Assessing good bankruptcy predictors: An Empirical Analysis.

Valentina Santoni

Ph.d student

D.I.S.e.a –Department of Economics and Business- University of Sassari

e-mail: santoniv@tiscali.it

phone 3887436829

Please refer to santoniv@tiscali.it for any communications concerning the manuscript.

Abstract With this paper we intend to extend the existing literature in Credit Risk management, comparing three different methodologies in order to define which one is the most reliable : Zscore analysis, Logit model and Random Forest.

The aim is to establish if Altman's Z-score, a widely used tool to evaluate the financial health of a company, is still an efficient methodology to predict bankruptcy or financial stress conditions. Several other forecasting methods have been developed over the years, most of them based on logistic regression.

Here we present a methodology based on a machine learning algorithm (Random Forest) to analyse and predict the bankruptcy of 3000 Italian companies.

We performed the same analysis with Altman's Z-score and with the Logit model. According to our results, Random Forest obtained the best performance, with an accuracy of 99,85%.

Our results show that applications of machine learning based methods to predict bankruptcy might

overcome pre-existing methodologies and be more efficient to identify companies that may become insolvent and unable to repay loans.

Introduction

The main objective of this survey is to detect and recognize premonition's signs of corporate crisis and bankruptcy that can help the management to apply corrective action as soon as possible.

Furthermore financial institutions and shareholders express a keen interest in the future outlook of the company, showing a special attention to default prediction. The interest of Financial Institutions in failure forecasting is directed to the identification of those companies which may become insolvent and unable to repay loans.

In the literature, several studies have been devoted to this topic and, in particular, to the dissection of the causes of bankruptcy's firms. Indeed, various empirical tests proved the usefulness of accounting data in order to predict possible failures.

The main purpose of our research is to trace, examine and compare these works to establish who contributed the most to crisis prediction.

The earliest studies, dating back to the '30s, showed that "healthy" firms (i.e. capable to pay loans) have different financial ratios compared with “sick” (insolvent) ones. The first pioneers in the research of

premonitory signs of economic crisis have been Beaver (1966) and Altman (1968). They developed models to predict firms' failure consisting of indices derived from financial statements.

Beaver published a paper based on univariate approach, based on 30 ratios whose selection was based on popularity in the literature and good performance. He also chose the best predictive ratio, *Cash flow to total debt.* But the limitation of Beaver's study was the adoption of an univariate approach, in wich each ratio is autonomously analysed.

A few years later, in 1968, Altman employed the Zscore model with Multivariate Discriminant Analysis (MDA) to identify financial ratios statistically associated with future bankruptcy. Particularly, he analysed a sample of 66 companies, divided into two groups (failed and non failed firms) and applied the multiple dicriminant analysis on 5 variables. The advantage of this model was that it tested the correlation among different ratios.

Since then, there has been an ongoing achievement in the literature to develop models with a greater predictive performance. Ohlson (1980) was one of the first authors who used a logit model to predict the financial crisis, Zmijewski (1984) adopted a probit approach which is also based on accounting data. More recently, Abidali and Harris (1955) developed a model to predict construction company failure using both financial ratios and variables related to an inadequate management; Shumway (2001) has proposed a model to predict a firm's bankruptcy using both accounting and market variables; Fuertes

and Kalotychou (2006) found that the logit model should be preferred to alternative competing models; Ishwaran and Kogalur (2006) introduced a non parametric approach based on a machine learning method (Random Survival Forest); Fantazzini et al. (2009) applied both the logistic model and Random Survival Forest obtaining better results with the classical logit model.

In this paper we intend to research and extend the existing literature in credit risk mangement for small medium firms, using an empirical research in which three competing methodologies are applied: Zscore analysis, Logit model and Random Forest. Our purpose is to compare their performances and establish the best predictive methodology.

Methods

So far, research on failure prediction has been more focused on large companies while small firms have been less studied, as in Keasey et al. (1987)(1993)(1995); Huyhebaert et al. (2003); and Pompe et al. (2005).

Therefore our empirical analysis is based on annual accounting data (2006) for 3.000 Italian small-medium manufacturing firms. The sample has been collected through a survey made by the Unicredit institute (Unicredit Corporate Analysis, 2011). We focused on the manufacturing sector for its relevance and significance within the Italian context.

Given our available dataset, we estimated a set of 22 financial ratios (Table 1) that have been chosen on the basis of their popularity in the literature (Becchetti et al.(2003), De Andrés et al.(2012), Du Jardin (2010), Li et al.(2011), Ravi et al.(2007), Sun et al.(2007), Thomas et al. (2011), Yeh et al.(2012). The ratios can be classified into these broad categories:

1) **Liquidity ratios** calculate a firm's ability to meet its short term financial commitments;

2) **Profitability ratios** measure the performance in terms of Return on Assets (Roa), Return on equity (Roe);

3) **Solvency indexes** measure the firm's financial methods, by debt and shareholders' funds;

4) **Structural indexes** calculate firm size, flexibility and rigidity of capital asset.

Table 1. Financial Ratios

Category	**Financial Index**
Liquidity Index	1) Profit quality index = Net operating cash flow / operational profit
	2) Net operating cash flow over current liabilities = Net operating cash flow /current liabilities
	3) Net operating cash flow over total liabilities = Net operating cash flow /total liabilities
	4) Total cash flow ratio = Net operating cash flow /(fund raising cash flow in + investment cash flow out)
	5) Total assets cash recovery ratio = Net operating cash flow /(Total Assets)

	6) Sales cash flow ratio = Net operating cash flow / turnover ratio
Profitability Index	7) Return on equity = Net profit / share holder's equity
	8) Return on asset = Net profit / total assets
Structural Index	9) Firm Size = Log (total assets)
	10) Fixed assets ratio = Fixed asset / total assets
	11) Current assets ratio = Current asset / total assets
Solvency Index	12) Debt to asset ratio = Total liabilities / total assets
	14) Long-term debt to assets ratio = Long term debt / total assets
	15) Interest cover ratio = Earnings before interest and taxes / interest expense
Solvency Index	16) Current ratio = Current assets / current liabilities
	17) Quick ratio = (Current assets - inventory) / current liabilities
Altman's index	18) x1 = Working capital/ total assets
	19) x2 = Retained earnings / total assets
	20) x3 = Earnings before interest and taxes / total assets
	21) x4 = Equity / total liabilities
	22) x5 = sales/ total assets

Variables used in the widely used Altman's Zscore model are also included in our analysis. They represent five standard ratio categories: liquidity, profitability, leverage, solvency and activity ratios.

Firm's health is represented by a dependent binary variable Y = (0,1) for “healthy” and “sick”, respectively. In our analysis we define as “sick” firms, those in state of bankruptcy and of winding up, and as “healthy”, those not subject to any procedure.

In total we have 200 sick and 2800 healthy, we created five random subsets of 160 sick and 160 healthy firms for the "training" of the prediction methods. Then we used the complementing 5 subsets of 40 sick and 40 healthy firms as "test set" to evaluate the respective predictive power with cross validation (Kohavi,1995; Picard et al., 1984).

Zscore model

Altman's Zscore model (1968) (1994) is based on MDA (Duda et al., 2001). Even being one of the earliest prediction methodology, it is still the most popular technique and the predominant statistical method for corporate failure prediction (Balcaen et al., 2006) . Furthermore, it continues to be used in a variety of business situations involving the prediction of bankruptcy and financial stress conditions. Commercial banks use Z-Score model as part of the regular loan review process, and investment bankers use it in security and portfolio analysis (Grice et al., 2001).

Altman developed his model on a linear combination of variables discriminating between bankrupt and non bankrupt firms. Specifically, this method projects multidimensional data on a single dimension maximizing the absolute distance between the means of the two classes (bankrupt and non bankrupt) while minimizing the variance within each class.

Firms' solvency index is known as Zscore and it is defined as:

$$Z_j = v_1 x_1 + v_2 x_2 + ... + v_n x_n$$

where Z_j provides an indication of firm's financial healty; $v_1, v_2, .., v_n$ are discriminator coefficients evaluated via MDA on the training set and $x_1, x_2, \ldots, x_n$ are the independent variables described in Table 1.

Logit model

The logit model (Berkson, 1944) assumes that there is a cause-effect relationship between accounting data (x_1, x_2 ,.. x_n) and the company's health (Y). More formally, the perspective (expected value) of a company failure is determined by E[Y_j | x] and modelled by the logistic function defined between $(-\infty,+\infty)$ and range (0,1):

$$E[Y \mid X] = \frac{1}{e^{(-\beta_0+\beta_1 x_1+...+\beta_n x_n)}+1}$$

Although logistic regression approach is very popular, logistic regression models are sensitive to multicollinearity (Ooghe et al., 1994).

Random Forest

Random Forest (RF) is a machine learning method developed by Leo Breiman (2001). It is essentially an ensemble classifier consisting of many decision trees which are calculated on random subsets of the data. A decision tree is a predictive model aiming to predict the value of a target variable based on several input variables (Shih, 2011).Observing the simplified decision tree in Figure 1, each node corresponds to one input variable and each leaf represents a value of the target variable.

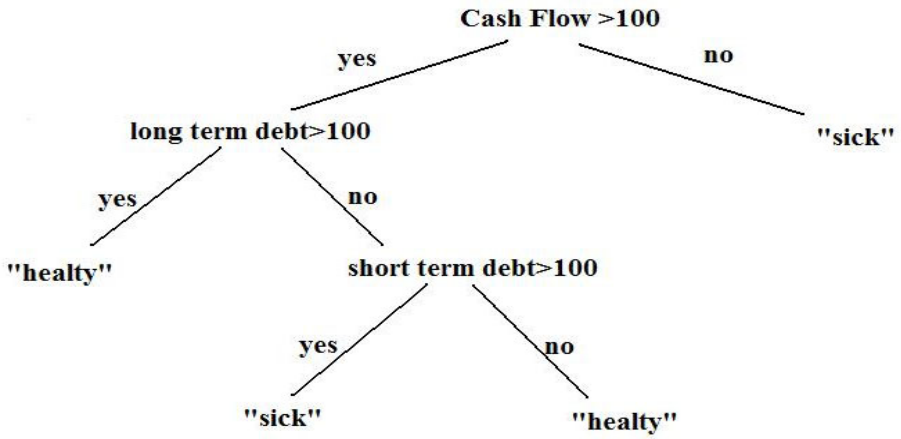

Figure 1. A simplified decision tree

During the training, RF constructs many decision trees by setting the thresholds at each node through the minimization of an "impurity" function. Intuitively the impurity, according to Gini (Shmueli et al., 2011), expressed as **i (t) = P (1 | t) * P (0 | t)** where t is a tree fed by a bootstrapped sample of the training set, measures the probability of incorrectly label a firm (sick or healthy) based on the current label distribution. It is minimized when all firms are assigned to the same label.

The prediction is achieved by using the trained "forest" of trees on the test set and collecting the decisions of each terminal node in a voting system.

For our analysis, we use the **Sensitivity** (number of true positives/ (number of true positives+number of false negatives)) and the **Specificity** ((number of true negative/ (number of true negative+number of false positive)) where:

- True positives = firms correctly identified as sick;

- False negatives = firms incorrectly diagnosed as healthy;
- True negatives = firms correctly identified as healthy;
- False positives = firms incorrectly diagnosed as sick.

Results

Tables 2 and 3 show the results obtained from logistic regression and Random Forest.

Table 2. Logistic Regression

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.425e+07	1.888e+07	-2.343	0.019122*
Profit quality index	3.022e-03	4.130e-03	0.732	0.464327
Net operating cash flow /current liabilities	1.658e-01	8.877e-02	1.868	0.061813 .
Net operating cash flow/total liabilities	-6.485e-02	6.918e-01	-0.094	0.925321
Total cash flow ratio	1.523e-03	3.919e-03	0.389	0.697577
Total assets cash recovery ratio	-1.731e+00	1.090e+00	-1.587	0.112414
Sales cash flow ratio	1.229e-10	7.938e-11	1.549	0.121436
Return on equity	-1.677e-03	1.176e-03	-1.426	0.153989
Return on asset	- 4.779e-01	2.577e+00	-0.185	0.852844
Firm Size	-1.219e-01	6.189e-02	-1.969	0.048951 *

Fixed assets ratio	-3.588e-02	2.479e-02	-1.447	0.147778
Current assets ratio	-1.675e+04	1.773e+04	-0.945	0.344817
Debt to asset ratio	4.494e+05	1.887e+05	2.382	0.017223 *
Short-term debt to assets ratio	9.830e+03	1.030e+04	0.955	0.339771
Long-term debt to assets ratio	-6.920e+03	8.775e+03	-0.789	0.43337
Interest cover ratio	5.320e-06	4.436e-06	1.199	0.230400
Current ratio	1.996e-03	1.046e-03	1.908	0.056434 .
Quick ratio	-1.526e-01	1.510e-01	-1.010	0.312339
X1	1.675e+06	1.773e+06	0.945	0.344818
X2	-1.713e+00	9.043e-01	-1.894	0.058195 .
X3	-6.905e+00	1.964e+00	-3.516	0.000438 ***
X4	4.425e+07	1.888e+07	2.343	0.019122 *
X5	-2.292e-01	1.747e-01	-1.312	0.189551

Table 3.Random Forest

Sales cash flow ratio	0.0803117563
Net operating cash flow/total liabilities	0.0373900579

Net operating cash flow /current liabilities	0.0337649242
Total assets cash recovery ratio	0.0268030430
Profit quality index	0.0113119403
Firm Size	0.0089734359
Return on equity	0.0039267832
Short-term debt to assets ratio	0.0036384218
Current ratio	0.0032999754
Long-term debt to assets ratio	0.0022240949
Return on asset	0.0022103035
x3	0.0021104102
x4	0.0015482753
x5	0.0013264341
Quick ratio	0.0012533269
Fixed assets ratio	0.0009155139
Current assets ratio	0.0008658904
Interest cover ratio	-0.0007301545
Debt to asset ratio	0.0007016611
x2	0.0006254522
Total cash flow ratio	-0.0001354440

x1	-0.0001155913

The logistic regression identify the following variables as statistically significant for the prediction: *Earnings before interest and taxes over total assets (Altman's index-x3), Equity over total liabilities (Altman's index x4), Debt to asset ratio, Size, Net operating cash flow over current liabilities, Current ratio and, finally, Retained earnings over total assets (Altman's index-x2).* It is worth to note that two significant variables in the logistic regression analysis are also Altman's variables confirming the relationship between the two methodologies.

On the other hand, Random Forest gives more relevance to the variables: *Sales cash flow ratio, Net operating cash flow /current liabilities, Net operating cash flow /total liabilities,* and *Total assets cash recovery ratio*.

The only common significant variable is *Net operating cash flow/current liabilities*.

Here we reports the comparison of the predictive power of the Zscore analysis, Logit model and Random Forest (Table 4). The following table shows sensitivity and specificity results for each implemented methodology.

Table 4. Performance Comparison of Zscore analysis, Logit model and Random Forest on the test set

	ZScore	Logit	Random Forest
Sensitivity	0.62	0.80	0.83
Specificity	0.67	0.85	0.87

Random Forest clearly performs better than the other methods, both in terms of Specificity and Sensitivity. Moreover the accuracy of Random Forest is 99.85%, the logistic function scored 99.83% and the Zscore scored 64%. Hence we conclude that Altman's model, ranking third in our comparison, is not the most reliable methodology to predict bankruptcy or financial stress condition.

4. Conclusions

Corporate crisis prediction is an extensively widely discussed topic for its implication on credit risk management and portfolio analysis.

With this paper we intend to research and extend the existing literature in Credit risk management for Small-Medium companies. To this aim, we applied three competing methods (Zscore analysis, Logit model and Random Forest) to predict the failure probability of Italian manufacturing firms from their respective annual accounting data.

All three methodologies obtained appreciable results on failure prediction but, despite its widely usage, the Altman's model may not be the best ranking method. Our results clearly state that Breiman's Random Forest is a superior tool for firm's failure prediction, even better than logistic regression that, nevertheless, has a better performance than the Altman's ZScore. Indeed three out of five Altman's variables (x2, x3, x4) are considered as significant in the logistic regression analysis that seems to extend the Zscore predictive power from Altman's initial choice of parameters, mostly centred on balance sheet and incoming statement data (Altman, 1968). Conversely, Random Forest seems to assign more relevance to cash flow indexes that better illustrate the dynamics of a firm's evolution.

Remarkably, the results obtained in our empirical analysis contradict what stated by Fantazzini et al. (2009), who analyzed the probability of a credit risk default over 1003 German firms. Starting from a set of 16 variables based on balance sheet and incoming statement data, they concluded that logistic regression had a slightly superior predictive power than Random Forest. We could ascribe the inconsistency with our results to the number of firms taken in account (3000 vs 1003), to the different financial context (Italian vs. German firms) and the lack of cash flow variables in Fantazzini's analysis.

The present work is limited by the variables considered here that don't include macroeconomic data or management information which may be useful predictors of financial difficulties.

Future applications of the Random Forest methodology would be certainly improved by modelling it on also macroeconomic data and variables related to management.

References

Abidali A.F., Harris F.(1995), "*A methodology predicting failure in the construction industry*", Construction Management and Economics, Vol.13, 189-196.

Altman E.I. (1968), "*Financial ratios, discriminant analysis and the prediction of corporate bankruptcy*", The Journal of Finance, Vol. 23, N.4, pp.589–609.

Altman E.I., Haldeman R. G., Narayanan P. (1977), "*Zeta analysis. A new model to identify bankruptcy risk of corporations*", Journal of Banking and Finance, N. 1, pp. 29-54.

Balcaen S., Ooghe H.(2006), *"35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems*", The British Accounting Review, Vol.38, pp.63-93.

Beaver W.H. (1966), "*Financial Ratios as Predictors of Failure*", Journal of Accounting Research, Vol. 4, Supplement, pp.71-111.

Becchetti L., Sierra J. (2003), "*Bankruptcy risk and productive efficiency in manufacturing firms*", Journal of Banking & Finance, Vol. 27, pp. 2099–2120.

Berkson J. (1944), *"Application of the logistic function to bio-assay"*, Journal of the American Statistical Association, Vol.39, pp. 357–365.

Breiman L. (2001), *"Random Forests"*, Machine Learning, Vol.45, pp.5–32.

De Andrés J., Landajo M., Lorca P. (2012), "*Bankruptcy prediction models based on multinorm analysis: An alternative to accounting ratios*", Knowledge-Based System, Vol.30, pp.67-77.

Duda R, Hart P, Stork D (2001), *Pattern Classification, Second Edition.* New York, NY, USA, John Wiley and Sons.

Du Jardin P. (2010), "*Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy*", Neurocomputing, Vol. 73, pp.2047–2060.

Fantazzini D., Figini S. (2009), "*Random Survival Forests Models for SMECredit Risk Measurement*", Methodology and Computing in Applied Probability, Vol.11, pp.29-45.

Fuertes A.M , Kalotychou E.(2006), *"Early warning systems for sovereign debt crises: the role of heterogeneity*", Computational Statistics & Data Analysis, Vol.51, pp.1420–1441.

Grice J.S., Ingram R.W.(2001), "*Tests of the generalizability of Altman's bankruptcy prediction model*", Journal of Business Research, Vol.54, pp. 53–61.

Huyhebaert N. R.W., Gaeremynck A., Roodhooft F., Van de Gucht L. M. (2003), "*New Firm Survival: The Effects of Start-up Characteristics"*, Journal of Business Finance & Accounting, Vol.27, pp.523-784.

Ishwaran H., Kogalur U.B.(2006), "*Random survival forests for R*". Rnews, 25–31.

Keasey K., Watson R. (1987), "*Non-financial symptoms and the prediction of small company failure: A test of Argenti's hypotheses*", Journal of Business & Accounting, pp.335-354.

Keasey K., Watson R. (1993), "*The Management of small firms, Ownership, Finance and Performance*", Blackwell, Oxford.

Keasey K., Watson R.(1995), "*The Pricing of small firms Bank finance*", Applied Economics Letter, 2, pp.208-210.

Kohavi R. (1995), "*A study of cross-validation and bootstrap for accuracy estimation and model selection*", Proceedings of the Fourtheenth International Joint Conference on Artificial Intelligence, pp.1137-1143.

Li H., Lee Y.C. ,Zhou Y.C., Sun J.(2011), "*The random subspace binary logit (RSBL) model for bankruptcy prediction"*, Knowledge-Based System, Vol.24, 1381-1388.

Ohlson J.A. (1980), "*Financial ratios and the probability of bankruptcy"*, Journal of Accounting Research, Vol. 18, No. 1, pp. 109-131.

Ooghe H., Joos P., De Vos D., De Bourdeaudhuij C., (1994), "*Towards an improved method of evaluation offinancial distress models and presentation of their results*", Working Paper.

Picard R., Cook D. (1984), "*Cross-Validation of Regression Models*", Journal of the American Statistical Association, pp.575-583.

Pompe P.M., Bilderbeek J.(2004), "*The prediction of bankruptcy of small- and medium-sized industrial firms"*, Journal of Business Venturing, Vol.20, N.6, pp.847-868.

Ravi Kumar P., Ravi V. (2007), "*Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review*", European Journal of Operational Research, Vol. 180, pp.1–28.

Shih S. (2011), "*Random Forests for Classification Trees and Categorical Dependent Variables: an informal Quick Start R Guide"*, http://www.stanford.edu/~stephsus/R-randomforest-guide.pdf.

Shumway T. (2001), "*Forecasting Bankruptcy More Accurately: A simple hazard model*", Journal of Business, Vol. 74, No. 1, pp. 101-124.

Shmueli G., Patel N.R, Bruce P.C. (2011), "*Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner"*, John Wiley & Sons.

Sun L., Shenoy P. (2007), "*Using Bayesian networks for bankruptcy prediction: Some methodological issues"*, European Journal of Operational Research, Vol. 180, pp. 738–753.

Thomas N.G. S., Wong J. M.W., Zhang J. (2011), "*Applying Z-score model to distinguish insolvent construction companies in China"*, Habitat International, Vol.35, pp.599-607.

Yeh C.C., Lin F., Hsu C.Y. (2012), "*A hybrid KMV model, random forests and rough set theory approach for credit rating"*, Knowledge-Based System.

Zmijewski M.E. (1984), "*Methodological Issues Related to the Estimation of Financial Distress Prediction Models,*" Journal of Accounting Research, Vol.22, pp.59-82.