DEPARTMENT OF ENGINEERING MANAGEMENT

Bankruptcy prediction for SMEs using relational data

Ellen Tobback, Julie Moeyersoms, Marija Stankova & David Martens

UNIVERSITY OF ANTWERP
Faculty of Applied Economics

City Campus
Prinsstraat 13, B.226
B-2000 Antwerp
Tel. +32 (0)3 265 40 32
Fax +32 (0)3 265 47 99
www.uantwerpen.be

FACULTY OF APPLIED ECONOMICS

DEPARTMENT OF ENGINEERING MANAGEMENT

Bankruptcy prediction for SMEs using relational data

Ellen Tobback, Julie Moeyersoms, Marija Stankova & David Martens

RESEARCH PAPER 2016-004
MARCH 2016

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration – room B.226
phone: (32) 3 265 40 32
fax: (32) 3 265 47 99
e-mail: joeri.nys@uantwerpen.be

The research papers from the Faculty of Applied Economics are also available at www.repec.org (Research Papers in Economics - RePEc)

D/2016/1169/004

Bankruptcy prediction for SMEs using relational data.

Ellen Tobback[a], Julie Moeyersoms*[a], Marija Stankova*[a] and David Martens[a]

[a]Department of Engineering Management, University of Antwerp

March 1, 2016

Abstract

Bankruptcy prediction has been a popular and challenging research area for decades. Most prediction models are built using traditional data such as financial figures, stock market data and firm specific variables. We complement such *dense* data with *fine-grained* data by including information on the company's directors and managers in the prediction models. This information is used to build a network between Belgian enterprises, where two companies are related if they share or have shared a director or high-level manager. We start from two possibly related assumptions: (i) if a company is linked to many (or only) bankrupt firms, it will have a higher probability of becoming bankrupt and (ii) the management has an influence on the performance of the company and incompetent or fraudulent managers can lead a company into bankruptcy. The weighted-vote relational neighbour (wvRN) classifier is applied on the created network and transforms the relationships between companies in bankruptcy prediction scores, thereby assuming that a company is more likely to file for bankruptcy if one of the related companies in its network has failed. The more related companies have failed, the higher the predicted probability of bankruptcy. The relational model is then benchmarked against a base model that contains only structured data such as financial ratios. Finally, an ensemble model is built that combines the relational model's output scores with the structured data. We find that this ensemble model outperforms the base model when detecting the riskiest firms, especially when predicting two-years ahead.

1 Introduction

Bankruptcy prediction is a widely studied topic due to its importance for the banking sector. The current volume of outstanding debt to non-financial firms in Belgium is about 122 billion euros, which is 123% of GDP as measured in the first quarter of 2015 [28]. The size of corporate lending makes sound lending decisions a matter of national interest. To counter the adverse effects of these high exposures, Basel II and III have introduced capital requirements that are more sensitive to risk. For many Small to Medium Enterprises (SMEs) this implies that banks are charging a higher risk premium [3]. Investing in improved bankruptcy prediction models is therefore in the interest of both the banks and the clients, as better predictions will reduce risk and lower the forthcoming risk premia.

Research on bankruptcy prediction has largely focused on traditional data such as financial ratios, stock data or macroeconomic data [7, 27, 39]. However, it is often noted that the (in)competence of the managerial team has a great influence on a company's chance of survival [32]. To measure a business manager's or board member's competence, one could take

*Julie and Marija contributed equally.

a look at the business history of this person. When a person was involved in a bankruptcy case in the past, banks will be reluctant to grant this person a loan for the start-up of a new firm. Notwithstanding the clear importance of the management's competence and historical success/failure, most research on bankruptcy prediction ignores this kind of data. In this paper, we intend to fill this research gap and try to predict bankruptcy using both the traditional, financial data and fine-grained data on person-related relationships.

We exported data from Belfirst, a database containing financial reports and statistics on Belgian and Luxembourg companies. This data can be categorized into traditional data and relational data. The traditional, dense data are mostly financial ratios such as the current ratio, debt ratio and return on assets; and firm-specific data such as the company's age and sector. The relational data captures the links between companies at the board and management level. Using relational data in a bankruptcy setting we start from two possibly related assumptions: (i) if a company is linked to many (or only) bankrupt firms, it will have a higher probability of becoming bankrupt and (ii) the management has an influence on the performance of the company and incompetent or fraudulent managers can lead a company into bankruptcy. The latter assumption has already been investigated and proven by researchers [4, 32], the former assumption is a possible derivation from the latter, however we do not exclude the fact that there could be other causes leading to the first assumption.

Figure 1 illustrates our methodology in line with the approach of Stankova et al [38]. We project a bipartite graph of links between companies and their directors/managers into a weighted unigraph that links companies to each other. Relational learners are applied to the network of companies. The network model is then compared to a base model containing only financial data and an ensemble model that combines the network scores with the financial ratios.

The contributions to the literature are

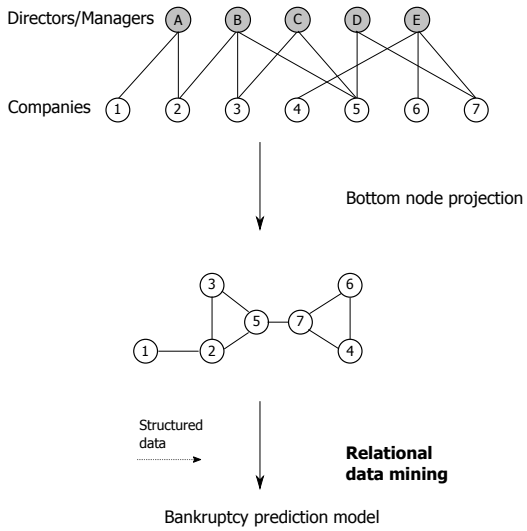

Figure 1: We create a weighted projection from the bigraph. The board members/managers are the top nodes and the companies are the bottom nodes. Relational learners are applied to the resulting network of companies. The network model is then compared to a base model containing financial data and an ensemble model.

three-fold. Empirical research on corporate bankruptcy focuses mainly on innovations in modeling techniques and only to a lesser extent on innovations in the feature space. To our best knowledge, we are the first to use fine-grained data about relationships between companies to predict bankruptcy. Secondly, whereas most studies focus on a sample of companies from a certain country, we used all Belgian SMEs that publish financial statements leading to a data set of approximately 400,000 companies. Finally, we use a completely out-of-time set-up and take into account that healthy companies in the training set can become bankrupt companies in the test set and apply a tailored 'leave-many-out' procedure to deal with the double occurrences.

The major findings of our research can be di-

vided into two categories: the empirical results and the insights that can be derived from them. Concerning the empirical results we find that (i) the relational model has some predictive power, though cannot be used separately; (ii) the network scores add complementary predictive power to our base model using traditional, financial data and (iii) the difference in lift for the top segment between the base and ensemble model increases in favour of the latter when predicting bankruptcy two-years ahead. From these results the following insights can be deduced: (i) companies related to many bankrupt firms have a higher probability of failure, especially when these companies find themselves in a bad financial position; (ii) combining the network scores of the relational model with the financial data improves the detection of the companies that are most likely to fail and (iii) by adding the network scores to the base model, the model becomes more forward-looking.

The remainder of this paper is organized as follows. Section 2 defines the research question and provides the reader with a concise overview of the relevant literature and progress. Section 3 describes the financial performance indicators and relational data used in this study. Section 4.1 provides a detailed description of our methodology and Section 4.2 summarizes and analyses the empirical results. Finally, Section 5 discusses the conclusions and provides insights for future research.

2 Literature overview

2.1 Definition and terminology

In this study the term bankruptcy is used interchangeably with failure and default, where the notion of bankruptcy refers to the legal status of an entity when it cannot repay its owed debt. In Belgium, bankruptcy is part of the commercial law, which implies that only merchants can go bankrupt. Bankruptcy is declared by the Chamber of Commerce. According to Article 2 in bankruptcy law, the directors of a commercial company that has durably ceased making payments or that has lost its creditworthiness, are legally obligated to petition for the company's bankruptcy. When bankruptcy is declared, the chairman of the Chamber of Commerce appoints at least one temporary administrator and at least one trustee, dependent on the size of the company and the magnitude of the bankruptcy case. From the moment bankruptcy is declared, the company's directors lose the right to control its assets. During the bankruptcy settlement, all company assets are liquidated and the proceedings are distributed among the creditors.

2.2 Why do firms go bankrupt?

A firm goes bankrupt when it is no longer able to fulfil its financial obligations [5]. Filing for bankruptcy is usually a voluntary decision, however, in rare cases a firm can be forced to file bankruptcy after court decision. Although quite often unexpected, bankruptcy is rarely a sudden event. There are many possible root causes for firm failure. The reasons can be internal, such as mismanagement, fraud, insufficient capital and bad credit management or external, such as unforeseen legal changes, international competition and worsening economic conditions [10, 32]. Whatever the reason, companies eventually fail because there is a mismatch (in time or magnitude) between their cash in- and outflow [36]. Because bankruptcy is usually the result of a gradual process of deterioration, credit lenders and auditors can search for warning signs of an upcoming bankruptcy in the annual report and financial statements. The many performance indicators can show symptoms of an approaching bankruptcy, such as a shortage of cash [36].

Another reason for a firm to go bankrupt is fraud. In a bust-out scheme the company builds up a good credit reputation to eventually obtain loans and goods without the intention to repay them. When payment is due, the company declares bankruptcy. A possible warning sign of bankruptcy fraud are serial bankruptcy cases, i.e. similar businesses are incorporated

near the time period of bankruptcy filing and will go bankrupt not long after [10]. A representation of relationships between companies in a network can help the detection of fraudulent companies.

2.3 Data mining and bankruptcy prediction

There is a vast amount of research on bankruptcy prediction, going all the way back to the 1960's. The earliest research applied a univariate approach, comparing one historical ratio at a time [5]. The multivariate approach to bankruptcy prediction was first introduced by Altman and Ohlson. The former used multivariate discriminant analysis to find the linear function that distinguishes between healthy and bankrupt firms resulting in the famous Z-score [2], while the latter used logistic regression to estimate the probability of bankruptcy for each firm [29]. Both added financial ratios as inputs to their prediction models.

Since the 1990's the focus has shifted towards artificially intelligent expert models, such as neural networks and Support Vector Machines. Multilayer neural networks are reported to significantly outperform both logistic regression [44, 14] and Multivariate Data Analysis (MDA) [43, 17] and a number of studies have successfully applied Support Vector Machines (SVM) for corporate bankruptcy prediction [26, 37] and shown that they are competitive with MDA [25] and logistic regression [26, 40, 25]. The performance improvement of bankruptcy prediction with these intelligent techniques indicates that the influence of financial ratios on a firm's health has non-linear properties, however the choice of non-linear, black-box models decreases the comprehensibility of the bankruptcy predictions. Hence, in a practical setting discriminant analysis and logit models remain dominant.

Empirical research on corporate bankruptcy focuses mainly on innovations in modelling techniques and only to a lesser extent on innovations in the features space. The most frequently used features are firm or industry specific information and performance indicators. Amongst the performance indicators, the current ratio and the Return on Assets ratio are the most commonly used factors [7]. More recent studies have investigated the predictive power of market/stock data [39] and macroeconomic variables [27, 39]. What all the aforementioned studies have in common, is that they focus on dense, structured data (mainly financial ratios). Recent advances in the use of sparse fine-grained data[1] have shown that they add incremental predictive power to the models. Hence, this kind of data can be useful in bankruptcy predictions as well. Various studies highlighted the management's role in the failure process. Ooghe and De Prijcker [32] distinguish three types of shortcomings that may cause corporate bankruptcy: (i) a lack of competences and skills (ii) insufficient motivation and (iii) certain personal characteristics such as risk affinity, over-optimism and haste and Baldwin et al. [4] found that managerial weakness was the main cause of small business bankruptcy in Canada. We can account for these influences on bankruptcy by adding sparse relational data as features to our prediction models.

2.4 Challenges and success of relational data

Relational data is data that defines relationships between two entities. The use of relational data has already proven to be successful in other domains such as targeted advertising [16], fraud detection [19] and customer retention [41]. Two major categories of relational data can be distinguished: real network data and pseudo-network data. A seminal paper using real network data was written by Hill et al [16] and described the use of call data to predict product and service adoption. Verbeke et al [41] used similar data to successfully predict churn and Domingos [12] used social network data for viral marketing. However, quite often no real network

[1] See e.g. [23] for transactional data, [42] for behavioural data and [12] for social network data.

data is available, in which case two nodes can be linked through similar interests or activities: e.g. if they have watched the same videos [42], visited the same places [35]or paid to the same entities [23]. In this research we create an implied network by linking companies based on the shared board members/managers.

The nature of relational data requires a different approach than the traditional financial data. One of the main challenges of using network data is the transformation from its rough form, i.e. a list of managers per company, to a structured form, i.e. a weighted sparse matrix where the weights denote the strength of the link between two entities (here companies). The sparsity of the data set requires a large sample that contains all relevant neighbours for each entity in the data set. Next, specifically tailored learners have to be used to obtain a prediction score for each entity with unknown class in the network. Relational learners are a powerful tool and can handle the low event rate of most network data sets[2]. Section 4.1 explains in detail how we processed the relational data.

3 Data

We gathered data from Belfirst on $+400,000$ Belgian SMEs, covering the time-period of 2011-2014. The classification of companies as SMEs complies with the definition of the Basel II capital accords, where companies are granted an SME-status if the reported yearly sales for the consolidated group the firm belongs to are less than EUR 50 million [31]. We exported financial ratios, the name and unique identifier of the current and past directors and managers, the date of incorporation, the NACEBEL industry code and information about the state of the company (bankrupt or active).

[2]In our training set we have a '1%' event rate.

3.1 Financial performance indicators

Generally, the ratios can be divided into three categories: solvency, liquidity and profitability. Insufficient solvency and liquidity as well as low profitability are all factors that can lead to bankruptcy if not resolved by management. It is therefore important to include at least one indicator from each category in the bankruptcy prediction model. Table 1 lists the financial ratios that are used in this study. We chose a selection of financial performance indicators that covers all categories (liquidity, solvency and profitability). The first category, solvency, represents the company's ability to meet its long-term obligations and can be represented by the equity ratio. This ratio measures the part of total assets that is financed by investor's equity and gives an indication of the long-term debt burden of a firm. The lower the equity ratio, the more leveraged the firm is and the higher the risk of insolvency. The second category, liquidity, indicates a company's ability to meet its short-term obligations. The 'current' ratio is a good indicator of a firm's liquidity as it represents the company's ability to quickly convert assets into cash without any loss. The ideal current ratio is two-to-one, which means the current assets are double the current liabilities [15, 33]. It is important to measure both solvency and liquidity to assess a firm's performance, as they represent a company's long-term and short-term chance of survival, respectively. The last category, profitability, measures the economic viability of a company. Over the long term, the firm must be profitable to ensure that both liquidity and solvency are maintained. Return on assets (ROA) measures the management's ability of converting its assets into profit. A higher ROA indicates that the company is able to generate more earnings with less investment. Return on Equity (ROE) measures the profit the company generates with the shareholder's equity. Finally, the cash flow to equity ratio indicates the company's capacity to create gross income, independent of the use [15, 33]. Insufficient liquid-

Table 1: Financial performance indicators

Variables	Category	Used by
Debt to total assets ratio	Leverage/Solvency	[1, 18, 21, 25, 34]
Current ratio	Liquidity	[9, 11, 36]
Cash flow to equity ratio	Profitability	[20]
Return on equity	Profitability	[6, 9, 20]
Profit/Loss	Profitability	[30]
Return on total assets	Profitability	[6, 36, 45]

ity, insolvency and low profitability (or loss) are warning signs of a possible future bankruptcy, however, the prediction is not perfect. It might be that the company chooses not to publish its financial statement in periods of financial stress, that the financial statement is manipulated, that the company's behaviour is fraudulent or that - due to the delay in publication - the deterioration is not noted on time. Combined with the fact that mismanagement can lead to a company's failure, we supplement financial performance indicators with relational data.

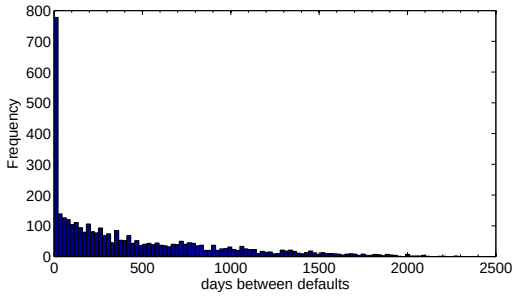

Figure 2: The number of days between the defaults of two linked firms.

3.2 Relational data

We define relational data in this context as data containing information about companies and entities that connect them. In a bipartite graph, the connecting entities would be the top nodes and the SMEs the bottom nodes. There exists a large variety of entities that can be used to connect two firms, from directors and managers to suppliers and clients. In this paper, we use information about past and current directors and managers to link companies. Hence, we create a network of Belgian SMEs where two companies are linked if they share or have shared a member of the board of directors and/or the management board. Using this data builds upon the assumption that being linked to a bankrupt company increases your own probability of default. Figure 2 displays the 'days between default' of two linked SMEs. Note that a large amount of linked firms goes bankrupt on the exact same day. These firms cannot be used in the model. Due to our ensemble set-up which is further explained in Section 4.1, we have a one-year gap between the network training set and our test set. The companies that defaulted in 2013 are not added to the network, since they are used in the ensemble model's training set. If a bank decides to predict bankruptcy for 2015, it can/should update the network to include the bankrupt firms of 2013. We try to predict default one year ahead. However, as Figure 2 shows, the influence of a defaulted company on its linked companies can be delayed. The prediction scores of the relational learners can be interpreted as warning signs as well, i.e. companies with a high score are linked to many or only bankrupt firms and should be closely monitored. As the first histogram in Figure 3 illustrates, most firms in the data set are

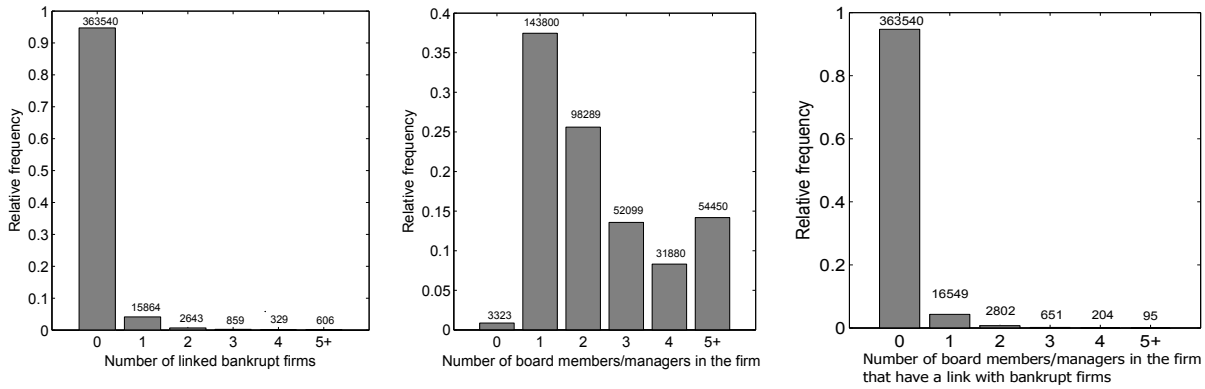

Figure 3: Relational data of the training set. The number of bankrupt firms the company is linked to, the number of board members in the firm and the number of board members with a link to a bankrupt company.

not linked to a bankrupt firm. For the entities that have a link with a bankrupt firm, it is often only one board member/manager that is responsible for this link.

4 Methodology and results

4.1 Methodology

The financial data is used to build a base model. As input variables, we chose a selection of financial ratios that have been shown to be predictive of bankruptcy. Due to the delay on the publication of the financial statements, we are obliged to consult the financial statement of a year prior to the observation date. This means that if we want to predict if the companies active on January 1 2014 will go bankrupt within the next year, we will have to use the financial ratios of 2012.

Some firms have missing values for all financial ratios. There are two major reasons for these missing values: (i) the firm did not publish a financial statement and (ii) the firm was founded in 2012. It can be expected that missing values due to a missing financial statement are a predictive factor for bankruptcy. To distinguish between these firms and newly founded firms, we add a dummy variable that has value 1 if the firm was founded in the respective year and 0 otherwise. All missing values are replaced by the average value of the training set and accompanied by a 'missing values'-dummy. To control for age- and industry-specific effects, we included the normalised number of years since the foundation of the company and the 21 dummy-encoded NACEBEL sections (A-U). We train an SVM with a linear kernel, a technique that is both powerful and comprehensible, thus rendering it an appropriate choice for the modelling problem at hand. We tuned the cost parameter on an in-time, out-of-sample random validation set.

For the relational data, we need to use tailored learners. The relationships can be represented in a bipartite graph, with the directors/managers as top nodes and the companies as bottom nodes. To this bipartite graph, we applied the three-step framework as proposed by Stankova et al [38]. Most of the relational learners are defined for the more general case of graphs with only one type of nodes and we want to make use of them. Hence, we need to first transform the bipartite graph to a weighted unigraph projection, where companies are linked if they share at least one member of the board of directors or managing

board. Next, we apply a relational learner for unigraphs to create the network scores. In accordance with the empirical results, we calculate the edge weight w_{ij} in the projection using the hyperbolic tangent function for the top nodes k with degree d_k and sum of shared nodes as an aggregation function (see Equation 1). The hyperbolic tangens downweights managers that are associated with a large amount of firms, based on the assumption that they will be less discriminative. The choice of this particular weighting scheme is justified by the superior empirical results on a wide range of diverse data sets as reported by [38].

$$w_{ij} = \sum_{k \in N(i) \cap N(j)} tanh(\frac{1}{d_k}) \tag{1}$$

As a relational learner, we apply the weighted-vote Relational Neighbor (wvRN) classifier [22]. It is a simple, yet powerful classifier that uses the network structure to calculate a bankruptcy probability score $P(l_i = c|N(i))$ for a company as a weighted average of its j neighbours' probability scores (see Equation 2). The classifier is based on the property of assortativity (also known as homophily in social networks theory [24]), as it makes the assumption that the connected companies are similar and therefore more likely to belong to the same class. This is aligned with our premise that the companies related through the same managers/directors are likely to exhibit similar bankruptcy behavior due to the incompetence or fraudulent intentions of the managerial team.

$$P(l_i = c|N(i)) = \frac{1}{Z} \sum_{j \in N(i)} w_{ij} P(l_j = c|N(j))$$

where the normalization factor Z is equal to

$$Z = \sum_{j \in N(i)} w_{ij} \tag{2}$$

To build the network, we use the companies active at the beginning of 2013 as training set. Each company is assigned a label 0 if it remained active in 2013 or 1 if it went bankrupt in the course of 2013. To give a complete view of the network, we add to the training set the linked companies that went bankrupt *before* 2013. This network is used to calculate the network score for the companies in the test set, i.e. the companies observed at the beginning of 2014. Figure 4 gives an example of an actual test node and its links to the companies in the training set. The test node is grey, the companies that defaulted before 2013 are black and the companies that were still active at the end of 2012 are white. Companies that were active in 2012 and 2014 are present in both the network training and test set. For these observations, we apply the logic behind 'leave-one-out cross-validation', i.e. we build a network on the entire 2012 training set excluding the values of this company in 2012 and use the values of the company in 2014 as test instance to estimate the

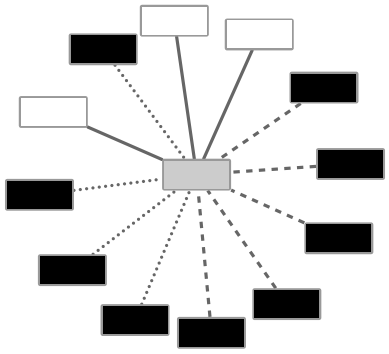

Figure 4: Example of a test node in 2014 (gray) and its connections with healthy companies (white) and companies that defaulted before 2013 (black). The different line types represent the different managers through which the test node is connected to the surrounding nodes. One manager (full lines) is responsible for the connections with the three healthy companies, the two remaining managers (striped and dotted lines) connect the company to bankrupt firms.

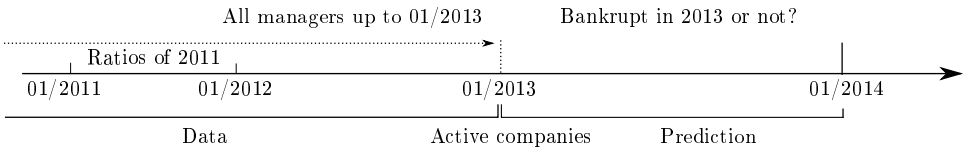

Figure 5: Train set: we use the labels in 2013 (bankrupt +1 and active 0), the ratios of 2011 and all the managers that are/have been part of the companies that are still active on the prediction date (01/2013).

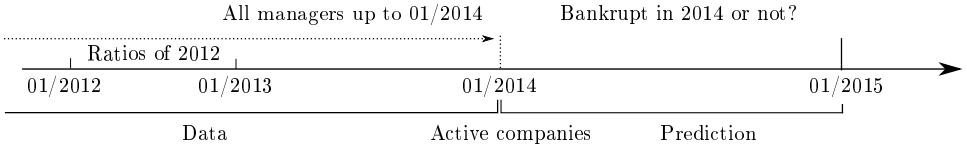

Figure 6: Test set: we want to predict the labels of 2014 (bankrupt +1 and active 0) using the ratios of 2012 and all the managers that are/have been part of the companies that are still active on the prediction date (01/2014). Companies that went bankrupt in 2013 are not part of the test set and are not added to the network either.

network score.

Financial data and relational data are heterogenous types of data that have different modelling requirements. Since it is not possible to use both kinds of data in one model, we combine both models in an ensemble model. The network scores for the training and test set are added as extra variable to the base model. To define the added value of using relational learners, we compare the ensemble model to a 'base plus dummy' model, i.e. a model that adds a dummy variable as extra variable to the base model with value 1 if the company has at least one link with a bankrupt firm. As regards the modelling technique, we chose a linear SVM to keep the comprehensibility that is required in credit risk predictions, while preserving predictive performance.

Figures 5 and 6 facilitate the understanding of our out-of-time ensemble set-up. In the training set, we use the labels for all companies that were active on the 1st of January 2013. These labels received the value +1 if the company went bankrupt in the course of 2013 and 0 if the company was still active at the end of 2013. Concerning the inputs to our ensemble model, the network scores of the companies are calculated using all managers that were part of the company up to January 2013 and the financial ratios are calculated using the financial statement of 2011[3] Figure 6 illustrates the set-up of our test set. To predict bankruptcy one year ahead for all firms active on 1 January 2014, we use the financial statement of 2012 and the network scores calculated using all managers up to 1 January 2014. A 'leave-one-out' procedure is applied here as well, and is further explained in Figure 7.[4] The un-

[3] As mentioned earlier, due to the delay on publication, we cannot use the statement of 2012.

[4] To decrease computational time, instead of exclud-

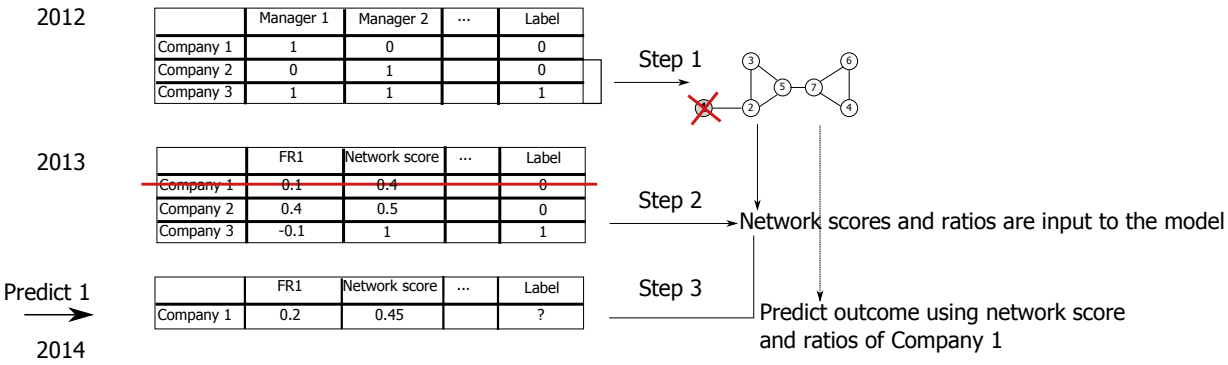

Figure 7: The modeling procedure for the ensemble model when double observations occur. We want to predict whether Company 1 will go bankrupt in the course of 2014. First, we build the network using the 'leave-one-out' procedure, i.e. we build a network using the data of 2012, excluding the information on Company 1. This procedure is used for all double occurrences, leading to multiple networks. Next, we train our ensemble model on the 2013 data set. As input features we use the financial ratios and the network scores for all companies except Company 1. The network scores are estimated using the 2012 networks and the updated list of managers in 2013. We finally predict the probability of failure for Company 1 using the ensemble model, the company's financial ratios and network score (estimated using the 2012 network that excludes Company 1 and the updated list of managers for Company 1 in 2014).

balanced distribution of the classes in the data set, necessitates an undersampling of the negative class (active firms) in the training set for both the base model and the ensemble model. We reduced the non-event rate to 50:50 in the training set. The relational learners are a powerful tool and are able to manage the issue of unbalanced classes [19]. The network scores are therefore calculated using the entire network and not a sample. The results are calculated on the complete test set of 400,203 firms.

4.2 Results

We compare the results of the network, base, base plus dummy and ensemble model using the Area under the ROC-curve characteristic [13] and the lift [8]. The results in Figure 8 show that the relational data on its own is insufficient, however with an AUC of 57.74% it still has reasonable predictive power. Adding the network scores to the base model, slightly increases the AUC from 82.45% to 83.06%. A larger increase can be seen at the beginning of the lift curve in Figure 9. The ensemble model has a 15.07 times higher bankruptcy detection rate than the average bankruptcy rate of 1.5% for the 0.1% highest scores, i.e. the top 400 companies. In this segment, the ensemble model has a 22.66% higher lift than the base model. The ensemble model's lift is comparable with the base model's at percentiles higher than 3%. This result confirms that the highest ranked companies in the ensemble model, those connected to many (or only) bankrupt firms, have indeed a higher probability of going bankrupt. However, it also shows that one should still consider their financial situation. When replacing the network score with

ing one company at a time, we exclude companies in chunks of 1000.

Table 2: Ranking of input features

	Base model	Coefficient	Ensemble model	Coefficient
Positive	Missing ROE	0.8553	Network scores	0.9084
	Missing Equity Ratio	0.3210	Missing ROE	0.8450
	Water supply/sewerage (sector E)	0.1943	Missing Equity ratio	0.3053
	Construction (sector F)	0.1583	Missing ROA	0.1369
	Missing ROA	0.1209	Water supply and sewerage (sector E)	0.1519
Negative	Company age	-1.2039	Company age	-1.1210
	Newly founded	-1.0446	Newly founded	-1.0380
	Human health/social work (sector Q)	-0.6036	Human health and social work (sector Q)	-0.6141
	Equity ratio	-0.5963	Equity ratio	-0.5826
	Real estate activities (sector L)	-0.4444	Real estate activities (sector L)	-0.4828

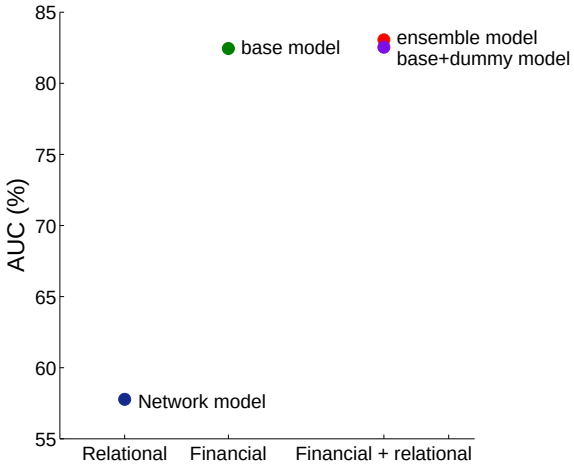

Figure 8: AUC results for the relational data, financial data and the ensemble model.

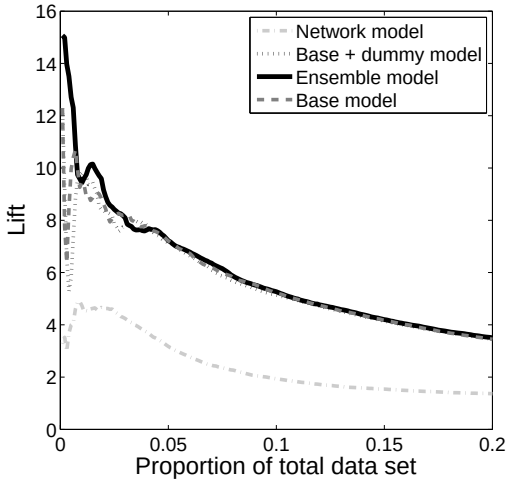

Figure 9: Lift curve of the base model, relational model (wvRN), the base plus dummy model and the ensemble model for defaults in 2014 (one-year horizon).

a dummy variable, the performance lowers to 82.54% and the lift decreases to the base model's level of 12.28. The real added value of the network scores lies in its forward-looking nature. Figure 10 shows the lift curve for all four models when predicting two-years ahead, i.e. the companies most likely to default in the next two years. For the 0.1% highest scores, the ensemble model maintains a high lift of 12.54, while the base model's lift decreases to 7.27. This means that adding network scores to the model, results in a 72% increase of the bankruptcy detection rate for the top segment.

Table 2 shows the ranking of the top five and bottom five input features. The top five input features have a positive coefficient and are pre-

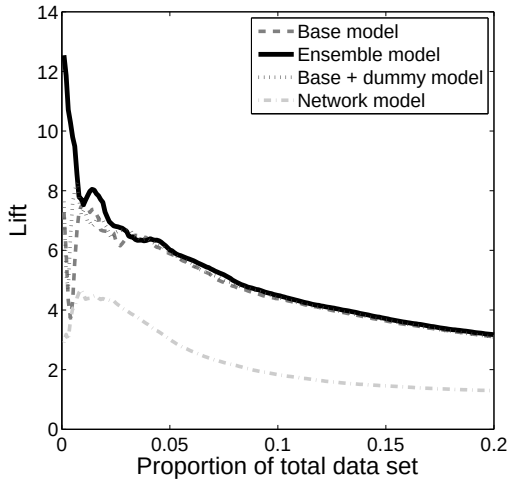

Figure 10: Lift curve of the base model, relational model (wvRN), the base plus dummy model and the ensemble model for defaults in 2014 and 2015 (a two-year horizon).

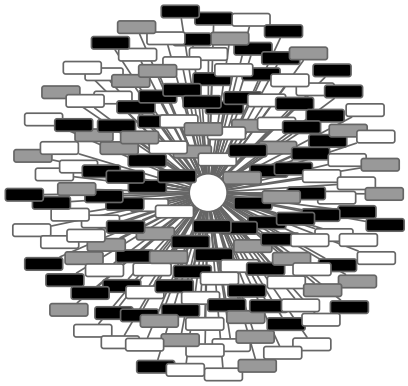

Figure 11: Example of a bipartite graph. The middle round node is the manager. Surrounding rectangular nodes are companies the manager is involved in. Grey nodes depict companies in an ongoing bankruptcy case, white nodes depict companies that are still active at the moment of observation (June 2015) and black nodes depict bankruptcies that were closed without discharge. The large amount of non-excusable bankruptcies indicates that this network is likely fraudulent.

dictive for bankruptcy. The bottom five features have a negative coefficient and are predictive for the non-event. The base and ensemble model's coefficients differ slightly in magnitude. This could indicate an interaction between the influence of a company's network and its financial situation. There are, however, no coefficients that change sign when the network score is added. For the ensemble model, the most predictive variable for bankruptcy is the network score. With a coefficient of 0.9084, the size of the network score is almost linearly transferred to the prediction score.

The application of relational data in a corporate setting is not restricted to bankruptcy prediction only. Similar data can be used for the more refined prediction of loan default, which is defined as a 90 days delay on loan repayments and is therefore not necessarily followed by bankruptcy. Another promising research topic is fraud prediction using relational data. Figure 11 shows an existing bipartite graph where the middle round node is a business manager and the surrounded nodes are the firms this manager is or was involved in. Note the large amount of grey and black nodes, indicating firms that went bankrupt. The black nodes are bankruptcies that were closed without discharge, which usually means that there are suspicions of fraud. In a fraud prediction setting similar networks can be used to detect fraudulent managers and as a consequence likely fraudulent firms.

5 Conclusion

In this paper, we investigated the potential of relational data for bankruptcy prediction. We showed that linking companies based on their managers/board members adds complementary predictive power to the traditional bankruptcy

prediction. Combining the relational data with financial data results in a higher lift in the first 5% segment, confirming the assumption that companies linked to many (or only) bankrupt firms have a higher probability of bankruptcy.

The proposed methodology can be extended to different applications such as loan default prediction and fraud detection. Certain extensions to the methodology can still be made. The network scores are based on links with all bankrupt firms, no distinction is made between a bankruptcy in 1990 and a bankruptcy in 2011. A topic of future research is the addition of a discount factor that diminishes the influence of bankrupt firms on their linked partners over time. We have also included all board members/managers that are/have been part of the firm right up to the prediction date. A discount factor for previous managers/board members could be a fruitful addition as well. A third topic for future research is the calculation of the weights given to the top nodes. In this paper, we have applied a hyperbolic tangent function to the top nodes, following the popular assumption that top nodes with many links are less discriminative. However, when calculating the weight of a particular top node in this particular corporate bankruptcy setting, it may be important to take into account the amount of other top nodes the companies are linked to. A person that is the sole manager of 50 firms will have a larger influence on the performance of these firms than a manager that is just one of the many managers in 50 firms.

6 Acknowledgements

The authors would like to thank the Flemish Research Council (FWO) for financial support (Grant G.0827.12N and the fellowship for Ellen Tobback).

References

[1] B. Ahn, S. Cho, and C. Kim. The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert systems with applications*, 18(2):65–74, 2000.

[2] E. I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

[3] R. AYADI. The new basel capital accord and sme financing. *Center for European Policy Studies, Brussels*, 2005.

[4] J. R. Baldwin. Failing concerns: business bankruptcy in canada. *Failing Concerns: Business Bankruptcy in Canada*, 1998.

[5] W. H. Beaver. Financial ratios as predictors of failure. *Journal of accounting research*, pages 71–111, 1966.

[6] T. B. Bell. Neural nets or the logit model? a comparison of each model's ability to predict commercial bank failures. *Intelligent Systems in Accounting, Finance and Management*, 6(3):249–264, 1997.

[7] J. L. Bellovary, D. E. Giacomino, and M. D. Akers. A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial education*, pages 1–42, 2007.

[8] M. J. Berry and G. S. Linoff. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2004.

[9] H. Bian and L. Mazlack. Fuzzy-rough nearest-neighbor classification approach. In *Fuzzy Information Processing Society, 2003. NAFIPS 2003. 22nd International Conference of the North American*, pages 500–505. IEEE, 2003.

[10] J. Brown, B. Netoles, S. T. Rasnak, and M. Tighe. Identifying bankruptcy fraud. Technical report, Credit Research Foundation, 1999.

[11] A. Cielen, L. Peeters, and K. Vanhoof. Bankruptcy prediction using a data envelopment analysis. *European Journal of Operational Research*, 154(2):526–532, 2004.

[12] P. Domingos. Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20(1):80–82, 2005.

[13] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[14] D. Fletcher and E. Goss. Forecasting with neural networks: an application using bankruptcy data. *Information & Management*, 24(3):159–167, 1993.

[15] E. A. Helfert. *Techniques of financial analysis: a guide to value creation*. McGraw-Hill Professional, 2002.

[16] S. Hill, F. Provost, and C. Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statist. Sci.*, 21(2):256–276, 05 2006.

[17] H. Jo, I. Han, and H. Lee. Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications*, 13(2):97–108, 1997.

[18] S. Jones and D. A. Hensher. Predicting firm financial distress: A mixed logit model. *The Accounting Review*, 79(4):1011–1038, 2004.

[19] E. Junqué de Fortuny, M. Stankova, J. Moeyersoms, B. Minnaert, F. Provost, and D. Martens. Corporate residence fraud detection. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1650–1659. ACM, 2014.

[20] K. C. Lee, I. Han, and Y. Kwon. Hybrid neural network models for bankruptcy predictions. *Decision Support Systems*, 18(1):63–72, 1996.

[21] M. Leshno and Y. Spector. Neural network prediction analysis: The bankruptcy case. *Neurocomputing*, 10(2):125–147, 1996.

[22] S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. *The Journal of Machine Learning Research*, 8:935–983, 2007.

[23] D. Martens and F. Provost. Pseudo-social network targeting from consumer transaction data. Technical Report CEDER-11-05, New York University, 2011.

[24] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[25] J. H. Min and Y.-C. Lee. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert systems with applications*, 28(4):603–614, 2005.

[26] S.-H. Min, J. Lee, and I. Han. Hybrid genetic algorithms and support vector machines for bankruptcy prediction. *Expert systems with applications*, 31(3):652–660, 2006.

[27] C. W. Nam, T. S. Kim, N. J. Park, and H. K. Lee. Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies. *Journal of Forecasting*, 27(6):493–506, 2008.

[28] N. B. of Belgium. Central credit register: total of credits granted to resident non-financial corporations, 2015. [Online; accessed 3-July-2015].

[29] J. A. Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, pages 109–131, 1980.

[30] D. L. Olson, D. Delen, and Y. Meng. Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, 52(2):464–473, 2012.

[31] B. C. on Banking Supervision. International convergence of capital measurement and capital standards: a revised framework. Technical report, Bank for International Settlements, 2006.

[32] H. Ooghe and S. De Prijcker. Failure processes and causes of company bankruptcy: a typology. *Management Decision*, 46(2):223–242, 2008.

[33] H. Ooghe and C. Van Wymeersch. *Handboek financiële analyse van de onderneming*. Intersentia nv, 2008.

[34] S. Piramuthu, H. Ragavan, and M. J. Shaw. Using feature construction to improve the performance of neural networks. *Management Science*, 44(3):416–430, 1998.

[35] F. Provost, D. Martens, and A. Murray. Finding similar mobile consumers with a privacy-friendly geo-social design. *Information Systems Research*, In Press, 2015.

[36] S. Sharma and V. Mahajan. Early warning indicators of business failure. *The Journal of Marketing*, pages 80–89, 1980.

[37] K.-S. Shin, T. S. Lee, and H.-j. Kim. An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, 28(1):127–135, 2005.

[38] M. Stankova, D. Martens, and F. Provost. Classification over bipartite graphs through projection. Technical report, University of Antwerp Working Paper, 2015.

[39] M. H. Tinoco and N. Wilson. Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis*, 30:394–419, 2013.

[40] T. Van Gestel, B. Baesens, J. Suykens, M. Espinoza, D.-E. Baestaens, J. Vanthienen, and B. De Moor. Bankruptcy prediction with least squares support vector machine classifiers. In *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on*, pages 1–8. IEEE, 2003.

[41] W. Verbeke, D. Martens, and B. Baesens. Social network analysis for customer churn prediction. *Applied Soft Computing*, 14, Part C(0):431 – 446, 2014.

[42] I. Weber, V. R. K. Garimella, and E. Borra. Inferring audience partisanship for youtube videos. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 43–44. International World Wide Web Conferences Steering Committee, 2013.

[43] R. L. Wilson and R. Sharda. Bankruptcy prediction using neural networks. *Decision support systems*, 11(5):545–557, 1994.

[44] G. Zhang, M. Y. Hu, B. E. Patuwo, and D. C. Indro. Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European journal of operational research*, 116(1):16–32, 1999.

[45] M. E. Zmijewski. Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting research*, pages 59–82, 1984.