$ and $SN_D(0.1, 2, 0.5)$.

Since the proportions (4.1) computed by applying the minimum of the MEL curve are equal to one for all the pairs of simulated samples, we do not report these values in Table 2. This means that the minimum of the MEL curve always shows the same ordering of the classifier performances of the scoring models.

From Table 2 we can deduce that our proposal is preferable to the minimum of PW loss for most of the couples of generated samples. When the proportion of defaulters is 0.05 and we simulate from normal distributions, both the methods show similar performance. Compliant with the expectations, by decreasing the

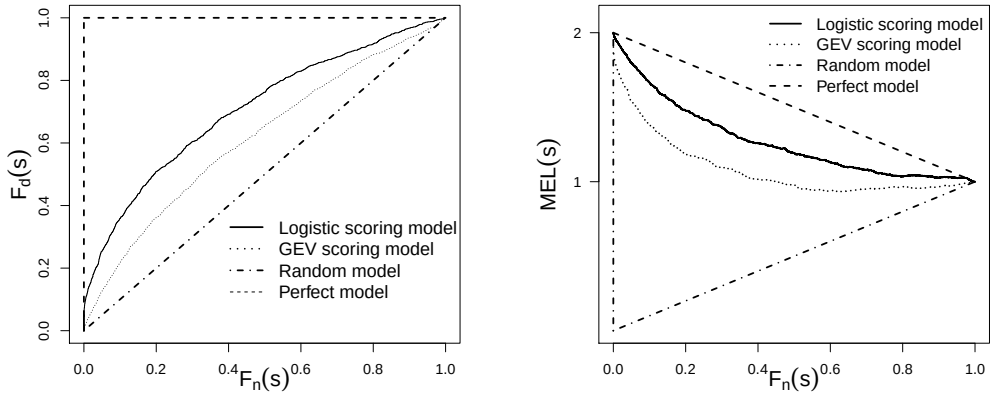
proportion of defaulters the minimum of the PW loss shows inadequate performance. Similarly, the results for the skew normal distributions show worse performance of the minimum of the PW loss when the proportion of defaulters is low (0.01). It is important to underline that when the score distributions are skewed, the minimum of the PW loss shows inadequate performance even when the proportion of defaulters is 0.05. We obtain the same result in the following section. We can thus conclude that our proposal is robust for different numbers of defaulters.

5 Empirical evidence

SMEs play a very important role in the economic system of many countries and particularly in Italy (about 90% of Italian firms are SMEs (Vozzella, Gabbi 2010)). Furthermore, Basel II (BCBS, 2004) establishes that banks should develop credit risk models specifically addressed to SMEs. Only a few studies consider SMEs (e.g. Altman and Sabato, 2007; Altman et al. 2010; Ciampini and Gordini, 2008; Vozzella and Gabbi, 2010) since the gathering of SMEs data is quite difficult.

Data used in our analysis comes from AIDA-Bureau van Dijk, a large Italian financial and balance sheet information provider. We consider Italian defaulted and non-defaulted SMEs over the years $2005 - 2009$. In particular, since the default probability is one-year forecasted, the covariates concern the period of time $2004 - 2008$. The database contains accounting data of approximately 210,000 Italian firms with total assets below 10 million euros (Vozzella and Gabbi, 2010). From the sample we exclude the firms without the necessary information on the covariates.

We consider a default occurred when a specific firm enters a bankruptcy procedure as defined by the Italian law (Altman and Sabato, 2007). In accordance with Altman and Sabato (2007) we apply a choice-based or endogenous stratified sampling on this dataset. In this sampling scheme, data are stratified by the values of the response variable. We randomly draw the observations within each stratum defined by the two categories of the dependent variable (1=default, 0=non-default) and we consider all the defaulted firms. Then, we select a random sample of non-defaulted firms over the same year of defaults in order to obtain a percentage of



(a) ROC curve.

(b) MEL curve.

Figure 3: Plots on data on Italian SMEs (165 defaults and 3,300 non-defaults) over the years $2005 - 2009$.

defaults in our sample as close as possible to the default percentage (5 %) for Italian SMEs (Cerved Group, 2011).

We apply the logistic regression model (McCullagh and Nelder, 1989) and the Generalized Extreme Value (GEV) regression model proposed by Calabrese and Osmetti (2011) to forecast the probability of default.

In order to model the default event, we choose the independent variables that represent the financial and economic characteristics of firms according to the recent literature (Vozzella and Gabbi, 2010; Ciampi and Gordini, 2008; Altman and Sabato, 2007). These covariates cover the most relevant aspects of firm's operations: leverage, liquidity and profitability. By applying the GEV model, 7 variables are significant at the level of 5% for the PD forecast: *Solvency ratio* (the ratio of a company's income over the firm's total debt obligations); *Return on investment* (the ratio of the returns of a company's investments over the costs of the investment); *Turnover per employee* (the ratio of sales divided by the number of employees); *Added value per employee* (the enhancement added to a product or service by a company divided by the number of employees); *Cash flow* (the amount of cash generated and used by a company in a given period); *Bank loans over turnover* (short and long term debts with banks over sales volume net of all discounts and sales

taxes); *Total personnel costs over added value* (the ratio of a company's labor costs divided by the enhancement added to a product or service by a company).

Since the developed models may overfit the data, resulting in over-optimistic estimates of the predictive accuracy, the validation is applied on a sample (3,465 SMEs), called out-of-sample sample, which is different from that used in estimating the model parameters (31,600 SMEs). The out-of-sample is randomly drawn.

Since the GEV model is proposed to classify correctly the defaulters (Calabrese and Osmetti, 2011), the (global) classifier performance of the GEV model is worse than the one of the logistic model for every cut-off, as both the ROC and the MEL curve show in Figure 3. This result is also shown by the AUC that is equal to 0.615 for the GEV model and 0.708 for the logistic model.

Models	*Minimum of the MEL curve*	*Minimum of the PW loss*
GEV method	0.9976	0.0935
logistic method	0.9304	0.0998

Table 3: The minima of the MEL curve and of the PW loss function for the logistic and the GEV models on 3,115 Italian SMEs.

We compute the minima of the MEL curve and of the PW loss curve by considering a ratio of misclassification error costs equal to 2 and we report these values in Table 3. Even if the difference between the AUCs for the two models is high, the minimum of the PW loss function shows incorrectly that the classifier performance of the GEV model is better than the one of the logistic regression model.

6 Conclusions

In this work we overcome some main problems of the validation of scoring models. At first, we propose the MEL curve to represent the discriminatory power of rating models whose shape depends on the ratio of the misclassification error costs. Our proof shows that the ROC dominance is equivalent to the MEL dominance. Moreover, we derive that the normalized area under the MEL curve is the Gini index.

In coherence with the MEL representation, we suggest a measure to evaluate the classifier performance that is not affected by the number of defaults in the portfolio. We derive also the probability density function of the suggested discriminatory power index. Monte Carlo simulations show that our proposal is definitely preferable to the minimum of the weighted-probability loss for skewed score distributions. Finally, the same result is obtained by an empirical analysis on Italian SMEs.

This work is important since we suggest classifier performance assessments that allows to monitor credit scoring models for different numbers of defaulters. Since simulation studies on the validation of rating models concern only symmetric credit score distributions, another innovative aspect of this model is that the Monte Carlo simulations are performed by drawing from skewed distributions of the credit scores. Finally, a further relevant contribution of this paper is the application of the methodological proposals to data on Italian SMEs.

References

Adams, N. M., Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. Pattern Recognition, 32, 1139-1147.

Altman, E., Haldeman R., Narayanan P. (1977). ZETA analysis: a new model to identify bankruptcy risk of corporations. Journal of Banking & Finance, 1, 2954.

Altman, E., Sabato, G. (2007). Modeling Credit Risk for SMEs: Evidence from the US Market, ABACUS, 19(6), 716-723.

Altman, E., Sabato, G., WIlson, N. (2010). The value of non financial information in small and medium-sized enterprise risk management. Journal of Credit Risk, 6(2), 95-127.

Azzalini, A. (1985). A class of distributions which includes the normal ones. Scandinavian Statistical Journal, 12, 171-178.

Basel Committee on Banking Supervision (2005). Studies on the Validation of Internal Rating Systems. Working paper 14. Basel, BIS.

Basel Committee on Banking Supervision (2004). International Convergence of Capital Measurement and Capital Standards: A Revised Framework. June, Basel, BIS.

Beling, P., Covaliu, Z., Oliver R. M. (2005) Optimal Scoring Cutoffs Policies and Efficient Frontiers. Journal of Operational Reserach Society, 56 (9), 1016-1029.

Calabrese, R., Osmetti, S. A. (2011) Generalized Extreme Value Regression for Binary Rare Events Data: an Application to Credit Defaults. Working paper. Geary Institute. University College Dublin.

Cerved Group (2011). Caratteristiche delle imprese, governance e probabilit di insolvenza. Report. Milan, February.

Ciampi, F., Gordini, N. (2008). Using Economic-Financial Ratios for Small Enterprize Default Prediction Modeling: an Empirical Analysis. Oxford Business & Economics Conference, Oxford.

Christodoulakis, G. A., Satchell, S. E. (2006). Assessing the Accuracy of Credit R.O.C. Estimates in the Presence of Macroeconomic Shocks. Working Paper.

Crook, J.N., Edelman, D.B., Thomas, L.V. (2007). Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183, 1147-1465.

Dryver,A. L., Sukkasem, J. (2009). Validating risk models with a focus on credit scoring models. *Journal of Statistical Computation and Simulation* 79, 181-193.

Engelmann, B., Hayden, E., Tasche, D. (2003). Testing rating accuracy. *Risk* 16, 82-86.

Fantazzini, D., Figini, S. (2008) Random Survival Forest models for SME Credit Risk Measurement. *Methodology and computing in applied probability*, 11, 29-45.

Fawcett, T. (2003). ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. *HP Laboratories Working Paper.*

Hand, D.J., Henley W.E. (1997). Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society, Ser A 160, 523-541.

Hand, D.J., Vinciotti, V. (2003). Local versus Global for Classification Problems: Fitting Models Where it Matters. The American Statistician, 57(2), 124-131.

Herbert, A. D., Nagaraja, H. N.(2003) Order Statistics. Wiley

Kraznowski, W. J., Hand, D. J. (2009). ROC Curves for Continuous Data. Taylor & Francis, Inc. Boca Raton.

Mann, H. B., Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics. 18, 50-60.

McCullagh P., Nelder J.A. (1989). Generalized Linear Model, Chapman Hall, New York.

Oliver, R.M., Wells, E.R. (2001) Efficient frontier cutoff policies incredit portfolios. Journal of Operational Reasearch Society, 52, 1025-1033.

Oliver, R.M. and Thomas, L.C. (2009) Optimal score cutoffs and pricing in regulatory capital in retail credit portfolios. Southampton, UK, University of Southampton, Discussion Papers in Centre for Risk Research.

Provost, F., Fawcett, T. (2001). Robust Classification for Imprecise Environment. Machine Learning 42 (3), 203-231.

Satchell, S., Xia, W. (2007). Analytic Models of the ROC Curve: Applications to Credit Rating Model Validation. Woking paper.

Sobehart, J., Keenan, S. (2001). Measuring default accurately. Credit Risk Special Report, Risk, 14, 31-33.

Stein, R. M. (2002). Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation. Moody's KMV Technical Report.

Stein, R. M., Jordao, F. (2003). What is a more powerful model worth? Technical Report, Moody's KMV, New York.

Stein, R. M. (2005). The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. Journal of Banking & Finance. 29, 1213-1236.

Tasche, D. (2002). Remarks on the monotonicity of defaults probabilities. Deutsche Bundesbank Working paper.

Tasche, D. (2006). Validation of internal rating systems and PD estimates. Deutsche Bundesbank Discussion Paper.

Thomas, L., Edelman, D., Crook, J.C. (2002). Credit Scoring and Its Applications. Society for Industrial and Applied Mathematics, Philadelphia.

Vozzella, P., Gabbi G. (2010). Default and Asset Correlation: An Empirical Study for Italian SMEs. Working Paper.