Tax payment default prediction using genetic algorithm-based variable selection

Henrik Höglund[*]

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

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%.

Key words: tax default, discriminant analysis, genetic algorithms, variable selection

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

[*] Corresponding author. Tel.: +358 (0)40 3521768
E-mail address: henrik.hoglund@hanken.fi

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 et al., 1996; Lensberg et al., 2006) whereas variable selection based on empirical results might suffer from shortcomings related to various statistical issues, such as multicollinearity (Gilbert et al., 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 et al.,1996; Brabazon & Keenan, 2004; Ravisankar et al., 2010; Oreski & Oreski, 2014; Gordini, 2014). 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 that more sophisticated models typically yield rather small marginal improvements (Hand, 2004; Balcaen & Ooghe, 2006). 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 et al., 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 et al. (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 et al.,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 (Mann et al., 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 is a central genetic algorithm parameter associated with the cross-over mechanism. It defines the ratio of chromosome pairs that will be subjected to cross-over and chromosome pairs that will be copied into the next generation. Another central genetic algorithm parameter is the mutation rate. That is, the rate at which each new chromosome is subjected to random mutation. The purpose of the random mutation is to prevent the genetic algorithm from converging prematurely. The cross-over probability usually ranges between 0.6 and 1.0 whereas typical values for the mutation rate are below 0.1 (Mann et al., 1996). When the new generation has been formed, the fitness function for the chromosomes is evaluated and the selection process starts all over. The evolutionary process then continues until a satisfactory solution to the problem is found or some predetermined condition is met.

2.4 Variable selection and genetic algorithms

The ability to generalize to new data is a crucial measure when developing predictive models. One of the major challenges in the model development process is the selection of the optimal set of explanatory variables from an often large number of available variables. Including too many variables in a model might lead to problems such as overfitting, collinearity, inclusion of irrelevant variables and an increased risk of missing data. These problems will often reduce the predictive ability of the models. The best method for selecting the optimal set of variables is to perform an exhaustive search. This is, however, not always feasible as the number of subsets is growing exponentially with the number of possible variables (Reunanen, 2003).

Several methods for selecting an optimal subset of variables have been suggested in prior research. Both Guyon & Elisseeff (2003) and Saeys et al. (2007) provide extensive overviews of various issues regarding variable selection. In general, the methods for variable selection can be categorized as filter, wrapper or embedded methods (Guyon & Elisseeff,2003). The filter methods selects variables regardless of the model whereas wrapper and embedded methods evaluate the subsets of variables using the model. GA-based variable selection belongs to the wrapper category and it has been used in many previous studies. In GA-based variable selection the chromosomes comprise bit strings and the length of the bit strings equal the total number of available variables. If a specific bit is set to 1 it signifies that the corresponding variable is used when evaluating the fitness function. After the initial population has been created, the GA is run until it is stopped based on predetermined conditions. Once the GA has been stopped, the variables which are selected in the chromosome with the highest value for the fitness function are used in the final model.

According to Leardi et al. (1992) variable subsets selected by genetic algorithms are more efficient than subsets obtained by more traditional methods as they generally produce better results and comprise fewer variables. Paterlini & Minerva (2010) compared forward and backward stepwise regression with genetic algorithm approaches for selecting both variables and variable transformations. The results in their study showed that the genetic algorithm approaches outperformed the stepwise regression models. Although not frequently, GAs have been used in previous studies for selecting optimal subsets of financial statement-based variables for various types of models. Back et al. (1996) used a GA for selecting a subset of variables for bankruptcy prediction among 31 available financial statement-based variables. The variables were selected for various time periods and underlying models and the resulting overall classification accuracy was high. They did not, however, employ alternative variable selection methods for comparison. Brabazon & Keenan (2004) compared the performance of LDA- and neural network-based bankruptcy prediction models. The variables for the LDA-based models were selected using a forward stepwise method whereas the variables for the neural network-based models were selected using a genetic algorithm approach. The models were compared using data up to three years prior to bankruptcy, and during the entire time period the neural network-based models outperformed the LDA-based models. Ravisankar et al. (2010) used a similar approach for selecting optimal variables for predicting failures of dotcom firms. Their results showed that using genetic algorithms for both variable selection and classification outperformed a number of other techniques.

3 Research methodology

3.1 Research task

The purpose of this study is to develop a decision support tool for predicting tax payment defaults. In the first step of the model development, a GA is used for finding an optimal set of variables for a LDA model. The GA is then run multiple times after which the most commonly occurring variables are identified. In the second step of the model development, these commonly occurring variables are used to build a LDA model for predicting tax defaults. The remainder of this main section first describes how the tax default and financial statement data is collected and organized, after which the GA-based variable selection and the strategy of analysis is described.

3.2 Data collection method

The dataset consists of Finnish limited liability firms that have defaulted on employer contribution taxes or on value added taxes in 2014. During the year examined, a total of 3036 firms defaulted on these taxes. Out of these 3036 firms, 1118 firms had submitted financial statements for the years 2012 and 2013 at the time when the financial statement database was compiled. Another 545 firms were dropped due to missing data for some or all of the 36 variables. Finally, 161 firms were left out of the dataset as they had sales or total assets below 10 k€ during the two years prior to the tax default. Thus, the total number of defaulting firms in the dataset is 384. The defaulting firms were matched with non-defaulting firms on year, industry and pre-default year total assets. Combining the defaulting and non-defaulting firms, the final dataset comprises 768 firms. The information regarding the tax defaults has been retrieved from the official journal of Finland and the source of the financial statement data is the Voitto+ database[1] (the data file is available for download at https://goo.gl/52cK41).

For each observation, there is data for two years prior to defaulting for 17 financial statement based ratios (see Table 1 for variable description). In addition, there is also data about industry payment default and industry bankruptcy risk one year prior to defaulting, making a total of 36 variables. Some previous studies have indicated that financial statement ratios are not normally distributed (Deakin, 1976; Ezzamel et al., 1987). This might reduce the accuracy of a LDA-based tax default prediction model as the underlying assumption for LDA is that the data is normally distributed. Therefore, a second dataset with the original 36 variables as well as with signed square root transformations of these variables is created. There are several alternative non-linear transformations and the reason why the square root transformation was selected in this study is that Deakin (1976) showed that it normalizes the distribution of several financial statement ratios.

[1] A financial statement database compiled by Suomen Asiakastieto Oy (a major Finnish credit rating and financial information company).

Table 1. Variable description

Variables	Period			
	t	t-1	±√t	±√t-1
Industry risk of payment defaults	x		x	
Industry risk of bankruptcy	x		x	
Sales / total assets	x	x	x	x
Total assets	x	x	x	x
Change in sales	x	x	x	x
Gross result / sales	x	x	x	x
Operating margin	x	x	x	x
Operating income	x	x	x	x
Quick ratio	x	x	x	x
Current ratio	x	x	x	x
Return on investment	x	x	x	x
Return on assets	x	x	x	x
Equity ratio	x	x	x	x
Net gearing	x	x	x	x
Debt / sales	x	x	x	x
Working capital /sales	x	x	x	x
Inventory turnover ratio	x	x	x	x
Collection period of trade receivables	x	x	x	x
Payment period of trade payables	x	x	x	x

Period t = 2014, period t-1 = 2013

3.3 Genetic algorithm-based variable selection

In the first step of the model development process a GA is employed for selecting variables for predicting tax payment defaults with a LDA model. The chromosomes in the genetic algorithm are represented by bit strings where each bit represents a predictive variable. If a bit is set to 1, the corresponding variable will be included in the LDA model. If the probability for each bit to be set to 1 would be 0.5, the average number of variables selected would be about the half of the total number of variables. However, according to Leardi et al. (1992) a good value for the average number of selected variables per chromosome in the initial population is 10% of the total number of available variables. Also, according to Chtioui et al. (1998), a lower number of bits set to 1 in the initial population reduces the number of generations required. Thus, in the initial population, the chromosomes are generated randomly so that the number of bits set to 1 is on average 4 and 7 for the 36 variable and 72 variable datasets, respectively. The number of chromosomes in the population is set to 50 and the fitness function value for the chromosomes is measured as the average ratio of correctly classified defaulting and non-defaulting firms (see Table 2 for a summary of GA parameters). More closely, the fitness function value is determined as the average value in a five-fold cross-validation, where 80% of the data set is used for estimation and 20% of the data set for evaluation. Once the initial population has been generated, chromosome pairs are selected pairwise with the roulette wheel selection to form two new chromosomes into the next generation. For the selected chromosome pair, the cross-over probability is set to 0.8 and the cross-over mechanism employed is a single random point cross-over. Furthermore, each new chromosome is subjected to random mutation with the mutation probability set to 0.001.

Selecting the GA parameters is a difficult task and there are no straight forward methods for it. The parameters in this study are mainly selected based on the generic recommendations of De Jong (1975). To increase the speed of the GA search, the cross-over probability has been set somewhat higher than what was recommended by DeJong. To counter the increased risk of losing good chromosomes, the chromosome with the highest value for the fitness function is copied into the next generation without being subjected to mutation.

Once the next generation has been completed, the breeding process continues until the termination criteria are met. That is, until 250 generations have been evaluated or when the population average fitness function has not improved for 25 generations. When the GA has terminated, the variables that are selected in the best chromosome are registered. After this the GA is run again until it has run 100 times. The final outcome in this first step of the model development process is a frequency of occurrence list of variables. Separate lists are created for the 36 variable dataset and the 72 variable dataset.

Table 2. Genetic algorithm parameters

Parameter	Value
Population size	50
Cross-over probability	0,8
Mutation probability	0,001
Elitism	1
Maximum number of generations	250

In general, regarding the parameter settings of a GA, there are no simple answers or universal rules (De Jong, 2007; Schoenauer, 2015). However, in a study by Mills et al. (2015) the relative importance of different GA settings were assessed by running over 60 numerical optimization problems. The results showed that the most important parameters were cross-over, mutation rate and population size. Based on these findings, the robustness of the results in this study will be assessed by running the GA with alternative values for these three parameters.

3.4 Strategy of analysis

In the second step of the model development process the frequency of occurrence lists are used for creating LDA models for predicting tax defaults. The reason the relatively simple LDA is used as the underlying model is that more sophisticated models typically yield rather small marginal improvements (Balcaen & Ooghe, 2006). The top five most commonly occurring variables from both variable datasets are selected and an exhaustive search for the best variable combination is employed. As the number of variables has been narrowed down to five, a search of the best model among only 32 combinations becomes a feasible option. Once all combinations have been evaluated, the different combinations are ranked according to the average ratio of correctly classified defaulting and non-defaulting firms. As in the first step, the predictive accuracy of each combination is determined using a five-fold cross-validation. Finally, the top three best performing models for both datasets are presented and discussed more in detail. The classification accuracy of these models is also compared with the classification accuracy of the financial distress prediction model developed by Prihti (1975). Prihtis model is based on LDA and comprises three variables measuring profitability, liquidity and solvency. It can be considered a suitable benchmark models as it was developed using Finnish data.

4 Results and discussion

4.1 Descriptive statistics

The descriptive statistics for the 768 firms in the dataset is presented in Table 3. Data for the defaulted firms is presented in panel A and data for firms matched on year, industry and total assets is presented in panel B. Based on total assets, the average (median) size of the firms is about 243 k€ (82 k€). The largest firms in the dataset have total assets over 12 500 k€ whereas the smallest firms have total assets just above 10 k€. As the defaulting and non-defaulting firms have been matched on total assets, there are no significant differences in this measure between the two groups of firms. The sales figures are also similar in both groups, although no matching has been done on this variable. Both return on assets (ROA) and the equity rate, however, show considerable differences between defaulting and non-defaulting firms. The defaulting firms have a negative average equity ratio and a negative average ROA, implying that they are financially distressed. The non-defaulting firms, on the other hand, show a good financial performance with the average (median) equity ratio and ROA at 0.206 (0.500) and 0.072 (0.035), respectively. The magnitude of the tax debt for the defaulting firms range between 5 k€ and 327.7 k€. For the average firm, this equals about 5% of sales and 14% of total assets.

Table 3. Descriptive statistics

		Panel A: Defaulting firms			
	Total assets		Equity		Tax
n = 384	(k€)	Sales (k€)	ratio	ROA	debt (k€)
Mean	242,9	359,7	-0,474	-0,020	14,2
Median	82,0	179,0	-0,005	0,006	8,8
StDev	508,4	622,8	1,515	0,216	21,1
Min	10,0	11,0	-12,824	-0,964	5,0
Max	4 185,0	7 408,0	0,918	0,597	327,7
		Panel B: Non-defaulting firms			
	Total assets		Equity		
n = 384	(k€)	Sales (k€)	ratio	ROA	
Mean	243,1	408,8	0,206	0,072	
Median	82,0	163,5	0,500	0,035	
StDev	510,1	859,6	1,237	0,243	
Min	11,0	11,0	-9,000	-0,633	
Max	4 198,0	12 570,0	0,999	1,911	

All data except for the tax debt is based on year 2013 financial statements. The tax debt data is based on information from from tax defaults that have occured during 2014.

4.2 Assessing the performance of the models

The GA is run with two datasets of variables comprising 36 and 72 variables and with both datasets the GA is run 100 times. The statistics for the GA performance for both datasets are presented in Table 4. The termination criteria for the GA are either when 250 generations have been evaluated or when the population average prediction accuracy has not improved for 25 generations. On average, the statistics show that the GA has terminated after 112 generations with the 36 variable dataset and after 120 generations with the 72 variable dataset. The maximum number of generations in a single run is 223, indicating that the maximum allowed number of generations is sufficient. The time required for evaluating one generation is about 7 seconds, making the total time for completing 100 runs around 80 000 seconds or close to one day. In the initial population, the average number of variables selected per chromosome is 4 for the 36 variable dataset and 7 for the 72 variable dataset. Once the GA has terminated, the average number of variables selected for the fittest chromosome for the 36 and 72 variable datasets is 5.5 and 10.9, respectively.

Table 4. Genetic algorithm performance

			Panel A: 36 variable dataset			
	Generations	Variables		Prediction accuracy		
			Pop. average	Total	Defaulting	Non-defaulting
Average	111,2	5,5	0,716	0,733	0,796	0,670
Median	105,5	5,0	0,717	0,734	0,800	0,671
St. dev.	41,5	1,5	0,010	0,006	0,025	0,032
Min	50,0	3,0	0,686	0,712	0,734	0,560
Max	223,0	10,0	0,736	0,747	0,865	0,732
			Panel B: 72 variable dataset			
	Generations	Variables		Prediction accuracy		
			Pop. average	Total	Defaulting	Non-defaulting
Average	119,1	10,9	0,724	0,742	0,740	0,745
Median	116,0	11,0	0,725	0,741	0,740	0,747
St. dev.	40,0	3,6	0,012	0,007	0,030	0,027
Min	35,0	2,0	0,688	0,730	0,643	0,669
Max	216,0	19,0	0,747	0,762	0,794	0,831

The GA fitness function is measured as the predictive accuracy of the underlying LDA model. That is, the higher the average percentage of correctly classified defaulting and non-defaulting firms, the higher the value for the fitness function. The average prediction accuracy of the entire GA population is just above 0.700 for both the 36 variable and the 72 variable datasets. When looking at the best chromosome, the average prediction accuracy for the 36 variable dataset is 0.733 whereas the average prediction accuracy for the 72 variable dataset is 0.741. In other words, the average predictive accuracy is only marginally higher when including signed square root transformations of the variables. There are, however, differences in predictive accuracies between the two variables datasets when examining the defaulting and non-defaulting firms separately. With the 36 variable dataset the predictive accuracy is considerably higher for the defaulting firms whereas with the 72 variable dataset the predictive accuracies for the defaulting and non-defaulting firms are almost equal.

The GA is run 100 times with both the 36 variable and the 72 variable datasets and for each run the selected variables for the chromosome with the highest value for the fitness function are counted on a frequency of occurrence list. The top 5 most commonly occurring variables for both datasets are presented in Table 5. With the 36 variable dataset the most frequently occurring variable is the equity ratio one year prior to default, whereas the second most commonly occurring variable is the payment period of trade payables two years prior to default. These two variables are clearly the two most commonly occurring variables, being included in 81 and 69 runs out of 100. When examining the frequency of occurrence of variables with the 72 variable dataset, it is noteworthy that all top 5 variables have been subjected to the signed square root transformation. This suggests that there is value in including non-linear transformations of the financial statement ratios.

Table 5. Frequency of occurrence of variables

36 variable dataset		72 variable dataset	
Variable	Frequency	Variable	Frequency
Equity ratio (t)	81	±√Equity ratio (t)	65
Payment period of trade payables (t-1)	69	±√Payment period of trade payables (t)	41
Quick ratio (t)	53	±√Equity ratio (t-1)	39
Current ratio (t)	38	±√Current ratio (t)	37
Change in sales (t-1)	35	±√Payment period of trade payables (t-1)	37

t = one year prior to default, t-1 = two years prior to default, ±√ = signed square root transformation

In the final step, LDA models are developed based on the variables selected by the GA. The models are developed with an exhaustive search including the five most commonly occurring variables. The predictive ability of the models is measured as the average correctly classified defaulting and non-defaulting firms using a five-fold cross-validation. Table 6 shows the performance and the coefficients of the LDA models for both the 36 variable dataset (Panel A) and the 72 variable dataset (Panel B). For both datasets, the three highest performing variable combinations are presented. The highest performing model with the 36 variable dataset comprises four variables and has a predictive accuracy of 73.6%. The second and third highest performing models have similar predictive accuracies of 73.6% and 72.5%, respectively. All three models with the 36 variable dataset clearly outperform models where the variables have been randomly selected. The average predictive accuracy when randomly selecting three or four variables out of the 36 variable dataset is 61.2%. When the correctly classified defaulting and non-defaulting firms are examined separately, a considerable imbalance is evident. For all three models in Panel A the percentage of correctly classified defaulting firms is higher than the percentage of correctly classified non-defaulting firms. In the LDA models a function score above zero indicates a non-defaulting firm whereas a function score below zero indicates a defaulting firm. Consistent with this, most coefficients in the three models in Panel A show the expected signs. The only exception is the coefficient for change in sales two years prior to defaulting. It would be expected that a positive change in sales would reduce the risk for defaulting, but here the coefficient has a negative sign. The explanation for this is that two years prior to the tax default the defaulting firms actually have a higher positive change in sales than the non-defaulting firms. The final LDA model is built using the unstandardized coefficients whereas the standardized coefficients are used for measuring the relative impact of the variables in the LDA model. In the three models presented in Panel A, liquidity measures quick ratio and current ratio clearly show the highest impact while the equity ratio has the second highest impact.

Table 6. LDA model performance

Panel A: 36 variables dataset						
	1st		2nd		3rd	
	Average prediction accuracy					
All firms	0,732		0,732		0,725	
Defaulting firms	0,781		0,807		0,823	
Non-defaulting firms	0,682		0,656		0,628	
	LDA Coefficients					
	Unst.	*Stand.*	*Unst.*	*Stand.*	*Unst.*	*Stand.*
Constant	-0,168		-0,202		-0,307	
Change in sales (t-1)	-0,252	-0,391				
Quick ratio (t)	0,195	1,169	0,192	1,155	-0,015	-0,090
Current ratio (t)					0,209	1,366
Equity ratio (t)	0,278	0,745	0,279	0,748	0,269	0,720
Payment period of trade payables (t-1)	-0,001	-0,518	-0,001	-0,492	-0,001	-0,449
Panel B: 72 variables dataset						
	1st		2nd		3rd	
	Average prediction accuracy					
All firms	0,738		0,733		0,730	
Defaulting firms	0,763		0,753		0,729	
Non-defaulting firms	0,714		0,714		0,732	
	LDA Coefficients					
	Unst.	*Stand.*	*Unst.*	*Stand.*	*Unst.*	*Stand.*
Constant	-0,993		-0,845		-0,844	
±√Current ratio (t)	0,933	1,497	0,908	1,457	0,845	1,355
±√Equity ratio (t)					0,688	1,179
±√Equity ratio (t-1)	0,580	0,901	0,557	0,866		
±√Payment period of trade payables (t)			-0,032	-0,534		
±√Payment period of trade payables (t-1)	-0,034	-0,586	-0,018	-0,309	-0,036	-0,613

1st = best performing model, 2nd = 2nd best performing mode, 3rd = 3rd best performing model. Both unstandardized and standardized coefficients are presented.

Overall, the LDA models based on the 72 variable dataset (Panel B) perform somewhat better than the models based on the 36 variable dataset. The highest performing model comprises four variables out of the 72 variables available and has a predictive accuracy of 73.8%. The other two models have four and three variables with the predictive accuracy of 73.3% and 73.0%. This clearly outperforms models with three or four randomly selected variables where the average predictive accuracy is 61.7%. Apart from performing marginally better than the models based on the 36 variable dataset, the models based on the 72 variable dataset show a considerably lower imbalance between correctly classified defaulting and non-defaulting firm. As with the models based on the 36 variable dataset, a function score above zero for the 72 variable dataset based models indicate a non-defaulting firm and a function score below zero a defaulting firm. All coefficients in the models presented in Panel B have the expected signs. Similar to the models based on the 36 variable dataset, the highest relative impact is shown by liquidity measures and the second highest impact by solvency measures.

Finally, the performance of LDA models is compared to that of the financial distress prediction model developed by Prihti (1975). On average, Prihtis model shows a predictive accuracy of 71.2% which is lower than the predictive accuracy of all six models presented in Table 6. Furthermore, Prihtis model shows a clear imbalance between correctly classified defaulting and non-defaulting firms with the predictive accuracies of 66.9% and 75.5%. The imbalance is of about the same magnitude as with the models based on the 36 variable dataset, but different in that the predictive accuracy is lower for defaulting firms and higher for non-defaulting firms.

4.3 Discussion

The results of this study are twofold. First, the results provide information about the importance of large number of variables in predicting a tax default. The most commonly selected variables by the GA are well in line the theory and previous research. For example, equity ratio, which is the most commonly occurring variable with both the 36 variable and the 72 variable dataset, is often used in various financial distress prediction models. Also, as expected, liquidity variables such as quick ratio and current ratio are among the commonly selected variables. Profitability ratio variables are, however, among the least commonly selected variables. This is somewhat surprising as various profitability ratios are often used in predicting financial distress. When looking at the frequency of occurrence of the variables with the 36 variable dataset, the only variables that are not selected in any of the GA runs are the current and lagged variables of return on assets (ROA) and return on invested capital (ROI). With the 72 variable dataset the profitability ratios and their non-linear transformations do occur in some of the GA runs, but they are still among the less commonly selected variables.

Second, the findings from the GA are used to create simple and intuitive LDA-based models for predicting tax defaults. The three best performing models for both the 36 variable dataset and the 72 variable dataset comprise three or four variables. Overall, there are only marginal differences between the predictive accuracy of the models that are based on the 36 variable dataset and models that are based on the 72 variable dataset. This would suggest that the non-linear transformation of the variables is not required and that the models based on the 36 variable dataset are to be preferred. However, when examining the correctly classified defaulting and non-defaulting firms separately, there is one advantage with including non-linear transformations of the variables that stands out. The imbalance between correctly classified defaulting and non-defaulting firms is considerably lower with the 72 variable dataset than it is with the 36 variable dataset. Thus, assuming that it is equally important to classify correctly both defaulting and non-defaulting firms, the use of the models developed with datasets including non-linear transformations of variables is to be preferred. This is in line with previous research (Deakin, 1976; Ezzamel et al.,1987), which state that that financial statement ratios are not normally distributed. The classification accuracy of the GA-based models can be considered good as they outperform both models based on a random set of variables as well as a, in a Finnish context, commonly used financial distress prediction model (Prihti, 1975). The best performing GA-based model has a classification accuracy of 73.8%, compared with 61.7% for the models with a random set of variables and 71.2% for Prihti's model. Furthermore, the classification accuracy of the GA-based models is more than 10 percentage points higher than what was shown by Marghescu et al. (2010) who also studied Finnish tax defaulting firms. A limitation with the variable selection approach in this study is the time consumption of running the GA. Using an average performance computer, it takes almost 24 hours to run the GA 100 times with a relatively small data set. Increasing the data set and the complexity of the GA would further increase the time required to determine the optimal set of variables. However, once the optimal set of variables for the LDA model has been determined, the GA needs to be run only when new data becomes available. When dealing with financial statement data, an update of the model would be necessary only a few times a year at the most.

Finally, the robustness of the model development process was assessed with some additional tests. First, the GA was run with alternative settings for population size, cross-over rate and mutation rate. Second, the restriction on the average number of selected variables in the initial population was lifted so that the probability for each bit to be set to 1 was 0.5. The frequency of occurrence list showed some minor differences in the order of the variables but none of these additional test had a significant impact on the predictive accuracy of the models. Doubling the population size resulted in an increase in the duration for evaluating one chromosome from about 7 seconds to about 12 seconds. This did, however, not translate in to an increased accuracy of the models.

5 Conclusion

The purpose of this study was to develop a GA-based decision support tool for predicting tax payment defaults. The output of the system is both a list providing information about the importance of various variables in predicting tax defaults and a LDA-based tax default prediction model. The results show that variables measuring solvency, liquidity and payment period of trade payables are important variables in predicting tax defaults. The results also indicate that employing a non-linear transformation on the variables improve the predictive accuracy of the LDA models. The best performing model comprises three non-linearly transformed variables and has a predictive accuracy of 73.8%.

The decision support tool developed in this study has several practical implications. It can, for example, be used by tax authorities when selecting firms for tax audits, by auditors as an assisting tool in the audit process or by suppliers and creditors when assessing risk before extending credit. In general, having an early warning system for and impending tax default would undoubtedly benefit most corporate stakeholders. One major advantage with the system is that it produces an intuitive and simple LDA model which is easy to implement and use, even without the need of a computer. The system is also flexible in that it can easily be applied in other contexts than the Finnish one. From a research point of view, the main contribution of this study is that it adds to the knowledge of factors that predict financial distress. For example, contrary to previous studies, the results in this study indicate that profitability ratios have a relative low importance in predicting tax defaults.

Finally, this study could be extended in several ways. First, the dataset containing the available variables could be augmented with other financial as well as non-financial variables. Furthermore, various types of non-linear transformations could be applied to both the financial and non-financial variables. Second, more sophisticated types of models, such as logistic regression and neural networks, could be used for classification. Third, more emphasis could be placed on tuning the GA parameters. Particularly, this could include using dynamic parameter strategies. Lastly, considering the time consumption of running the GA in this study, one strand for future research would be to develop a faster and more efficient GA to reduce the computation time.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, 23(4), 589-609.

Beaver, W. H. (1966). Financial ratios as predictors of failure. Journal of accounting research, 71-111.

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. The British Accounting Review, 38(1), 63-93.

Back, B., Laitinen, T., & Sere, K. (1996). Neural networks and genetic algorithms for bankruptcy predictions. Expert Systems with Applications, 11(4), 407-413.

Brabazon, A., & Keenan, P. B. (2004). A hybrid genetic model for the prediction of corporate failure. Computational Management Science, 1(3), 293-310.

Chtioui, Y., Bertrand, D., & Barba, D. (1998). Feature selection by a genetic algorithm. Application to seed discrimination by artificial vision. Journal of the Science of Food and Agriculture, 76(1), 77-86.

De Jong, K. (1975). An analysis of the behavior of a class of genetic adaptive systems. Ph. D. Thesis, University of Michigan.

De Jong, K. (2007). Parameter setting in EAs: a 30 year perspective. In Parameter setting in evolutionary algorithms (pp. 1-18). Springer Berlin Heidelberg.

Deakin, E. B. (1976). Distributions of financial accounting ratios: some empirical evidence. The Accounting Review, 51(1), 90-96.

Dimitras, A. I., Zanakis, S. H., & Zopounidis, C. (1996). A survey of business failures with an emphasis on prediction methods and industrial applications. European Journal of Operational Research, 90(3), 487-513.

Du Jardin, P. (2010). Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. Neurocomputing, 73(10), 2047-2060.

Du Jardin, P., & Séverin, E. (2011). Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of a financial failure model. Decision Support Systems, 51(3), 701-711.

Ezzamel, M., Mar-Molinero, C., & Beech, A. (1987). On the distributional properties of financial ratios. Journal of Business Finance & Accounting, 14(4), 463-481.

Gilbert, L. R., Menon, K., & Schwartz, K. B. (1990). Predicting bankruptcy for firms in financial distress. Journal of Business Finance & Accounting, 17(1), 161-171.

Gordini, N. (2014). A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy. Expert Systems with Applications, 41(14), 6433-6445.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

Hand, D. J. (2004, May). Marginal classifier improvement and reality. In Symposium on Data Mining, Ghent University (Belgium), May.

Hanlon, M., & Heitzman, S. (2010). A review of tax research. Journal of Accounting and Economics, 50(2), 127-178.

Holland, J. H. (1992). Adaptation in natural and artificial systems, University of Michigan Press, Ann Arbor, MI.

Laitinen, E. K. (1991). Financial ratios and different failure processes. Journal of Business Finance & Accounting, 18(5), 649-673.

Leardi, R., Boggia, R., & Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. Journal of chemometrics, 6(5), 267-281.

Lennox, C., Lisowsky, P., & Pittman, J. (2013). Tax aggressiveness and accounting fraud. Journal of Accounting Research, 51(4), 739-778.

Lensberg, T., Eilifsen, A., & McKee, T. E. (2006). Bankruptcy theory development and classification via genetic programming. European Journal of Operational Research, 169(2), 677-697.

Lobo, F. G., & Lima, C. F. (2005, June). A review of adaptive population sizing schemes in genetic algorithms. In Proceedings of the 7th annual workshop on Genetic and evolutionary computation (pp. 228-234). ACM.

Man, K. F., Tang, K. S., & Kwong, S. (1996). Genetic algorithms: concepts and applications [in engineering design]. IEEE transactions on Industrial Electronics, 43(5), 519-534.

Marghescu, D., Kallio, M., & Back, B. (2010). Using Financial Ratios to Select Companies for Tax Auditing: A Preliminary Study. Organizational, Business, and Technological Aspects of the Knowledge Society, 393-398.

Mills, K. L., Filliben, J. J., & Haines, A. L. (2015). Determining relative importance and effective settings for genetic algorithm control parameters. Evolutionary computation, 23(2), 309-342.

Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert systems with applications, 41(4), 2052-2064.

Paterlini, S., & Minerva, T. (2010, June). Regression model selection using genetic algorithms. In Proceedings of the 11th WSEAS international conference on nural networks and 11th WSEAS international conference on evolutionary computing and 11th WSEAS international conference on Fuzzy systems (pp. 19-27). World Scientific and Engineering Academy and Society (WSEAS).

Prihti, A. (1975). Konkurssin ennustaminen taseinformaation avulla: Summary The prediction of bankruptcy with published financial data. Helsingin kauppakorkeakoulu.

Ravisankar, P., Ravi, V., & Bose, I. (2010). Failure prediction of dotcom companies using neural network–genetic programming hybrids. Information Sciences, 180(8), 1257-1267.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. Journal of Machine Learning Research, 3(Mar), 1371-1382.

Schoenauer, M. (2015). Evolutionary Algorithms. In Handbook of Evolutionary Thinking in the Sciences (pp. 621-635). Springer Netherlands.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507-2517.

Serrano-Cinca, C., & GutiéRrez-Nieto, B. (2013). Partial least square discriminant analysis for bankruptcy prediction. Decision Support Systems, 54(3), 1245-1255.

Wilson, R. L., & Sharda, R. (1994). Bankruptcy prediction using neural networks. Decision support systems, 11(5), 545-557.