Expert Systems With Applications 88 (2017) 368–375

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Tax payment default prediction using genetic algorithm-based variable selection

Henrik Höglund

Hanken School of Economics, Biblioteksgatan 16, 65101 Vasa, Finland

ARTICLE INFO

Article history:
Received 23 April 2017
Revised 15 July 2017
Accepted 15 July 2017
Available online 17 July 2017

Keywords:
Tax default
Discriminant analysis
Genetic algorithms
Variable selection

ABSTRACT

According to the statistics from the Finnish tax authorities, about 12% of all active firms in Finland had unpaid taxes at the end of year 2015. In monetary terms, this translates to over 3 billion euros in unpaid taxes. This is a highly significant amount as the total amount of taxes collected during 2015 was 49 billion euros. Considering the economic significance of the unpaid taxes, relatively little research has been done on identifying tax defaulting firms. The objective of this study is to develop a genetic algorithm-based decision support tool for predicting tax payment defaults. More closely, a genetic algorithm is used for determining an optimal or near optimal subset of variables for a linear discriminant analysis (LDA) model that classifies the examined firms as either defaulting or non-defaulting. The tool also provides information about the importance of various variables in predicting a tax default. The dataset consists of Finnish limited liability firms that have defaulted on employer contribution taxes or on value added taxes and the total number of available variables is 72. The results show that variables measuring solvency, liquidity and payment period of trade payables are important variables in predicting tax defaults. The best performing model comprises three non-linearly transformed variables and has a predictive accuracy of 73.8%.

1. Introduction

According to the statistics from the Finnish tax authorities, about 12% of all active firms in Finland had unpaid taxes at the end of year 2015. In monetary terms, this translates to over 3 billion euros in unpaid taxes. This is a highly significant amount as the total amount of taxes collected during 2015 was 49 billion euros. Furthermore, it is estimated that only 20% of these unpaid taxes will be recovered. Considering the economic significance of the unpaid taxes, relatively little research has been done on identifying tax defaulting firms. Both tax authorities as well as other stakeholders would undoubtedly benefit from being able to predict tax defaults. For example, tax authorities could use such tools when selecting firms for tax audits or creditors could make risk assessments based on the likelihood of tax defaults. Although the development of models for tax default prediction has been given relatively little attention, prediction of financial distress in general has been a popular research topic for decades. In a Finnish setting, tax debts are directly collectible and information about unpaid tax debts is publicly available. It can, therefore, be argued that defaulting on taxes is a strong indication of financial distress. Thus, modelling issues in financial distress prediction can, to a large extent, be applied to tax default prediction. A large number of different variables have been used in various financial distress prediction models but there is no consensus on which variables are best suited for the task (Balcaen & Ooghe, 2006). The subset of variables used in the models can be selected both on theoretical considerations and based on empirical results. The drawback with selecting variables based on theory is the limited theoretical framework (Dimitras, Zanakis, & Zopounidis, 1996; Lensberg, Eilifsen, & McKee, 2006) whereas variable selection based on empirical results might suffer from shortcomings related to various statistical issues, such as multicollinearity (Gilbert, Menon, & Schwartz, 1990). In line with this, Du Jardin (2010) showed that there is a significant improvement in failure prediction models when they are designed using appropriate variable selection techniques instead of relying on common methods from the financial literature.

The best way for determining the optimal subset of variables for predictive models is to perform an exhaustive search of different variable combinations. This is, however, often not feasible as the number of subsets grows exponentially with the number of available variables. Genetic algorithms (GA) (Holland, 1975) are an efficient method for solving various complex optimization problems and they have frequently been used for feature or variable selection in the context of determining the financial health of companies (e.g. Back, Laitinen, & Sere, 1996; Brabazon

E-mail address: henrik.hoglund@hanken.fi

http://dx.doi.org/10.1016/j.eswa.2017.07.027
0957-4174/

& Keenan, 2004; Gordini, 2014; Oreski & Oreski, 2014; Ravisankar, Ravi, & Bose, 2010). The objective of this study is to develop a GA-based decision support tool for predicting tax payment defaults. More closely, a GA is used for determining an optimal or near optimal subset of variables for a linear discriminant analysis (LDA) model that classifies the examined firms as either defaulting or non-defaulting. Although a LDA model has its drawbacks, previous research show that more sophisticated models typically yield rather small marginal improvements (Balcaen & Ooghe, 2006; Hand, 2004). Furthermore, a LDA model is relatively easy to use and interpret. Therefore, this study focuses on a LDA model instead of more advanced approaches. In addition to the LDA model, the GA also provides a frequency of occurrence list of all the variables used for developing the model.

The remainder of this study is organized as follows. Related literature on tax defaults, financial distress prediction with financial statement data, genetic algorithms and variable selection is covered in Section 2. The dataset description, the research design and the strategy of analysis is presented in Section 3 and the results from the study in Section 4. Section 5 concludes the study.

2. Related literature

2.1. Tax defaults and the Finnish setting

Limited liability firms in Finland pay taxes on their taxable income with a corporate tax rate of 20%. The corporate taxes are paid based on annual tax reports which are based on the annual financial statements. In addition to the annual tax reports, firms are also required to file semi-annual tax reports for value added taxes, employer contribution taxes and a number of other, less common, taxes. If tax debts are not paid by the due date, the tax debt recovery process is initiated. The unpaid tax debt is subject to a late-payment interest and within three weeks it is sent to the enforcement authorities. As the tax debt is enforceable without a court decision, the recovery may begin immediately. Furthermore, information about unpaid taxes becomes publicly available in the tax debt register if the amount exceeds 10 000 €. The tax debt register does, however, not contain information about the amount of the tax debt. Unpaid value added taxes and employer contribution taxes together with their amounts are also published in the official journal of Finland.

There are several studies that have examined various aspects of tax non-compliance. These studies comprise both tax reductions by legal means (see Hanlon (2010) for an extensive review) as well as tax fraud (Lennox, Lisowsky, & Pittman, 2013). However, only a small number of studies have examined tax defaults and how to predict them. A tax default differs from general tax non-compliance in that tax defaults are rarely planned events. In one of the few studies dealing with tax defaults, Marghescu, Kallio, and Back (2010) analyzed to what extent financial statement ratios can be used for predicting tax defaults in a Finnish setting. Using a binomial logistic regression model with four variables, they showed a rather low classification accuracy of 61.6%.

2.2. Predicting financial distress using financial statement data

There are several reasons why firms experience financial distress and eventually even fail, but in general the problem lies in that the sales are too low or that the costs are too high which results in a poor profitability. The poor profitability leads to insufficient cash flows and eventually to a weak liquidity. To be able to meet its obligations, the firm is forced to resort to external debt financing which in turn weakens its solvency. If the firm is unable to improve its profitability by increasing the sales or by altering the cost structure, it will ultimately fail. Predicting financial distress and failure with financial statement data have been popular topics in accounting research since the studies by Beaver (1966) and Altman (1968).

When using financial statement data in predicting financial distress, the assumption is that the distress process is characterized by deteriorating values of financial statement based variables (Laitinen, 1991). The selection of suitable variables for the models is usually based on either empirical findings or on theory (Balcaen & Ooghe, 2006). Although there are large numbers of models for predicting financial distress, the variables used in them are quite similar. Dimitras et al. (1996) examined 59 models in 47 papers and found that the most commonly occurring variable was working capital divided by total assets. Other commonly occurring variables that were identified were total debt divided by total assets, current assets divided by current liabilities and EBIT divided by total assets. These commonly used variables coincide well with the general failure process described earlier. That is, they measure profitability, liquidity and solvency, which may all be indicators of financial distress when deteriorating.

A commonly used modelling method when predicting financial distress is LDA. There LDA-based models do, however, have shortcomings when used together with financial statement data. One of the major problems is that the LDA has an assumption of multivariate normal distribution of variables, whereas studies have shown that financial statement ratios are not normally distributed (Deakin, 1976; Ezzamel, Mar-Molinero, & Beech, 1987). This has led to several more sophisticated modelling methods, such as neural networks (Wilson & Sharda, 1994), self-organizing maps (Du Jardin & Séverin, 2011) and partial least square discriminant analysis (Serrano-Cinca & Gutiérrez-Nieto, 2013), being employed for predicting financial distress. Hand (2004) does, however, argue that the LDA-based models are not obsolete as they can achieve over 90% of the predictive accuracy of more complex models and as they are also less likely suffer from problems with overfitting of data.

2.3. Operating principles of genetic algorithms

Genetic algorithms, first presented by Holland (1975), are an optimization technique based on models of natural selection and evolution. The starting point when dealing with genetic algorithms is the initial population. The population consists of a number of chromosomes (individuals) that each represents a solution to the problem. Each chromosome comprises a number of genes that are typically coded as binary numbers. The optimal size of the population is dependent on the complexity of the problem to be solved. Generally, a population too small might lead to poor solutions, whereas a population too large will waste unnecessary computational resources (Lobo & Lima, 2005). Once the size of the initial population (first generation) has been determined, the chromosomes are typically randomly generated. The next step is to evaluate the fitness function for each chromosome in the first generation. Based on their fitness values, chromosomes are selected to create the next generation through breeding. There are several methods for selecting the parent chromosomes for the breeding process, but one of the most commonly used is the roulette wheel selection (Man, Tang, & Kwong, 1996). With the roulette wheel selection, a proportion of the wheel is assigned to each chromosome based on their fitness value and the higher the fitness value, the higher the probability of being selected. When two chromosomes have been selected, they are combined with a cross-over mechanism to form two new chromosomes. A typical cross-over mechanism is to select a random point in the chromosome after which the genes to the right of that point are swapped between the parent chromosomes. The selection and cross-over are then repeated until a new generation has been created. The cross-over probability