
Fifty years since Altman (1968): Performance of financial distress prediction models

Surbhi Bhatia¹ and Manish K. Singh^{2,3}

¹Independent Researcher

²Department of Humanities and Social Sciences
Indian Institute of Technology Roorkee, Roorkee, India

³XKDR Forum, Mumbai, India

How to cite this paper? Bhatia, S. and Singh, M. K. (2022). Fifty years since Altman (1968): Performance of financial distress prediction models, Working Paper 12, XKDR Forum, Mumbai, India.

Available from: <https://xkdr.org/research.html>

Fifty years since Altman (1968): Performance of financial distress prediction models

Surbhi Bhatia^a, Manish K. Singh^{b,c}

^a*Independent Researcher*

^b*Department of Humanities and Social Sciences
Indian Institute of Technology Roorkee, Roorkee, India*

^c*xKDR Forum, Mumbai, India*

Abstract

Using bankruptcy filings under the new Insolvency and Bankruptcy Code (2016), we investigate the effect of firm characteristics and balance sheet variables on the forecast of one-year-ahead default for Indian manufacturing firms. We compare traditional discriminant analysis and logistic regression models with state-of-the-art variable selection technique-the least absolute shrinkage and selection operator, and the unsupervised techniques of variable selection-to identify key predictive variables. Our findings suggest that the ratios considered as important by Altman (1968) still hold relevance for the prediction of default, no matter the technique applied for variables selection. We find cash to current liability (a liquidity measure) as an additional robust and significant predictor of default. In terms of predictive accuracy, the reduced-form multivariate discriminant analysis model used in Altman (1968) performs at par with the more advanced econometric specification for both in-sample and full-sample default prediction.

Keywords: firm distress, default prediction, Altman z-score, discriminant analysis, logistic regression, LASSO, elastic net

JEL: C53, G17, G32, G33

Email addresses: surbhibhatia1906@gmail.com (Surbhi Bhatia), mks@hs.iitr.ac.in (Manish K. Singh)

Contents

1	Introduction	4
2	Bankruptcy prediction using accounting data	4
3	Bankruptcy prediction in the India	5
4	The dataset	6
4.1	Firm characteristics and accounting ratios	7
5	Selection of relevant accounting variables	8
5.1	Analysis of variance	10
5.2	Logistic regression	10
5.3	LASSO variable selection	13
5.4	Principal Component Analysis	14
5.5	Summary	17
6	Relative performance of distress prediction models	18
6.1	Univariate logit models	19
6.2	Multivariate analysis	22
6.3	Performance comparison	27
6.4	Robustness checks	27
7	Conclusion	28
	Appendix A List of variables	31

1. Introduction

For over 50 years, researchers have analysed firm data and tried to tease out models that can predict bankruptcy at the firm level. Innovations have arisen, in statistical models, in statistical modelling techniques and in the economics of bankruptcy.

In India, a fundamental constraint has impeded the development of bankruptcy models, and that is the lack of data. Traditionally, there was no systematic way to construct a dataset of bankrupt firms as there was no formal bankruptcy process. A major event in this field took place in May 2016, when the Parliament enacted the Insolvency and Bankruptcy Code (IBC). This created the first wave of publicly visible bankruptcy filings and has made possible the systematic study of firm characteristics and bankruptcy.

Using this dataset, we systematically undertake the statistical analysis for the purpose of identifying accounting variables that are relevant for bankruptcy prediction, and for selecting models that fare well at utilising these variables for the purpose of bankruptcy prediction. We find that net working capital to total asset, retained profit to total asset, profit to total asset, total sales to total asset, and the cash to current liability are statistically robust. Interestingly, four out of these five ratios were also considered significant by Altman in his seminal paper fifty years back, when analysing data for the US.

When comparing the model performance in terms of their one-year predictive accuracy, the reduced form model selected via the discriminant analysis (also used in [Altman \(1968\)](#)) performs at par with the more advanced econometric specification in in-sample and full-sample default predictions.

The remainder of this paper is organized as follows. In section 2, we provide a brief overview of the existing literature. In section 3, we discuss the Indian institutional setting and how it made it extremely difficult to construct a trusted dataset about firm defaults. In section 4, we discuss our data collection procedure. In section 5, we first discuss the matching technique applied to figure out a sample of non-defaulted firms similar in characteristics to the defaulted firms in our dataset. We then robustly estimate the most relevant firm characteristics and accounting ratios using the matched sample and different econometric methods. In section 6, we present and discuss different default prediction models and their predictive accuracy. Our conclusions are presented in section 7.

2. Bankruptcy prediction using accounting data

At the dawn of quantitative analysis of firm default were univariate studies on the association between corporate bankruptcy and certain accounting ratios ([Bureau of Business Research \(1930\)](#); [Patrick \(1932\)](#); [Smith and Winakor \(1935\)](#)). They compared location estimators of accounting ratios of failed firms with those of successful firms ([Bellovary et al. \(2007\)](#)). [Beaver \(1966\)](#) expanded this work along two dimensions. He tested the ability of financial ratios to classify firms into bankrupt and non-bankrupt firms. And, he suggested that multiple ratios considered simultaneously might have higher predictive ability than ratios considered in isolation.

[Altman \(1968\)](#) implemented this idea using discriminant analysis and developed a five-factor model to predict the default of manufacturing firms. Altman's z-score model predicted

default if the firm's z-score fell below a certain threshold. It had a very high in-sample predictive ability (95% accuracy). However, the predictive ability dropped considerably as balance sheet information became stale. The model accuracy dropped to 72% when he used two-year-old accounting information and dropped further to 48%, 29%, and 36% when he used three, four, and five year-old financial data. When tested on a holdout sample, the out-of-sample predictive ability was 79%.

This has been an extremely influential paper. It presently has about 22,000 citations on google scholar. In the following years, there was a vast escalation in the sophistication of the statistical and economic thinking that was brought to bear on the question.

In the 1980s, [Ohlson \(1979\)](#) brought logistic regressions to bear on default prediction. This was thought to improve upon discriminant analysis by relaxing modelling assumptions and incorporating some kinds of non-linearity.

In recent decades, new tools of statistics and machine learning have been developed for addressing 'the classification problem', i.e. the problem of bankruptcy prediction using accounting data. This includes methods such as artificial neural networks ([Shah and Murtaza \(2000\)](#)), support vector machines ([Horta and Camanho \(2013\)](#)), random forests, decision trees and classification and regression trees (CART). The methods for assessing out-of-sample model performance were improved and permitted large-scale comparisons of rival models [Li et al. \(2010\)](#). Novel ideas like bagging and boosting have been brought to bear on improving the models (see [Barboza et al. \(2017\)](#)).

Alongside these statistical developments, economists wrote models of the foundations of bankruptcy, drawing on option theory (see [Black and Scholes \(1973\)](#); [Merton \(1974\)](#)). These approaches are visible in [Shumway \(2001\)](#); [Bauer and Agarwal \(2014\)](#), where market-based measures such as stock returns, corporate bond yields, and other market-traded debt securities have better forward-looking properties and are better suited for predicting distress. Hazard models with time-varying regressors perform better than single period forecasting models. [Campbell et al. \(2008\)](#) also builds on the theory by including market-based variables to model default and find greater explanatory power in their model (also see [Sreedhar and Shumway \(2008\)](#); [Agarwal and Taffler \(2008\)](#)).

In this paper, we compare the performance of a diverse array of methods for the problem of bankruptcy prediction in India. For the purpose of this paper, we focus on accounting data and not stock market measures, owing to the paucity of data for firms with actively traded shares. Similarly, corporate bonds and CDS are not meaningfully traded in India, which rules out the lines of research which exploit these.

3. Bankruptcy prediction in the India

Thousands of large firms exist in India, and there has been high interest in quantitative models of bankruptcy for decades. In the industry, some applications were visible from the late 1970s and early 1980s, which utilised estimates for the Altman model based on US data.

The key bottleneck faced in *estimating* bankruptcy prediction models in India has been the lack of a bankruptcy dataset. Traditionally, firm default was a secret. Firms were comfortable defaulting to some lenders while paying others. This made it extremely difficult

to construct a trusted dataset about firm default (for a detailed discussion, refer to [Sengupta et al. \(2016\)](#)).

In the existing literature on bankruptcy prediction in India, researchers have engaged in various kinds of plausible proxies (for example, see [Kulkarni et al. \(2005\)](#)). Some authors classified firms as defaulted if the interest coverage ratio went below one. Similarly, some authors consider firms downgraded to default by a rating agency as defaulted firms. [Sharma et al. \(2019\)](#) considers a list of defaulters as per the list provided by the Reserve Bank of India.

The major milestone in this journey was the Insolvency and Bankruptcy Code (“IBC”), which was enacted in May 2016. For the first time, it constructed the legal machinery through which any one aggrieved lender could initiate a bankruptcy process, that was visible in the public eye, after which a Committee of Creditors would seize control of the firm away from its shareholders.

Through this, systematic observation of firms which went into bankruptcy, defined as a formal legal step, commenced. This makes it possible to estimate models of bankruptcy prediction based on firm data. An early literature has examined this data [Foo and Pathak \(2016\)](#); [Sharma et al. \(2019\)](#). In this paper, we undertake the comprehensive statistical research based on the new bankruptcy data to create new knowledge about bankruptcy prediction in India.

4. The dataset

The bankruptcy dataset here was hand-collected based on the information on these orders available on the National Company Law Tribunal (NCLT) and the National Company Law Appellate Tribunal (NCLAT) website, which are guaranteed to have a comprehensive list of the firms which went into bankruptcy. It covers the first wave of firms which were brought into the bankruptcy process, with filings from December 2016 to November 2017. The dataset has been released into the public domain as the Finance Research Group (FRG) Insolvency Cases dataset and is available at the http://ifrogs.org/dms/IBC/nclt_data.html¹

For each defaulting firm, the dataset includes the case identification number, bench at which the case is filed, name and type of filing party, date of filing, date of disposal, whether the case was admitted or dismissed. If dismissed, it captures the reason for dismissal.

A first examination of the cases, and their status as of June 30, 2018, suggests that 6678 cases have been filed. In 3379 cases (roughly half), a decision has been pending regarding whether to admit the case or dismiss it based on the validity of the information provided. Four hundred seventy-seven cases have been withdrawn. We do not know if they were

¹[Chatterjee et al. \(2017\)](#), introduces this dataset and answers questions on the operationalization of the IBC in its first six months. They track the uptake of the law by different stakeholders and the functioning of the NCLT in this period. Besides, they observe the number of cases filed and disposed of and the average duration of time between case filing and disposal. For this reason, they highlight the gaps between the expected outcomes of the law and its implementation, as seen through the empirical evidence.

settled out of court, or withdrawn for other reasons in these cases. Roughly 2,225 cases were dismissed (not admitted for various reasons).

Records were deleted by us when the case was dismissed, withdrawn, pending. Records with missing or duplicate data were deleted. Records for firms where the bankruptcy process had begun under older institutional arrangements (BIFR, CDR, DRT) were deleted. This left a dataset of 335 firms. This is an Indian defaults dataset, a list of 335 firms where we are certain that a default took place.

As with all bankruptcy datasets, lenders try to negotiate settlements out of court. In many cases, firm failure leads to a negotiated settlement, and does not generate records in the bankruptcy process. This limitation is present in our dataset also. We know for certain that there was a default for these 335 firms; we do not know about firms that defaulted outside of this set but these certainly exist. This problem is, of course, not unique to India. As an example, consider the US. Based on the data from the Federal Judicial Center’s (FJC) Bankruptcy Petition Database and the Census Bureau’s Business Dynamics Statistics reports from 2000-2016, [Greenwood et al. \(2020\)](#) show that 91.7% of all distressed firms ‘renegotiate and continue to operate’ or ‘stop operating without filing for bankruptcy’ in the United States. Only in 8.3% of the cases do distressed firms show up in bankruptcy court.

The standard database for obtaining a broad range of firm characteristics is the CMIE Prowess database. When we join the defaults dataset into the CMIE database, we find 174 matches of which 41 are manufacturing firms. Alongside this, there is data for 4,312 non-defaulted manufacturing firms.

4.1. Firm characteristics and accounting ratios

We draw upon the extensive bankruptcy prediction literature to create an exhaustive list of accounting and identity-based measures for our analysis. This includes the five ratios which were considered in the Altman z-score model, which are Working capital/Total assets (R1 for liquidity), Retained earnings/Total assets (R2 for profitability), Earnings before Interest and Tax (EBIT)/Total assets (R3 for efficiency), Market value of equity/Book value of total liabilities (R4 for leverage) and Sales/Total Assets (R5 for asset turnover).

The precise measure of working capital as is used in [Altman \(1968\)](#) does not port into Indian accounting data. Hence, we employ three alternative definitions of working capital from CMIE:

1. R1A: Gross working capital (based on the cost of sales method)/Total assets
2. R1B: Net working capital (based on the cost of sales method)²/Total assets, and
3. R1C: Net working capital³/Total assets.

Similarly, we define the leverage ratio for R4 as the ratio of non-current liabilities to the total assets.

²After deducting current liabilities from current assets calculated using the cost of sales method, we get the amount of networking capital as per cost of sales method.

³Defined as current assets over current liabilities.

We include 15 additional financial ratios in our analysis as a proxy for a firm’s asset turnover, profitability, liquidity, and tangibility to account for asset utilization, return on investments, and short-term liquidity.

Based on conversations with market participants and our judgment, we consider the interest cover ratio, current ratio, quick ratio, and raw material turnover ratio. We have reported a complete list of these ratios in Table 1.

Next, we create controls for entity type (government-owned, private, or foreign), industry sector, ownership type (affiliation to business houses), and public ownership (listed on a public exchange) using a vector of dummy variables. For example, if the firm is registered as a public limited company, it is assigned as 1. If it is registered as a private limited company, it is assigned as 2. Similarly, if the firm is associated with a business house, it is assigned 1; otherwise, 0. If the firm is listed either on the National Stock Exchange or the Bombay Stock Exchange, it is assigned 1; otherwise, 0. For industry categories, we include dummy variables for food and agro-based products, textiles, chemicals, and chemical products, consumer goods, construction materials, metal and metal products, transport equipment, and miscellaneous and diversified.

[Appendix A](#) documents the complete list of identity and accounting-based variables used in our analysis.

Once we have selected the variables, our empirical estimation consists of two parts. First, in Section 5, we attempt to identify the variables that are consistently significant, irrespective of the choice of methodology for estimation. We use five approaches: analysis of variance, logistic regression, LASSO, elastic net approach, and PCA approach. Then in Section 6, we use the significant variables in four different distress prediction models and compare their relative performance to select the best fit model to use to predict financial distress.

5. Selection of relevant accounting variables

In this section, we identify the variables that are significant predictors of default for Indian firms. However, given the limited number of defaulted firms arising from the challenges associated with the identification of defaulted companies and the limited availability of financial data, we first try to balance our sample by conducting a one-to-one nearest neighbor matching. For every defaulted firm i in our sample, we select a non-defaulted firm j based on attributes of size (proxied here by Total Sales),⁴ industry type, affiliation, entity type, and listing. The linear propensity score matches closely resembling firms on these characteristics. The propensity score is derived using predicted values from a logit regression.⁵

⁴We also match defaulted firms with non-defaulted firms using two different proxies for the size of the firms: (1) By considering Total Assets; and (2) By considering Total Sales and Total Assets together. The quality of the match is significantly better when we use Total Sales compared to these other proxies. The results are not shown here to save space but are available upon request.

⁵Let T_i be a default indicator equal to 1 when a firm default and 0 otherwise and $e(X_i)$ be the propensity score. Then by definition,

Table 1: Accounting ratios

Description	Measure of	Unit
Altman z-score indicators		
R1A = Gross working capital/Total assets (Cost of sales method)	Liquidity	
R1B = Net working capital/Total assets (Cost of sales method)	Liquidity	R1B
R1C = Net working capital/Total assets	Liquidity	R1C
R2 = Retained profits/Total assets	Profitability	R2
R3 = (PBDITA - Depreciation and amortisation of fixed assets)/Total assets	Efficiency	R3
R4 = Non current liabilities/Total assets	Leverage	Long term debt ratio
R5 = Total sales/Total assets	Asset turnover	R5
Additional financial ratios		
Raw material turnover	Asset turnover ratio	sa_rawmat_stores_spare_avg_stk_rawmat_stores
Total income/Compensation to employees	Asset utilisation ratio	sa_total_inc_net_of_pe_compensation_to_empl
Total sales/Net fixed asset	Asset utilisation ratio	sa_sales_net_repairs_avg_nfa_net_of_reval
Profit after tax/Total income	Net profit margin	sa_pat_net_of_pe_pc_tot_inc_net_of_pe
PBDITA ^a /Total sales	Operating profit	sa_pbdita_net_of_peoifi_pc_sales_chg_in_stk
PBDITA/Total income ^b	Profitability	sa_pbdita_net_of_peoifi_pc_tot_inc_net_of_pe
PBDITA/Net Fixed Assets	Return over investments	sa_pbdita_net_of_peoifi_pc_avg_nfa_net_of_reval
Profit after tax/Total assets	Returns on total assets	sa_pat_net_of_pe_pc_avg_tot_asset_net_menow_reval
Quick ratio	Short-term liquidity	sa_quick_ratio
Cash profit/Total income	Short-term liquidity	sa_cash_profit_net_of_pe_pc_tot_inc_net_of_pe
Current ratio	Short-term liquidity	sa_cash_profit_ratio
Interest coverage ratio	Short-term liquidity	sa_interest_cover
Cash to current liabilities	Short-term liquidity	sa_cash_n_bank_no_fd_security_current_liab_n_prov
Total outside liabilities/Total net worth	Tangibility	sa_tot_tnw
Total term liabilities/Tangible net worth	Tangibility	sa_tot_tnw

Note: Ratios such as interest coverage ratio and other tangibility ratios were also considered but had to be dropped on account of the non-availability of accounting data.

^aProfit before depreciation, interest, tax and amortization

^bIt excludes prior period and extra-ordinary transactions.

Table 2 shows the quality of the resulting matched sample. As measured by the standardized mean differences, the initial imbalance is quite far from zero (column 4), suggesting the need for matching. When we do the one-to-one nearest neighbor matching, the standardized mean differences (last column) drop to near zero suggesting an adequate balance. We also compare match quality using variance ratios and empirical cumulative density function statistics. Note that the following analyses are based on our matched data.

5.1. Analysis of variance

To identify the financial ratios where defaulted and non-defaulted firms differ, we first perform a univariate one-way analysis of variance test (ANOVA) on our matched data. ANOVA compares the means from the two groups using the F-distribution. The null hypothesis is that the two means are equal. Therefore, a significant result means that the two means are unequal. Table 3 shows the accounting variables that are statistically significantly different between defaulted and non-defaulted firms.

Our findings suggest that all the ratios used in Altman (1968) are statistically different for the sample of defaulted firms who filed for bankruptcy and the correspondingly matched non-defaulted firms in our dataset. Additionally, two other ratios related to the cash holding of the firms (cash to current liability and cash profit to total income) are significantly different. It suggests that cash holdings might be an important differentiator for firms in India. Apart from this, all profitability measures play an important role. We also found that long-term liabilities and turnover ratios (Total sales to net fixed assets) differed across defaulted and non-defaulted firms.

5.2. Logistic regression

Next, we assess the goodness of fit for each financial ratio individually, that have been considered important in default literature using logistic regression. We consider all indicators that are mentioned in Table 1. We test them individually to check whether standalone they hold information about a firm going bankrupt or not. We use the binary logistic regression while controlling for firm-specific dummies, i.e., business group, entity type, listing, and industry type, using :

$$e(X_i) \equiv Pr(T_i = 1|X_i) = E(T_i|X_i) \quad (1)$$

Each firm from the first group is matched with one in the second based on the minimum distance in their scores. We match not on the predicted probabilities (bounded between zero and one) but on the linear predictor $\hat{\mu} = X\hat{\beta}$. Matching this linear predictor avoids compression of propensity scores near zero and one. This linear predictor is also more nearly normally distributed, giving equal percent bias reduction results if the propensity score is matched with other covariates.

Since the propensity score is unknown, we estimate it using the following:

$$Pr(T_1, T_2, T_3, \dots, T_N | X_1, X_2, X_3, \dots, X_N) = \sum_{i=1}^N e(X_i)^{T_i} (1 - e(X_i))^{(1-T_i)}$$

Table 2: Assessing the quality of match

Note: We use the one-to-one nearest neighbor matching on the propensity score. The *distance* variable shows the estimated propensity score, a one-dimensional summary of all the included covariates, computed as the predicted probability of being the treated group given the covariates implemented using the logistic regression. D(.) denotes the dummy used for different categorical variables.

	Initial imbalance			After matching	
	Means		Std. Mean	Means	Std. Mean
	(Treated)	(Control)	Diff.	(Control)	Diff.
Distance	0.02	0.01	0.97	0.02	0.00
Total Sales	14718.76	10804.31	0.09	9682.94	0.11
Entity type controls					
D(Government_owned)	0.93	0.59	1.29	0.93	0.00
D(Private)	0.07	0.41	-1.29	0.07	0.00
Controlling for business group ownership					
D(ownership)	0.32	0.25	0.15	0.32	0.00
Controlling for public listing					
D(Publicly listed)	0.71	0.41	0.66	0.68	0.05
Controlling for different industries					
D(Chemicals and chemical products)	0.12	0.21	-0.26	0.12	0.00
D(Construction materials)	0.07	0.04	0.12	0.07	0.00
D(Consumer goods)	0.02	0.06	-0.22	0.02	0.00
D(Diversified)	0.07	0.03	0.15	0.07	0.00
D(Food and agro based products)	0.05	0.12	-0.35	0.05	0.00
D(Machinery)	0.17	0.15	0.06	0.15	0.06
D(Metals and metal products)	0.29	0.14	0.35	0.32	-0.05
D(Miscellaneous)	0.00	0.04	-0.21	0.00	0.00
D(Textiles)	0.10	0.11	-0.04	0.10	0.00
D(Transport equipment)	0.10	0.10	-0.02	0.10	0.00

$$P(Y_i = 1|X_i) = \frac{e^{\beta_0 + \beta_1' X_i}}{1 + e^{\beta_0 + \beta_1' X_i}} \quad (2)$$

where X_i is one of the financial ratios mentioned in Table 1, and β_0 and β_1 are scalar parameters. The dependent variable Y_i is the indicator of default, which equals 1 for the defaulted firm, and is 0 otherwise. Using maximum likelihood, we estimate the coefficients β_0 and β_1 on the matched data. By plugging back these estimates into the model for $P(X_i)$ yields a number close to 1 for defaulted firms and a number close to 0 for non-defaulted firms. The significance of β_1 will tell us the list of financial ratios and firm characteristics that are relevant for default prediction.

Table 4 lists all the financial ratios which are individually significant in predicting default. Note that each line represents the result of a separate regression done using the respective

Table 3: Univariate analysis of difference in means

Financial ratio	p-value
R1C = Net working capital / Total assets	0.0031***
R2 = Retained profits / Total assets	0.0007***
R3 = (PBDITA - Depreciation and amortisation of fixed assets) / Total assets	0.0007***
R4 = Non current liabilities / Total assets	0.0114**
R5 = Total sales / Total assets	0.0001***
Cash to current liabilities	0.0182**
Cash profit / Total income	0.0332**
Profit after tax / Total assets	0.0004***
Profit after tax / Total income	0.0560*
PBDITA / Net Fixed Assets	0.0510*
PBDITA / Total income (excludes prior period and extra-ordinary transactions)	0.0330**
Total sales / Net fixed asset	0.0125**

financial ratio while controlling for a business group, entity type, listing, and industry. Also, note that we only document financial ratios which are individually significant. We document β_1 , the standard error, and the p-value for each regression.

Table 4: Significant predictors of default using logistic regression

Variable	Estimate	Std. Error	Pr(> z)
R1C = Net working capital / Total assets	-0.0420***	0.0120	0.0004
R2 = Retained profits / Total assets	-0.0915***	0.0299	0.0022
R3 = (PBDITA - Depreciation and amortisation of fixed assets) / Total assets	-0.0973***	0.0299	0.0011
R4 = Non current liabilities / Total assets	0.0321**	0.0132	0.0149
R5 = Total sales / Total assets	-0.0199***	0.0053	0.0002
Cash to current liabilities	-7.7820**	3.0880	0.0117
Cash profit / Total income	-0.0479***	0.0164	0.0035
Current ratio	-2.7475***	0.6722	0.0000
Quick ratio (Acid test ratio)	-0.9135**	0.4622	0.0481
Profit after tax / Total assets	-0.1071***	0.0318	0.0008
Profit after tax / Total income	-0.0340***	0.0126	0.0069
PBDITA / Net Fixed Assets	-0.0073*	0.0039	0.0609
PBDITA / Total income (excludes prior period and extra-ordinary transactions)	-0.0339**	0.0145	0.0189
Total sales / Net fixed asset	-0.1772***	0.0686	0.0098
Total term liabilities / Tangible net worth	0.1072*	0.0633	0.0904

The results show that all five ratios used in Altman (1968) are individually significant at 1% level. These are R1C, R2, R3, R4, and R5. Additional ratios which we found significant using ANOVA are also significant using logit regression. We also found quick ratio and current ratio-both measures of short-term liquidity to be statistically significant and negatively associated with default. It supports our previous findings that liquidity ratios are pretty different between bankrupt and non-bankrupt firms in the Indian setting.

Additional ratio measuring the long-term liability to tangible net worth becomes significant.

5.3. LASSO variable selection

In this subsection, we use the least shrinkage and selection operator (LASSO)-a supervised machine learning process-also called penalized regression, to help select the most relevant subset of variables for default prediction (for detailed methodology, refer to [Tibshirani \(1996\)](#)). This is a multivariate approach where the methodology involves putting a constraint on the sum of absolute values of the model parameters. By imposing a penalty (upper bound), the method picks up variables whose coefficients shrink below the pre-defined threshold. It enables us to select a parsimonious set of variables.

Following [Shaonan et al. \(2015\)](#), for the discrete hazard model in Eq.(3), we obtain LASSO parameter estimates by minimizing the negative log-likelihood function with a roughness penalty placed on the sum of the absolute value of the covariate parameters - the so-called ' l_1 penalty' or ' l_1 constraints'

$$\sum_{i=1}^n (-Y_{i,t+1} (\beta_0 + \beta_1' X_{i,t}) + \log(1 + \exp(\beta_0 + \beta_1' X_{i,t}))) \quad (3)$$

subject to,

$$\sum_{k=1}^p |\beta_k| \leq s$$

or equivalently,

$$\sum_{i=1}^n (-Y_{i,t+1} (\beta_0 + \beta_1' X_{i,t}) + \log(1 + \exp(\beta_0 + \beta_1' X_{i,t}))) - \lambda \sum_{k=1}^p |\beta_k|$$

where n is the number of firms and p is the number of predictive variables used in the hazard model. We control the amount of shrinkage through the roughness penalty tuning parameter s or λ . Note that a smaller value of s usually leads to a more parsimonious set of the selected predictive variables.

When the penalty parameter λ is 0, s is infinite, and no parameters are eliminated. This is akin to a simple OLS regression called the Ridge regression. As λ increases, more and more coefficients are shrunk to 0. This technique is called elastic net. When λ is very large, s is 0, and all coefficients forcibly become 0, this is the case of a LASSO regression. Results from all three altering values of s and λ are reported in Table 5. We have shown a graphical depiction of the magnitude and relevance of these coefficients under the three regimes in Figure 1.

We observe that eight variables are important under the most strict regime of $s = 0$, i.e., the LASSO. Three of these-R1C (Networking capital/total assets), R4 (Non-current liabilities/total assets), and R5 (Total sales/total assets)-are a subset of the set of original ratios used by Altman to predict default. Three additional measures - Cash to current liabilities, cash profits to total income, and interest coverage ratio, all related to the firms' cash holdings become a significant predictor of default. Instead of R2 and R3, another

Table 5: Variable selection using LASSO/Elastic Net

Variable	LASSO	Elastic Net	Ridge
R1A = Gross working capital / Total assets (Cost of sales method)			-0.0002
R1B = Net working capital / Total assets (Cost of sales method)			-0.001
R1C = Net working capital / Total assets	-0.005	-0.001	-0.003
R2 = Retained profits / Total assets		-0.003	-0.007
R3 = (PBDITA - Depreciation and amortisation of fixed assets) / Total assets		-0.002	-0.009
R4 = Non current liabilities / Total assets	0.007		0.006
R5 = Total sales / Total assets	-0.006	-0.003	-0.002
Cash to current liabilities	-0.596	-0.133	-0.390
Cash profit / Total income	-0.00002		-0.0002
Current ratio			-0.027
Interest cover ratio	0.001		0.0004
Quick ratio (Acid test ratio)			0.006
Profit after tax / Total assets	-0.021	-0.009	-0.009
Profit after tax / Total income			-0.0001
PBDITA / Net Fixed Assets			-0.0002
PBDITA / Total sales			0.00002
PBDITA / Total income (excludes prior period and extra-ordinary transactions)			-0.0004
Asset turnover ratio			-0.0003
Total sales / Net fixed asset			-0.016
Total outside liabilities / Total net worth			0.0004
Total income / Total salary			0.002
Total term liabilities / Tangible net worth	0.001		0.001

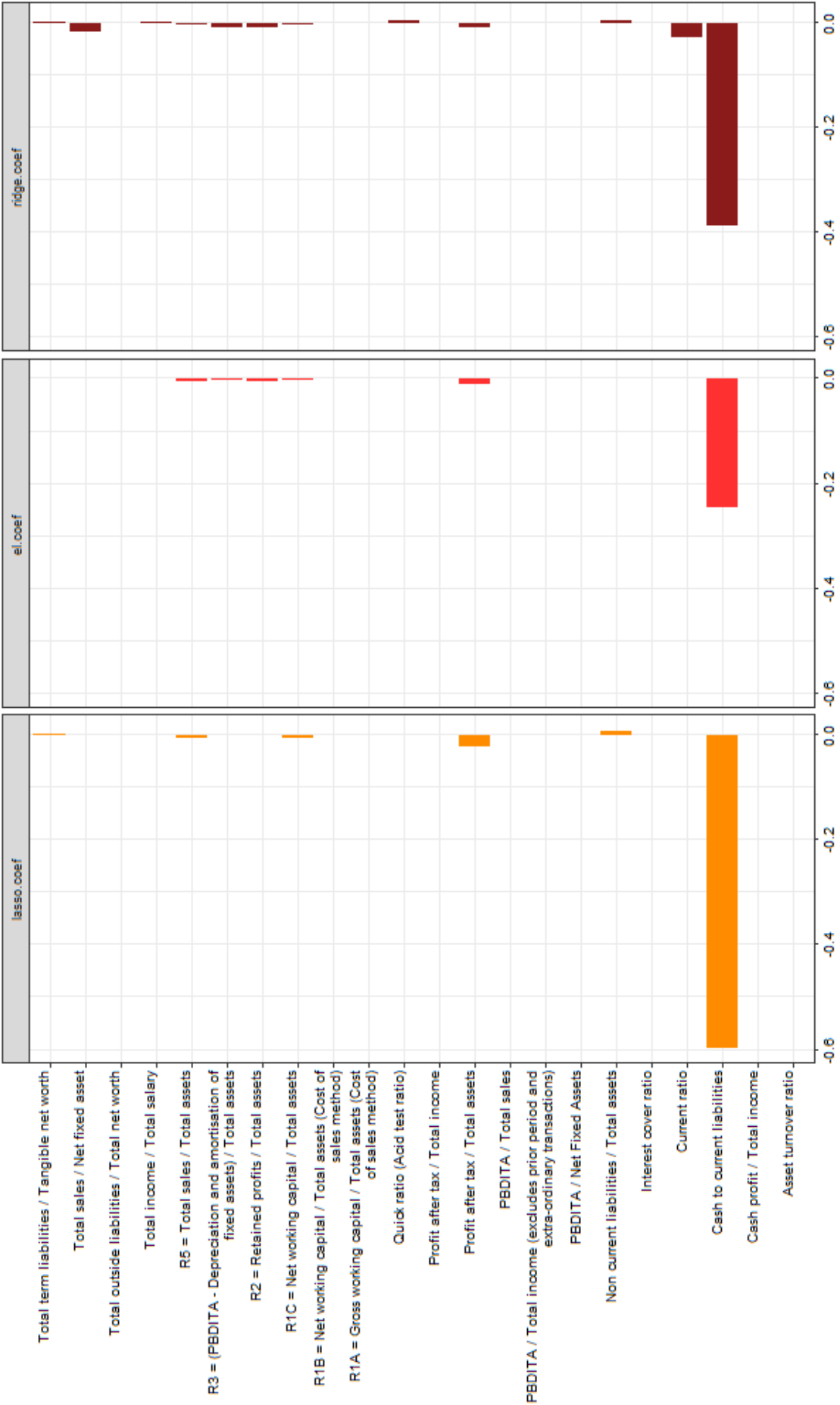
measure of profitability-Profit after tax to total assets becomes significant. In line with our previous findings, we again see term liabilities (Non-current liabilities to total assets and total term liabilities to tangible net worth) as significant.

The results change considerably when we move under the elastic net regime of $s = 0.5$. All ratios except R4 of Altman (1968) become significant together with the cash to current liability and profit after tax to total assets. These additional variables were also significant predictors under LASSO. The relaxed penalty of ridge regression, $s = 1$, emphasizes that all variables are important, barring a few whose coefficients are beaten down to almost 0. Here, liquidity measures, tangibility, and profitability emerge to be important, as seen in the case of quick ratio, current ratio, total term liabilities/tangible net worth, and profit after tax/total assets.

5.4. Principal Component Analysis

Given multiple ratios that hold similar information on liquidity, profitability, and tangibility, the PCA looks for a linear combination of variables that explain most of the variation in the larger dataset. In our case, for the set of 22 accounting ratios, when we construct a matrix of the components, it is observed that the first principal component explains 26.2%

Figure 1: Variable selection using LASSO/Elastic Net



of the variation, which drops down to 3% upon reaching the 10th component (see Table 6). Cumulatively, however, the top 10 principal components explain 88% of the variation in the ratios.

Table 6: Variation explained by first ten principal components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	2.324	1.628	1.541	1.440	1.377	1.186	1.089	0.967	0.915	0.812
Proportion of Variance	0.246	0.120	0.108	0.094	0.086	0.064	0.054	0.042	0.038	0.030
Cumulative Proportion	0.246	0.366	0.474	0.568	0.654	0.718	0.772	0.815	0.853	0.883

To test if principal components can explain the default status of a firm, we run a probit model with principal components as explanatory variables. Each principal component is tested standalone and together to test its significance. In addition, we also control for firm characteristics by adding dummies for the entity type, industry type, listing, and affiliation. The results are reported in Table 7.

Table 7: Regression results (Principal components as explanatory variables)

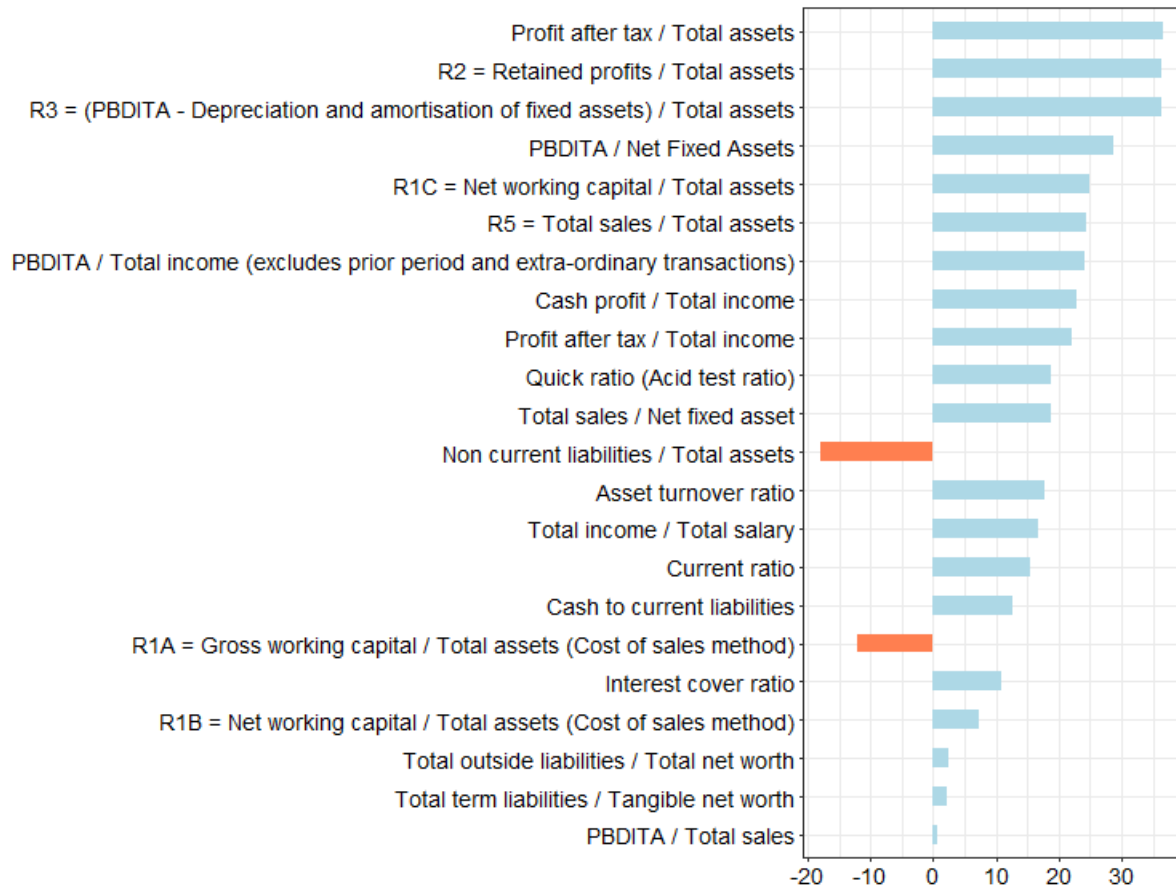
<i>Dependent variable: Default status</i>					
	(1)	(2)	(3)	(4)	(5)
PC1	-0.400*** (0.112)	-0.408*** (0.113)	-0.414*** (0.113)	-0.447*** (0.124)	-0.507*** (0.130)
Ownership	No	Yes	Yes	Yes	Yes
Listed	No	No	Yes	Yes	Yes
Entity type	No	No	No	Yes	Yes
Industry type	No	No	No	No	Yes
Observations	82	82	82	82	82
Log Likelihood	-46.249	-45.969	-45.715	-45.108	-43.235
Akaike Inf. Crit.	98.499	99.939	101.429	102.216	114.471

Note: *p<0.1; **p<0.05; ***p<0.01 The coefficients and standard error for other principal components is not reported here to save space but is available upon request to authors.

We found only the first principal component (PC1) to be significant. Since PC1 is solely significant and explains most of the variation, we investigate the elements of this component. For this, we decompose PC1 into its eigenvalues and plot the same (see Figure 2). Here again, we find that four out of the five z-score ratios-R2, R3, R1C, and R5-come within the top six ratios by eigenvalues. Additionally, the measure of profitability (Profit after tax to total assets, PBDITA to net fixed assets, and PBDITA to total income) also

has high weights. Liquid cash holding of firms (cash profit to total income and quick ratio) also shows significant predictive ability.

Figure 2: Eigenvalues for principal component one



5.5. Summary

All the five ratios considered significant in Altman (1968) - R1C, R2, R3, R5, and non-current liabilities to total assets (a close proxy of R4) - prove to be significant in the Indian data.

The bulk of ratios related to the firm's profitability - profit after tax to total assets, profit after tax to total income, PBDITA to net fixed assets, and PBDITA to total income are also significant. Additionally, we found short-term liquidity measures proxied by the cash holding of the firms as significant. Cash to current liabilities, cash profit in total income, and quick ratio are statistically significant predictors of default. Total sales by a net fixed asset - a measure of asset productivity also came significant. Total term liability to tangible net worth, a measure of leverage closely similar to R4, is also significant. In total, we found

14 financial ratios to be a statistically significant predictor of default based on at least two variable selection models.⁶

In the following section, we will only consider these 14 financial ratios to evaluate the performance of various distress prediction models.

6. Relative performance of distress prediction models

We now compare the classification accuracy of rival default prediction models, on a one-year horizon. We show the confusion matrix for each model, showing Type I and Type II errors. Type I errors are the misclassification of non-bankrupt firms as bankrupt. Type II errors are the reverse - bankrupt firms misclassified as non-bankrupt. We also show the Receiver Operating Curve (ROC) and report the model performance using the area under the curve (AUC). The ROC curve plots sensitivity (true positive rate) against specificity (false positive rate), displaying this trade-off graphically. The further the curve is away from the diagonal line, the more the area is covered under it. The more the area covered under it, the better the curve is at differentiating true positives from true negatives.

To train the models, we split the data into two parts: (1) Training sample: It consists of 70% of the matched data where bankrupt and non-bankrupt firms are in equal proportion (50-50) to train our models, and (2) Test sample: The rest 30% of the data is set aside to check the accuracy of the models. We also report the model's performance on the entire universe of manufacturing firms for each default prediction model. The confusion matrix showing Type-I and II errors is reported for the training sample, test sample, and the full sample of manufacturing firms, together with the ROC curve and AUC estimates.

The choice of reporting the entire confusion matrix and not just the AUC or accuracy ratio is intentional. In most default prediction datasets, the number of firms filing for bankruptcy is far less than the number of firms not filing for bankruptcy (in our case, it is around 1%). It makes a random guess of not bankrupt quite an accurate guess. In our case, 99% of the time, we will be right if we randomly forecast a firm not to be bankrupt. This is simply the statistical property of the input data. This also makes most accuracy measures relatively imprecise and strongly driven by the classification of non-bankrupt firms. Also, note that it is generally agreed upon that Type I errors are very costly for a firm compared to Type II errors. It can result in loss of business (audit clients), damage to a firm's reputation, and potential lawsuits/court costs. Therefore, the choice of the model will also be strongly driven by Type I error. The ultimate user and the purpose of the model will also help guide the choice of model.

The rest of the section proceeds as follows. First, we document the predictive accuracies of individual financial ratios using logistic regression. Secondly, we consider all 14 financial ratios in a multivariate setting and use linear discriminant analysis, multivariate logit, LASSO, elastic net, and principal component analysis to estimate the predictive accuracies of these models. Finally, we compare and contrast the performance of univariate and multivariate models.

⁶Note that for the PCA, we consider only the first principal component and the ratios that have the highest eigenvalues (top 11 ratios).

6.1. Univariate logit models

We begin by first checking the performance of individual financial ratios to predict default using logistic regression. Table 8 documents the confusion matrix, while Figure 3 shows the ROC curve for individual financial ratios based on the training, test, and full sample forecast accuracies. The AUCs are reported in Table 9.

When we compare the Type I error in the training sample (Table 8 column 1 second row), R2, profit after tax to total income, PBDITA to total income, non-current liabilities by total assets, and total term liabilities to tangible net worth have Type I error above 0.45 suggesting very high probability to classify non-bankrupt firms as bankrupt. On the other hand, quick ratio, cash to current liability, and total sales to net fixed assets have Type I error less than 0.25. However, when we look at their Type II error, it is around half. This suggests inferior accuracy in classifying bankrupt firms as bankrupt. The remaining-R1C, R3, R5, profit after tax to total assets, PBDITA by net fixed assets, and cash profit to total income-seem like good candidates for default prediction.

Next, we compare the Type II error for these remaining ratios. Cash profit to total income at 0.14 is followed by R1C and R3 at 0.21, profit after tax by total assets, and PBDITA to net fixed assets at 0.24. R5 has the highest Type II error at 0.34. When we look at the test sample forecasts, all these ratios have Type I error higher than 0.5 except R5, which has a Type I error of 0.42. It shows inferior predictive ability in out-of-sample forecasts.

Another way to filter financial ratios having high predictive accuracy is to think of which all financial ratios have both Type I and II error below the 50% threshold in out-of-sample forecasting. We found only two ratios that satisfy these criteria: R5 (0.42/0.33) and non-current liabilities by total assets (0.33/0.25). Using the same criteria for full sample forecasts, we found R1C, R2, R3, R5, profit after tax by total assets, PBDITA by net fixed assets, non-current liabilities by total assets, and cash profit by total income as significant.

If we look at the AUC for individual financial ratios (Table 9), we found the training sample AUC to be relatively high (around 75%). Nevertheless, this drops more than 20% for almost all financial ratios when we forecast default for the test sample. In the training sample, profit after tax to total income and cash profit to total income have the highest AUC at 0.83, dropping to 0.52 and 0.58 in the holdout sample. Some financial ratios' low Type I error mainly drives the recovery we see in the full sample. Non-current liabilities to total assets is the only financial ratio maintaining its accuracy ratio in all samples.

In summary, we found the forecasting ability of individual financial ratios to be quite limited. R5 and non-current liabilities to total assets are better and more consistent candidates, with AUC at 0.68 and 0.75 for out-of-sample forecasting. However, the Type I error for R5 and non-current liabilities to total assets at 0.42 and 0.33 are significantly high.

Table 8: Univariate logit models performance: Confusion matrix

	Actual	Training sample Predicted		Test sample Predicted		Full sample Predicted	
		Yes	No	Yes	No	Yes	No
Financial Ratio							
R1C = Net working capital / Total assets	Yes No	0.79 0.34	0.21 0.66	0.67 0.50	0.33 0.50	0.76 0.39	0.24 0.61
R2 = Retained profits / Total assets	Yes No	0.79 0.45	0.21 0.55	0.75 0.50	0.25 0.50	0.87 0.46	0.13 0.54
R3 = (PBDITA - Depreciation and amortisation of fixed assets) / Total assets	Yes No	0.79 0.38	0.21 0.62	0.67 0.50	0.33 0.50	0.83 0.41	0.17 0.59
R5 = Total sales / Total assets	Yes No	0.66 0.31	0.34 0.69	0.67 0.42	0.33 0.58	0.76 0.34	0.24 0.66
Profit after tax / Total assets	Yes No	0.76 0.34	0.24 0.66	0.58 0.58	0.42 0.42	0.84 0.41	0.16 0.59
Profit after tax / Total income	Yes No	0.90 0.45	0.10 0.55	0.92 0.75	0.08 0.25	0.92 0.54	0.08 0.46
PBDITA / Net Fixed Assets	Yes No	0.76 0.31	0.24 0.69	0.67 0.67	0.33 0.33	0.80 0.41	0.20 0.59
PBDITA / Total income (excludes prior period and extra-ordinary transactions)	Yes No	0.90 0.48	0.10 0.52	0.83 0.75	0.17 0.25	0.92 0.56	0.08 0.44
Non current liabilities / Total assets	Yes No	0.72 0.52	0.28 0.48	0.75 0.33	0.25 0.67	0.72 0.46	0.28 0.54
Total term liabilities / Tangible net worth	Yes No	0.93 0.62	0.07 0.38	0.92 0.58	0.08 0.42	0.81 0.61	0.19 0.39
Cash to current liabilities	Yes No	0.34 0.17	0.66 0.83	0.17 0.00	0.83 1.00	0.31 0.12	0.69 0.88
Cash profit / Total income	Yes No	0.86 0.38	0.14 0.62	0.83 0.67	0.17 0.33	0.92 0.46	0.08 0.54
Quick ratio (Acid test ratio)	Yes No	0.52 0.14	0.48 0.86	0.33 0.17	0.67 0.83	0.46 0.15	0.54 0.85
Total sales / Net fixed asset	Yes No	0.48 0.21	0.52 0.79	0.42 0.25	0.58 0.75	0.50 0.22	0.50 0.78

Figure 3: ROC curve

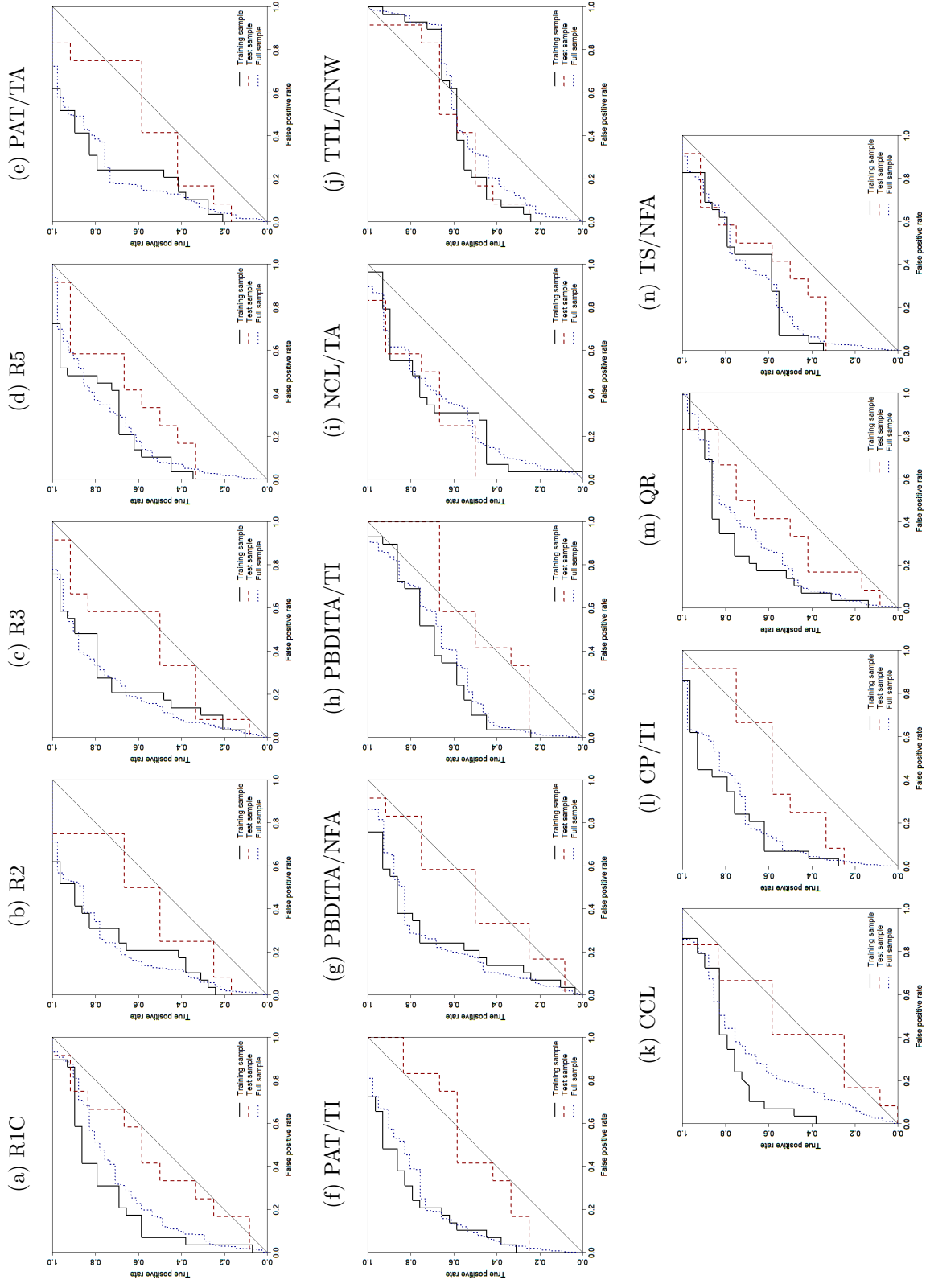


Table 9: Goodness of fit - Univariate logit models: Area under the curve (AUC)

Financial_Ratio	Sample		
	Training	Test	Full
R1C = Net working capital / Total assets	0.78	0.56	0.72
R2 = Retained profits / Total assets	0.80	0.60	0.81
R3 = (PBDITA - Depreciation and amortisation of fixed assets) / Total assets	0.78	0.60	0.80
R5 = Total sales / Total assets	0.81	0.68	0.79
Profit after tax / Total assets	0.80	0.58	0.80
Profit after tax / Total income	0.83	0.52	0.81
PBDITA / Net Fixed Assets	0.76	0.53	0.77
PBDITA / Total income (excludes prior period and extra-ordinary transactions)	0.70	0.47	0.68
Non current liabilities / Total assets	0.71	0.75	0.70
Total term liabilities / Tangible net worth	0.58	0.60	0.54
Cash to current liabilities	0.80	0.52	0.72
Cash profit / Total income	0.83	0.58	0.80
Quick ratio (Acid test ratio)	0.78	0.62	0.73
Total sales / Net fixed asset	0.73	0.65	0.73

6.2. Multivariate analysis

Here we consider all 14 financial ratios together in various multivariate settings-linear discriminant analysis, multivariate logit, the least absolute shrinkage, and selection operators (elastic net and LASSO) together with the unsupervised technique of variable selection (PCA)-to estimate the predictive accuracies of these models.

Linear discriminant analysis (LDA): Following [Altman \(1968\)](#), based on the known classification of the firms in two groups: N (non-bankrupt firms) and D (bankrupt firms) in the training sample, we define the discriminant function as follows:

$$f(X_i) = (m_N - m_D) W^{-1} \left(X_i - \frac{(m_N + m_D)}{2} \right) \quad (4)$$

Where X_i is the financial characteristic measured by various financial ratios of firm i , m_D is the average point of the group of bankrupt firms, m_N is the average point of the group of the non-defaulted firms, and W is the matrix of intra-group variance-covariance. The decision rule is as follows: $f(X_i) > s_i$, the company is allocated to group N ; otherwise to group D . Since our training sample is evenly split between bankrupt and non-bankrupt firm, the score function $s_i = 0.5$.⁷

⁷This model is slightly different from Altman's z-score model. In [Altman \(1968\)](#), using the firms' financial ratios, a linear combination of financial ratios with different weights is defined to calculate the z-score. The weights are estimated using LDA. For a z-score above a certain threshold, firms were considered safe. For z-score below a certain threshold, they were classified as risky and expected to default in the coming year. In between, Altman defines a grey area where we must be concerned and, if possible, take preventive measures.

By virtue of its simplicity and effectiveness, this is an appropriate statistical framework to begin building models for the early detection of firm distress.

Table 10 shows the weights assigned to different financial ratios using LDA. All of Altman's ratios - R1C, R2, R3, R4, and R5 turn out to be significant at 5%. Additionally, we find profit after tax by the total asset to be significant at 1%. However, the sign of R3 is not in line with what we should have expected. In addition, two ratios related to the cash holdings of the firm - cash to current liabilities and cash profit to total income - are statistically significant. Total sales to net fixed asset - a productivity measure also turns out to be significant at the 5% confidence. The coefficients are in sync with our expectations. With a higher cash margin and sales, the firms are less likely to go bankrupt. All ratio related to the firm's profitability turns out to be significant. Higher leverage measured by the non-current liability by total asset is also positively associated with default, suggesting that the model estimates a higher probability of going bankrupt for firms with accumulated term liabilities.

Table 10: Linear discriminant analysis (LDA)

Description	LD1	Pr(>F)
R1C = Net working capital / Total assets	-0.0038***	0.0031
R2 = Retained profits / Total assets	-0.0085***	0.0007
R3 = (PBDITA - Depreciation and amortisation of fixed assets) / Total assets	0.0152***	0.0007
R4 = Non current liabilities / Total assets	0.0064**	0.0114
R5 = Total sales / Total assets	-0.0050***	0.0001
Cash to current liabilities	-1.3546**	0.0182
Cash profit / Total income	-0.0018**	0.0332
Quick ratio (Acid test ratio)	0.3812	0.1462
Profit after tax / Total assets	-0.0385***	0.0004
Profit after tax / Total income	0.0014*	0.0560
PBDITA / Net Fixed Assets	0.0013*	0.0510
PBDITA / Total income (excludes prior period and extra-ordinary transactions)	-0.0025**	0.0330
Total sales / Net fixed asset	-0.0275**	0.0125
Total term liabilities / Tangible net worth	0.0048	0.2651

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 11 (Part A) documents the predictive accuracy of the LDA model. For the training sample, Type I errors is 24%. However, when we compare this with the classification accuracy for the holdout sample, Type I error rises to 58%. As we have pointed out before, Type I classification error is a severe concern in default prediction models. Classifying a non-bankrupt firm as bankrupt might lead to severe damage to the firm's reputation. The full sample forecast is relatively better, with a Type I error of 37%.

Type II error for the LDA model is relatively low at 21%. Even for the holdout sample,

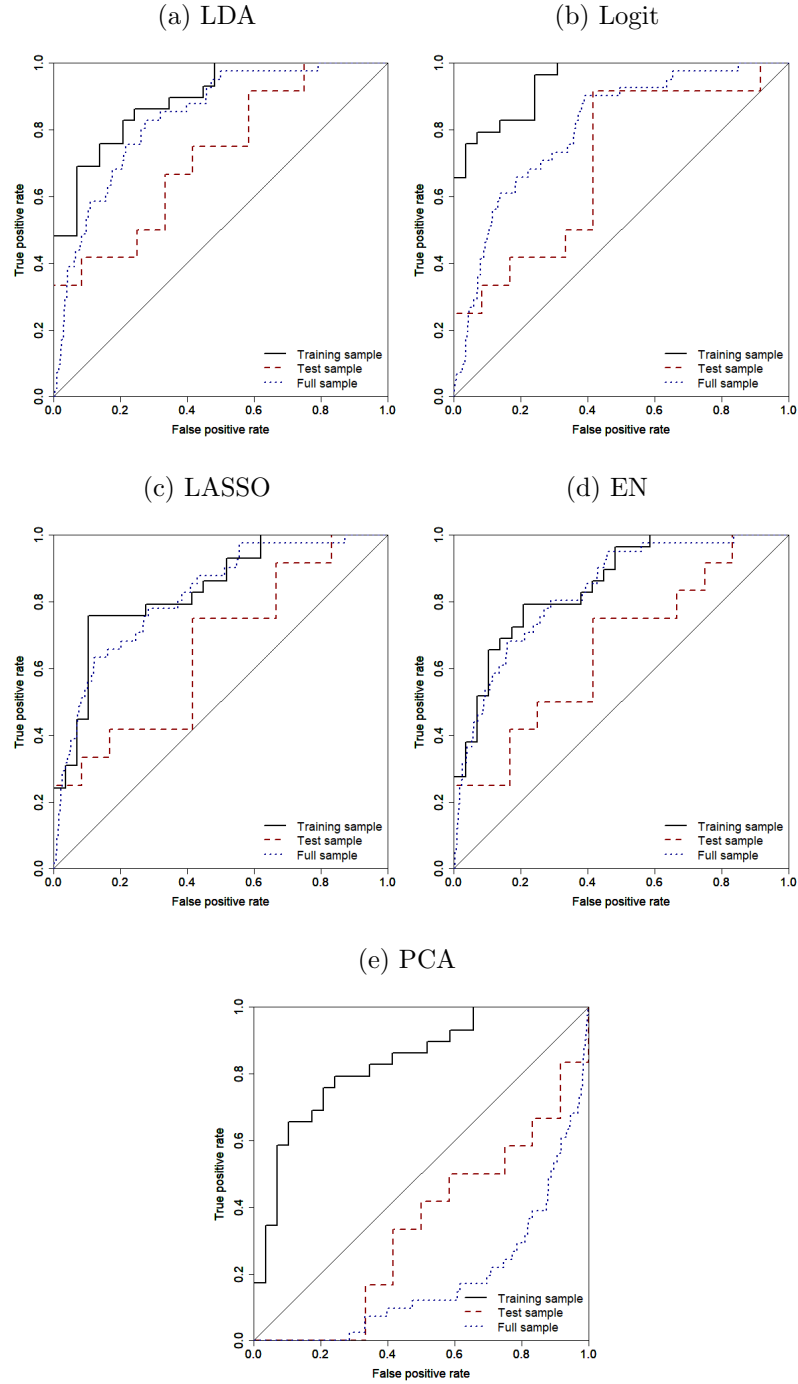
the classification error is 8%. For the full sample, it is 17%. When we look at the ROC (Figure 4), the model performs quite well for the training sample with the area under the curve at 0.89 (see Table 12). Given its large Type I error, it drops to 0.72 for the test sample. In the full sample, the AUC is 0.84.

Table 11: Multivariate model results

		Predicted				
		Training sample	Test sample		Full sample	
Part A: Linear Discriminant Analysis (LDA)						
Actual	Yes	No	Yes	No	Yes	No
Yes	0.79	0.21	0.92	0.08	0.83	0.17
No	0.24	0.76	0.67	0.33	0.37	0.63
Part B: Multivariate logistic regression						
Actual	Yes	No	Yes	No	Yes	No
Yes	0.86	0.14	0.67	0.33	0.76	0.24
No	0.21	0.79	0.58	0.42	0.32	0.68
Part C: LASSO						
Actual	Yes	No	Yes	No	Yes	No
Yes	0.72	0.28	0.67	0.33	0.81	0.19
No	0.24	0.76	0.58	0.42	0.34	0.66
Part D: Elastic Net (EN)						
Actual	Yes	No	Yes	No	Yes	No
Yes	0.79	0.21	0.83	0.17	0.84	0.16
No	0.28	0.72	0.58	0.42	0.37	0.63
Part E: Principal Component Analysis (PCA)						
Actual	Yes	No	Yes	No	Yes	No
Yes	0.79	0.21	0.25	0.75	0.28	0.72
No	0.28	0.72	0.42	0.58	0.78	0.22

Multivariate logit model: We estimate the multivariate logistic regression using all the 14 financial ratios found significant in the previous section. Table 11 - Part B shows the classification accuracy calculated using the multivariate logit model. The correct classification rate of healthy companies is approximately 79%, with Type I error of

Figure 4: ROC curve



21% and Type II error of 14% in the training sample. The accuracy drops sharply in the test sample, with Type I error at 58% and Type II error at 33%. For the full

Table 12: Goodness of fit statistics: Multivariate models

Models	Training sample	Test sample	Full sample
LDA	0.89	0.72	0.84
Multivariate Logit	0.95	0.70	0.81
LASSO	0.84	0.66	0.82
Elastic Net	0.85	0.66	0.83
PCA	0.83	0.33	0.18

sample, the Type I error at 32% is relatively better than what we observe for the LDA model (Type I error of 37%). On the other hand, Type II error is relatively large at 24% compared with 17% for LDA.

Comparing the results with the work of [Jabeur and Fahmi \(2018\)](#), [Du Jardin \(2015\)](#), and [Tinoco and Wilson \(2013\)](#), we find the accuracy of the logit model to be in line with findings in other country studies. However, when compared with the LDA (see [Jabeur and Fahmi \(2018\)](#), [Lin \(2009\)](#), [Lennox \(1999\)](#) and [Li et al. \(2010\)](#)), our findings do not support the claim that logit performs better than LDA. We find logit models have lower Type I error, but the Type II error is significantly higher than the LDA model. If we compare the goodness of fit using AUC (Table 12), we do find the logit model (0.95) performing better than LDA (0.89) in the training data. However, LDA performs better for the holdout sample and full sample (0.72 to 0.70 in the test sample and 0.84 to 0.81 in the full sample). These findings are more in line with [Bardos \(1998\)](#) where better rankings are obtained using LDA than from the logit models.

LASSO and Elastic Net: Table 11 - Part C shows the results for the reduced-form model with LASSO-selected variables. According to Table 11, in the training sample, the rate of correct classification of healthy companies is 76%, with a Type I error of 24% and Type II error of 28% for LASSO. The LASSO model does not perform well in the test sample, and the Type I error reaches 58%, while the Type II error rises to 33%. For the full sample, the Type I error drops to 34% and the Type II error to 19%. The overall model accuracy measured using AUC is 0.84 for the training sample, 0.66 for the test sample, and 0.82 for the full sample.

Table 11 - Part D shows the results for the reduced-form model with elastic net-selected variables. For the elastic net model, the model predicts 79% of the firm failure one year ahead in the training sample. It increases to 83% and 84% in the test and full samples, respectively. However, for the non-defaulted firms, the Type I error is 28% in the training sample that rises to 58% in the test sample. In the full sample, 63% of the time, the model classifies non-bankrupt firms correctly.

When we compare the performance of LASSO with elastic net, we find that LASSO has lower Type I error, while elastic net has lower Type II error. In terms of AUC, elastic net marginally outperforms LASSO in both - in-sample and out-sample predictions. Compared with the logit model, logit performs better than both LASSO and elastic

net in correctly classifying non-bankrupt firms (lower Type I error). In the training sample, logit also performs better incorrectly classifying the bankrupt firm. However, both these models perform better than the logit model in out-of-sample predictions (holdout sample and full sample), incorrectly classifying bankrupt firms.

Principal Component Analysis: In Table 11 - Part E, we show the results for the unsupervised technique of variable selection using principal component analysis. Since only the first principal component was significant, we predict a one-year ahead default using the first principal component in the logit regression. In the training sample, the correct classification rate of healthy companies is 72%, with a Type I error of 28%. For bankrupt firms, the classification accuracy is 79%, with the Type II error rate at 21%. However, for the test sample, the Type II error rises above 50%. For the holdout sample, the Type II error is 75%, while in the full sample, it is 72%. We also found similarly high Type I error rates in the out-of-sample predictions. When we look at the ROC (Figure 4) and AUC (Table 11), we found the area under the curve below 0.5, suggesting worse than the random classification.

6.3. Performance comparison

So far, the results obtained do not indicate a robust and efficient model that outperforms all other models. LASSO performs best when it comes to having lower Type I error, while LDA and elastic net perform significantly better in reducing Type II error. When we compare performance based on AUC, for the training sample, multivariate logit performs significantly better with an AUC of 0.95. LDA comes next with an AUC of 0.89, followed by elastic net (0.85), LASSO (0.84), and PCA (0.83). However, when we look at the out-of-sample predictive accuracies, LDA performs better than logit in the holdout and full samples. LASSO and elastic net come close third in the holdout sample but performs at par with logit in the full sample. PCA performs extremely poorly with AUC below 0.5 for out-of-sample predictions.

Comparing the results of the multivariable models with the performance of single variables, we observe that by far, the LDA model (proposed by Altman (1968)) has the best predictive accuracy. The AUC is highest for LDA in both test and full samples at 0.72 and 0.84. However, when we compare the Type I errors, single variable forecast using R5 and non-current liabilities to total assets perform at par with LDA. In the holdout sample, the Type I error for R5 is 0.42, and that for non-current liabilities to total assets is 0.33 compared to 0.67 for LDA. In the full sample, the Type I errors for R5 and non-current liabilities to total assets are 0.34 and 0.46, respectively, compared to 0.37 for LDA. Regarding Type II error, LDA outperforms single variables R5 and non-current liabilities to total assets.

6.4. Robustness checks

To validate the performance of default prediction models, we do two additional tests: First, based on the Jones (1987) suggestion, we change the size of our holdout sample. We randomly select 50%, 40%, 20%, and 10% of our sample as a holdout sample. Based on the remaining sample, we develop our model and then apply this to the holdout sample

to acquire a more robust measure of the model’s predictive accuracy. Secondly, we use the Lachenbruch method, where one observation is withheld from the estimation sample, and its classification is predicted. This process is repeated until each observation has been withheld and predicted. Our predictive accuracies fluctuate, but qualitatively the performance of models remains consistent. Our finding remains robust, and we do not find a robust and efficient model that outperforms all other default prediction models in the context of IBC.

7. Conclusion

This paper falls in a long tradition of a literature that has developed methods that utilise firm data for the purpose of bankruptcy prediction. In our knowledge, this is the first time this has been done in the Indian data using a sound dataset about defaulted firms, where we were able to harness the information that emerged from filings under the newly enacted bankruptcy code. The contribution of this paper lies in bringing a sound Indian defaults dataset to a comprehensive set of statistical techniques of the bankruptcy prediction literature.

Given the immense development of statistical techniques and economic thinking on bankruptcy, and given the fact that the Indian locale is quite different from the US, it was a surprise to see a familiar echo from Altman’s work from 1968 for the US. We found that R1C, R2, R3, R5, and the cash to current liability are statistically robust predictors of default. Four of these five are from the 1968 Altman model.

Cash holding and other liquidity measures associated with in-hand cash are also significant predictors of default. This is a unique feature of the Indian experience. We conjecture that this may also be a feature of other emerging markets with under developed financial systems.

In terms of model accuracy, the reduced form model selected via the linear discriminant analysis (also used in [Altman \(1968\)](#)) performs at par with the more advanced econometric specification in in-sample and full-sample default predictions. Here also, it was surprising to see an echo of Altman’s 1968 work survive the challenges of immensely more sophisticated methods developed in the following decades.

Acknowledgement

We thank Ajay Shah, Susan Thomas, Anjali Sharma, Sourish Das, Abhishek Seth, and Swayamshree Barik for valuable comments. The views expressed in this paper belong to the authors and not their employer. All remaining errors are our own.

References

- Agarwal, V., Taffler, R.J., 2008. Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking and Finance* 32, 1541–1551.
- Altman, E., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance* 23, 589–609.
- Barboza, F., Kimura, H., Altman, E., 2017. Machine learning models and bankruptcy prediction. *Expert Systems with Applications* 83, 405–417.
- Bardos, M., 1998. Detecting the risk of company failure at the banque de france. *Journal of Banking & Finance* 22, 1405–1419.
- Bauer, J., Agarwal, V., 2014. Are hazard models superior to traditional bankruptcy prediction approaches? a comprehensive test. *Journal of Banking & Finance* 40, 432–442.
- Beaver, W.H., 1966. Financial ratios as predictors of failure. *Journal of Accounting Research* 4, 71–111.
- Bellovary, J.L., Giacomino, D.E., Akers, M.D., 2007. A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial Education* 33, 1–42.
- Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–654.
- Bureau of Business Research, 1930. A test analysis of unsuccessful industrial companies. Urbana, University of Illinois. URL: <https://books.google.co.in/books?id=F9fcPQAACAAJ>.
- Campbell, J., Hilscher, J., Szilagyi, J., 2008. In search of distress risk. *The Journal of Finance* 63, 2899–2939.
- Chatterjee, S., Shaikh, G., Zaveri, B., 2017. Watching India's insolvency reforms: a new dataset of insolvency cases. Indira Gandhi Institute of Development Research, Working Paper WP-2017-012.
- Du Jardin, P., 2015. Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research* 242, 286–303.
- Foo, S.L., Pathak, S., 2016. Financial health and corporate performance: A comparison of manufacturing companies in China and India. *Journal of Asian Development* 2, 18.
- Greenwood, R., Iverson, B., Thesmar, D., 2020. Sizing up corporate restructuring in the covid crisis. Technical Report. National Bureau of Economic Research.
- Horta, I., Camanho, A., 2013. Company failure prediction in the construction industry. *Expert Systems with Applications* 40, 6253–6257.
- Jabeur, S.B., Fahmi, Y., 2018. Forecasting financial distress for french firms: a comparative study. *Empirical Economics* 54, 1173–1186.
- Jones, F., 1987. Current techniques in bankruptcy prediction. *Journal of Accounting Literature* 6, 131–164.
- Kulkarni, A., Mishra, A.K., Thakker, J., 2005. How good is Merton model at assessing credit risk? Evidence from India. National Institute of Bank Management.
- Lennox, C., 1999. Identifying failing companies: a re-evaluation of the logit, probit and DA approaches. *Journal of economics and Business* 51, 347–364.
- Li, H., Sun, J., Wu, J., 2010. Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems with Applications* 37, 5895–5904.
- Lin, T.H., 2009. A cross model study of corporate financial distress prediction in Taiwan: Multiple discriminant analysis, logit, probit and neural networks models. *Neurocomputing* 72, 3507–3516.
- Merton, R.C., 1974. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance* 29, 449–470.
- Ohlson, J.A., 1979. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18, 109–131.
- Patrick, P., 1932. A comparison of ratios of successful industrial enterprises with those of failed companies. *Certified Public Accountant* 2, 598–605.
- Sengupta, R., Sharma, A., Thomas, S., 2016. Evolution of the insolvency framework for non-financial firms in India. Indira Gandhi Institute of Development Research, Working Paper WP-2016-018.
- Shah, J.R., Murtaza, M.B., 2000. A neural network based clustering procedure for bankruptcy prediction. *American Business Review* 18, 80–86.

- Shaonan, T., Yan, Y., Hui, G., 2015. Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance* 52, 89–100.
- Sharma, A., Theresa, L., Mhatre, J., Sajid, M., 2019. Application of Altman z-score to RBI defaulters: Indian case. *Asian Journal of Research in Business Economics and Management* 9, 1–11.
- Shumway, T., 2001. Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business* 74, 101–124.
- Smith, R., Winakor, A., 1935. Changes in financial structure of unsuccessful industrial corporations. Urbana: University of Illinois.
- Sreedhar, T., Shumway, T., 2008. Forecasting default with the Merton distant to default model. *Review of Financial Studies* 21, 1339–1369.
- Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)* 58, 267–288.
- Tinoco, M.H., Wilson, N., 2013. Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis* 30, 394–419.

Appendix A. List of variables

Table A.1: List of variables

Description	Prowess fieldname
Unique identity of firm in ProwessIQ dataset	sa_finance1_cocode
Company Name	sa_company_name
Financial year	sa_finance1_year
Balance-sheet characteristics	
Total assets	sa_total_assets
Total sales	sa_sales
Net fixed assets	net_fixed_asset
Gross working capital	sa_gross_working_capital
Net working capital	sa_net_working_capital
Net working capital (based on cost of sales method)	sa_net_working_capital_cosm
PBDITA	sa_pbdita
Depreciation and amortisation of fixed assets	sa_dep_owned_ast
Long term borrowings	sa_long_term_borrowings
Short term borrowings	sa_short_term_borrowings
Identity variables	
Industry sub-classification within manufacturing	co_industry_gp_code
Entity type = 1 if the firm is registered as public limited and 2 if private limited	entity_type_code
Business House = 1 if the firm belongs to any business group and 0 otherwise	owner_gp_code
Listed = 1 if the firm is listed on an exchange and 0 otherwise	nse_symbol and bse_code