REVSTAT – Statistical Journal
Volume 14, Number 2, April 2016, 157–169

RISK ANALYSIS AND RETROSPECTIVE UNBALANCED DATA

Authors: Francesca Pierri
– Department of Economics,
Statistical Section University of Perugia, Italy
francesca.pierri@unipg.it

Elena Stanghellini
– Department of Economics,
Statistical Section University of Perugia, Italy
elena.stanghellini@unipg.it

Nicoló Bistoni
– Graduate of Department of Economics,
University of Perugia, Italy
nicobistoni@yahoo.it

Received: October 2015 Revised: February 2016 Accepted: February 2016

Abstract:

- This paper considers three different techniques applicable in the context of credit scoring when the event under study is rare and therefore we have to cope with unbalanced data. Logistic regression for matched case-control studies, logistic regression for a random balanced data sample and logistic regression for a sample balanced by ROSE (Random OverSampling Examples, Lunardon, Menardi and Torelli, 2014) are tested. We applied the methods to real data: balance sheets indicators of small and medium-sized enterprises and their legal status are considered. The event of interest is the opening of insolvency proceedings of bankruptcy.

Key-Words:

- *bankruptcy; case-control studies; data augmentation; logistic regression; ROSE method; unbalanced data.*

AMS Subject Classification:

- 62J05, 62M20, 62P20, 91G40.

1. INTRODUCTION

In recent years, mainly because of the economic crisis that involves several European countries, the measurement of credit risk plays an important role; it simply concerns classifying out-of-sample units into two categories, bad and good, but it is crucial for its implications. The different classes of risk such as Probability of Default, Loss Given Default, Exposure at Default, Expected or Unexpected Loss are subjects of special attention from financial institutions which are making more and more frequently use of quantitative tools in decision-making. Credit quality is in fact crucial to the profitability and stability of banking systems.

An approach to estimate the probability of default is represented by statistical models, known as *credit scoring* techniques, and logistic regression is widely used in this context (Stanghellini, 2009). Frequently we have data where one of the two events is rare, so even in the case of all categorical explanatory variables, contingency tables will have very low or zero frequencies in the cells related to this event. Things get more extreme when there is at least one continuous variable in the set of the explanatory variables. In this situation, estimation of the logistic regression model may lead to high classification errors of rare units (King and Zeng, 2001). The aim of this paper is to compare different techniques which allow accurate estimation under these conditions.

The study is carried out on data selected from the *AIDA* database, concerning balance sheet indicators of companies in the Tuscany region of Italy which contains a large number of small and medium-sized enterprises. The event of interest is the opening of insolvency proceedings for bankruptcy which, luckily from an economic point of view, can be considered rare.

In order to face this problem we applied logistic regression to a retrospective data collection, using different sampling techniques: case-control sampling, balanced random sampling and random oversampling (ROSE method). From the full dataset we built a training and a hold-out sample: the first one forms the basis of data for the implementation of the different methodologies, and the second is used to compare the three classification methods on the basis of the Receiver Operating Characteristic (ROC) Curve (Fawcett, 2006). The theoretical illustration of the three methodologies (Section 2, 3 and 4) is followed by a brief description of the data (Section 5) and their application. The three models are then compared, based on the area underlying the ROC Curve.

2. THE LOGISTIC MODEL WITH BALANCED DATA

A prospective study often involves a long follow-up period and a large sample and therefore many investigations relay on a retrospective technique. The default status is regarded as a fixed variable, while variables specifying risk factors are viewed as random conditional on the default status. A retrospective study draws separate samples of cases (the bankruptcy event occurred) and controls (good firms) and therefore a smaller total sample size is usually required in comparison to a prospective study. Mantel and Haenszel (1959) and Mantel (1973) provide discussions of retrospective studies and their relationship with prospective ones. The logistic model is widely used in the analysis of retrospective studies, but it is necessary to ensure that the retrospective sample includes a representative sample of cases and controls from the population.

Let Y be the Bernoulli random variable taking value 1 when the event occurs (bankruptcy) and 0 otherwise (good firm). Let $\mathbf{x}' = (x_1, x_2, ..., x_p)$ be a covariate vector representing risk factors thought to be related to event under study. Assume the suitability of the retrospective sample and that $P(y|\mathbf{x})$ is represented by the logistic model

$$P\big(Y=1|\mathbf{x}\big) \;=\; \frac{e^{\alpha+\boldsymbol{\beta}'\mathbf{x}}}{1+e^{\alpha+\boldsymbol{\beta}'\mathbf{x}}}\,, \tag{2.1}$$

where α is an unknown scalar parameter and $\boldsymbol{\beta}' = (\beta_1, \beta_2, ..., \beta_p)$ is an unknown vector of coefficients. Now consider the hypothetical population to which (2.1) refers and let the marginal distribution of the covariates be denoted by $P(\mathbf{x})$. We draw a random retrospective sample of size n, with n_1 cases $(Y=1)$ and n_0 controls $(Y=0)$, in such a way that the marginal distribution of Y in the retrospective sample has M good cases for each bad one.

Let Z be a binary variable which takes the value 1 if a unit is included in the sample and 0 otherwise; moreover define $K_1 = P(Z=1|Y=1)$ the probability to extract a default unit and $K_0 = P(Z=1|Y=0)$ the complementary probability, both independent of the p dimensional vector of $\mathbf{x}$ covariates.

Let $P(\cdot|\mathbf{x}, Z=1) = P^*(\cdot|\mathbf{x})$ represent the distribution which is conditional on being observed in the retrospective sample. The probability distribution of Y given $\mathbf{x}$, conditional on being observed, is the following:

$$P^*\big(Y=1|\mathbf{x}\big) \;=\; \frac{K_1 P\big(Y=1|\mathbf{x}\big)}{K_1 P\big(Y=1|\mathbf{x}\big) \,+\, K_0 P\big(Y=0|\mathbf{x}\big)}\,, \tag{2.2}$$

$$\log \frac{P^*\big(Y=1|\mathbf{x}\big)}{P^*\big(Y=0|\mathbf{x}\big)} \;=\; \log \frac{K_1}{K_0} + \log \frac{P\big(Y=1|\mathbf{x}\big)}{P\big(Y=0|\mathbf{x}\big)}\,,$$

and substituting in the second term the logistic model expression, we have:

$$\log \frac{P^*\big(Y=1\,|\,\mathbf{x}\big)}{P^*\big(Y=0\,|\,\mathbf{x}\big)} \;=\; \log \frac{K_1}{K_0} + \alpha + \beta_1 x_1 + \cdots + \beta_p\, x_p \;, \tag{2.3}$$

where α and $\beta_j,\ j = 1, 2, 3, ..., p$, are the unknown parameters. It then follows that the logistic model for the retrospective balanced data has different intercepts but equal slopes and inference about α would require knowledge of $\frac{K_1}{K_0}$.

If $\alpha^\star$ denotes the intercept in the logistic model in the population with $Z = 1$, it follows that:

$$\alpha \;=\; \alpha^\star - \log \frac{K_1}{K_0} \;.$$

We indicate with the distribution $P(\mathbf{x}\,|\,y)$ the conditional distribution of the covariates given the response and with $P^*(\mathbf{x}\,|\,y)$ the same distribution conditional on being in the sample. As the sampling is independent of the covariates, the two distributions should be the same. The likelihood function for the retrospective sample can then be written:

$$\prod_{i=1}^{n} P\big(\mathbf{x}_i\,|\,y_i\big) \;. \tag{2.4}$$

Let $P^*(y_i\,|\,\mathbf{x}_i)$ be the conditional distribution of the response given that the covariates in a unit i are observed in the sample. Furthermore, let $P^*(y)$ denote the distribution of y and $P^*(\mathbf{x})$ represents the distribution of $\mathbf{x}$ conditional on being in the sample. Then from Bayes's rule (2.4) can be written as:

$$\prod_{i=1}^{n} \frac{P^*\big(y_i\,|\,\mathbf{x}_i\big)\, P^*(\mathbf{x}_i)}{P^*(y_i)} \;. \tag{2.5}$$

By the sampling scheme we know $P^*(Y=1)$ and $P^*(Y=0)$ are respectively equal to $\frac{n_1}{n}$ and $\frac{n_0}{n}$. For maximum likelihood inference (V.T. Farewell, 1979) we maximize (2.5) subject to the constraint

$$\sum_{\mathbf{x}} P^*\big(Y=1\,|\,\mathbf{x}\big)\, P^*(\mathbf{x}) \;=\; \frac{n_1}{n} \;, \tag{2.6}$$

where we have assumed that $\mathbf{x}$ is discrete. Anderson (1972) shows that the constrained maximum likelihood estimates of α^* and $\boldsymbol{\beta}$ are algebraically equivalent to the unconstrained estimates which maximize

$$\prod_{i=1}^{n} P^*\big(y_i\,|\,\mathbf{x}_i\big) \;, \tag{2.7}$$

while R.L. Prentice and R. Pyke (1979) show that the constrained estimation of (2.5) is a reparametrization of a likelihood based on the population model, where the constraints are defined in terms of the population value of $P(Y=1)$.

3. THE LOGISTIC MODEL FOR MATCHED CASE-CONTROL STUDIES

The logistic regression model for matched case-control studies, developed and widely used in epidemiology, may be considered as a refinement of the logistic modelling for balanced data. This method stratifies subjects on the basis of variables believed to be associated with the outcome. Again, we assume that the population model is logistic, as in (2.1). Within each stratum, samples of cases ($Y=1$) and controls ($Y=0$) are chosen, according to a $1-1$ design or $1-M$ design, where M is usually no more than five (Hosmer, Lemenshow and Sturdivant, 2013, p. 243). Let K be the number of strata, n_{1k} and n_{0k} respectively the cases and the controls within the k-th stratum, where $k = 1, 2, ..., K$. The stratum-specific logistic regression model for a unit in the sample, is

$$P\big(Y=1 \,|\, \mathbf{x},\, K=k\big) \;=\; \pi_k(\mathbf{x}) \;=\; \frac{e^{\alpha_k+\boldsymbol{\beta}'\mathbf{x}}}{1+e^{\alpha_k+\boldsymbol{\beta}'\mathbf{x}}}\,, \tag{3.1}$$

where α_k represents the contribution of all constant terms within the k stratum (i.e. stratification variable or variables) and $\boldsymbol{\beta}$ the vector of the p coefficients $\boldsymbol{\beta}' = (\beta_1, \beta_2, ..., \beta_p)$. From (2.3), it follows that the relationship between α and the stratum-specific parameters α_k varies among strata. Therefore, α_k are nuisance terms and should be eliminated from the set of parameters on which we want to make inference. The conditional likelihood method gives consistent and asymptotically normally distributed estimates of the β_j slope coefficients (Prentice and Pyke, 1979). The conditional likelihood for the k-th stratum is the probability that the observed case and control configuration is verified, conditioned on the stratum total and total number of observed case. Denoting $n_k = n_{1k}+n_{0k}$ as the number of subjects, the conditional likelihood for each stratum gives the probability to observe the data, conditioned on all possible assignment of cases n_{1k} and controls n_{0k}. The number of possible assignments of case status to n_{1k} among the n_k subjects is given by:

$$c_k \;=\; \binom{n_k}{n_{1k}} \;=\; \frac{n_k!}{n_{1k}!\,(n_k-n_{1k})!}\,.$$

Let the subscript j denote any one of these c_k assignments; moreover let subjects from 1 to n_{1k} correspond to the cases and subjects $n_{1k}+1$ to n_k to the controls. Any assignment is indexed by i for the observed data and by i_j for the j^{th} possible assignment. The conditional likelihood for the k-stratum is

$$l_k(\boldsymbol{\beta}) \;=\; \frac{\displaystyle\prod_{i=1}^{n_{1k}} P\big(\mathbf{x}_i \,|\, Y_i=1\big) \prod_{i=n_{1k}+1}^{n_k} P\big(\mathbf{x}_i \,|\, Y_i=0\big)}{\displaystyle\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} P\big(\mathbf{x}_{ji_j} \,|\, Y_{i_j}=1\big) \prod_{i_j=n_{1k}+1}^{n_k} P\big(\mathbf{x}_{ji_j} \,|\, Y_{i_j}=0\big) \right\}} \tag{3.2}$$

and the full conditional likelihood over the K strata would be given by the product:

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K} l_k(\boldsymbol{\beta}) \,. \tag{3.3}$$

Assuming (3.1) is true and applying Bayes's rule to each $P(\mathbf{x}|y)$ term, we can write equation (3.2) as follows:

(3.4)

$$l_k(\boldsymbol{\beta}) = \frac{\displaystyle\prod_{i=1}^{n_{1k}} \left[\frac{P\big(Y_i=1|\mathbf{x}_i\big)\, P(\mathbf{x}_i)}{P\big(Y_i=1\big)} \right] \prod_{i=n_{1k}+1}^{n_k} \left[\frac{P\big(Y_i=0|\mathbf{x}_i\big)\, P(\mathbf{x}_i)}{P\big(Y_i=0\big)} \right]}{\displaystyle\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} \left[\frac{P\big(Y_{i_j}=1|\mathbf{x}_{ji_j}\big)\, P(\mathbf{x}_{ji_j})}{P\big(Y_{i_j}=1\big)} \right] \prod_{i_j=n_{1k}+1}^{n_k} \left[\frac{P\big(Y_{i_j}=0|\mathbf{x}_{ji_j}\big)\, P(\mathbf{x}_{ji_j})}{P\big(Y_{i_j}=0\big)} \right] \right\}} \,.$$

Remembering that $P(Y_i=1|\mathbf{x}_i) = \pi(\mathbf{x}_i)$ and $P(Y_i=0|\mathbf{x}_i) = 1-\pi(\mathbf{x}_i)$ we can write:

$$l_k(\boldsymbol{\beta}) = \frac{\displaystyle\prod_{i=1}^{n_{1k}} \left[\frac{\pi(\mathbf{x}_i)\, P(\mathbf{x}_i)}{P\big(Y_i=1\big)} \right] \prod_{i=n_{1k}+1}^{n_k} \left[\frac{\big[1-\pi(\mathbf{x}_i)\big]\, P(\mathbf{x}_i)}{P\big(Y_i=0\big)} \right]}{\displaystyle\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} \left[\frac{\pi(\mathbf{x}_{ji_j})\, P(\mathbf{x}_{ji_j})}{P\big(Y_{i_j}=1\big)} \right] \prod_{i_j=n_{1k}+1}^{n_k} \left[\frac{\big[1-\pi(\mathbf{x}_{ji_j})\big]\, P(\mathbf{x}_{ji_j})}{P\big(Y_{i_j}=0\big)} \right] \right\}}$$

and also

(3.5)

$$l_k(\boldsymbol{\beta}) = \frac{\displaystyle\prod_{i=1}^{n_{1k}} \left[\frac{\frac{e^{\alpha_k+\beta'\mathbf{x}_i}}{1+e^{\alpha_k+\beta'\mathbf{x}_i}}\, P(\mathbf{x}_i)}{P\big(Y_i=1\big)} \right] \prod_{i=n_{1k}+1}^{n_k} \left[\frac{\frac{1}{1+e^{\alpha_k+\beta'\mathbf{x}_i}}\, P(\mathbf{x}_i)}{P\big(Y_i=0\big)} \right]}{\displaystyle\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} \left[\frac{\frac{e^{\alpha_k+\beta'\mathbf{x}_{ji_j}}}{1+e^{\alpha_k+\beta'\mathbf{x}_{ji_j}}}\, P(\mathbf{x}_{ji_j})}{P\big(Y_{i_j}=1\big)} \right] \prod_{i_j=n_{1k}+1}^{n_k} \left[\frac{\frac{1}{1+e^{\alpha_k+\beta'\mathbf{x}_{ji_j}}}\, P(\mathbf{x}_{ji_j})}{P\big(Y_{ij}=0\big)} \right] \right\}} \,.$$

Moreover, collecting common terms of the form

$$\frac{1}{1+e^{\alpha_k+\boldsymbol{\beta}'\mathbf{x}}}$$

we can write (3.5) as:

$$l_k(\boldsymbol{\beta}) = \frac{\displaystyle\prod_{i=1}^{n_{1k}} \Big[e^{\alpha_k+\beta'\mathbf{x}_i} \Big] \prod_{i=1}^{n} \left[\frac{1}{1+e^{\alpha_k+\beta'\mathbf{x}_i}} \right] \prod_{i=1}^{n_k} \left[\frac{P(\mathbf{x}_i)}{P(Y_i)} \right]}{\displaystyle\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} \Big[e^{a_k+\beta'\mathbf{x}_{ji_j}} \Big] \prod_{i_j=1}^{n} \left[\frac{1}{1+e^{\alpha_k+\beta'\mathbf{x}_{ji_j}}} \right] \prod_{i_j=1}^{n_k} \left[\frac{P(\mathbf{x}_{ji_j})}{P(Y_{i_j})} \right] \right\}} \,. \tag{3.6}$$

Further algebraic simplification leads to the following:

$$l_k(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_{1k}} e^{\boldsymbol{\beta}'\mathbf{x}_i}}{\sum_{j=1}^{c_k}\prod_{i_j=1}^{n_{1k}} e^{\boldsymbol{\beta}'\mathbf{x}_{ji_j}}}, \tag{3.7}$$

where $\boldsymbol{\beta}$ is the only unknown parameter.

In a $1-1$ matched design each case is matched to a single control. Let $\mathbf{x}_{1k}$ and $\mathbf{x}_{0k}$ respectively denote the data vector for the case and the control in the k-th stratum, the conditional likelihood for the k-th stratum is

$$l_k(\boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_{1k}}}{e^{\boldsymbol{\beta}'\mathbf{x}_{1k}} + e^{\boldsymbol{\beta}'\mathbf{x}_{0k}}} \tag{3.8}$$

given specific value for $\boldsymbol{\beta}$, $\mathbf{x}_{1k}$ and $\mathbf{x}_{0k}$, equation (3.8) is the probability that the unit identified as the case is in fact the case. If data for case and control are identical, $\mathbf{x}_{1k}= \mathbf{x}_{0k}$, it follows from equation (3.8) that $l_k(\boldsymbol{\beta}) = 0.5$ for every value of $\boldsymbol{\beta}$ and the stratum will be considered as *uninformative* meaning that the covariates do not discriminate cases from controls.

In a $1-1$ matched data design with a binary explanatory variable X, the conditional maximum likelihood estimator is the log of the ratio of discordant pairs (see Breslow and Day, 1980). It follows that it is advisable to classify (in a 2×2 table) cases versus controls for each dichotomous variable to verify the presence of discordant pairs: the absence of both types of pairs ($x_{1k} = 1$, $x_{0k} = 0$) and ($x_{1k} = 0$, $x_{0k} = 1$) gives rise to an undefined estimator.

In a $1-M$ matched design each case is matched to M controls, so there are $M + 1$ units in each stratum. Letting $M = 4$ and denoting by $\mathbf{x}_{k1}$ the case and by $\mathbf{x}_{k2}$, $\mathbf{x}_{k3}$, $\mathbf{x}_{4k}$, $\mathbf{x}_{5k}$ the controls in the k^{th} stratum, the contribution to the likelihood for this stratum of matched subjects from equation (3.7) is

$$l_k(\boldsymbol{\beta}) = \frac{e^{\beta'\mathbf{x}_{k1}}}{e^{\boldsymbol{\beta}'\mathbf{x}_{k1}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k2}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k3}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k4}} + e^{\boldsymbol{\beta}'\mathbf{x}_{k5}}}. \tag{3.9}$$

Given the coefficients' values (3.9) gives the probability that the unit with the observed data $\boldsymbol{x}_{k1}$ is the case relative to four controls with data $\mathbf{x}_{k2}$, $\mathbf{x}_{k3}$, $\mathbf{x}_{k4}$, and $\mathbf{x}_{k5}$. If the four covariates have the same value, then $l_k(\boldsymbol{\beta}) = 0.20$ for each $\boldsymbol{\beta}$ value. Hence, for each covariate at least one control should have a value different from the case, otherwise the stratum would be considered *uninformative*.

4. THE LOGISTIC MODEL FOR "ROSE" DATA

Random OverSampling Examples (Lunardon, Menardi and Torelli, 2014) is a new procedure developed in the R language (R Development Core Team, 2015), based on the generation of new artificial data according to a smoothed bootstrap approach (Efron and Tibshirani, 1993). Let $P(\mathbf{x}) = f(\mathbf{x})$ be the probability density function on X. Let $n_j < n$ be the size of Y_j, $j = 0, 1$. A new sample is generated by the following three steps:

1. select $y = Y_j$, $j \in \{0, 1\}$, with probability $\frac{1}{2}$;
2. select $(\mathbf{x}_i, y_i)$ in the sample such that $y_i = y$ with probability $p_i = \frac{1}{n_j}$;
3. sample the vector of covariates $\mathbf{x}$ from the kernel probability distribution $K_{\boldsymbol{H}_j}(\cdot, \mathbf{x}_i)$, centered on $\mathbf{x}_i$ and depending on the matrix of smoothing parameters $\boldsymbol{H}_j$.

According to the ROSE method, in the training sample you extract a unit belonging to one of the two classes with the same probability. Then a new sample is generated in its neighborhood of width determined by $\boldsymbol{H}_j$. Generally $K_{\boldsymbol{H}_j}$, is chosen as symmetric and unimodal. Therefore the generation of new samples for the class Y_i according to ROSE corresponds to the generation of data from the kernel density estimate of $f(\mathbf{x}, Y_j)$, with matrix of smoothing parameters $\boldsymbol{H}_j$:

$$\hat{f}\big(\mathbf{x}|y=Y_j\big) = \sum_{i=1}^{n_j} p_i \, P\big(\mathbf{x}|\mathbf{x}_i\big) = \sum_{i=1}^{n_j} \frac{1}{n_j} \, P\big(\mathbf{x}|\mathbf{x}_i\big) = \sum_{i=1}^{n_j} \frac{1}{n_j} \, K_{\boldsymbol{H}_j}(\mathbf{x} - \mathbf{x}_i) \,, \tag{4.1}$$

see Menardi and Torelli (2014).

5. DATA ANALYSIS

Data are drawn from the *AIDA* database[1], one of the most important Italian databases containing historical balance sheets as well as financial, commercial and demographic information on more than one million Italian firms. We selected all the firms in the Tuscany region having positive revenues of sales in the year 2006 and for these we extracted revenues, profits, fixed assets, financial indicators, indexes of resultant profits and current management. On May 2010, we verified their legal status: the database provides the legal status of each firm, which is periodically updated without indicating the reference date; therefore, we do not know the exact time at which a firm was declared bankrupt. The selected time interval of four years is due to the delay between the bankruptcy event and the availability of the company balance sheet. The distribution according to the legal status is shown in Table 1.

[1]The database is distributed by Bureau Van Dijk s.p.a.; `https://aida.bvdinfo.com/`

Table 1: Companies' distribution by state law in May 2010.

Legal status	**Frequency**	**Percentage**
Active	33798	89.23
Bankruptcy	537	1.42
Liquidation	2800	7.39
Not active	744	1.96

Due to the lack of information on the causes of inactivity and liquidation, we included in the analysis only the firms that are active (33798) and bankrupt (537). Data clearly shows the rareness of the default event (1.56%) and therefore the inadequacy of a logistic regression model due to the unbalanced data. We built a training sample, to implement the methods, and a randomly selected hold-out sample consisting of 10% of the whole sample. Since the aim of the study is to compare three different methodologies, from among the balance sheet indicators we selected those that were found to be most informative in a previous case-control study (Pierri, 2013) on the same data.

Logistic regression is estimated on three different data sets: a balanced sample with 2505 observations where the frequencies of Y_0 and Y_1 are respectively 501 and 2004; a stratified sample (2440 observations) with strata formed on Legal Form and the first two numbers of the ATECO code (industry sector) jointly considered, where the frequencies of Y_0 and Y_1 are 488 and 1952, respectively; and a ROSE data set of 68000 observations where the frequencies of Y_0 and Y_1 are respectively 33671 and 34329. We used the ROSE routine included in R software to generate data based on the ROSE method.

In the multivariable logistic model we considered as explanatory variables Net Profit (NP), Asset Coverage Index (AC), Liabilities index (L), Quick Ratio (QR), Debt Ratio (DR), Asset Turnover (AT) and EBITDA value. The linearity in the odds of these variables was checked following the methodology proposed by Hosmer, Lemenshow and Sturdivant (2013, Ch. 4): transformation of variables, applied where necessary, led to a final model including two quadratic forms. For a detailed implementation, see also Pierri, Burchi and Stanghellini (2013). We refer to Table 2 for a summary of our main results. The logistic model on the balanced sample indicates that for Asset Turnover the quadratic form is not statistically significant. The same holds for the ROSE sample where Net Profit is also not significant. The same table also displays the estimated coefficients for the models considering only the significant ($p < 0.05$) covariates (Balanced2 and ROSE2). Balanced and case-control methods produce close estimates, while ROSE method gives smaller values. This may due to the use of artificial data. The economic interpretation of the model is also consistent with the expected results: Tuscany region is characterized by small and medium-sized enterprises as the 98% of the Italian firms. In this context the financial structure plays an

important role, because they are often under-capitalized. From data in Table 2 we can notice that the forms of debt that a company chooses is of great importance in determining the probability of default: the negative value of the Debt Ratio combined with a positive Asset Coverage Index, show that the probability of having a healthy company increases if you prefer forms of internal financing. Moreover companies with a positive Quick Ratio and a negative Liabilities Index are less exposed to the risk of default as more able to obtain long-term funding, while short-term debt may compromise the health of a company.

Table 2: Estimates of the coefficients applying the three different methods.

Explanatory Variables	**Balanced**	**Balanced2**	**ROSE**	**ROSE2**	**Case Control**
NP	0.00063	0.00064	2.22e−07*	—	0.00241
AC	0.12323	0.12541	0.05821	0.05813	0.10371
AC^2	−0.01069	−0.01082	−0.00636	−0.00636	−0.01176
L	−1.05589	−0.98879	−0.61510	−0.61631	−0.88630
QR	0.59219	0.60493	0.44750	0.44740	0.72559
DR	−0.00583	−0.00590	−0.00174	−0.00174	−0.00472
AT	0.47878	0.3058	0.26130	0.26927	1.02132
AT^2	−0.05038*	—	−0.00289*	—	−0.17471
EBITDAV	0.01522	0.01520	0.00747	0.00747	0.00574

(* *p-value* > 0.1)

We compared the predictive and discriminatory ability of the three methods looking at the ROC curves built with the hold-out sample. In Figure 1 we notice that the logistic model on balanced data (AUC = 0.7955) has the greatest capability to discriminate between good and bad firms while ROSE (AUC = 0.7645)

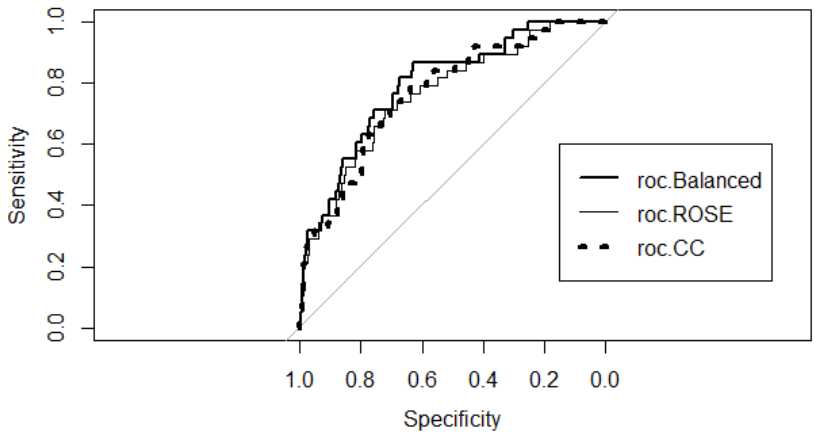

Figure 1: Estimated ROC curve in the three models using hold-out sample: Balanced (AUC = 0.7955); ROSE (AUC = 0.7645); Case Control (AUC = 0.7686).

and Case-Control (AUC = 0.7686) methods exhibit very similar results to each other. Testing the difference between their AUC, we find a significant difference ($p\text{-}value < 0.05$) only between Balanced and both Case Control and ROSE areas. We achieve similar results if we consider ROC curves built for Balance2 (AUC = 0.7911) and ROSE2 (AUC = 0.7685) models.

6. DISCUSSION

Three different methodologies have been compared. On the basis of the data and the model applied, the oversampling (ROSE) and case-control studies methods seem to give very similar results, on the other hand logistic regression on balanced data show the best predictive capabilities. We underline some particularities: ROSE allows only for continuous covariates; in case-control studies, confidence intervals are generally narrower than in standard logistic regression, but does not produce the predicted probability of bankruptcy; standard logistic regression, applied over a random balanced sample, is very easy and quick to implement. Future developments of this study will test whether stepwise model selection procedures applied to the different datasets will lead to different models.

REFERENCES

[1] Anderson, J.A. (1972). Separate sample logistic discrimination, *Biometrika*, **59**(1), 19–35.

[2] Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research, Vol. 1: The Analysis of Case-Control Studies*, International Agency of Cancer, Lyon, France.

[3] Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*, Chapman & Hall/CRC, Boca Raton London New York Washington, D.C..

[4] Farewell, V.T. (1979). Some results on the estimation of logistic models based on retrospective data, *Biometrika*, **66**(1), 27–32.

[5] Fawcett, T. (2006). An introduction to ROC analysis, *Pattern Recognition Letter*, **27**(8), 861–874.

[6] Hosmer, D.W.; Lemeshow, S. and Sturdivant, R.X. (2013). *Applied Logistic Regression*, Wiley, New Jersey.

[7] King, G. and Zeng, L. (2001). Logistic Regression in Rare Events Data, *Political Analysis*, **9**(2), 137–163.

[8] Lunardon, N.; Menardi, G. and Torelli, N. (2014). ROSE: a package for Binary Imbalanced Learning, *R Journal*, **6**(1), 79–89.

[9] Mantel, N. (1973). Synthetic retrospective studies and related topics, *Biometrics*, **29**(3), 479–486.

[10] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Biometrics*, **29**(3), 479–486.

[11] Menardi, G. and Torelli, N. (2014). Training and Assessing Classification Rules with Imbalanced Data, *Data Mining and Knowledge Discovery*, **28**(1), 92–122.

[12] Pierri, F. (2013). *Valutazione del rischio di default nelle piccole e medie imprese attraverso uno studio caso-controllo*. In "Gli Analytics come motore per i big data, la ricerca ed il sistema paese" (A. Di Ciaccio and W. Lanzani, Eds.), Aracne Editrice, Roma (Italy), 57–68.

[13] Pierri, F.; Burchi, A. and Stanghellini, E. (2013). La capacitá predittiva deglii indicatori di bilancio delle PMI, *Piccola Impresa Small Business*, **1**, 85–106.

[14] Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies, *Biometrika*, **66**(3), 403–411.

[15] R Development Core Team, (2015). R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*, Vienna, Austria, `http://www.R-project.org`.

[16] Stanghellini, E. (2009). *Introduzione ai metodi statistici per il credit scoring*, Springer Verlag (in Italian).

[17] Zhang, B. (2006). Prospective and retrospective analyses under logistic regression models, *Journal of Multivariate Analysis*, **97**, 211–230.