

Beyond the Cox Model: Applying Machine Learning Techniques with Time-to-Event Data

Francesca Pierri[1(✉)], Damiano Perri[2], and Chrys Caroni[3]

[1] Department of Economics, University of Perugia, Perugia, Italy
francesca.pierri@unipg.it
[2] Department of Mathematics and Computer Science, University of Perugia, Perugia, Italy
damiano.perri@unipg.it
[3] Department of Mathematics, National Technical University of Athens, Athens, Greece
ccar@math.ntua.gr

Abstract. The enhancement of data storage capacity over the years and the advancement of various technologies that enable data acquisition and support big data have increased interest in collecting information and its variation over time. In many different disciplines, the aim of monitoring this kind of observations is to analyse and predict the time until the occurrence of an event of interest.

For 50 years, Cox's semi-parametric proportional hazards regression model has been the statistical methodology most widely used for survival analysis (the analysis of time-to-event data). However, nowadays, machine learning techniques are becoming more popular and several studies assert their supremacy from the predictive point of view.

The purpose of the present research is the application of new methodologies related to the Cox model and their comparison with the classic survival model in the field of corporate default risk prediction. In particular, *scikit survival* [1], a Python module for survival analysis designed to be compatible with *scikit-learn* [2] Application Program Interface (API), will be evaluated as a machine learning technique and the results compared with those obtained by applying classical survival analysis techniques implemented with the SAS® PHREG procedure.

Keywords: Survival Analysis · Machine Learning · Penalised methods · Prediction · Bankruptcy · scikit-survival

1 Introduction

The increased ability to acquire and store data has enhanced over time the interest in longitudinal studies in the most diverse fields of application, from medicine and engineering to economics, with the aim of predicting the time of the occurrence of the event object of study. In such studies you fix a start and an end of the observational period along with collecting for all the subjects the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14815, pp. 412–427, 2024.
https://doi.org/10.1007/978-3-031-65154-0_25

information about the variables of interest; however, only for some, due to various reasons (duration of study, loss of contact with the subject), will you observe an event. When you are in the presence of censored data it is not appropriate to straightaway apply predictive algorithms using standard statistical or machine learning approaches.

For about 50 years, Cox's semi-parametric proportional hazards regression model [3] and its extensions (for a detailed review see [4,5]) have been applied to analyse survival data and handle censored data problems. Over the last few decades, machine learning techniques have become increasingly popular and comparison among the classic and new models is of interest [6]. In the medical field, Cruz and Wishart [7] and Kourou et al. [8] discuss applications in cancer prediction and provide a comparison of several machine learning techniques. Wan et al. [9] assess the effectiveness of machine learning methods for time-to event recurrence prediction and Spooner [10] states that the latter may be a valid alternative to Cox proportional hazards model in the presence of high dimensional data. The purpose of the present research is to expand previous researches in the field of corporate default risk prediction, applying machine learning methodologies developed for survival analysis and comparing results with the classical survival model. In particular, *scikit survival*, a Python module for survival analysis designed to be compatible with *scikit-learn* API, will be evaluated as a machine learning technique and the results compared with those obtained by applying classical survival analysis techniques implemented with the SAS® PHREG procedure. It would also have been possible to apply SAS using Python libraries, which would have allowed all calculations to be performed with a single script. Instead, we preferred to use SAS ® PHREG procedure as in our workflow its report was easily obtainable and required no additional code. The rest of the paper is structured as follows: a brief review of the Cox proportional hazards model and Cox model-based penalised regression approaches; description of data used for the analysis; results and discussion.

2 Survival Analysis

Survival analysis is designed to connect covariates with the time of an event. The concept of survival analysis is derived from clinical research [11–14], where the primary goal is to predict the time of death, which is survival. In engineering, it is also referred to as reliability analysis [15,16], which is applied to assess the time-to-failure of mechanical items or products. In economics and finance, it is used in the credit scoring field [17–23] to understand the causes of financial distress and identify the main predictors of the probability of business failure within a given time frame. In survival analysis, unlike traditional regression, parts of the training data are only partially observed due to censoring. The survival function is a key measure that relates time to the probability of surviving past a certain time point.

Let T denote a continuous non-negative random variable corresponding to a subject's survival time. The survival function $S(t)$ returns the probability of

survival beyond time t and is defined as $S(t) = P(T > t)$. The survival function decreases monotonically with time t, and its initial value at $t = 0$, when all observed subjects are alive, is 1.

The cumulative distribution function $F(t)$, which represents the probability that the event of interest occurs before t, is defined as $F(t) = 1 - S(t)$.

The hazard function $h(t)$ representing the rate of event at time t given that no event occurred before time t has the form shown in Eq. 1.

$$h(t) = \lim_{\delta t \to 0} \frac{P[t < T \leq t + \delta t | T > t]}{\delta t} = \frac{f(t)}{S(t)} \tag{1}$$

2.1 Cox's Semi-parametric Regression Model

Cox's semi-parametric proportional hazards model defines the dependence of the hazard function on the values of a vector of covariates $\mathbf{x}$ as shown in Eq. 2,

$$h(t|\mathbf{x_i}) = h_0(t) e^{\mathbf{x_i}'\boldsymbol{\beta}} \tag{2}$$

where $h_0(t)$ is the baseline hazard function. The parameter β in the hazard ratio expression ($e^{\mathbf{x_i}'\boldsymbol{\beta}}$) is estimated through the partial likelihood method shown in Eq. 3,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{e^{\mathbf{x_i}'\boldsymbol{\beta}}}{\sum_{j \in R(t)} e^{\mathbf{x_j}'\boldsymbol{\beta}}} \right\} \tag{3}$$

which requires no specification about h_0.

2.2 Penalised Methods

Correlation among variables and high-dimensional data often suggests the use of penalised methods, that include Lasso [24], Ridge [25], and Elastic Net [26]. The mathematical problem caused by matrix inversion in the presence of high correlation is solved by introducing a penalty that shrinks the coefficients to 0. The Lasso (Least Absolute Shrinkage and Selection Operator) penalty allows the selection of a small subset of features that are most predictive of the event object of study. The so called $l1$ penalty is given by Eq. 4.

$$\arg\max_{\boldsymbol{\beta}} \quad l(\boldsymbol{\beta}) - \frac{\alpha}{2} \sum_{i=1}^{n} |\beta_i| \tag{4}$$

In Eq. 4, $l(\boldsymbol{\beta})$ is the log likelihood partial function of the Cox model, $\beta_1, ...\beta_k$ are the k parameters for the k features and $\alpha \geq 0$ is a hyper-parameter that controls the amount of shrinkage. Therefore, when α is set to 0 you obtain the standard Cox model.

The Ridge method solves the mathematical problem in the presence of correlation; the penalty used, known as $l2$, is given by the following Eq. 5.

$$\arg\max_{\boldsymbol{\beta}} \quad l(\boldsymbol{\beta}) - \alpha \sum_{i=1}^{k} \beta_i^2 \tag{5}$$

The Elastic Net penalty uses a weighted combination of the $l1$ and $l2$ penalty by solving Eq. 6.

$$\arg\max_{\boldsymbol{\beta}} \quad l(\boldsymbol{\beta}) - \alpha(r \sum_{i=1}^{k} |\beta_i| + \frac{1-r}{2} \sum_{i=1}^{k} \beta_i^2) \tag{6}$$

In Eq. 6, $r \in [0; 1]$ is the relative weight of the $l1$ and $l2$ penalty. Consequently, it is clear that the subset selection property of the Lasso is combined with the regularisation strength of the Ridge penalty.

3 Data

Data used in this paper were extracted from the Orbis [27] database, which offers a collection of information and tools for searching and analysing over 400 million private companies. The same data were used in previous research [28] where different variable selection methods in binary regression model are compared. The sample extracted consist of 37875 Italian firms operating in the manufacturing sector during the period 2000–2018.

For each firm we know the financial data, the legal form, the current legal status and the geographical location. From the whole sample we selected the firms still active at the end of the observational period (34046) and those closed as a result of bankruptcy (3312). The low incidence of the default event (8.86%) in the total sample, suggested/required a data balancing operation, randomly choosing a number of active firms (3312) equal to the ones closed (3312) due to bankruptcy. Therefore, the pre-processing phase resulted in reducing the sample size from 37358 to 6624 observations, taking all the bankrupt firms and an equal number of active enterprises chosen through random selection without replacement. The final data, consisting of 6624 observations, was then split randomly into training (80% of the total sample, 5299 observations) and validation (20% of the total sample, 1325 observations) data sets. The first is used to develop the model and the second to test its predictive ability.

From the financial data extracted from the database, a series of indexes with economic and financial relevance were computed, according to their deemed appropriateness and literature evidence [29–32]. A detailed description of the statistical distribution for the indexes representative of the overall firms' performance status including profitability, efficiency, solvency and liquidity, can be found in the appendix, specifically, in Table 5, Table 6 and Table 7 (data in the appendix is organised into three tables to improve readability).

The firms' distribution (in the training set) by geographical area (Table 1) and Legal form (Table 2) shows that the bankruptcy event is more frequent in the

Table 1. Distribution of firms in the training set, by geographical area

		North West	North East	Centre	South	Insular	Total
Active	N	1064	858	1728	220	74	2657
	%	20.08	16.19	8.32	4.15	1.40	50.14
Bankruptcy	N	956	744	489	325	113	2642
	%	18.04	14.04	9.51	6.13	2.13	49.86
	Column %	36.18	46.44	53.33	59.63	60.43	
Total	N	2020	1602	945	545	187	5299
	%	38.12	30.22	17.23	10.28	3.53	100.00

South (59.63%) and Insular (60.43%) areas and in Private Limited Companies (51.33%).

Table 2. Distribution of firms in the training set, by legal form (LC = limited company)

		Partnerships	PrivateLC	PublicLC	Total
Active	N	58	2188	411	2957
	%	1.09	41.29	7.76	50.14
Bankruptcy	N	46	2308	288	2642
	%	0.87	43.56	5.43	49.86
	Column %	44.23	51.33	41.20	
Total	N	104	4496	699	
	%	1.96	82.87	15.17	100.00

4 Results

4.1 Classic Survival

We performed regression analysis of survival data based on the Cox proportional hazards model applying the SAS® PHREG procedure to the training data set. First of all, correlation problems among indexes were checked and among those indexes with correlation higher than 0.70 we included only one. The selection of one index rather than another is influenced by its broader applicability or its lack of correlation with other indexes.

The stepwise model selection method was applied: variables are entered into and removed from the model in such a way that each forward selection step can be followed by one or more backward elimination steps [33].

The resulting model incorporates significant variables such as Legal Form, Geographical Area, and 17 indexes (Table 3). It is noted that concerning the geographical area, the variable holds overall significance; however, the significant

comparisons are those excluding the number 1 within the HR confidence interval. As regards the Legal Form, Private Limited Companies are more susceptible to bankruptcy (HR = 1.387) compared to Partnerships, which have a lower risk (HR = 0.689). Indexes with negative estimated coefficients are associated with a decreased probability of bankruptcy, whereas those with positive coefficients are linked to an increased risk. The interpretation of the estimated hazard ratio (HR) is the rate of progression to bankruptcy among firms. For example, a one-unit increase in ind007 (current ratio index) results in the probability of bankruptcy rising by 2.762 times, with a 95% confidence interval ranging from 2.386 to 3.198. Additionally, a one-unit increase in ind029 (short-term debt ratio) causes the same probability to increase more than tenfold. Conversely, ind042 (equity ratio) and ind068 (Net income/Total Debts) are associated with a protective effect against the default event, reducing its likelihood by factors of 0.994 (HR = 0.006) and 0.983 (HR = 0.017), respectively.

The concordance index for the training and validation datasets is 0.7925 and 0.7818, respectively. The area under the ROC curve (AUC) increases from 0.7215 to 0.8025 over a period of 2 to 10 years in the validation data, whereas in the training dataset, it rises from 0.7846 to 0.8245.

4.2 Penalised Methods

The subsequent analyses are performed using *scikit-survival* [1], a Python module specifically designed for survival analysis. It is constructed on top of the widely-used open-source *scikit-learn* library [2], which incorporates a broad array of cutting-edge machine learning algorithms for medium-scale supervised and unsupervised problems.

Lasso. The use of the Lasso penalty in the training set returns a model with concordance index equal to 0.8052. This index experiences a slight decrease to 0.7947 in the validation set. At first, we calculated an α value that does not take into account any features, and then gradually decreased it until it was only about 1% of the initial value. The final $l1$ value, which is used for selecting the optimal model, is not returned by the software. In Fig. 1 you see that this method for larger α values selects a smaller subset of features. On the left you may notice among the most important features ind042 and ind082.

Ridge. Various α values were evaluated, and the minimum viable value was determined to be 0.5; any value lower than this caused the system to crash. When training the model with $\alpha = 0.5$, we achieve the highest concordance index of 0.8141. The validation results also show a satisfactory concordance index (0.7943), which reaches its best value (0.7954) for $\alpha = 8$. In Fig. 2, it is evident that as the penalty weight increases, all coefficients are reduced to zero. Additionally, Fig. 3a illustrates that ind042 and ind082 are once again deemed significant predictors of bankruptcy. The selected indexes for $\alpha = 0.5$ and $\alpha = 8$ are displayed in Fig. 3a and Fig. 3b respectively. Here, only a change in the first five significant features can be observed.

Table 3. Estimated regression coefficients, Hazard Ratio (HR) and 95% Confidence Interval (CI) in survival analyses. Reference category: *Private Limited Companies, **Centre

	Est. Coefficient	HR	95% HR CI
Legal Form *			
Partnerships	−0.373	0.689	0.51–0.929
PublicLC	0.327	1.387	1.215–1.583
Geographical Area **			
North West	0.020	1.020	0.915–1.138
North East	−0.069	0.934	0.832–1.047
South	0.252	1.286	1.116–1.482
Insular	0.181	1.199	0.973–1.476
ind001	−0.026	0.974	0.960–0.989
ind079	−0.076	0.926	0.879–0.977
ind007	1.016	2.762	2.386–3.198
ind083	−0.098	0.907	0.880–0.935
ind084	0.060	1.062	1.038–1.087
ind042	−5.071	0.006	0.004–0.010
ind092	−0.281	0.755	0.686–0.832
ind094	−0.003	0.997	0.995–0.999
ind055	0.305	1.357	1.223–1.506
ind116	−0.015	0.985	0.982–0.989
ind060	0.470	1.600	1.254–2.041
ind117	−0.012	0.988	0.980–0.997
ind124	−0.495	0.610	0.513–0.724
ind132	0.389	1.475	1.343–1.621
ind029	2.329	10.268	7.838–13.451
ind068	−4.049	0.017	0.010–0.029
ind082	5.280	196.461	80.239–481.027

Elastic Net. The weighted combination of $l1$ and $l2$ regularisation displays a coefficient plot (Fig. 4) with a shape closely resembling those in Fig. 1 and Fig. 2. However, the most significant features differ (Fig. 5). The top-performing model achieved a concordance index of 0.8006 in the training set and 0.7943 in the

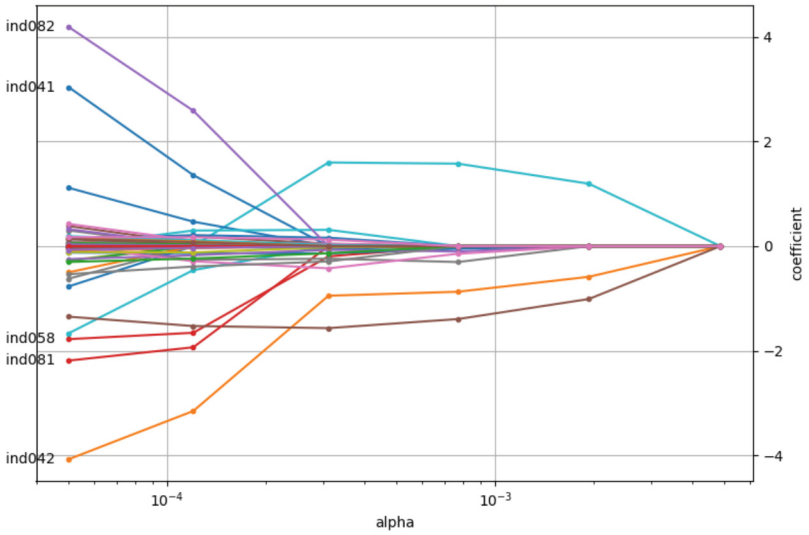


Fig. 1. Lasso coefficient plot

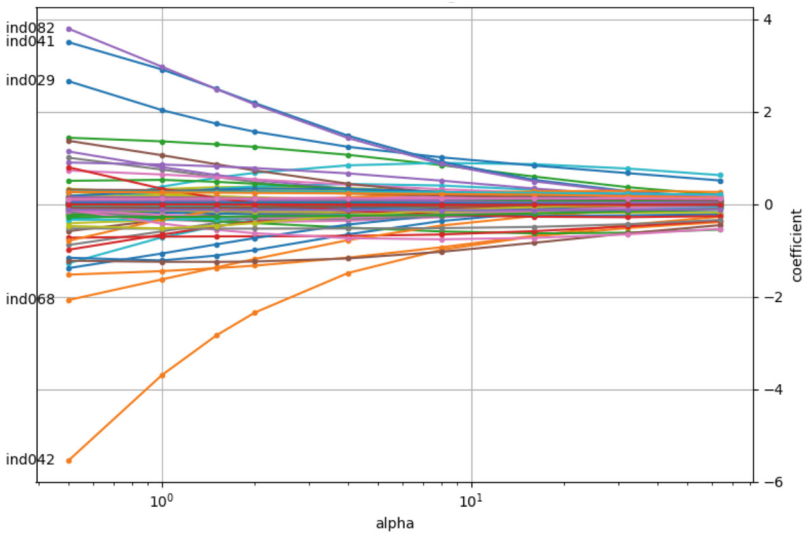
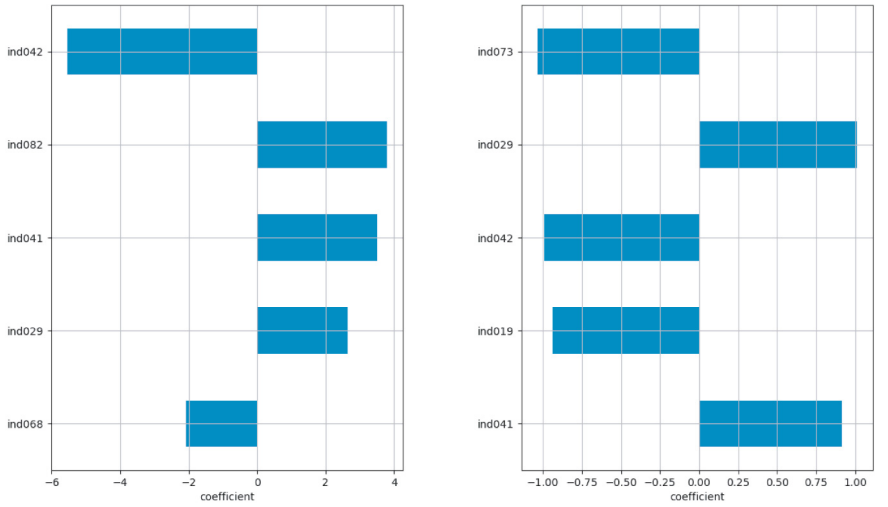


Fig. 2. Ridge coefficient shrinkage for varying α

validation set. As in Lasso in the output, besides the most important features, the value chosen for the penalty parameters are not shown.



(a) Ridge's best coefficients for $\alpha = 0.5$ (b) Ridge's best coefficients for $\alpha = 8$

Fig. 3. Comparison of Ridge's best coefficients for different α values

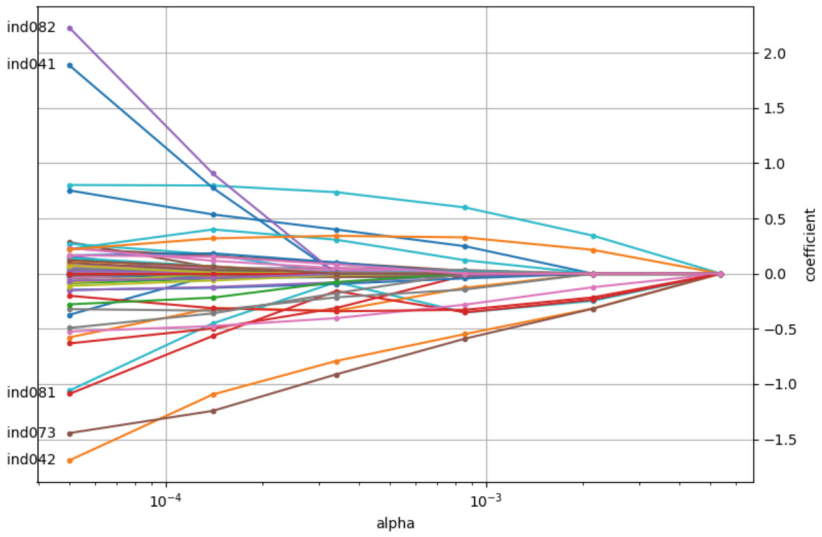


Fig. 4. Elastic Net Coefficients

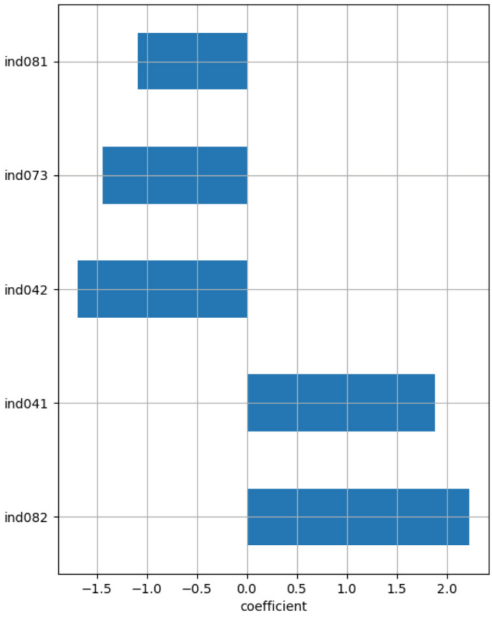


Fig. 5. Elastic Net Features importance

5 Conclusions

The growing interest in machine learning applications has led to the adoption of a user-friendly package such as *scikit-learn* [2], which now includes a module specifically for survival analysis. This module is being employed across diverse research domains, as evidenced by [34–37] and in this paper, we have implemented it in the context of predicting corporate default risk. Utilising the same dataset, divided into training and validation sets, we developed a classical Cox proportional hazards model using SAS software. Subsequently, employing *scikit-survival*, we focused on penalised methods to perform a comparative analysis.

Examining the concordance index within the validation dataset, the optimal results are achieved with ridge penalisation, registering a score of 0.7954 when the penalty parameter is set to 8. Fixing an α value equal to 0.5 Lasso and Ridge methods select the same five significant features, the first with a slightly lower concordance index in the training set, while there is almost parity in the validation set, with the difference being only in the fourth decimal place.

Among the penalised methods, the Elastic Net yields the lowest concordance index in the training set (0.8006) incorporating only a subset of the features chosen by the other two models. In the validation set, the difference occurs at the fourth decimal place.

On the other hand, the classical survival model has the lowest concordance index in the validation set (0.7818) compared to others, yet it demonstrates an increasing predictive ability over time, achieving an AUC value of 0.8025 at 10

years. Additionally, the PHREG output (Table 3) clearly presents coefficients' values, their hazard ratios, and confidence intervals. Moreover, while penalised methods focus solely on indexes, classical survival analysis also emphasises the significance of Legal Form and Geographical Area, in line with prior research on similar data in Europe [38] and Italy [39].

Table 4. Description of the indexes chosen as predictors of bankruptcy events through the four analysed methods

ID	Formula	ID	Formula
ind001	ln (EBITDA)	ind079	Quick Assets/Sales
ind007	Cash flow/Current Liabilities	ind081	Return On Assets
ind019	Ln(Sales)	ind082	EBIT/Total assets
ind029	Current Liabilities/Total Liabilities	ind083	Return On Equity
ind041	EBITDA/Total Assets	ind084	Net Income /Shareholder's Equity
ind042	Total sharholder's Equity/Total Assets	ind092	Sales/Total Assets
ind055	Long Term Debts/Sales	ind094	Shareolders Funds/Capital
ind058	Long Term Liabilities /Total Assets	ind116	EBIT/Interest Paid
ind060	(Long Term Debt + Loans)/Total Assets	ind117	Long Term Debts/Equity
ind068	Net Income/Total Debts	ind124	Receivables/Current Assets
ind073	Other shareholders fund/Total assets	ind132	Equity/Sales

For a clearer understanding of the results, Table 4 provides a description of the indexes selected by the four methods analysed. The Results section shows agreement on the significance of ind042, which indicates the firm's capacity to manage its debt. However, the coefficient's value, representing its relative importance, differs across models: it is higher in the classical survival model (-5.071) and in Ridge penalisation with $\alpha = 0.5$ (< -5.5) and less than half in the alternative methods. Another index consistently present among the penalised models is ind041, which assumes a value of approximately 3 in both Lasso and Ridge for $\alpha = 0.5$ while its importance decreases to 1.5 in Elastic Net and to less than 1 in Ridge when $\alpha = 8$. In the classical survival model, the information provided by ind041 is echoed by indexes ind082 and ind116, albeit with distinct nuances.

In conclusion, penalised methods appear to be more concise and provide an immediate insight into the most significant variables. The classical model offers more detail and may be easier to interpret, yet it has in this case a marginally lower concordance index. Given that there are no significant differences in predictive ability between the models, the selection of one method over another should not be predetermined but should always be based on a comparative analysis.

Appendix

Table 5. Indexes evaluated as potential predictors of the bankruptcy event

Variable	Mean	Coeff. Var	Std Dev	Min	Max
Age	14.11	88.24	12.45	0.01	123.07
ind001	3.98	94.05	3.74	−11.32	13.53
ind004	18.05	354.38	63.97	0.08	1138.55
ind006	813.21	1629.90	13254.49	−15697.00	820789.00
ind007	0.11	338.08	0.37	−11.05	8.08
ind009	0.20	617.46	1.25	−9.01	16.24
ind010	0.03	382.14	0.11	−1.06	0.85
ind011	0.20	615.29	1.26	−9.01	16.24
ind012	0.03	214.60	0.07	−0.76	0.63
ind013	0.07	247.45	0.17	−2.43	2.74
ind019	7.91	18.05	1.43	0.65	16.16
ind020	8.08	16.83	1.36	−0.02	15.28
ind021	78.44	118.99	93.33	0.00	2534.44
ind022	13.99	1257.93	176.00	−170.20	12440.85
ind023	0.94	105.40	0.99	0.08	43.20
ind024	0.71	29.47	0.21	0.04	1.01
ind025	0.89	65.98	0.59	0.01	17.76
ind027	13.04	257.69	33.60	−16.96	1115.02
ind028	0.57	39.55	0.23	0.01	1.20
ind029	0.73	24.92	0.18	0.02	1.00
ind030	5.74	340.46	19.55	−25.38	891.18
ind031	1.74	213.32	3.71	0.06	98.83
ind032	114.35	118.74	135.79	−0.03	3849.12
ind033	0.32	118.74	0.38	0.00	10.69
ind034	605.57	1666.80	10093.66	−108113.60	582932.00
ind036	0.38	677.09	2.60	−79.24	48.97
ind037	0.02	541.31	0.11	−0.89	0.77
ind038	3.16	1447.60	45.70	−885.89	2569.31
ind040	0.06	175.71	0.11	−0.86	0.98
ind041	0.06	130.37	0.08	−0.62	0.55
ind042	0.23	90.34	0.20	−0.52	0.96
ind043	0.49	199.35	0.98	−0.34	23.30
ind044	2.32	370.51	8.60	−7.84	284.47
ind048	0.29	70.68	0.21	−0.01	0.96
ind049	0.31	65.54	0.21	0.00	0.98
ind050	0.32	152.15	0.49	0.00	12.36
ind051	0.22	76.26	0.17	0.00	0.94
ind052	−23.81	−7835.09	1865.48	−135728.00	3989.01
ind053	0.13	111.69	0.15	0.00	0.90

Table 6. Indexes evaluated as potential predictors of the bankruptcy event

Variable	Mean	Coeff. Var	Std Dev	Min	Max
ind055	0.11	255.97	0.28	−0.02	7.69
ind056	7.37	1890.81	139.33	−1.51	9708.55
ind057	0.06	147.73	0.10	−0.03	0.72
ind058	0.20	72.85	0.15	0.00	0.98
ind059	0.24	92.42	0.23	0.00	0.99
ind060	0.20	96.91	0.19	0.00	0.90
ind062	0.07	422.38	0.30	−4.32	4.13
ind063	1.32	5115.09	67.58	−625.97	4833.35
ind064	0.00	−3091.34	0.15	−3.43	1.14
ind065	0.04	5563.24	2.14	−83.31	21.87
ind066	−0.25	−435.83	1.11	−9.67	7.03
ind068	0.02	743.22	0.16	−3.68	2.32
ind070	0.33	193.32	0.65	−0.39	26.10
ind071	0.55	252.35	1.38	0.00	61.30
ind072	0.32	161.58	0.52	0.00	11.80
ind073	0.17	117.88	0.20	−1.09	0.94
ind074	1.55	94.01	1.46	0.66	44.79
ind075	449.36	2125.39	9550.61	−68634.00	588170.00
ind076	0.10	367.33	0.37	−6.76	3.60
ind077	6335.94	931.93	59046.78	0.48	3116481.00
ind078	1.21	212.25	2.57	0.01	71.63
ind079	0.62	114.82	0.71	0.02	35.52
ind080	0.48	42.82	0.21	0.01	0.99
ind081	0.00	2102.16	0.07	−0.77	0.54
ind082	0.03	248.92	0.08	−0.63	0.52
ind083	−0.26	−433.47	1.11	−9.67	7.03
ind084	0.29	551.23	1.58	−9.33	25.60
ind085	−0.01	−921.59	0.11	−1.19	0.83
ind086	14888.96	1224.12	182258.40	1.93	10396912.00
ind087	2.61	73576.65	1920.59	−126093.49	36075.41
ind088	1.46	57.26	0.84	0.02	13.33
ind089	−114.15	−6547.40	7473.91	−518257.00	3572.02
ind090	98735.21	608.35	600660.45	−65395.79	27873248.41
ind091	13.05	377.57	49.29	−96.09	1385.69
ind092	0.98	53.70	0.53	0.02	6.24
ind093	−1.11	−25450.52	282.00	−17549.90	4834.21
ind094	20.05	631.81	126.65	−14.62	7432.27
ind097	17.62	220.42	38.83	−23.53	1128.00
ind098	1.46	102.97	1.51	0.16	43.32
ind100	1.13	108.41	1.22	0.01	36.87

Table 7. Indexes evaluated as potential predictors of the bankruptcy event

Variable	Mean	Coeff. Var	Std Dev	Min	Max
ind101	16.62	233.37	38.78	−24.53	1127.00
ind102	0.77	27.07	0.21	0.02	1.52
ind104	−2537.90	−732.76	18596.80	−542334.00	387454.07
ind105	2328.83	1591.49	37062.97	−153109.15	2368921.00
ind107	0.74	502.24	3.71	−0.94	97.83
ind108	0.14	439.03	0.60	−7.55	11.52
ind109	0.13	187.05	0.25	−0.83	0.93
ind113	15.91	223.10	35.49	−21.37	722.96
ind114	−0.15	−879.21	1.30	−9.96	9.43
ind115	1076.49	1293.82	13927.92	−82402.99	752554.00
ind116	21.38	384.14	82.15	−207.69	994.97
ind117	1.01	408.01	4.11	−0.52	131.60
ind118	0.55	265.23	1.46	−0.34	43.79
ind119	0.23	90.97	0.21	−0.52	0.98
ind121	1.08	103.46	1.12	0.04	33.59
ind122	0.94	105.40	0.99	0.08	43.20
ind123	13.03	257.37	33.54	−16.96	1115.02
ind124	0.36	71.83	0.26	0.00	0.98
ind132	0.33	194.94	0.64	−0.39	26.10
ind136	2.75	44.67	1.23	0.00	9.04

References

1. Pölsterl, S.: Scikit-survival: a library for time-to-event analysis built on top of scikit-learn. J. Mach. Learn. Res. **21**(212), 1–6 (2020). http://jmlr.org/papers/v21/20-729.html
2. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
3. Cox, D.R.: Regression models and life-tables. J. Roy. Stat. Soc.: Ser. B (Methodol.) **34**(2), 187–202 (1972)
4. Kalbfleisch, J.D., Schaubel, D.E.: Fifty years of the cox model. Annu. Rev. Stat. Appl. **10**(1), 1–23 (2023)
5. Andersen, P.K.: Fifty years with the Cox proportional hazards model: history, influence, and future. J. Roy. Stat. Soc. Ser. A Stat. Soc. qnad114 (2023). https://doi.org/10.1093/jrsssa/qnad114. eprint: https://academic.oup.com/jrsssa/advance-article-pdf/doi/10.1093/jrsssa/qnad114/51361152/qnad114.pdf. https://doi.org/10.1093/jrsssa/qnad114. ISSN 0964-1998
6. Perri, D., Simonetti, M., Lombardi, A., Faginas-Lago, N., Gervasi, O.: Binary classification of proteins by a machine learning approach. In: Gervasi, O., et al. (eds.) ICCSA 2020. LNCS, vol. 12255, pp. 549–558. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58820-5_41

7. Cruz, J.A., Wishart, D.S.: Applications of machine learning in cancer prediction and prognosis. Cancer Inform. **2**, 117693510600200030 (2006)
8. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. **13**, 8–17 (2015)
9. Wan, G., et al.: Prediction of early-stage melanoma recurrence using clinical and histopathologic features. NPJ Precis. Oncol. **6**(1), 79 (2022)
10. Spooner, A., et al.: A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci. Rep. **10**(1), 20410 (2020)
11. Fleming, T.R., Lin, D.Y.: Survival analysis in clinical trials: past developments and future directions. Biometrics **56**(4), 971–983 (2000)
12. Jänne, P.A., et al.: Twenty-five years of clinical research for patients with limited-stage small cell lung carcinoma in North America: meaningful improvements in survival. Cancer **95**(7), 1528–1538 (2002)
13. Hosmer, D.W., Jr., Lemeshow, S., May, S.: Applied Survival Analysis: Regression Modeling of Time-to-Event Data. Wiley, Hoboken (2008)
14. Collett, D.: Modelling Survival Data in Medical Research. Chapman and Hall/CRC, Boca Raton (2023)
15. Huang, X., Aslett, L.J.M., Coolen, F.P.A.: Reliability analysis of general phased mission systems with a new survival signature. Reliab. Eng. Syst. Saf. **189**, 416–422 (2019)
16. Karim, M.R., Islam, M.A., et al.: Reliability and Survival Analysis. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-9776-9
17. 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, Oxford (1992)
18. Shumway, T.: Forecasting bankruptcy more accurately: a simple hazard model. J. Bus. **74**(1), 101–124 (2001)
19. Bellotti, T., Crook, J.: Credit scoring with macroeconomic variables using survival analysis. J. Oper. Res. Soc. **60**(12), 1699–1707 (2009)
20. Lee, M.-C.: Business bankruptcy prediction based on survival analysis approach. AIRCC's Int. J. Comput. Sci. Inf. Technol. **6**, 103–119 (2014)
21. LoPucki, L.M., Doherty, J.W.: Bankruptcy survival. UCLA L. Rev. **62**, 969 (2015)
22. 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)
23. Zelenkov, Y.: Bankruptcy prediction using survival analysis technique. In: 2020 IEEE 22nd Conference on Business Informatics (CBI), vol. 2, pp. 141–149. IEEE (2020)
24. Tibshirani, R.: The lasso method for variable selection in the Cox model. Stat. Med. **16**(4), 385–395 (1997)
25. Hoerl, R.W.: Ridge regression: a historical context. Technometrics **62**(4), 420–425 (2020)
26. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. Ser. B Stat. Methodol. **67**(2), 301–320 (2005)
27. Orbis. Orbis. Bureau van Dijk. https://orbis.bvdinfo.com/. Accessed June 2020
28. Pierri, F.: Variable selection in binary logistic regression for modelling bankruptcy risk. In: Kitsos, C.P., Oliveira, T.A., Pierri, F., Restaino, M. (eds.) ICRA 2022, pp. 143–155. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-39864-3_12

29. Beaver, W.H.: Financial ratios as predictors of failure. J. Account. Res. **4**, 71–111 (1966)
30. Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Finance **23**(4), 589–609 (1968)
31. Mossman, C.E., Bell, G.G., Swartz, L.M., Turtle, H.: An empirical comparison of bankruptcy models. Financ. Rev. **33**(2), 35–54 (1998)
32. 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). Subspace Learning/Selected papers from the European Symposium on Time Series Prediction. ISSN 0925-2312. https://doi.org/10.1016/j.neucom.2009.11.034. https://www.sciencedirect.com/science/article/pii/S0925231210001098
33. SAS. SAS/STAT® 9.22 User's Guide. https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#phregtoc.htm. Accessed 15 Mar 2024
34. Belounnas, A., Brissaud, F., Jardon, G.: Analysis of reliability factors of gas transport industrial assets using supervised machine learning (2022)
35. Ashfaq, A., et al.: Survival analysis for pediatric heart transplant patients using a novel machine learning algorithm: a UNOS analysis. J. Heart Lung Transplant. **42**(10), 1341–1348 (2023)
36. Peng, Z.-H., Huang, Z.-X., Tian, J.-H., Chong, T., Li, Z.-L.: The application of time-to-event analysis in machine learning prognostic models. J. Transl. Med. **22**(1), 146 (2024)
37. Cherukuri, L.S., Rajput, R.S., Neema, S.: Unveiling the impact of demographic factors on disease survival: a multifaceted examination across diverse medical conditions. Int. J. Health Sci. **7**(2), 1–14 (2024)
38. Caroni, C., Pierri, F., Restaino, M.: A comparison of variable selection methods in competing risks models for business failures. Statistica Applicazioni (2023). https://doi.org/10.26350/999999_000053
39. 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)