UCD GEARY INSTITUTE DISCUSSION PAPER SERIES

Cost-sensitive classification for rare events: an application to the credit rating model validation for SMEs

Raffaella Calabrese
Dynamics Lab
Geary Institute
University College Dublin

Geary WP2011/34
November 2011

Cost-sensitive classification for rare events: an application to the credit rating model validation for SMEs

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

November 25, 2011

Abstract

Receiver Operating Characteristic (ROC) curve is used to assess the discriminatory power of credit rating models. To identify the optimal threshold on the ROC curve, the iso-performance lines are used. The ROC curve and the iso-performance line assume equal classification error costs and that the two classification groups are relatively balanced. These assumptions are unrealistic in the application to credit risk. In order to remove these hypotheses, the curve of Classification Error Costs is proposed. Coherent with this curve, a methodology to identify the optimal threshold is suggested. Monte Carlo simulations that reproduce similar characteristics to the empirical credit scoring models for SMEs show that our proposal performs better that the iso-performance line. Finally, we apply the suggested methodologies to empirical data on Italian Small and Medium Enterprises (SMEs).

1 Introduction

A significant innovation to 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.

The credit risk model adequacy is usually based on the ability of rating or scoring models to discriminate *ex ante* between defaulting and non-defaulting borrowers that represents the discriminatory power of the credit model. The assessment of the discriminatory power of a credit model represents the validation process. The most popular validation methodologies are the Cumulative Accuracy Profile (CAP) and the Receiver Operating Characteristic (ROC) curves (Kraznowski and Hand, 2009). In order to evaluate the accuracy of internal scoring models, banks and supervisors should apply methodologies whose results do not depend on sample characteristics, so they can compare the accuracy of different scoring models. Since the CAP curve depends on the sample frequency of defaulters, we analyse only the ROC curve in this article. A discriminatory power index coherent with the ROC curve is the Area Under the Curve (AUC). The threshold of optimal classification accuracy on the ROC curve can be found by using iso-performance tangent lines, which are based on the accuracy measure.

The above-mentioned methodologies rely upon two unrealistic assumptions in credit risk analysis. The use of accuracy assumes that the two classification groups are relatively balanced (Vinciotti and Hand, 2003). By applying the validation methodologies to scoring models, the two groups are represented by the defaulting and non-defaulting borrowers. In a large population of debtors the percentage of defaulting is very low (Hand and Henley, 1997). Since the credit default is a rare event (Calabrese and Osmetti, 2011), the distribution of defaults is very skewed (Provost and Fawcett, 2001). As the default distribution becomes more skewed, by using the iso-performance line the evaluation of the optimal cut-off on the ROC curve breaks down, as explained in section 2.

The second drawback of these validation methodologies is the assumption of equal classification error costs. For banks 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. This means that the first classification cost is much higher that the latter, e.g. Altman et al. (1977) obtain a ratio of the misclassification costs equal to 31.

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 that represents the discriminatory power of a scoring model and a method to compute the optimal threshold. In particular, we propose the curve of *Classification Error Costs* (CEC) that represents graphically the discriminatory power when the cut-off changes and its shape depends on the ratio of the classification error costs. Coherent with the CEC curve, we suggest a methodology to obtain the optimal cut-off that depends on the ratio of the classification error costs, but not on the sample frequency of defaulters. Both these characteristics are important in the application to credit risk and the latter one allows banks to compare the accuracy of different rating models evaluated on different samples. Furthermore, the ratio of classification error costs is usually known, unlike the misclassification costs (Adams and Hand, 1999).

Two important theoretical results obtained in this work are that the ordering of the ROC curves is preserved by the ordering of CEC curves and that the normalized area under the CEC curve is equal to the Gini index. Finally, the slope of the CEC curve is obtained.

For the method proposed for the calculation of the optimal cut-off, in this work we derive the probability density function of the optimal threshold. This method is compared with the iso-performance line 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 optimal cut-off by drawing from skewed score distributions and by considering a high ratio of misclassification costs and low relative frequency of defaulters. In these cases the simulation results show that our proposal exhibits much better performance than the one of the iso-performance line.

Another innovative aspect of this paper is the application of the methodological proposals to Italian Small and Medium Enterpriese (SMEs). Basel II (BCBS, 2004) establishes that banks should develop credit risk models specifically addressed to SMEs. To the authors' knowledge, no 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 exhibits much lower classification error costs than the iso-performance line. Therefore, our proposal leads to a better classification of defaulters that represents a pivotal aim for banks.

The present paper is organized as follows. Section 2 analyzes the ROC curve, the AUC index and the method to identify the optimal threshold based on the classification accuracy. In section 3 the CEC curve and a methodology to compute the optimal cut-off are suggested. In the following section we compare the properties of our proposal to those of the iso-performance line by simulations. Successively, Section 5 presents the database of Italian SMEs, to which we apply the suggested methodologies, Particularly, we present the main empirical results and we compare our proposal with the iso-performance line. Finally, the last section is devoted to conclusions.

2 Validation methodologies for default risk models

It is assumed that two random variables are associated. The first variable S represents the score on a continuous scale that is assigned to the borrower and which is intended to forecast the borrower's creditworthiness. The distribution function of the random variable S is denoted by $F(s)$ and its probability density function is $f(s)$. Since the second variable B represents the borrower's future state at the end of a fixed time-period, it is a Bernoulli random variable

$$B = \begin{cases} 1, & \text{the borrower's state is default } (d); \\ 0, & \text{the borrower's state is non default } (n). \end{cases}$$

The population of borrowers is divided in two groups: the future defaulters and the borrowers who remain solvent in the future. Hence, borrowers with $B = 1$ belong to the population of defaulters and borrowers with $B = 0$ belong to the population

of non-defaulters. The conditional distribution functions of S given a value of B are denoted respectively by $F_d(\cdot)$ and $F_n(\cdot)$. Analogously, the conditional probability density functions of S given a value of B are indicated by $f_d(\cdot)$ and $f_n(\cdot)$. Therefore, the distribution function of the score S is given by

$$F(s) = pF_d(s) + (1-p)F_n(s)$$

where p is the probability of default $p = P[B = 1]$.

The institution's intention with the score variable S is to forecast the borrower's future state B by relying on the information on the borrower's creditworthiness that is summarized in S. This means that the scoring and rating models are used to decide which debtors will survive during the next period and which debtors will default. One possibility for the decision-maker would be to introduce a *cut-off* value s^* and to classify each debtor with a score lower than s^* as a potential defaulter and each debtor with a score higher than s^* as a non-defaulter.

The errors of the scoring model are given by $1 - F_d(s^*)$ and $F_n(s^*)$ that 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-defaults (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.

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

Table 1: Contingency table or confusion matrix.

The total number of defaults in the sample 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

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

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 sample relative frequency of predicted defaults. For different cut-off values, any model would exhibit different performances; thus, contingency tables could be used as a mean of assessment of 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 representations are the Cumulative Accuracy Profile (CAP) and the Receiver Operating Characteristic (ROC) curves (BCBS, 2005).

2.1 ROC curve

The ROC curve is defined as the plot of the non-diagonal element 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 2, 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. In the case that

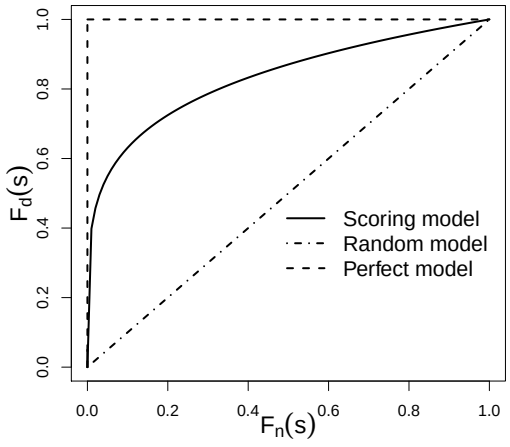


Figure 1: The Receiver Operating Characteristic (ROC) curve.

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 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 the conditional probability density functions $f_d(\cdot)$ and $f_n(\cdot)$ for a given score s (Tasche, 2002). From this property Tasche (2002) proves that the concavity of the CAP and ROC curves is equivalent to the property of the conditional probabilities of default given the underlying scores is a decreasing function of the scores. Hence, non-concavity indicates sub-optimal use of information in the specification of the score function (BCBS, 2005 and Tasche, 2002).

The application of these curves to validate scoring models causes two main problems. The shape of the CAP curve, unlike the ROC curve, is affected by the sample relative frequency of defaulters (Stein, 2005). This means that only ROC curves relating different samples are comparable.

A drawback of both the CAP and ROC curves is that they assume equal classification 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. 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 a 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 overcome these drawbacks.

2.2 The discriminatory power index: AUC

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 u is close to 0, implying the conditional default probability being close to 1 for low scores (Tasche, 2006), and the weaker the slope of the respective curve for u is close to 1, implying the conditional default probability being close to 0 for high scores, the conditional 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 considered.

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.

The area between the ROC curve and the axis of abscissa represents a probability

$$\begin{aligned} AUC &= \int_0^1 F_d[F_n^{-1}(u)]du = \int_{-\infty}^{+\infty} F_d(s)f_n(s)ds = \int_{-\infty}^{+\infty} P\{S_d < s\}f_n(s)ds \\ &= P\{S_d < S_n\}. \end{aligned} \tag{2.1}$$

From this representation, the non-parametric Mann-Whitney test (Mann and Whitney, 1947) for the hypothesis that the conditional distribution $F_n(s)$ first-order stochastically dominates $F_d(s)$ can be applied as a test.

By normalizing the AUC index, it is obtained

$$\frac{AUC - AUC_R}{AUC_P - AUC_R} = \frac{AUC - 0.5}{0.5} = 2AUC - 1 = G \tag{2.2}$$

where AUC_P and AUC_R are, respectively, the indexes for the perfect and the random models and G is the Gini index. The equation (2.2) shows that the AUC is a linear transformation of the Gini index G, as Engelman and al. (2003) show. This means that both statistics contain the same information. As a consequence of this observation, the higher AUC - the higher the discriminatory power of the rating system under consideration - as is the case for G.

2.3 Optimal cut-off on ROC curve

Every point of the ROC curve corresponds to a binary classifier, this means that the performance of the credit model can be analyzed for a given threshold. From such curves, the optimal cut-off C can be deduced and the confusion matrix of Table 1 can be defined. The most used methodology to identify the optimal cut-off considers the classification accuracy (AC) as a quality measure, defined as

$$AC = \left(\frac{D}{D+ND}\right)\frac{TD}{D} + \left(\frac{ND}{D+ND}\right)\frac{TN}{ND} = \frac{TD+TN}{D+ND}. \tag{2.3}$$

From the previous equation the classification accuracy could be regarded as the weighted mean of the true positive and true negative rates and their weights are proportional to the sample sizes of defaults and non-defaults.

In order to apply the classification accuracy as a quality criterion for the choice of the optimal threshold, the equation (2.3) becomes

$$AC = \hat{p}\,\hat{F}_d(s^*) + (1-\hat{p})(1-\hat{F}_n(s^*)) \tag{2.4}$$

From the result (2.4), the straight line connecting the points with equal classification accuracy is obtained, known as the iso-performance or iso-parametric line (Fawcett, 2003)

$$\hat{F}_d(s^*) = \frac{1-\hat{p}}{\hat{p}}\hat{F}_n(s^*) + \frac{1}{\hat{p}}(AC + \hat{p} - 1) \quad \text{for ROC curve,} \tag{2.5}$$

where $\hat{p} = \dfrac{D}{ND+D}$ is the sample relative frequency of defaults in the sample.

Using (2.5), a set of parallel lines is obtained that represents different classification accuracy. The best one goes through the upper left corner and the worst one goes through lower right corner. Hence, from both the CAP and ROC curves the optimal threshold can be deduced by intersecting the iso-performance tangent and the respective curve (Vuk and Curk, 2006). Hong (2009) obtains that the scores corresponding to the intersection of the first tangent line in (2.5) and the CAP curve is the same as that corresponding to the intersection of the second tangent line (2.5) and the ROC curve.

This methodology to compute the optimal threshold shows drawbacks in the application to scoring models. At first, equal classification error costs are assumed. Moreover, since the credit default is a rare event, the distribution of defaults is very skewed. As above-mentioned, Provost and Fawcett (2001) show that as the distribution of scores becomes more skewed, evaluation based on accuracy breaks down. For example, by considering a sample relative frequency of default equals 0.01, a simple rule - always classify as the maximum likelihood class - gives a 99.9% accuracy.

The methodology to identify the optimal threshold proposed in this paper aims at overcoming these drawbacks.

3 Methodological proposals

3.1 Curve of Classification Error Costs

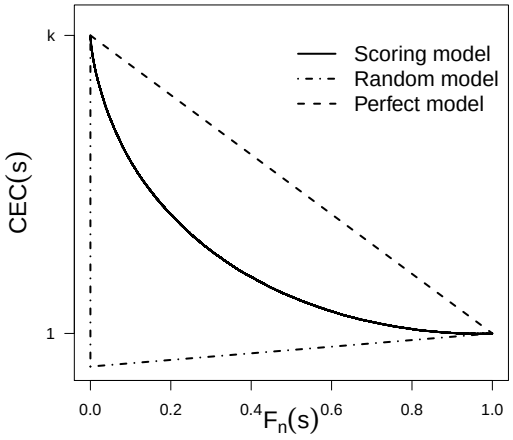


Figure 2: The Classification Error Costs (CEC) Curve.

In section 2.1 the properties of the CAP and ROC curves are analyzed. The main aim of this section is to propose a curve that does not depend on the default probability (a sample characteristic) but depends on the classification error costs. In terms of the distribution of scores, the curve of Classification Error Costs (CEC)

is proposed, defined as

$$CEC(u) = \frac{C(FN)}{C(FD)}\{1 - F_d[s]\} + F_n[s] = k\{1 - F_d[s]\} + F_n[s] \quad s \in (-\infty, +\infty) \quad (3.1)$$

where k is the ratio $C(FN)/C(FD)$ of the costs corresponding to Type I error for FP and Type II error for FD, respectively. Hence, the ratio of the costs $C(FN)$ and $C(FD)$ is assumed to be constant for all borrowers. For a given cut-off C, $1 - F_d(C)$ represents Type I error for FN and $F_n(C)$ indicates Type II error for FD. 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, on the contrary the latter depends on the interest spread. This means that the costs ratio $k = C(FN)/C(FD)$ is often higher than 1.

Unlike the Bayesian error rate (Tasche, 2006), Type I and II errors are not weighted for the prior probabilities p and $1 - p$, since it would imply the underestimation of the total cost for Type I error when p is too small. Furthermore, it is better for the supervisors if the control methodologies of the discriminatory power of the scoring models are independent from the sample characteristic, as the relative frequency of default $\hat{p}$, analogous to our proposal and unlike the Bayesian error rate.

For that reason a simple mean is considered and not a weighted mean of Type I and II errors. In order to understand the difference between these approaches, the same example of Section 2.3 is analyzed. The sample relative frequency of defaulters is 0.01 and the simple rule - always classify as the maximum likelihood class. On the one hand, the accuracy, defined in equation 2.3, is 0.999. On the other hand, the classification costs result 0.5.

From Figure 2, it is highlighted that all the CEC curves pass for the points $\left(0, \frac{C(FN)}{C(FD)}\right)$ and $(1, 1)$. In particular, the CEC curve of the random model, with zero discriminatory power, is given by the dotted line that joins the points $\left(0, \frac{C(FN)}{C(FD)}\right)$ and $(1, 1)$. On the contrary, the CEC curve of the perfect credit model is given by two dotted lines, the first joining the points $\left(0, \frac{C(FN)}{C(FD)}\right)$ and (0,0), the second one the points (0,0) and $\left(1, \frac{C(FD)}{C(FD)}\right)$. Any other model with some predictive power is given by a curve positioned between the two extreme cases.

Analogous to the CAP and ROC curves, in the case that the CEC curve of a particular model lies uniformly above the CEC curve of a competing model, the latter exhibits superior discriminatory power for all possible cut-off points. In the case that the two curves intersect, it is not clear which model has the higher discriminatory power.

Proposition 3.1. *The ordering of the ROC curves in terms of the discriminatory power is preserved by the ordering of the CEC curves.*

Proof. In order to prove this statement, at first two ROC curves are considered with one lying uniformly above the other one

$$F_d[F_n^{-1}(u)] > F_d^*[F_n^{-1}(u)] \quad \forall\, u \in [0, 1]. \quad (3.2)$$

The condition (3.2) can be written also as

$$1 - F_d[F_n^{-1}(u)] < 1 - F_d^*[F_n^{-1}(u)] \quad \forall\, u \in [0,1]. \tag{3.3}$$

By multiplying both the sides of the inequality (3.3) for the costs ratio $\frac{C(FN)}{C(FD)}$ and by summing up the probability u, it is obtained

$$\frac{C(FN)}{C(FD)}\{1 - F_d[F_n^{-1}(u)]\} + u < \frac{C(FN)}{C(FD}\{1 - F_d^*[F_n^{-1}(u)]\} + u \quad \forall\, u \in [0,1]. \tag{3.4}$$

that the CEC curve with lower discriminatory power is uniformly superior to the other one. □

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

$$\frac{\partial CC[F_n(s)]}{\partial F_n(s)} = -\frac{C(FN)}{C(FD)}\frac{f_d(s)}{f_n(s)} + \frac{C(FD)}{2}. \tag{3.5}$$

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

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

it results

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

By applying the equation (3.6), it is derived

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

By considering the equation (3.7) it is obtained

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

□

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

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

In order to understand the behavior of the CEC 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 creditworthiness, 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 CEC curve is decreasing for scores lower than the one at which the CEC curve has the minimum and it is increasing for scores higher, as Figure 3 shows.

3.2 The AUC as the area under the CEC curve

Analogously to the AUC, Figure 2 shows that the area between the CEC curve and the dotted line of the perfect model that joins the points $(0, c)$ and $(1, 1)$ can be considered a measure of discriminatory power.

The aim of this section is to show the relationship between the AUC and the area under the CEC curve.

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

Proof. The area under the CEC curve is

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

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

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

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

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

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

By normalizing the are under the CEC curve and by considering the result (2.2), the Gini index G is obtained

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

□

3.3 The choice of optimal threshold

In section 2.2 the optimal cut-offs for the ROC curve and its relationship with the accuracy are analyzed. Since the accuracy is the weighted average of the true positive and true negative rates, their weights are proportional to the sample sizes of defaults and non-defaults, P and N. This measure becomes inefficient when the ratio D/ND is too small or too large (Hong, 2009). Note that in a real population of borrowers the number of the defaults, P, is usually much less than that of the non-defaults, N.

To overcome this drawback, it is proposed to define the optimal threshold for the CEC curve as the value s that satisfies

$$\min_s \left\{ \frac{C(FN)}{C(FD)}[1 - F_d(s)] + F_n(s) \right\} = \max_s \left[\frac{C(FN)}{C(FD)} F_d(s) - F_n(s) \right]. \tag{3.12}$$

It is underlined that the proposal (3.12) to choose the optimal threshold overcomes the problems, analyzed in Section 2.3, that concern the iso-performance lines. In particular, the optimal cut-off does not depend on the sample frequency of defaulters. Moreover, this proposal is not affected y the problem attached to the skewness of defaulters' distribution. Finally, the difference between the classification costs of Type I and II errors is considered.

Proposition 3.4. *The probability density function of the optimal cut-off S^*that satisfies the equation (3.12) is*

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

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

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

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

$$F_Z(z) = \begin{cases} \frac{1}{k}\left(\frac{s^*}{2}+s^*+\frac{1}{2}\right), & -1 \le z < 0; \\ \frac{1=2z}{2k}, & 0 \le z < k-1; \\ \frac{1}{k}\left(-\frac{(s^*)^2}{2}+ks^*+\frac{-k^2+2k}{2}\right), & k-1 \le z < k; \\ 1 & z > k. \end{cases} \tag{3.16}$$

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 $C(FN)/C(FD) = 1$, 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.

In order to identify the optimal threshold, the expression (3.12) is equalized to zero, so when both the conditional distribution functions $F_d(s)$ and $F_n(s)$ are continuous, the following condition is obtained

$$\frac{f_d(s^*)}{f_n(s^*)} = \frac{C(FD)}{C(FN)},$$

so the optimal cut-off s^* corresponds to the score value for which the likelihood ratio $\frac{f_d(s^*)}{f_n(s^*)}$ is equal to the ratio $\frac{C(FD)}{C(FN)}$ of classification error costs.

4 Simulation results

Two random samples are generated from two random variables and they play the roles of defaulters and non-defaulters. Analogous to Hong (2009), the sample size of defaulters is 100 and the sample size of non-defaulters is 400. This means that the probability of default is 0.2 for this sample. Since the probability of default is lower in the empirical analyses (i.e. Cerved Group, 2011), a probability of default equal to 0.05[1] For this sample, it follows that the sample size of defaulters is 25 and the sample size of non-defaulters is 475.

In order to understand how the properties of these methodologies vary according to the number of observations, we consider also a random sample of size 1000 both with the probabilities of default 0.2 and 0.05.

Hong (2009) applies the cost function applied by Provost and Fawcett (2001)

$$Cost = C(FN)FN + C(FP)FP.$$

Since the misclassification error costs are often uncertain (Adams and Hand, 1999), but their ratio $C(FN)/C(FP)$ is usually known, we prefer to consider the following cost function

$$Cost^* = \frac{Cost}{C(FP)} = \frac{C(FN)}{C(FP)}FN + FP. \tag{4.1}$$

[1]In the following section of the empirical evidence Italian Small and Medium Enterprises are analyzed. The default percentage for Italian SMEs is 5%(Cerved Group, 2011).

Sample Size	PD	$N_D(0,1)$; $N_{ND}(2,1)$	k=2	k=10	k=30
500	0.2	ROC costs	92(0.005)	395(0.008)	1149 (0.009)
		CEC costs	123(0.010)	238(0.013)	313(0.005)
500	0.05	ROC costs	37(0.009)	175(0.018)	521(0.019)
		CEC costs	122(0.010)	274(0.005)	367(0.006)
1000	0.2	ROC costs	188(0.002)	794(0.006)	2311(0.006)
		CEC costs	241(0.013)	474(0.002)	641(0.002)
1000	0.05	ROC costs	73(0.005)	348(0.010)	1031(0.010)
		CEC costs	258(0.012)	526(0.006)	755(0.002)
Sample Size	PD	SN_D(0,2,-3) SN_{ND}(0,2,3)	k=2	k=10	k=30
500	0.2	ROC costs	201(0.005)	1001(0.001)	3001(0.003)
		CEC costs	402(0.005)	410(0.024)	430(0.070)
500	0.05	ROC costs	51(0.017)	484(0.018)	751(0.001)
		CEC costs	476(0.003)	251(0.003)	503(0.054)
1000	0.2	ROC costs	400(0.002)	2000(0.004)	6000(0.001)
		CEC costs	801(0.002)	809(0.012)	829(0.036)
1000	0.05	ROC costs	101(0.010)	959(0.010)	1501(0.006)
		CEC costs	951(0.002)	501(0.002)	979(0.029)
Sample Size	PD	SN_D(-4,2,3); SN_{ND}(4,2,-3)	k=2	k=10	k=30
500	0.2	ROC costs	27(0.014)	104(0.032)	298(0.036)
		CEC costs	33(0.027)	69(0.012)	112(0.011)
500	0.05	ROC costs	14(0.032)	65(0.069)	175(0.076)
		CEC costs	35(0.023)	60(0.026)	120(0.016)
1000	0.2	ROC costs	54(0.008)	212(0.018)	606(0.021)
		CEC costs	66(0.010)	146(0.017)	220(0.004)
1000	0.05	ROC costs	27(0.014)	152(0.030)	342(0.033)
		CEC costs	64(0.044)	117(0.031)	258(0.003)

Table 2: The results of Monte Carlo simulations on 1000 samples, where $N_D(0,1)$ and $N_{ND}(2,1)$ indicate the normal random variables with parameters $\mu = 0, 2$, respectively, and $\sigma^2 = 1$; SN_D(-4,2,3), SN_{ND}(-4,2,-3) and SN_{ND}(4,2,-3) indicate the skewed normal random variables with parameters $\mu = -4, 4$, respectively, $\sigma^2 = 2$ and $\lambda = 3, -3$, respectively.

As above-mentioned, the risk for FN is usually much higher than that for FD, so the costs $C(FN)$ are often much higher than $C(FD)$. For this reason Hong (2009) considers the ratio $C(FN)/C(FP)$ in the range [2,5]. Considerable empirical evidence shows that the ratio of the classification error costs is much higher, e.g. Altman et al. (1977) obtain a ratio equal to 31. For this reason, we compare the methodology based on the accuracy, and our proposal is compared by considering the ratios of the classification error costs equal to 2, 10, 30.

Hong (2009) considers two normal random variables to generate the scores of defaulters and non-defaulters. These distributions with the same parameters are

also considered in the following simulations. On the contrary, much empirical evidence shows asymmetric distributions of the scores for defaulters and non-defaulters (Christodoulakis and Satchell, 2006), even the score distributions in the empirical evidence of this paper show these characteristics. For this reason, the scores are generated from two skewed normals random variables, analogously to Christodoulakis and Satchell (2006).

In order to analyze the properties of the iso-performance line and of the methodology here proposed, by applying the expression (4.1), the average cost and the coefficient of variation of cost are computed on 1,000 random samples[2]. The results[3] are reported in Table 2.

From these results we can deduce that the performance of the methodologies depend on the default probability when the distributions of the scores are symmetric. For 20% of the defaults in the sample, our proposal is preferable to the accuracy method when the ratio of the classification error costs is equal to 10 or higher. On the contrary, when the probability of default is only 0.05, the CEC method is only preferable for high ratio of the classification error costs (k=30).

When k=30 our proposal exhibits the coefficient of variation of the costs lower than those of the accuracy method for each simulated sample. We obtain the same performance for most of the simulated samples when k=10 except for n=500 and PD=0.2.

When scores are simulated from a skewed random variable[4] (skewed normal distribution), our proposal shows better performance than the accuracy method for the ratio of the classification error costs at least equal to 10. We underline that when we sample from asymmetric distributions, the dispersion of the costs of our method increases and results higher than those of the costs of the accuracy method when k=30 for all the simulated samples.

In order to analyze the performance of our proposal we increase the discriminatory power of the scoring model by increasing the distance between the expected values of the scores S_D and S_{ND}, Also in this case our proposal shows better performance than the one of the accuracy method for both k=10 and k=30. By comparing these results with those of the previous case where the discriminatory power is lower, we can note that by increasing the discriminatory power the difference of the performances of the two methodologies becomes less relevant. This is an important characteristic for the credit risk application, since - when the discriminatory power is lower - which represents a riskier situation for banks, the performance of our method is definitely better than the accuracy method. Furthermore, the dispersion of the costs of our method is lower not only for c=30 (analogous to the case where the discriminatory power is lower) but also for c=10.

[2]Calabrese and Zenga (2010) consider the same number of replications.

[3]Since the high values of the costs are reported in Table 2, the values after the points are not considered.

[4]In order to generate these sets of data we use the R package "sn".

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, a large part of the literature (Altman, Sabato 2006; Ansell and al. 2009; Ciampi, Gordini 2008; Vozzella, Gabbi 2010) have focused on the special character of small business lending and the importance of relationship banking for solving information asymmetries. The informative asymmetries puzzle affects particulary SMEs for their difficulty to estimate and make known their fair value. Therefore, the lending to SMEs is riskier than to large corporates (Altman, Sabato 2006; Dietsch, Petey 2004; Saurina, Trucharte 2004). As a consequence, Basel II (BCBS, 2004) establishes that banks should develop credit risk models specifically addressed to SMEs. Only a few studies consider SMEs (Andreeva et al., 2011; Altman and Sabato, 2007; Altman et al. 2010; Hu and Ansell, 2007) since the gathering of SMEs data is quite difficult. Discriminant analysis and logistic regression have been the most widely used methods for constructing scoring systems for SMEs (e.g. Hand and Henley, 1997a, b; Hand and Niall, 2000).

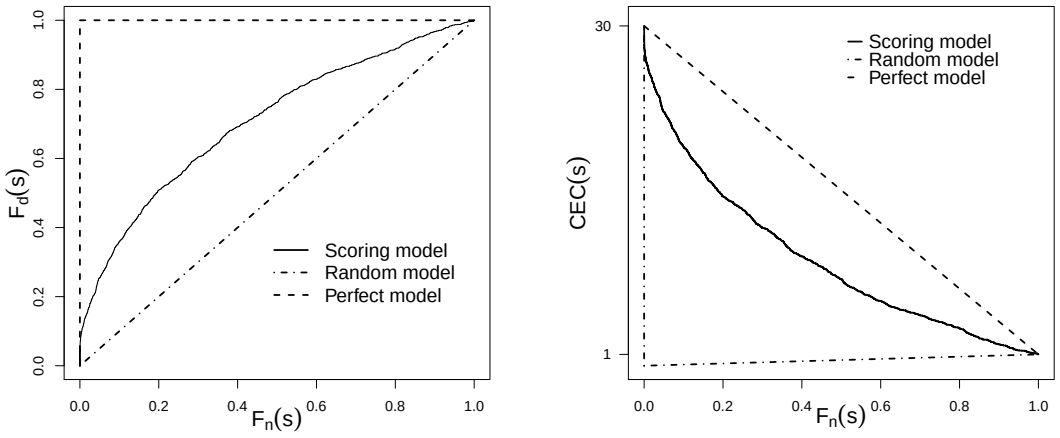
Data used in our analysis comes from AIDA-Bureau van Dijk, a large Italian financial and balance sheet information provider. We consider Italian defaulted and non-defaulted SMEs over the years $2005 - 2009$. In particular, since the default probability is one-year forecasted, the covariates concern the 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.

Often default definitions for credit risk models concern single loan defaults of a company versus a bank, as also emerges from the Basel II instructions. We consider a default occurred when a specific firm enters a bankruptcy procedure as defined by the Italian law (Altman and Sabato, 2005). In accordance with Altman and Sabato (2006) 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 defaults in our sample as close as possible to the default percentage (5 %) for Italian SMEs (Cerved Group, 2011).

In order to overcome the drawbacks of the logistic regression model for rare events (Calabrese and Osmetti, 2011; King and Zeng, 2001), we apply the Generalized Extreme Value (GEV) regression model proposed by Calabrese and Osmetti (2011) to forecast the probability of default (PD) estimate.

In order to model the default event, we choose the independent variables that represent the financial and economic characteristics a firms according to the recent literature (Vozzella and Gabbi, 2010; Ciampi and Gordini, 2008; Altman and Sabato, 2006). 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



(a) ROC curve.

(b) CEC curve.

Figure 3: Plots on data on Italian SMEs (1485 defaulters and 29700 non-defaulters) over the years $2005 - 2009$.

the developed models may overfit the data, resulting in over-optimistic estimates of the predictive accuracy, the validation is applied on a sample, called out-of-sample sample, which is different from that used in estimating the model parameters. We choose a out-of-sample sample size (3,115) of 10% of the sample size (31,185) used in estimating the model parameters.

	Accuracy method		*CEC method*	
	Actual default	Actual non-default	Actual default	Actual non-default
Default forecast	50	1,099	112	1,286
Non-default forecast	95	2,016	33	1,829
	145	3,115	145	3,115

Table 3: Contingency tables on data on 3,115 Italian SMEs.

In Figure 3 we represent the ROC and the CEC curves for Italian SMEs data. By applying our proposal and the accuracy method we compute the two optimal cutoffs. This means that we can represent the contingency tables for both the

methodologies, as Table 3 shows. The most important results shown by Table 3 are that the frequency of defaulters that are correctly classified by applying our proposal (112) is much higher than that obtained from the iso-performance line (50). This characteristic is very important for banks whose main aim is the correct classification of defaulters. Finally, by considering a ratio of classification error costs equal to 30 (Altman et al. 1977) and by applying the cost function defined in equation (4.1), we compute the classification error costs for our methodology and for the accuracy method.

	Classification Error costs
Accuracy method	3,949
CEC method	2,276

As reported in the previous table, our method shows Classification Error Costs much lower than those of the iso-performance line.

6 Conclusions

In this work we overcome some main problems of the validation of scoring models. At first, we propose the CEC 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 ordering of the CEC curves in term of discriminatory power is preserved by the ordering of the CEC curves. Moreover, we proof that the normalized area under the CEC curve is the Gini index. In coherence with the CEC representation, we suggest a method to compute the optimal cut-off that is not affected by the sample frequency of defaulters. We derive also the probability density function of the optimal cut-off. Monte Carlo simulations show that our proposal is preferable to the accuracy method for data with similar characteristics to empirical score distributions. Finally, we apply all the previous proposals to data on Italian SMEs.

This work is important since simulation studies on the validation od rating models mainly concern symmetric credit score distributions and low ratio of the misclassification errors costs. On the contrary, we analyze our proposals by drawing from skewed distributions of the credit scores and by considering the high ratio of the costs of misclassification errors. 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. (2006). Modeling Credit Risk for SMEs: Evidence from the US Market, ABACUS, 19(6), 716-723.

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.
Calabrese, R., Zenga, M. (2010) Bank loan recovery rates: Measuring and nonparametric density estimation. Journal of Banking and Finance 34 (5), 903-911.
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.
Herbert, A. D., Nagaraja, H. N.(2003) *Order Statistics*. Wiley
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.
Hong, C. S. (2009). Optimal Threshold from ROC and CAP Curves. *Communications in Statistics* 38, 2060-2072.
Gibbons J. D. (1971). *Nonparametric Statistical Inference.* McGraw-Hill, Inc. New York.
King G., Zeng L. (2001). Logistic Regression in Rare Events Data. Political Analysis, 9, 137-163.
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.
Provost, F., Fawcett, T (2001). Robust Classification for Imprecise Environment. *Machine Learning* 42 (3), 203-231.
Sobehart, J., Keenan, S. (2004). The score for credit. *Risk* 17, 54-58.
Stein, R. M. (2002). Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation. *Moody's KMV Technical Report.*
Stein, R. M. (2005). The relationship between default prediction and lending prof-

its: Integrating ROC analysis and loan pricing. *Journal of Banking & Finance.* 29, 1213-1236.
Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science* 240, 1285-1293.
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.*
Vinciotti, V., Hand, D.J. (2003). Scorecard construction with unbalanced class sizes. *Journal of the Iranian Statistical Society*, 2, 2, 189-205.
Vozzella P., Gabbi G. (2010). Default and Asset Correlation: An Empirical Study for Italian SMEs. Working Paper.
Vuk, M., Curk, T. (2006). ROC curve, lift chart and calibration plot. *Metodolo ki zvezki* 3, 89-108.
Zou, K. H. (2002). Receiver Operating Characteristic (ROC) *Literature Research*, Radiological Society of North America. Working paper.