

Variable Selection in Binary Logistic Regression for Modelling Bankruptcy Risk



Francesca Pierri

Abstract One of the most fascinating areas of study in the current economic and financial world is the forecasting of credit risk and the ability to predict a company's insolvency. Meanwhile, one major challenge in constructing predictive failure models is variable selection. Standard selection methods exist alongside new approaches. In addition, the huge availability of data often implies limitations due to processing time and new high-performance procedures provide tools that can take advantage of parallel processing. In the present paper, different variable selection techniques were explored in the context of applying logistic regression for binary data to a balanced data set including only firms active or in bankruptcy. Models deriving from stepwise selection, the Least Absolute Shrinkage and Selection Operator (LASSO) and an unsupervised method, based on the maximum data variance explained, were compared. Then a non-parametric approach was considered and the selection of variables coming from a single decision tree and a forest of trees is compared and discussed.

Keywords Variable selection · LASSO · Stepwise · Unsupervised methods · Decision trees · Logistic · Unbalanced data

1 Introduction

From 2005 onwards, credit risk forecasting and bankruptcy prediction have become among the most important and interesting topics in the modern economic and financial field. However, quantitative methods have long been applied for predicting the bankruptcy event. First, Beaver in 1966 [5] applied discriminant analysis, then Altman [1] in 1968 developed the well-known Z score. Later on, Ohlson [28] in 1980 used logistic regression which has become the most applied model

F. Pierri (✉)

Department of Economics, University of Perugia, Perugia, Italy
e-mail: francesca.pierri@unipg.it; <http://www.stat.unipg.it/~frc/>

in the credit scoring field. Subsequently, in 1992 Narain [27] approached the problem via survival analysis, examining the timing of failure instead of simply considering whether or not an event occurred within a fixed interval of time; since then, Cox's semi-parametric proportional hazard model and its extensions have been extensively proposed and adopted in economic, banking and financial fields [4, 6, 9, 20, 30, 38, 39].

However, whichever model is applied, one major challenge in constructing predictive failure models, as has been widely stated in the literature [2, 3, 7, 8, 15–19], is the effective selection of the most relevant variables from among those that have been collected because of their perceived importance or widespread use.

Besides the problem of correlations between variables that may affect the discriminant ability of a risk model [24], a crucial point remains the procedure chosen for making the selection [13, 45]. Beyond the traditional methods such as backward, forward and stepwise selection, and the use of criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), new approaches known as penalty driven methods (Least Absolute Shrinkage and Selection Operator (LASSO), Smoothly Clipped Absolute Deviation (SCAD) or bridge estimator) [21, 41–44] and machine learning techniques (decision trees and neural networks) [11, 23, 25, 40] have become prominent. Moreover, the increased availability of high-dimensional data, which may impose limitations due to processing time, has led to the development of new high-performance procedures employing tools that can take advantage of parallel processing [37].

In the present paper, based on an application to economic data, we try to provide an answer to the following research questions: (1) do different variable selection methods among standard, modern and those taking advantage of parallel processing, lead to the same choice of variables; (2) which method is better for predicting the future state of a firm.

The paper is structured in the following way: Sect. 2 presents the methodology that will be applied; Sect. 3 gives a brief description of the data; results of the analysis are shown in Sect. 4; and Sect. 5 presents the conclusions of the investigation.

2 Methodology and Study Design

The primary purpose of this paper is to apply different techniques in order to select significant variables for predictive purposes, applying as quantitative method the binary logistic regression model. While acknowledging that different causes may lead to the end of a firm's life, that alternative variables may influence these various events, and that the same variables may even have opposite effects (see [10] and [31]), a single adverse event—bankruptcy—was studied. The problem of overestimating the intercept coefficient in the logistic model [22] due to the relative lack of data on rare events, was overcome by applying one of the available solutions that we have previously applied in statistical analysis [32]. Thus a balanced data

set was built by randomly selecting for each bankrupt firm four controls (firms that did not fail). Training and holdout samples were built to develop and test the models, respectively. The variables selected as relevant by each method were used as explanatory variables in a logistic model. The Wald test was applied to test whether a candidate variable should be included in the model, with the p-value cutoff set at 0.05. Each model's adequacy and predictive capability were tested, through the holdout sample, measuring the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC).

Three parametric (forward-stepwise, LASSO, Maximum Data Variance (MDV) explained [36]) and two non-parametric methods (single and forest decision tree) were applied and compared, taking into account the number of selected variables and the AUC value in the holdout sample.

Focusing attention on SAS[®] software, which provides both standard and high performance (HP) procedures running in either single-machine mode or distributed mode, the following procedures were called upon: LOGISTIC [33] to apply the forward stepwise selection and to run and test all the logistic models; GLMSELECT [34], specifying the logit link, to perform the LASSO selection following the Efron et al. implementation [14]; HPREDUCE [37] to identify variables that jointly explain the maximum amount of data variance; and HPSPLIT [35] and HPFOREST [37] to build a single tree and a forest of trees, respectively.

3 Data Description

The data used in this study were extracted from Orbis [29], a global company database compiled by Bureau Van Dijk, one of the major publishers of business information. Orbis combines private company data with software for searching and analysing over 400 million companies.

The sample employed in the present analyses consists of 37,875 Italian firms operating in the manufacturing sector from 2000 to 2018. For each firm, the financial data for the last available year, its legal form, current legal status and geographical location were extracted. Following the classification of company status available in the Orbis database, three main categories of firms' inactivity were identified: closure, liquidation and bankruptcy (Table 1). As indicated earlier in the Introduction, only one of the adverse events, bankruptcy, was taken into account and, due to its rarity (8.74%), a balanced data set was built by randomly choosing four controls (active firms) for each event (bankrupt firm). The data obtained in this way (16,560 observations) were then split at random into training (80% of the total sample, 13,095 observations) and holdout samples (20% of the total sample, 3465 observations) in order to develop and test the models on independent samples.

Table 1 Firms' distribution by status

Status	N	%
Active	34,046	89.89
Closed	43	0.11
Winding-up	474	1.25
Bankruptcy	3312	8.74
Total	37,875	100

Table 2 Distribution of firms in the training set, by geographical area

		North West	North East	Centre	South	Insular	Total
Active	N	4305	3316	1728	853	274	10,476
	%	32.88	25.32	13.20	6.51	2.09	80.00
Bankruptcy	N	962	729	489	332	107	2619
	%	7.35	5.57	3.73	2.54	0.82	20.00
	Column %	18.3	18.0	22.1	28.0	28.0	
Total	N	5267	4045	2217	1185	381	13,095
	%	40.22	30.89	16.93	9.05	2.91	100.00

Table 3 Distribution of firms in the training set, by legal form (LC = limited company)

		Partnerships	PrivateLC	PublicLC	Total
Active	N	214	8562	1700	10,476
	%	1.63	65.38	12.98	80.00
Bankruptcy	N	43	2290	286	2619
	%	0.33	17.49	2.18	20.00
	Column %	16.7	21.1	14.4	
Total	N	257	10,852	1986	13,095
	%	1.96	82.87	15.17	100.00

The distribution of firms in the training data set by geographical area (Table 2) shows an increasing percentage of defaulting firms going from the North (18%) to the South (28%). Moreover, private limited companies (21%) seem to be more prone to the adverse event (Table 3).

For each firm indexes or ratios representative of its economic and financial situation were built, taking into account both their perceived importance and widespread use in the literature [1, 5, 12, 26] and the information availability required for the calculation. Correlation problems were solved by including only one of the ratios among those with correlation higher than 0.70. Finally, besides the firm's age, geographical area and legal form, 37 indexes were used (Table 4), including liquidity and solvency ratios, profitability and operating efficiency ratios.

Table 4 Indexes evaluated as potential predictors of the bankruptcy event

ID	Formula	ID	Formula
ind001	ln (EBITDA)	ind079	Quick Assets/Sales
ind004	Operating Revenue/Inventories	ind080	Quick Assets/Total Assets
ind007	Cash flow/Current Liabilities	ind083	Profit (Loss) for period/Shareholders' Funds
ind011	Cash flow/Shareholders' Funds	ind084	EBIT/Shareholders' Funds
ind020	Ln(Total Assets)	ind085	Profit (Loss) for period/Operating Revenue
ind021	(Creditors/Operating Revenue)*360	ind087	Sales/Cash flow
ind031	Current Assets/Current Liabilities	ind088	Sales/Current Assets
ind033	Debtors/Operating Revenue	ind089	Sales/EBIT
ind042	Shareholders' Funds/Total Assets	ind090	Sales/Equity ratio
ind044	Equity/Fixed Assets	ind092	Operating Revenue/Total Assets
ind050	Inventory/Sales	ind093	Sales/Working Capital
ind052	Inventory/Working Capital	ind094	Shareholders' Funds/Capital
ind055	Long Term Debts/Sales	ind104	Sales/Shareholders' Funds
ind056	Long Term Debts/Net Capital	ind105	Working Capital
ind058	Non Current Liabilities/Total Assets	ind116	EBIT/Interest paid
ind060	(Long Term Debt + Loans)/Total Assets	ind117	Long Term Debts/Equity
ind063	Net Income/Cash flow	ind124	Debtors/Current Assets
ind065	Net Income/Fixed Assets	ind132	Equity/Sales
ind072	Non-Current Liabilities/Sales		

Table 5 Variable selection comparison among stepwise, LASSO and maximum data variance explained methods

Variables	Stepwise	LASSO	MDV
N. selected	21	19	13
% In common	61.90	68.42	100
AUC training	0.9081	0.906	0.9040
AUC holdout	0.8908	0.8921	0.8903

4 Results

4.1 Stepwise, LASSO and Maximum Data Variance Selection Methods

The variable selection comparison between the stepwise, LASSO and maximum data variance (MDV) explained techniques, shows good performance of all three methods. Although the best performance in the holdout sample was given by the LASSO (AUC = 0.8921), AUC values under the other methods were extremely close (Table 5). The MDV method selected the smallest number of indexes (13), which in turn are also identified by the other two techniques. As shown in Table 5 the three approaches agree on the selection of more than 60% of the variables.

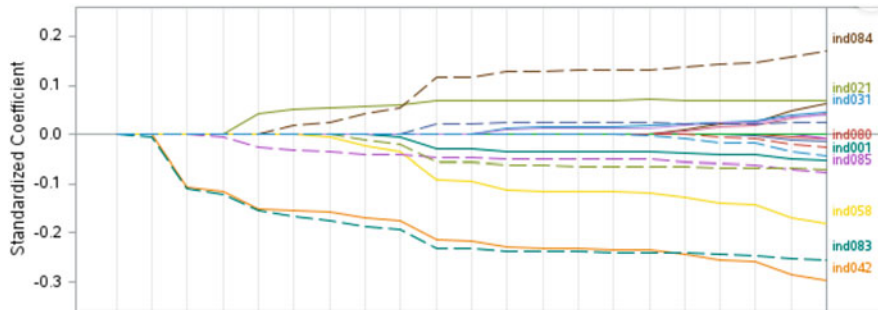


Fig. 1 Coefficient progression for response variable: output from GLMSELECT procedure

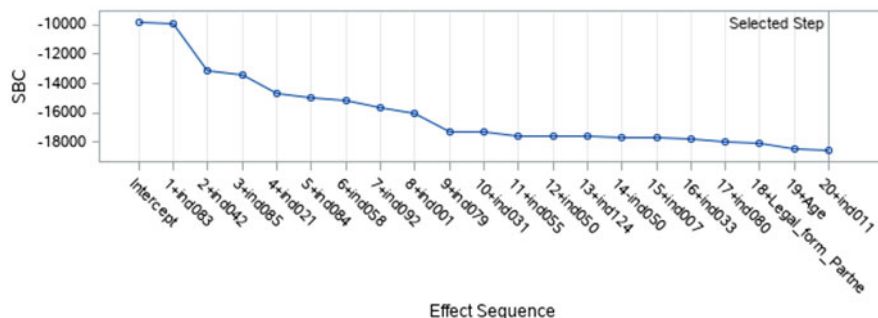


Fig. 2 Effect Sequence: output from GLMSELECT procedure

The LASSO output results from the GLMSELECT procedure include detailed graphs as an aid to interpretation. Figure 1 shows the coefficient progression for the response variable: the names of the most important indexes affecting bankruptcy appear on the right-hand side, with those above the zero line increasing the probability of the event under study when their value increases and those below the zero line decreasing it. Coefficients corresponding to effects that are not in the selected model at a step are zero and hence not observable. Figure 2, complementary to the previous graph, shows how the average square error used to choose among the examined models progresses. The initial model includes only one index (ind0042), then a second one (ind0085) is added and so on (Fig. 2). The procedure stops at the 20th step.

4.2 Single and Forest of Trees Methods

The two non-parametric approaches showed very similar results. The single tree and the forest of trees had in common 12 indexes, that is, respectively, 75% and 80% of the variables selected. Their performances in the holdout sample

Table 6 Variable selection comparison between the two non-parametric approaches

Variable	Single tree	Forest of trees
N. selected	16	15
% In common	75.00	80.00
AUC training	0.9061	0.9037
AUC holdout	0.8892	0.8888

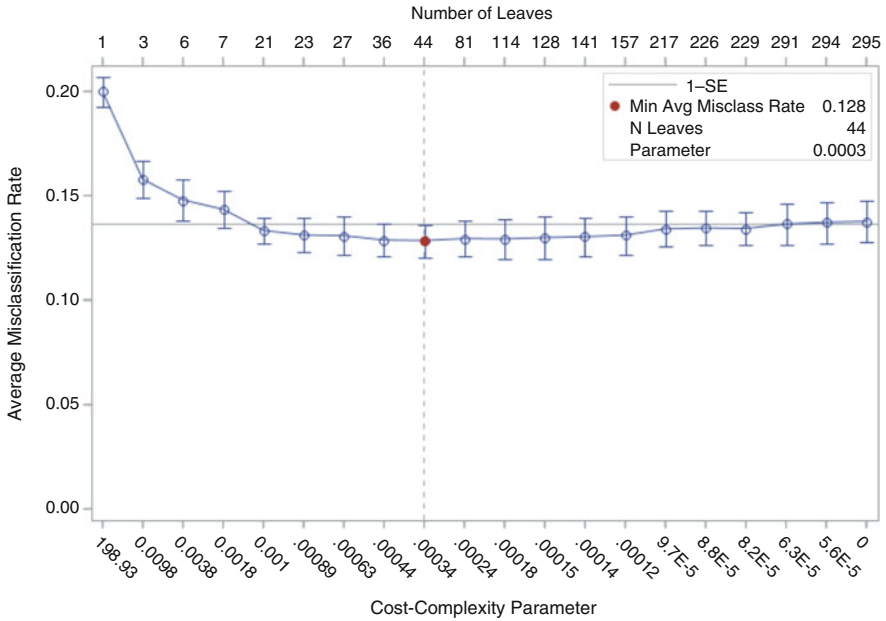


Fig. 3 Cost complexity analysis using cross-validation: PROC HPSPLIT output

were virtually identical (Table 6). HPSPLIT plots provide a tool for selecting the parameters that result in the smallest estimated Average Square Error (Fig. 3) and a classification tree (Fig. 4) that uses colours to aid understanding of where the higher percentage of active firms is found: blue for bankruptcy, and pink for active.

In Fig. 5 the subtree starting at node 0 shows important details regarding the indexes' values, that is, the cut-off at which they cause the separation into new leaves.

4.3 Comparison Between the Best Method of Each Group

Even though all the methods applied in this context lead to very similar results, the best of each group was selected (LASSO and single tree methods) with the aim of making a more detailed comparison among a parametric and non

Classification Tree for D3

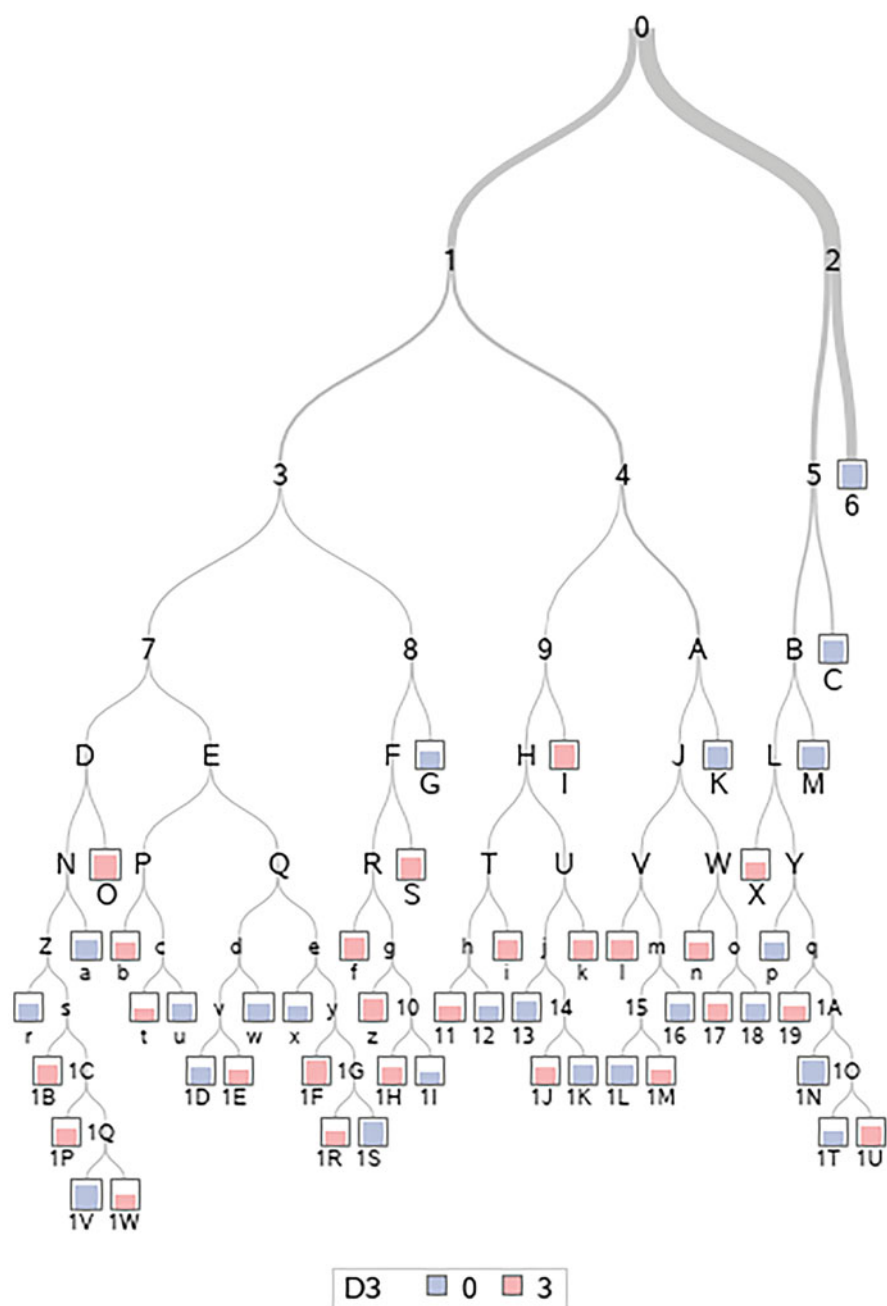


Fig. 4 Classification tree: PROC HPSPLIT output

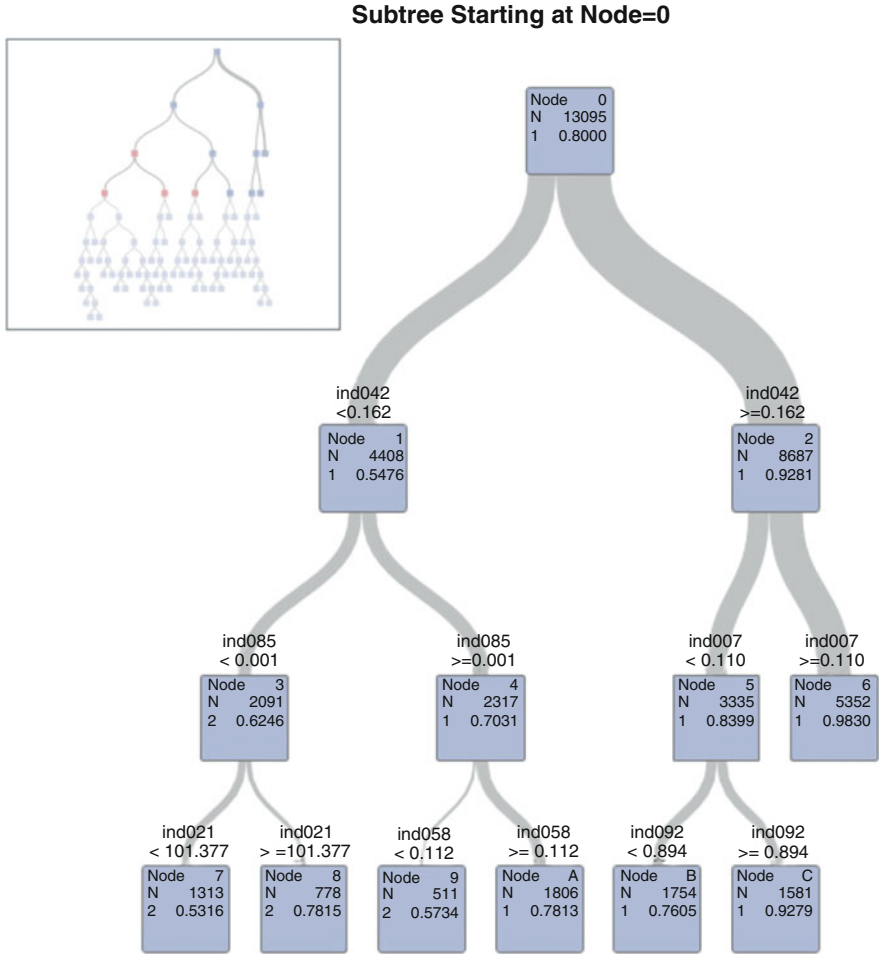


Fig. 5 Subtree starting at node 0: PROC HPSPLIT output

parametric technique (Table 7). The variable selection comparison, on the basis of the AUC value, showed a slight predominance of the first one, however, the difference was extremely small (0.891 against 0.8892). LASSO selected a slightly greater number of variables as predictors, most of which (14) were in common with the single tree method (73.68%). Table 8 shows the ratios that they had in common.

Table 7 Variable selection comparison between the best method in each group

Variable	LASSO	Single tree
N. selected	19	16
% In common	73.68	87.50
AUC training	0.9060	0.9061
AUC holdout	0.8921	0.8892

Table 8 Predictive variables in common between LASSO and single tree methods, in addition to Age and Legal Form. Increased values of variables above and below the horizontal line raise and reduce, respectively, the probability of bankruptcy

ID	Formula
ind021	(Creditors/Operating Revenue)*360
ind031	Current Assets/Current Liabilities
ind033	Debtors/Operating Revenue
ind060	(Long Term Debt + Loans)/Total Assets
ind084	EBIT/Shareholders' Funds
ind001	ln (EBITDA)
ind042	Shareholders' Funds/Total Assets
ind058	Non Current Liabilities/Total Assets
ind083	Profit (Loss) for period/Shareholders' Funds
ind085	Profit (Loss) for period/Operating Revenue
ind092	Operating Revenue/Total Assets
ind124	Debtors/Current Assets

5 Discussion

Variable selection techniques were evaluated within two main groups of methods and then the best of each group were compared further. The first group considered the standard and widely used forward stepwise selection method, the LASSO technique, and a procedure that conducts a variance analysis and reduces dimensionality by selecting the variables that contribute the most to the overall variance of the data. Among these, the models refitted and tested through logistic regression showed very stable results. The AUC values in the holdout sample were very close, with differences only in the third decimal point. The selection was most parsimonious using the third method which discarded variables that are included by both the stepwise and LASSO methods (Table 5), but the AUC value was slightly higher.

The non-parametric approach showed very slight differences between the single tree and the forest methods. Again the differences lay in the third decimal places of the AUC (in the holdout sample) and the number of selected variables was almost the same, with most of these in common.

The final comparison between LASSO and single tree selection methods highlighted that these different techniques led to models with high and stable predictive performance in the holdout sample, with a preference towards the first method for its slightly higher AUC value (0.8921 against 0.8892) and for its computational performance in terms of processing time (0.91 vs. 25.16 seconds). Moreover, the LASSO and single tree approaches selected almost the same predictive variables with a smaller number in the second. In particular both gave particular relevance

to variable ind042 reflecting the ratio of Shareholders' Funds to Total Assets: both LASSO and single tree selected it first, on the basis of the average square error and variable importance. This confirms the protection from bankruptcy provided by strong corporate capital structure, while the credit situation (ind021) and debt exposure (ind060) may play an opposite rule [31].

The SAS software procedures used (GLMSELECT and HPSPLIT) both provide very intuitive graphs although perhaps the LASSO ones seem easier to interpret for a wider and non-technical audience. On the other hand HPSPLIT is a high performance procedure that runs in either single-machine mode or distributed mode and can therefore take advantage of parallel processing.

Uniformity in the predictive capability of these selection methods may have been affected by data dimensionality, therefore in the future the same procedures will be applied to a smaller data set. Future developments will also include the extension to multinomial logistic analysis.

References

1. Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Finance* **23**(4), 589–609 (1968)
2. Amendola, A., Restaino, M., Sensini, L.: Variable selection in default risk models. *J. Risk Model Validation* **5**(1), 3 (2011)
3. Austin, P.C., Tu, J.V.: Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *J. Clin. Epidemiol.* **57**(11), 1138–1146 (2004)
4. Banasik, J., Crook, J.N., Thomas, L.C.: Not if but when will borrowers default. *J. Oper. Res. Soc.* **50**(12), 1185–1190 (1999)
5. Beaver, W.H.: Financial ratios as predictors of failure. *Journal of Account. Res.* **4**, 71–111 (1966)
6. Bonini, S., Caivano, G.: The survival analysis approach in Basel II credit risk management: modeling danger rates in the loss given default parameter. *J. Credit Risk* **9**(1), 101–118 (2013)
7. Bunea, F.: Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Stat.* **2**, 1153–1194 (2008)
8. Bursac, Z., Gauss, C.H., Williams, D.K., Hosmer, D.W.: Purposeful selection of variables in logistic regression. *Source Code Biol. Med.* **3**(1), 1–8 (2008)
9. Cao, R., Vilar, J.M., Devia, A., Veraverbeke, N., Boucher, J.P., Beran, J.: Modelling consumer credit risk via survival analysis. *SORT Stat. Oper. Res. Trans.* **33**(1), 31–47 (2009)
10. Caroni, C., Pierri, F.: Different causes of closure of small business enterprises: alternative models for competing risks survival analysis. *Electron. J. Appl. Stat. Anal.* **13**(1), 211–228 (2020)
11. Dreiseitl, S., Ohno-Machado, L.: Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inform.* **35**(5–6), 352–359 (2002)
12. Du Jardin, P.: Predicting bankruptcy using neural networks and other classification methods: the influence of variable selection techniques on model accuracy. *Neurocomputing* **73**(10), 2047–2060 (2010). <https://doi.org/10.1016/j.neucom.2009.11.034>, <https://www.sciencedirect.com/science/article/pii/S0925231210001098>, subspace Learning/Selected papers from the European Symposium on Time Series Prediction
13. Du Jardin, P.: The influence of variable selection methods on the accuracy of bankruptcy prediction models. *Bank. Mark. Invest.* **116**, 20–39 (2012)

14. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. *Ann. Stat.* **32**(2), 407–499 (2004)
15. Fan, J., Li, R.: Variable selection for Cox’s proportional hazards model and frailty model. *Ann. Stat.* **30**(1), 74–99 (2002)
16. Fan, J., Li, G., Li, R.: An overview on variable selection for survival analysis. In: *Contemporary Multivariate Analysis and Design of Experiments: In Celebration of Professor Kai-Tai Fang’s 65th Birthday*, pp. 315–336 (2005)
17. Fu, Z., Parikh, C.R., Zhou, B.: Penalized variable selection in competing risks regression. *Lifetime Data Anal.* **23**(3), 353–376 (2017)
18. Ghosh, K., Ramteke, M., Srinivasan, R.: Optimal variable selection for effective statistical process monitoring. *Comput. Chem. Eng.* **60**, 260–276 (2014)
19. He, Z., Tu, W., Wang, S., Fu, H., Yu, Z.: Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics* **71**(1), 178–187 (2015)
20. Kiefer, N.M.: Economic duration data and hazard functions. *J. Econ. Literature* **26**(2), 646–679 (1988)
21. Kim, J., Sohn, I., Jung, S.H., Kim, S., Park, C.: Analysis of survival data with group lasso. *Commun. Stat. Simul. Comput.* **41**(9), 1593–1605 (2012)
22. King, G., Zeng, L.: Logistic regression in rare events data. *Political Anal.* **9**(2), 137–163 (2001)
23. Kumar, A., Rao, V.R., Soni, H.: An empirical comparison of neural network and logistic regression models. *Mark. Lett.* **6**(4), 251–263 (1995)
24. Kundu, S., Mazumdar, M., Ferket, B.: Impact of correlation of predictors on discrimination of risk models in development and external populations. *BMC Med. Res. Methodol.* **17**(1), 1–9 (2017)
25. Meier, L., Van De Geer, S., Bühlmann, P.: The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **70**(1), 53–71 (2008)
26. Mossman, C.E., Bell, G.G., Swartz, L.M., Turtle, H.: An empirical comparison of bankruptcy models. *Financial Rev.* **33**(2), 35–54 (1998)
27. Narain, B.: Survival analysis and the credit granting decision. In: Thomas, L.C., Crook, J.N., Edelman, D.B. (eds.), *Credit Scoring and Credit Control*, pp. 109–122. Oxford University Press (1992)
28. Ohlson, J.A.: Financial ratios and the probabilistic prediction of bankruptcy. *J. Account. Res.* **18**(1), 109–131 (1980)
29. Orbis: Orbis. Bureau van Dijk, <https://orbis.bvdinfo.com/>. Accessed June 2020
30. Pierri, F., Caroni, C.: Bankruptcy prediction by survival models based on current and lagged values of time-varying financial data. *Commun. Stat. Case Stud. Data Anal. Appl.* **3**(3–4), 62–70 (2017)
31. Pierri, F., Caroni, C.: Analysing the risk of bankruptcy of firms: survival analysis, competing risks and multistate models. In: *Demography of Population Health, Aging and Health Expenditures*, pp. 385–394. Springer (2020)
32. Pierri, F., Stanghellini, E., Bistoni, N.: Risk analysis and retrospective unbalanced data. *Revstat-Stat. J.* **14**(2), 157–169 (2016)
33. SAS: SAS/STAT® 9.22 User’s Guide. https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#logistic_toc.htm. Accessed 19 Nov 2022
34. SAS: SAS/STAT® 9.22 User’s Guide. https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#glmselect_toc.htm. Accessed 19 Nov 2022
35. SAS: SAS/STAT® 9.22 User’s Guide. https://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_hpsplit_overview.htm. Accessed 19 Nov 2022
36. SAS: SAS® Enterprise Miner™: High-Performance Procedures. https://documentation.sas.com/doc/en/emhpprcrref/14.2/emhpprcrref_hpreduce_details01.htm (2016). Accessed 19 Nov 2022
37. SAS Institute Inc., Cary, NC: SAS® Enterprise Miner™ 15.2: High-Performance Procedures, last updated: August 18, 2022
38. Shumway, T.: Forecasting bankruptcy more accurately: a simple hazard model. *J. Bus.* **74**(1), 101–124 (2001)

39. Stepanova, M., Thomas, L.: Survival analysis methods for personal loan data. *Oper. Res.* **50**(2), 277–289 (2002)
40. Sun, K., Huang, S.H., Wong, D.S.H., Jang, S.S.: Design and application of a variable selection method for multilayer perceptron neural network with lasso. *IEEE Trans. Neural Netw. Learn. Syst.* **28**(6), 1386–1396 (2016)
41. Tang, Z., Shen, Y., Zhang, X., Yi, N.: The spike-and-slab lasso Cox model for survival prediction and associated genes detection. *Bioinformatics* **33**(18), 2799–2807 (2017)
42. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **58**(1), 267–288 (1996)
43. Tibshirani, R.: The lasso method for variable selection in the Cox model. *Stat. Med.* **16**(4), 385–395 (1997)
44. Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **73**(3), 273–282 (2011)
45. Zellner, D., Keller, F., Zellner, G.E.: Variable selection in logistic regression models. *Commun. Stat. Simul. Comput.* **33**(3), 787–805 (2004)