CART-based selection of bankruptcy predictors for the logit model

Arjana Brezigar-Masten*
University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technology, Glagoljaška 8, SI-6000 Koper, Slovenia, and Institute of Macroeconomic Analysis and Development, Gregorčičeva 27, SI-1000, Slovenia

Igor Masten*
University of Ljubljana, Faculty of Economics, Kardeljeva pl. 17, SI-1000 Ljubljana, Slovenia

Abstract

Balance-sheet data offer a potentially large number of candidate predictors of corporate financial failure. In this paper we provide a novel predictor selection procedure based on non-parametric regression and classification tree method (CART) and test its performance within a standard logit model. We show that a simple logit model with dummy variables created in accordance with the nodes of estimated classification tree outperforms both standard logit model with step-wise-selected financial ratios, and CART itself. On a population of Slovenian companies our method achieves remarkable rates of precision in out-of-sample bankruptcy prediction. Our selection method thus represents an efficient way of introducing non-linear effects of predictor variables on the default probability in standard single-index models like logit. These findings are robust to choice-based sampling of estimation samples.

JEL codes: G32, G33, C14, C25
Keywords: bankruptcy prediction, model selection, CART

*Corresponding author
Email addresses: arjanabm@gmail.com (Arjana Brezigar-Masten), igor.masten@ef.uni-lj.si (Igor Masten)
1. Introduction

The problem of predicting corporate failure is at the hart of risk management procedures in banks worldwide. The turmoil in financial markets after the outbreak of the financial crisis in 2008 has only re-emphasized the importance of proper risk assessment. Characteristics of a good bankruptcy prediction model would in this respect be: reliability and robustness, ease of implementation, high degree of prediction accuracy and clear interpretability and transparency of decision making process. The aim of this paper is to propose a bankruptcy prediction methodology that fulfills all of these characteristics.

Methodological approaches used in bankruptcy prediction can be broadly classified into statistical methods and artificial intelligence methods (Min and Jeong, 2009). The first group includes discriminant analysis (used in pioneering studies of Beaver (1996) and Altman (1968)) and binary-choice models like logit or probit. A common statistical property of these methods is that they are fully parametric. The artificial intelligence group comprises of methods that range from artificial neural networks (ANN) and genetic algorithms (GE) to classification and regression trees (CART) (see Paliwal and Kumar, 2009; and Li at al., 2010). A common statistical feature of this second class of models is a fully non-parametric specification of both the distributional form of variables and functional relations among them.

This paper attempts to bridge the two model classes and proposes a method of selecting predictor variables for the classic logit model based on the non-parametric classification and regression tree method. The spirit of our approach is the following. We apply CART to a large set of possible predictors to estimate a decision tree that partitions firms into bankrupt and healthy. In accordance with decision nodes given by the tree we construct a set of dummy variables that are used as predictors in the logit models. Such CART-determined dummy variables can enter the logit model both independently or as additional explanatory variables to a set of conventionally selected variables. Our dummy variable approach is different from the approach of Cho et al. (2010) who used the (untransformed) variables that CART method reports in the estimated tree as input variables in several models, including the logit model. The use of dummy variables is both simple and efficient because it preserves the nonlinearity in the relations among variables identified by CART also in the single-index models. It is especially the potential non-linearity in the effect of candidate predictor variables on the probability of financial distress that conventional selection methods or the approach of Cho et al. (2010) cannot capture.

Our results demonstrate that our CART-based selection procedure of bankruptcy predictors is indeed a very useful method for selection of bankruptcy predictors, which combined with the standard logit model provides an accurate prediction tool. Validity and robustness of this finding is corroborated by the fact that our comparisons of prediction accuracy are performed truly out of sample. In addition, we do not look only at overall prediction accuracy but put special emphasis to prediction accuracy of bankrupt and healthy firms separately. Moreover, our empirical findings are obtained on a population of Slovenian enterprises, ranging from very small private businesses to large international corporations. The final test of robustness of our findings is against the construction of the estimation sample. Namely, choice based sampling of the observations into estimation sample that equates the number of bankrupt and healthy firms in the estimation sample is a common approach in the literature and practice. While such an approach may be motivated by

\(^1\)Corporate balance sheets and income statements, which are the main data source for such application, can be used to compute numerous financial ratios, all of them in principle being candidate predictors of financial distress.
data availability and econometric considerations, it is definitely at odds with composition of data in reality and, consequently, bankruptcy prediction in practice. For this reason we test our method not only on a matched sample, but also on a large sample with population shares of bankrupt firms, which is the situation that banks face in real life.

The remainder of the paper is organized as follows. Section 2 introduces competing bankruptcy prediction models. Section 3 outlines our selection method and other methods with compare it to. Section 4 presents our data and the construction of estimation samples. Section 5 contains the results of predictor selection, while Section 6 reports the results of estimated logit models. Comparison of bankruptcy prediction is given Section 7. Finally, Section 8 concludes.

2. Prediction models

Our basic model of the probability of bankruptcy is the logistic regression. The logit model has been extensively applied in the literature (see for example, Li et al., 2010; Chen, 2011 and Min and Jeong, 2009, among others). There are several reasons for its use. First, the logit model has been widely used and taught. Second, it is relatively easy to understand and readily available in virtually all software packages. Finally, logit has resulted to be a fairly robust and reliable tool for forecasting financial distress.

We measure the incidence of bankruptcy with a binary random variable $y$ whose realizations can be represented as

$$y = \begin{cases} 1 & \text{if } \theta' X \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

where

$$P(y = 1|x) = h(\theta' X).$$

The probability that the binary dependent variable equals one given the covariates is equal to a probability transformation of the single index $\theta'x$. In principle, both the parameters of the single index $\theta$ and the probability transformation function $h$ need to be estimated. The logit model, however, is fully parametric in assuming a known form of $h$. In particular, $h$ is a logistic cumulative distribution function

$$h(\theta' X) = \frac{e^{\theta' X}}{1 + e^{\theta' X}}$$

With this assumption the parameter vector $\theta$ can be estimated consistently and efficiently by maximizing

$$L = \sum_{i=1}^{N} [y_i \ln (P_i) + (1 - y_i) \ln (1 - P_i)]$$

Our logit models differ in the choice of the predictor matrix $X$. We consider four different approaches to choosing $X$. The first approach is the standard in applications of the logistic regression: the step-wise procedure with pre-selection of data. Other approaches involve estimating classification trees.

Among the first to apply logit to the problem of bankruptcy were Santomero and Vinso (1977) and Martin (1977) who employed it to examine failures in the US banking sector. Ohlson (1980) applied it more generally to 105 bankrupt and 2,058 non-bankrupt firms. Notable applications that followed include Zmijewski (1984), and Wilson (1992). Accuracy of classification ranged from 76% in the work of Zmijewski (1984), where he employed probit and weighted exogenous sample likelihood models to investigate firms listed on the American and New York stock exchanges from 1972 to 1978, to 96% in the study by Pantalone and Platt (1987), where the authors use logit analysis to determine the causes of banks bankruptcy in the US after the deregulation.
Logit models are not the only bankruptcy prediction models we consider. While our paper focuses on the use of classification trees in selection of bankruptcy predictors for standard parametric models like logit, it is also straightforward to use the classification tree for bankruptcy prediction. The two pioneering studies using classification and regression trees for bankruptcy prediction are those of Frydman, Altman and Kao (1985), and Marais, Patell and Wolfson (1984) who employed it to assess loan classifications. The first mentioned study compared CART to the classification precision of two discriminant models. Overall the classification-tree models were found to perform best. On the other hand, Marais et al. (1984) compared their recursive partitioning results against those of a multinomial probit model. Interestingly, they concluded that in estimating loan classifications there was very little to choose between the two procedures. More recent applications of decision tree models in bankruptcy or corporate financial distress prediction in general include Chen (2011), Li and Wu (2010) and Min and Jeong (2009).

3. Variable selection

Selection of predictor variables is an important step in all bankruptcy prediction studies. To date no unified theory has been generally accepted. Most of the previous studies used a brute empirical approach of initial choice of variables (based also on expert knowledge) followed by a step-wise procedure to select the variables in the final logit or discriminant model. Such a procedure is not statistically rigorous. Different sequencing and/or initial ordering of variables need not result in a unique selection. As an attempt to overcome this deficiency some authors started using data mining techniques (Shirata, 1998; Cho et al., 2010). These are also better suited to capture potential nonlinearities in the relations between financial distress and predictor variables.

We propose a novel approach to using CART in selection of bankruptcy predictors and their subsequent use in prediction models. This approach is compared to more conventional methods of variable selection for the logit model. Both approaches are described in detail in the next two subsections.

3.1. CART selection approach

CART builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). Our reasons for selecting CART from the family of decision tree methods are similar to those of Li et al. (2010). From a practical point of view, one important advantage of decision trees in bankruptcy prediction is the ability to generate easily understandable decision rules. This feature is not shared by many artificial intelligence approaches.

The classic CART algorithm was popularized by Breiman et al. (1984) (see also Ripley, 1996). CART model is a flexible method for specifying the conditional distribution of a variable \( y \), given a vector of predictor values \( X \). Such models use a binary tree to recursively partition the predictor space into subsets where the distribution of \( y \) is successively more homogeneous. The terminal nodes of the tree correspond to the distinct regions of the partition, and the partition is determined by splitting rules associated with each of the internal nodes. By moving from the root node through to the terminal node of the tree, each observation is then assigned to a unique terminal node where the conditional distribution of \( y \) is determined. CART is nonparametric and can detect complex relationships between dependent variable and explanatory variables. Therefore CART is particularly suited for discovering nonlinear structures and variables interactions in datasets with a large number of potential explanatory variables.
In CART-based selection one needs to avoid over-fitting because it may lead to bad out-of-sample prediction accuracy. Namely, some of the lower branches in a tree may be strongly affected by outliers and other artifacts of the current data set. For this reason it is preferred to find a simpler tree. In out application the tree pruned to the best size was obtained with the process of cross validation (see Breiman et al., (1998) for details).

Our classification trees are estimated on two different sample (see details below), in each case starting from a dataset containing 64 different financial ratios. The results can be used for the purposes of predictor selection in two ways. Firstly, in the application of Cho et al. (2010) the variables identified by CART enter as they are into bankruptcy prediction models. We denote it the CHH approach hereafter.

Note, however, that the CHH strategy does not exploit the crucial information estimated classification trees give us: nonlinearity in the relation between predictors and probability of bankruptcy. In order to capture this information we propose a different use of CART results by constructing a set of dummy variables. We used the branches identified in CART analysis to create dummy variables that take value one if the values of the variable that defines a branch fall into the region above the CART threshold and zero otherwise. Because the classification tree effectively partitions the whole sample into two groups the full set of dummy variables is linearly dependent, which implies that the number of dummy variables created this way equals the number of branches.

Dummy variables are included to the logit model in two ways. First, we estimate the logit model that contains only the CART-determined dummy variables (our third approach denoted BM1). Second (our fourth approach denoted BM2), we include the CART-determined dummies in addition to the set of predictors selected by the conventional selection procedure described below.

The motivation for our procedure is quite straightforward. CART is by definition better suited to identify potential nonlinearities in the determinants of financial distress of firms. Including dummy variables that correspond to such nonlinearities may in this respect be a useful way to augment standard models by non-linearities identified with the algorithm that is well suited to identify them. Inclusion of dummies only gives allows us to evaluate their efficiency in bankruptcy prediction as such. Included in combination with conventionally selected predictor allows us to asses the relative merits of the two competing approached.

3.2. Conventional selection approach

We start by constructing 64 different financial ratios as potential predictors. In the first step, bivariate logistic regressions were run for each of the 64 ratios on ten randomly created matched subsamples of our dataset.\(^3\) Financial ratios that on average classified correctly at least 60 percent of bankrupt firms and 60 percent of non-bankrupt firms, were kept for further stages. This left us with a group of 27 financial ratios, 14 measuring profitability, 9 solvency and 4 liquidity of firms. Ratios that classify neither bankrupt nor non-bankrupt firms at 60 per cent accuracy are nine in number. Seven describe firm activity and two are profitability measures. The remaining 28 ratios classified at required precision either bankrupt or non-bankrupt firms, but not both. As such they were not considered in subsequent steps of variable selection.

The ratios that passed the first step were in the second step grouped into seven groups of highly correlated indicators, using 0.5 as the correlation threshold. From each of the groups we extracted one principal component. As a representative of each group we then took the variable with the

\(^3\)Matched sample have balanced shares of bankrupt and healthy observations.
largest loading to the principal component. We prefer to proceed in this way to using the principal component in prediction models in order to avoid the efficiency loss problem associated with generated regressors, and because principal components can be hardly given any direct economic interpretation.\(^4\)

In the last step logistic step-wise procedure was used to select the final variables. It starts by estimating parameters for variables forced into the model. Next, the procedure computes the adjusted chi-squared statistic for all the variables not in the model and examines the largest of these statistics. If it is significant at conventional levels, the variable enters into the model. One or more elimination steps follow each selection step, i.e. the variables already selected into the model do not necessarily stay. The step-wise selection process terminates if no further variable can be added to the model, or if the variable just entered into the model is the only variable removed in the subsequent elimination.

This variable selection approach is rather complex and time consuming relative to the CART-based approach. We deliberately put significant effort into it in order to get the toughest competitor to our proposed alternative approach. We experimented with alternative selection routes, but this one resulted in the highest in-sample fit and out-of-sample prediction precision. In the section that presents results we shall see that these are indeed very high.

4. Data and sample design

The data come from two databases of Slovenian companies. The first are data of annual financial statements of all Slovenian firms for the period 1995-2001 provided by Agency for public legal records and related services (AJPES).\(^5\) From the initial database we eliminated all observations with missing data. This resulted in 39005 observations on healthy firms in the sample. The second is the database of 592 bankruptcy cases for the same time period collected by I d.o.o. (Slovenian franchise of Dun&Bradstreet). These are data on all(!) bankruptcy cases filed with Slovenian legal authorities. Industries in the sample mainly cover the manufacturing sector with their size ranging from very small to large.

From the balance-sheet and income statement data we calculated 64 financial ratios as candidate predictors.\(^6\) Financial ratios can be broadly classified into four categories: liquidity, profitability, solvency and activity. Dependent variable is a binary variable that takes on value one if the firm operates in time \(t\), and zero if the firm filed for bankruptcy in time \(t\). All independent variables are dated \(t - 1\).

Common characteristic of all bankruptcy prediction cases is a very small population share of bankrupt firms. This implies that only a limited share of the overall probability mass is accounted for by bankruptcy cases, which may cause the estimated prediction models to describe well the characteristics of healthy firms, but have only limited prediction power for bankruptcy cases. For this reason many empirical applications balance the estimation sample by balancing the shares of healthy and bankrupt firms. Our first sample design is constructed accordingly. From the initial sample we created a sub-sample with 592 bankrupt firms and corresponding 592 non-bankrupt mates. Matching is based on the following characteristics: size (measured by total asset), industry

\(^4\)Chen (2011) reports prediction accuracy loss from using PCA.

\(^5\)Post-2001 data was not considered because of a change in accounting standards and because we did not dispose with the accurate record of bankruptcy filings after 2001.

\(^6\)Financial ratios, by their nature, have the effect of deflating statistics by size, implying that a their potential predictive power is not contaminated by firm size (Altman, 2000).
and year of bankruptcy. The last matching criterion ensures that financial statements of matched pairs are always of the same vintage. Because matching is primarily used to obtain a balanced sample of bankrupt and healthy firms the samples mainly consist of small and medium-sized companies, since the incidence of bankruptcy in the large-asset-size firm was quite rare.

We have to note, however, that matched samples are choice based, i.e. the probability of an observation entering the sample depends on the value of dependent variable, which violates the random sampling assumption. Choice-based sampling in general causes both parameter and probability estimates to be biased. This holds both for parametric models (see Zmijewski, 1984) and artificial intelligence models.

Banks are in real life faced with loan applicants that are drawn from a population, which may imply erroneous credit risk assessment if models are estimated on matched samples. For this reason our second sample design takes the data as they are, i.e. with considerably larger share of healthy firms in the sample. Very similar to the majority of studies, the share of bankruptcy cases in our sample is rather small, roughly 1.5%.

Evaluation of prediction precision is conducted in a pseudo out-of-sample context. To this end in both samples 75 percent of observations were allocated to the estimation sample and 25 percent to the evaluation sample, i.e. the subsample on which out-of-sample prediction accuracy was tested.

5. Results of variable selection

In this section we present the results of variable selection. The estimated classification tree on the matched sample is presented in Figure 1. Given the sample size it resulted optimal to estimate a simple tree with nodes determined by thresholds on two variables. The first node is determined by variable $ds$ that measures ?????, while the second is based on $kd$ that measures ??????. This implies that our logit model estimated on the matched sample will include $kd$ and $ds$ in the CHH approach, and two dummy variables in the approach that we propose.

On the sample with population share of bankrupt firms we estimated two trees with different degrees of pruning, namely three and four levels. A smaller tree is presented in Figure 2. While its top node is based on the same variable as on the matched sample, further branches are different. The first split is based on $dpro$ ?????, while at the third level a finer division between bankrupt and healthy firms is based on on a threshold at highly negative scores of return on equity ($roe$). In accordance with the tree’s nodes we created three dummy variables.

The final tree is a larger tree estimated on the full sample (see Figure 3). Compared to the previous tree it contains a refinement of the left branch based on variable $dprith$ ?????. Based on the same variable a threshold is found also on the right branch of the tree, which is further refined by using the information on $kf$ ??????. In the middle part of the tree after threshold based on the return on equity, there are two further branches based on $casher$, measuring liquidity ???, and $r6$, that measures ??????. Based on this tree we constructed eight linearly independent dummy variables.

The conventional variables selection approach based on statistical testing within the logit model resulted in a final set of four financial ratios. Two of the ratios measure liquidity $cf2d$ and $tfs$???, one solvency, $kol$, and one profitability, $pppo$ ????.

6. Estimated models

The estimates of the logit models are presented in Tables 1 and 2. Both tables are split into two panels, the left presents estimates on the matched sample while the right on the full sample. In
Figure 1: Estimated tree on the matched sample (N=888)

Figure 2: Estimated smaller tree on the full sample (N=29698)

Figure 3: Estimated larger tree on the full sample (N=29698)
In the logit models with CHH selection of predictors (see Table 2) the results show less convincing results about the significance of variables in the model. Both variables selected on the matched sample are significant in the logit model estimated on both samples, for other variables included on the full sample the results are less clear cut, with \( dpro \) absolutely insignificant.

Comparison of in-sample classification accuracy of the models is given in Table 3. Results will not be discussed in detail because all the main implications are essentially the same as in the case of out-of-sample prediction accuracy, which is the central object of our analysis.
7. Bankruptcy prediction

Comparison of models’ prediction accuracy is reported in Table 4. As explained in Section 4 both samples of data were divided so that 75 per cent of observations were used for estimation and 25 per cent for testing out-of-sample prediction accuracy.

The first observation that applies to all results is that compared to similar results in the literature we are generally able to achieve a high level of prediction accuracy. The lowest rates of classification rates obtained across all models on matched sample are 75.5 % for bankrupt firms, 62.8 % for healthy firms and 78.3 % overall. The latter two figures are obtained with CART. For the full sample the corresponding figure are 18.2 % (logit with CART variables), 95 % (CART) and 94.7 % (CART). Logit with conventional selection of predictors, which in no case performs the worst, therefore represents a tough competition for alternative bankruptcy prediction approach. Improving over its results can thus be taken as a signal of good forecasting performance.

Classifications trees result to be best suited for predicting bankruptcy cases. On matched sample the ratio of correct prediction is as high as 93.7 %. Matched samples, however, are not representative of population and can therefore lead to erroneous inference. But also on the large sample precision remains quite high, 71.8 % and 72.7 % respectively for the two trees of different sizes. The decrease in prediction accuracy of bankruptcy cases when moving from the matched to full sample is moreover the lowest for CART. All other methods exhibit larger reductions, which implies that CART methodology, with non-linearities it is suited to discover, is really useful for bankruptcy prediction.

CART, however, results to be the worst performing method for out-of-sample classification of good risks, which leads to the lowest predictive precision overall. Logit models that rely on
a parametric specification of conditional probability of bankruptcy perform significantly better, suggesting that a combination of the two methods - parametric specification of the probability function and variable selection that embeds non-linearities identified by CART - might provide a useful general tool for classification of loan applicants.

In evaluation of usefulness of CART as selector of predictors let’s first compare our approach (labeled BM1 and BM2) to the CHH approach. On the matched sample the CHH approach predicts healthy firms better, but not bankrupt firms. This should not be surprising given that our dummy-variable approach preserves the non-linearity identified by CART. On the full sample the difference becomes stark. While there is virtually no difference in prediction accuracy of healthy firms (or our approach even marginally outperforms), CHH approach yields considerably worse results for prediction accuracy of bankruptcies.

Final set of comments applies to the comparison of our proposed CART-based dummy variable method to conventional selection of predictors for logit. It is only on the non-representative matched sample that conventional logit outperforms in prediction of bankruptcies (not for identification of healthy firms). But note that our logit model in this case uses just two simple dummy variables as predictors. There is no financial ratio in the model. Already a combination of four financial ratios and two dummies yields the best performing model on matched sample overall.
On the full sample results are straightforward. Conventional logit predicts less accurately in every aspect. But more importantly, after including CART-based dummies into the model, there remain virtually no additional predictive power in conventionally selected predictors. Adding the four ratios to the dummy-variable model from the large tree adds absolutely nothing to prediction accuracy. In addition, the best performing model in absolute terms, BM1 model with large tree (logit with 8 dummies created from the large tree) on full sample exceeds 50% success in identifying bankruptcies and a 99% classification accuracy overall.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample</th>
<th>Matched</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit Healthy</td>
<td></td>
<td>84.8</td>
<td>99.9</td>
</tr>
<tr>
<td>Step-wise (S-W) Bankrupt</td>
<td></td>
<td>79.7</td>
<td>20.9</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>82.8</td>
<td>98.7</td>
</tr>
<tr>
<td>BM1 Healthy</td>
<td></td>
<td>91.0</td>
<td>99.8</td>
</tr>
<tr>
<td>Bankrupt</td>
<td></td>
<td>75.7</td>
<td>39.9</td>
</tr>
<tr>
<td>Overall</td>
<td>small tree*</td>
<td>83.3</td>
<td>98.9</td>
</tr>
<tr>
<td>BM2 Healthy</td>
<td></td>
<td>91.7</td>
<td>99.8</td>
</tr>
<tr>
<td>Bankrupt</td>
<td></td>
<td>82.4</td>
<td>41.9</td>
</tr>
<tr>
<td>Overall</td>
<td>small tree</td>
<td>87.0</td>
<td>98.9</td>
</tr>
<tr>
<td>BM1 Healthy</td>
<td></td>
<td>91.0</td>
<td>99.9</td>
</tr>
<tr>
<td>Bankrupt</td>
<td>logit</td>
<td>75.7</td>
<td>39.9</td>
</tr>
<tr>
<td>Overall</td>
<td>large tree**</td>
<td>83.3</td>
<td>98.9</td>
</tr>
<tr>
<td>BM2 Healthy</td>
<td></td>
<td>91.7</td>
<td>99.8</td>
</tr>
<tr>
<td>Bankrupt</td>
<td>logit</td>
<td>82.4</td>
<td>41.9</td>
</tr>
<tr>
<td>Overall</td>
<td>large tree*</td>
<td>87.0</td>
<td>98.9</td>
</tr>
<tr>
<td>CHH Healthy</td>
<td></td>
<td>94.5</td>
<td>99.9</td>
</tr>
<tr>
<td>Bankrupt</td>
<td>logit</td>
<td>73.0</td>
<td>21.6</td>
</tr>
<tr>
<td>Overall</td>
<td>small tree</td>
<td>83.6</td>
<td>98.7</td>
</tr>
<tr>
<td>CHH Healthy</td>
<td></td>
<td>94.5</td>
<td>99.9</td>
</tr>
<tr>
<td>Bankrupt</td>
<td>logit</td>
<td>73.0</td>
<td>21.6</td>
</tr>
<tr>
<td>Overall</td>
<td>large tree</td>
<td>83.6</td>
<td>98.7</td>
</tr>
<tr>
<td>CART Healthy</td>
<td></td>
<td>66.2</td>
<td>95.0</td>
</tr>
<tr>
<td>Bankrupt</td>
<td>small tree</td>
<td>89.9</td>
<td>68.9</td>
</tr>
<tr>
<td>Overall</td>
<td>CART tree</td>
<td>78.2</td>
<td>94.6</td>
</tr>
<tr>
<td>CART Healthy</td>
<td></td>
<td>66.2</td>
<td>95.0</td>
</tr>
<tr>
<td>Bankrupt</td>
<td>large tree</td>
<td>68.9</td>
<td>94.6</td>
</tr>
<tr>
<td>Overall</td>
<td>CART tree</td>
<td>68.9</td>
<td>94.6</td>
</tr>
</tbody>
</table>

*2 dummies on matched sample and 3 dummies on full sample.
** 8 dummies on full sample.

8. Conclusion

In the paper we propose a novel approach to selection of bankruptcy predictors for the logit model based on classification and regression tree method. Our selection approach is simple. It involves constructing dummy variables according to the estimated branches of the classification tree. Dummy variables in this respect map the nonlinearities identified by CART into a modeling framework that is otherwise ill suited to capture them from the data.

Our empirical analysis demonstrates that our method provides a good bridge between the non-parametric CART and fully parametric logit by yielding the highest overall score in prediction accuracy. It is worth emphasizing that we obtain these results on an exhaustive sample of Slovenian
firms and in a simulated out-of-sample context. In addition, prediction accuracy of our models is in general very high compared to several other applications the literature.

We see our method as a valuable practical tool for credit risk assessment for several reasons. First, selection of bankruptcy predictors rests on easily interpretable decision rules, which is a feature not shared by many artificial intelligence methods. The outcomes of the selection process can for this reason be easily compared against financial theory and expert knowledge. Moreover, the estimated prediction model is the logistic regression with high degree of familiarity among practitioners in banks. Finally, the method leads to very high prediction accuracy on the sample with population share of bankrupt firms. As such it represents a viable risk assessment tool for a realistic context in which banks operate.
References


