UNIVERSITY OF TARTU

Faculty of Social Sciences

School of Economics and Business Administration

Õie Renata Siimon

CORPORATE TAX ARREARS PREDICTION BASED ON MONTHLY TIME SERIES OF TAX ARREARS

Master's thesis

Supervisor: Senior Research Fellow Oliver Lukason (PhD)

Tartu 2020

Forwarded to defence ……..

(supervisor’s signature)

I have written this master's thesis independently. All viewpoints of other authors, literary sources and data from elsewhere used for writing this thesis have been referenced.

……..

(author’s signature)

Abstract

In this thesis, a model for predicting companies to have tax arrears next month, given their tax arrears in the preceding 12 months is proposed. Four machine learning (ML) methods – decision tree, random forest, k-nearest neighbours and multilayer perceptron – were used for building models with monthly tax arrears and other independent variables constructed from them. Data consisted of tax arrears of all Estonian SMEs (2011–2018). The best performing ML model was random forest trained on monthly tax arrears with aggregation of earlier months into period means (accuracy 84.5%). In order to reduce the high proportion (92%) of zero values, the model was built only for observations with previous tax arrears in at least two months. Accuracy of the final model comprising all test data, i.e. also observations with tax arrears in less than two months, was 95.3%. The novelty of this thesis is that, to the best knowledge of the author, monthly tax arrears have not been previously used in tax arrears prediction. Also, despite the economic importance of ensuring tax compliance, studies on predicting corporate tax arrears have so far been scarce, have only offered models making annual predictions, have nearly always used financial ratios as independent variables, and have achieved only moderate accuracy.

Keywords: tax arrears, SMEs, time series classification, machine learning, predictive models

CERCS: S180, S181, S184

1. Introduction

Taxes are an essential source of income for any government. Being able to detect companies that are likely to incur tax arrears as accurately as possible would enable tax authorities to better target their tax audits and preventive measures aimed at ensuring the timely payment of taxes. However, despite the high economic importance of ensuring tax compliance, studies on predicting corporate tax arrears have so far been scarce.

The main drawbacks of current studies (e.g. Marghescu *et al.* (2010), Höglund (2017) and Batista *et al.* (2012)) are that they have mostly concentrated on using financial ratios as predictors of tax arrears, and have only proposed models for predicting tax arrears next year. The disadvantage of using financial ratios is that they become available with a considerable time lag after the payment irregularities have already been going on for some time. Also, they cannot be used if financial reports are unavailable, which is much more likely to happen in case of financially distressed firms (Lukason and Andresson, 2019), which are also more likely to have tax arrears. In addition, accuracies of models using financial ratios have been moderate. The disadvantage of annual predictions is that they can only be made once a year, and only predict if a company will have tax arrears any time next year, which seems rather vague for practical purposes.

To the best knowledge of the author, there have been no studies where monthly time series of corporate tax arrears have been used for predicting tax arrears in the next month. This thesis intends to fill this research gap. Using data with monthly instead of annual frequency would have much higher practical value, since in carrying out their daily activities, tax authorities would need to be able to detect companies likely to incur tax arrears not only once a year and not only for the entire next year, but at any time and for a more immediate future, using the most recent information available.

The aim of this thesis is to explore which machine learning methods and types of independent variables work best in predicting companies to have tax arrears next month, given the time series of their tax arrears in the preceding 12 months. In addition to the 12 monthly amounts of tax arrears, two alternative types of variables constructed from them were considered. One of those types included statistical measures and counts of events, and the other one monthly amounts with aggregation of amounts in earlier months into

period means. The machine learning methods used were decision tree (DT), random forest (RF), k-nearest neighbours (KNN) and multilayer perceptron (MLP), which have also been applied in the related area of failure prediction (see, for example, Alaka *et al.* (2018), Ravi Kumar and Ravi (2007) and Barboza *et al.* (2017)). The two areas are related, since defaulting on taxes is a strong sign of financial distress, which in turn might eventually lead to bankruptcy (Höglund, 2017).

Data used for this thesis were corporate monthly tax arrears of the entire population of Estonian SMEs for the period 2011-2018, from which more than 2 million observations, i.e. company-12 month period pairs, were collected using moving window approach. In total, 49,156 companies were included.

A specific characteristic of monthly tax arrears is that they are rare events, as companies usually try to pay their taxes in time. However, learning from mostly zero-valued data is a difficult task for machine learning models. The approach used in this thesis for reducing the high proportion (92%) of zero values was to build machine learning models only for observations which had tax arrears in at least two among any of the 12 preceding months, while the rest of the observations, where tax arrears next month were very unlikely to occur and difficult to predict, were always predicted not to have tax arrears next month. Thus, two types of accuracies are reported in this thesis: a) accuracies based on different machine learning approaches to predict the presence of tax arrears for firms with at least two months with tax arrears, b) accuracy in all test data based on the best accuracy noted in a) and accuracy for other observations based on the previously described simple intuitive logic. While the accuracies of type a) express the performance of the machine learning methods in solving the classification task, calculating also the accuracy of type b) was necessary for measuring performance on the entire test set in order to make the results comparable to previous studies.

The models were implemented in Python programming language, using Keras neural networks library[1] for the MLP model, scikit-learn machine learning libraries[2] for other models and SciPy library[3] for statistical tests.

[1] https://keras.io/.
[2] Scikit-learn User Guide: https://scikit-learn.org/stable/user_guide.html.
[3] https://scipy.org/.

The rest of the paper is organized as follows. Literature review is provided in section 2. Section 3 contains description of the dataset, variables and methodology. The results, accompanied by related discussion, are presented in section 4. Section 5 contains conclusions, and in section 6, challenges for future research are laid out.

2. Literature review

Despite the high economic importance of ensuring tax compliance, studies on predicting corporate tax arrears have so far been scarce. To the best knowledge of the author, there have been no previous studies where the presence of tax arrears next month has been predicted based on previous monthly tax arrears.

There have been a handful of studies where the presence of tax arrears have been predicted based on financial ratios. Contrary to this thesis, all those studies make predictions for next year, instead of next month. For example, Marghescu *et al.* (2010) used data on 328 Finnish companies to predict the presence of arrears in employer contribution taxes using logistic regression. The classification accuracy of their model was very low (61.6%), and only exceeded the naïve baseline model of predicting none of the companies to have tax arrears by less than one percentage point. As the model was heavily underspecified, they suggested that more variables should be added.

Höglund (2017) used genetic algorithm based variable selection, followed by linear discriminant analysis (LDA) to predict tax arrears next year, using a dataset of 768 Finnish firms. The independent variables used in that study included 17 financial ratios and two industry related variables (bankruptcy risk and payment default risk). The accuracy of their best model was 73.8%.

Batista *et al*. (2012) used financial ratios to classify Portuguese real estate agencies as tax compliant (i.e. not having tax arrears), using discriminant analysis and logistic regression. They built separate models for each of the three years (2007–2009) included in the dataset, using data of ca 200 companies for each model, as their aim was also to compare results before and after the financial crisis. In addition to conventional financial ratios, independent variables in their model also included Taxation Effective Rate, which is an indicator associated with tax evasion. However, this independent variable turned out to

be statistically significant only in two out of the three years covered by the study. The accuracies of all their models were rather similar, with accuracy of 72.4% achieved by discriminant analysis model built for the year 2008 being the best.

The abovementioned three studies reveal that the performance of financial ratios as predictors of tax arrears is rather low. The likely reason for this is that annual accounts from which the ratios are calculated become available with a considerable time lag after the payment irregularities have already been going on for some time. In addition, it could be assumed that predicting tax arrears with monthly frequency, as is done in this thesis, using solely financial ratios would be even more difficult, because the presence of tax arrears may change monthly, while the ratios only change annually.

Another disadvantage of using financial ratios for predicting tax arrears is that the models cannot be used for cases where financial reports are unavailable. This, however, is much more likely to happen in case of financially distressed firms (Lukason and Andresson, 2019) which are therefore also more likely to have tax arrears. For example, when predicting tax arrears based on financial ratios, Höglund (2017) left as much as 63% of the companies with tax arrears out from the model for this reason alone. If tax authorities were to use such models in practice for selecting companies for tax auditing, they would not be able to predict tax arrears for a large proportion of companies that are likely to have them, which represents a serious drawback for the practical application of those models.

A methodologically entirely different approach for predicting tax debt was taken by Zhao *et al*. (2009), who used sequence classifiers, i.e. frequent pattern mining models where independent variables are temporally ordered, to predict social security debts. The independent variables were activity codes of 155 possible activities of ca 10,000 taxpayers in the Australian tax database, with each activity code being accompanied by taxpayer ID and date and time of the activity. They constructed ca 16,000 activity sequences from this data with the aim of predicting which sequences lead to social security debts. The accuracy of their best classifier was moderate (76.0%). Also, ’debt’ in their study had a different meaning than ’tax arrears’ in this thesis – instead of taxes left unpaid, by ’debt’ they meant overpayment of social security benefits by the tax

authorities. As the benefits depended on entries in the tax database, the issue they solved had some similarities to fraud discovery tasks.

Finally, Su *et al.* (2018) built an ensemble classification model composed of random forest, k-nearest neighbours, multilayer perceptron, extremely randomized trees, gradient tree boosting and XGBoost models, using data of 70,000 Chinese companies for training and 50,000 for testing to predict the presence of tax arrears next year. To the best knowledge of the author, this is the only study where previous tax arrears have been used to predict tax arrears in the future. However, in that study, tax arrears were only one independent variable among many others, which also included company specific parameters (incl. industry and region) and 17 financial statements items. The accuracy of the proposed model was excellent (90.58%). Contrary to this thesis, they used classification into three classes (no tax arrears, and tax arrears below or above the threshold of 5000 RMB) instead of binary classification. Also, contrary to this thesis, the model was built for making annual instead of monthly predictions.

Tax arrears prediction has some similarities to bankruptcy prediction. In both cases, the aim is to assess a company's ability to fulfil its obligations (Batista, 2006), i.e. whether it is in financial distress. In this regard, defaulting on taxes is a strong sign of financial distress which in turn might eventually lead to bankruptcy (Höglund, 2017). Therefore, tax arrears prediction can detect financial distress earlier than bankruptcy prediction, i.e. before the temporary tax payment difficulties (temporary insolvency) have evolved into bankruptcy (permanent insolvency). This is important in ensuring tax compliance, since, as shown by Kukalová *et al.* (2020), recovery rate of unpaid taxes in insolvency proceedings can be remarkably low. Since the two research topics are related, to a certain extent knowledge drawn from the bankruptcy prediction field is also applicable in the field of tax arrears prediction.

Given the above, studies where tax arrears have been used as independent variables in predicting bankruptcies might be relevant. For example, Lukason and Andresson (2019) compared the performance of tax arrears and financial ratios in bankruptcy prediction, and found that tax arrears were in fact better predictors of bankruptcy. They also noted that payment defaults can be a vital substitution for financial ratios in cases where annual reports are not available. The independent variables based on tax arrears used in their

study (maximum and median of tax arrears, number of month-ends with tax arrears and length of the longest sequence of month-ends with tax arrears) were also included among the initial independent variables considered in this thesis.

In their study, Kubicová and Faltus (2014) also tried to use tax obligations for bankruptcy prediction. Their approach was, however, quite different and experimental in nature, as they used ratios which had financial statement items related to income tax (e.g. total income tax, deferred income tax) in the numerator and own capital, sales and total assets in the denominator. They concluded, however, that such ratios are not suitable for predicting company defaults, at least not for Slovakian companies which were the object of their study.

As this thesis uses previous tax arrears, studies concerning previous payment behaviour might also be relevant. In this regard, there have been a few studies where previous payment behaviour has been used for predicting company defaults or credit risk. For example, Ciampi *et al.* (2020) used the numbers and values of more than 60 days past due and/or overdrawn exposures of bank loans, along with financial ratios, for bankruptcy prediction. Karan *et al*. (2013) used independent variables such as proportion of invoices paid late among all invoices, sum of days paid before deadline, total debt/total purchases, average amount paid, etc. among other independent variables for predicting credit risk that retailers pose for a wholesaler. And finally, Back (2005) used, *inter alia*, independent variables such as numbers of payment disturbances and delays for predicting financial difficulties of firms. In this thesis, however, independent variables combining information on both arrears and payments could not be used, because the dataset only contained the amounts of tax arrears, with no information on how long each individual exposure had lasted nor on the amounts of taxes paid. On the other hand, numbers of months with tax arrears have been also used in the models in this thesis.

When choosing the machine learning methods to be used in this thesis, methods that have previously been applied in the related area of bankruptcy prediction were considered. According to Veganzones and Severin (2020), those methods can be divided into three categories: traditional statistical methods, machine learning methods and ensemble methods (i.e. combinations of several methods), although some researchers (e.g. Jayasekera (2018)) place hazard models and neural networks in separate categories. When

comparing recent trends, Veganzones and Severin (2020) found that among bankruptcy prediction articles published in 2008–2017, only 13% use traditional statistical methods, while 36% use machine learning methods and 51% use ensemble methods. As noted by Domingos (2012), composing ensemble models has become a standard practice in machine learning, as they often provide better results than single models. A possible explanation for the high proportion of studies using ensemble models in bankruptcy prediction could also be that this field of research has already been thoroughly studied with a wide variety of standalone methods, which could be why researchers are now trying to increase predictive performance by combining the models in different ways.

Based on a review by Shi and Li (2019) of articles published in 1968–2017, traditional statistical methods used in bankruptcy prediction include logit (logistic regression) and probit, multivariate discriminant analysis (MDA) and hazard models. Among the machine learning methods, they identify neural networks, support vector machines (SVM), decision trees, genetic algorithm, fuzzy sets and rough sets as methods that have been applied in bankruptcy prediction literature already before 2007, while methods such as random forest, Adaboost, particle swarm optimization, naïve bayes and k-nearest neighbours (KNN)[4] have appeared only after 2007. According to du Jardin (2017), ensemble techniques widely used in bankruptcy prediction include bagging, boosting, rotation forest, Decorate and random subspace.

In general, it has been found that machine learning methods have higher accuracy in bankruptcy prediction than traditional statistical methods (Barboza *et al.*, 2017). For example, Alaka *et al.* (2018) found that across bankruptcy prediction articles published in 2010–2015, average accuracies of the most widely used machine learning methods (neural networks, SVM and decision tree) were all higher than those of the most widely used statistical methods (logit and MDA), with the average accuracy of neural networks being the highest, followed by SVM. The disadvantage of statistical methods is that they are subject to some restrictive assumptions. For example, MDA assumes variables to be normally distributed and have equal covariance matrices, logistic regression assumes absence of multicollinearity between independent variables (Sun *et al.*, 2014a), and probit assumes cumulative normal distribution (Balcaen and Ooghe, 2006). On the other hand,

[4] In bankruptcy prediction, KNN is often used in the framework of Case Based Reasoning (CBR).

machine learning methods have the advantage that they can deal with non-linear distributions and do not have stringent assumptions on the data (Veganzones and Severin, 2020). For the reasons above, traditional statistical methods were not used in this thesis.

As regards predictive performance of machine learning methods, there is no consensus on which one of them performs best in bankruptcy prediction (Barboza *et al.*, 2017), since no method performs consistently better than all others across different datasets (Chen *et al.*, 2016). Therefore, it was not possible to choose the methods for this thesis based on which methods have been established as best-performing in bankruptcy prediction.

The methods chosen for this thesis included decision tree (DT), k-nearest neighbours (KNN), multilayer perceptron (MLP) and random forest (RF), where DT and KNN are conventional machine learning methods, MLP is a neural networks method and RF is an ensemble method. While DT and neural networks are two out of the three most widely used machine learning methods in bankruptcy prediction (Alaka *et al.*, 2018), RF and KNN have appeared in the literature of this research field only recently (Shi and Li, 2019). SVM, which also belongs among the three most widely used machine learning methods (Alaka *et al.*, 2018), was not used in this thesis, since according to the information in the standard scikit-learn library for SVM[5], the time complexity of the algorithm makes its application impractical beyond sample sizes exceeding a few tens of thousands of observations.

All methods chosen for this thesis can handle well data where classes are not linearly separable, which also the case in this thesis. In addition, neural networks have the advantage that they can approximate any function, have excellent capability for finding patterns in complex data and have shown very good performance (Chen *et al.*, 2016). An advantage of KNN is that its classification decisions can be easily explained by offering examples of similar observations that the model has seen in the past (Ravi Kumar and Ravi, 2007). Random forest is based on decision tree models (Barboza *et al.*, 2017) and they therefore have similar advantages. These include easily interpretable results, high accuracy and identification of the importances of independent variables (Chen *et al.*, 2016).

[5] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

3. Data, variables and methodology

3.1. Data

The dataset used for this thesis included monthly amounts of tax arrears of Estonian SMEs during the period 2011–2018. The original dataset, which was obtained from the Estonian Tax and Customs Board, contained tax arrears of 419,210 legal entities at different monthly reporting dates, as well as the end-of-month figures. However, in this thesis only the end-of-month figures were used, because figures for other dates were not available for all the years. Also, using only end-of month figures allowed to disregard cases of less economic importance where taxes were paid just a few days late.

In order to increase homogeneity in the data, only SMEs that were going concerns and had at least some minimum level of economic activity at the time their tax arrears were recorded were considered. Therefore, on the one hand, companies that were too large to conform to the SME definition[6], that is, having annual turnover above 50 million EUR or total assets[7] above 43 million EUR, and on the other hand, companies with non-existing or very low level of economic activity, as evidenced by total assets below 2,500 EUR or turnover below 16,000 EUR[8], or by not having VAT registration status[9], were removed from the datasets for the periods they complied with those exclusion criteria. Also, companies in bankruptcy or liquidation or that had ceased their activities were removed starting from the date of their bankruptcy or liquidation notice or deletion date[10]. Finally, data on public entities and NGOs were also left out.

[6] As provided in Commission Recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises, OJ L 124, 20.5.2003, pp. 36–41.
[7] Data on total assets and turnover were obtained from a dataset of annual accounts entries from the Estonian Commercial Registry.
[8] The cut-off value of 16,000 EUR was chosen because, until the end of year 2017, it was also the threshold for being registered as a company having VAT liability in Estonia.
[9] Data on VAT registrations was obtained from a dataset from the Estonian Tax and Customs Board. In most cases, the data included either only the start date or only the end date of registration, which was also why additionally compliance with the minimum level of turnover (16,000 EUR) was checked.
[10] Deletion dates were obtained from a dataset on company statuses from the Estonian Commercial Registry, which, however, did not contain dates of other statuses than 'deleted'. Bankruptcy and liquidation dates were mostly obtained from the website of the Estonian Creditors Association (http://www.evul.ee/pankrotistunud-ettevotted/) and through automated queries to the API of Official Announcements (https://www.ametlikudteadaanded.ee/avalik/uriotsing), while the few remaining ones that could not be obtained this way, were manually retrieved from the Official Announcements (https://www.rik.ee/et/ametlikud-teadaanded) or the Commercial Registry (https://ariregister.rik.ee/).

After additionally removing a few outliers with tax arrears exceeding 2.5 million EUR, the dataset contained a total of 49,156 companies. The data for each company could also include non-full years, as the cut-off dates (VAT registration start and end dates, as well as bankruptcy, liquidation and deletion dates) could be any dates within a year. All consecutive 13-month periods for each company, i.e. 12 months for the independent variables and the last month for the dependent variable, were then collected into the final dataset using moving window approach.

The resulting dataset contained 2,078,408 company-13 month period pairs, with each company being included in the dataset on average 42 times (i.e. on average, 3.5 years of data was available for each company). The advantage of the moving window approach was that it allowed to capture the dynamics of tax arrears in the 12 months preceding the prediction while using a large amount of data. The disadvantage was that it did not allow to take multiannual dynamics into account.

A specific characteristic of the dataset was the sparsity of data, with an overwhelming proportion (91.82%) of the monthly tax arrears being zero. This shows that most of the time most companies do not have tax arrears, most likely because they try to avoid owing money to the government and pay their taxes in time. However, for any machine learning method, learning from mostly zero-valued data is a challenging task and is likely not to render good results. The reason for this is that the variation between observations may become dominated by noise (Kelleher *et al.*, 2015).

The approach used to reduce data sparsity was to only consider observations with tax arrears in at least any two months during the 12 months period preceding prediction (15% of the training data) for building the machine learning models, while the rest of the observations (85% of the training data), where the probability of tax arrears next month was very low (1.12%), were always predicted not to have tax arrears next month. Such dividing of the dataset into two parts achieved a considerable reduction in zero values (from 91.82% to 48.84%) in the part of the data used for the models. Also, this part of the data was indeed economically the most interesting, as a company could be expected to be much more likely to have tax arrears next month if it already had incurred tax arrears at least twice during the 12 preceding months.

12 month periods starting in any month in 2011–2016 (with dependent variable in 2012–2017) were used as training set and those starting in 2017 (with dependent variable in 2018) were used as test set. Using different periods for training and testing ensured independence of the training and test data. This was important because, as noted by Sun *et al.* (2014a), choosing the test data randomly may cause stochastic performance. Training and test set sizes, including their parts containing observations with zero, one and more than one months with tax arrears, along with the percentage of observations with tax arrears next month are presented in Table 1.

Table 1. Sizes of parts of the dataset and percentage of observations with tax arrears next month in each (composed by author)

Months with previous tax arrears	No. of observations			Tax arrears next month (%)		
	Train	Test	Total	Train	Test	Total
0	1,287,808	325,308	1,613,116	0.77%	0.65%	0.74%
1	134,427	24,674	159,101	6.69%	7.35%	6.79%
2–12	253,468	52,723	306,191	46.48%	50.51%	47.17%
Total	1,675,703	402,705	2,078,408	8.16%	7.59%	8.04%

In classification, a dataset is imbalanced if the number of observations in one class exceeds the number of observations in the other class (Sun *et al.*, 2014b). As shown in Table 1, the dataset as a whole was heavily imbalanced, with only 8.16% of observations in training set having tax arrears next month. However, the part of the dataset containing observations with 2–12 months with tax arrears, which was the only one for which machine learning models were to be built, was almost perfectly balanced, with 46.48% of observations in training set having tax arrears next month. Machine learning models tend to perform better if they are trained on balanced datasets *(Ibid.)*, where the sizes of the classes are equal. For balancing the dataset used for building the models, undersampling was used by randomly removing 17,844 observations without tax arrears next month from the part of the training set containing observations with 2–12 months with tax arrears. No balancing was performed for the test set, as this would have resulted in biased estimates on how well the models would perform on new real-life data.

3.2. Dependent variable

The dependent variable used in the machine learning models was a dummy variable 'tax arrears next month', the value of which was '1' for observations with tax arrears next month and '0' for observations without tax arrears next month. Due to the low economic significance of tax arrears below 100 EUR, a company was only considered to have tax arrears next month if its tax arrears next month exceeded 100 EUR[11].

3.3. Independent variables

Three types of independent variables were considered in this thesis: 12 monthly amounts of tax arrears without aggregation (M12) and with aggregation of amounts in earlier months into period means (M5), and counts of events and statistical measures (STATS) (see Table 2). As regards notation of the months in Table 2 and elsewhere in this thesis, month 1 is the earliest month of the period, and month 12 is the last month of the period (i.e. the month preceding the month for which predictions were made). The reason for considering other types of independent variables besides the 12 monthly figures was that it seemed uncertain whether predictive models would perform well with 48.84% of the independent variable values being zero. The added types of independent variables contained much lower proportion of zero values, and also helped to capture different aspects of the dynamics of tax arrears during the 12 months preceding the prediction.

The M12 type of independent variables (see Table 2) were just the 12 monthly amounts of tax arrears. The STATS type of independent variables contained counts of events (months with or without tax arrears) and statistical measures, which were included under a single type of variables, because otherwise both would have only had two variables in the final models. Independent variables corresponding to four STATS type of variables used in this thesis ('d max' and 'd med', 'd m in debt' and 'd longest') have previously been also successfully used in bankruptcy prediction by Lukason and Andresson (2019). Interestingly, in their research they found that tax arrears were in fact better predictors of bankruptcy than financial ratios.

[11] This is in line with §14(5) of the Estonian Taxation Act, according to which tax authorities are required to issue a certificate concerning the absence of tax arrears if tax arrears of the person requesting such certificate are below 100 EUR.

Table 2. Initial and final independent variables (composed by author)

Type	Independent variable	Included in final models	Description
Amounts without aggregation (M12)	month 1, ..., month 12	yes	Monthly tax arrears without aggregation (in EUR)
Amounts with aggregation of earlier periods (M5)	months 1–5	yes	Arithmetic mean of tax arrears in months 1-5 (in EUR)
	months 6–9	yes	Arithmetic mean of tax arrears in months 6-9 (in EUR)
	month 10	yes	Tax arrears in month 10 (in EUR)
	month 11	yes	Tax arrears in month 11 (in EUR)
	month 12	yes	Tax arrears in month 12 (in EUR)
Counts of events and statistical measures (STATS)	d first	yes	Number of consecutive months with tax arrears preceding the prediction (i.e. when were tax arrears first seen)
	d last	yes	Number of consecutive months without tax arrears preceding the prediction (i.e. when were tax arrears last seen)
	d m in debt	no	Total number of months with tax arrears
	d longest	no	Length of the longest sequence of consecutive months with tax arrears
	d med	yes	Median of monthly tax arrears (in EUR)
	d mean	no	Arithmetic mean of monthly tax arrears (in EUR)
	d max	no	Maximum of monthly tax arrears (in EUR)
	d std	yes	Standard deviation of monthly tax arrears (in EUR)

The motivation for using the M5 type of independent variables was an observation that Gini importances[12] extracted from a decision tree model[13] of all except the last four 12 monthly tax arrears were really low (below 1%) (see 'Gini before aggregation' in Table 3). Basically, Gini importances show the relative importance of each independent variable compared to other independent variables in making the decisions about the best splits in a decision tree model. Aggregating amounts in earlier months into period means allowed to increase the Gini importances of the resulting independent variables, and was therefore expected to also increase the performance of the models.

[12] According to scikit-learn documentation, Gini importances are calculated as normalized total reduction of the splitting criterion (in this case, Gini impurity) brought along by each independent variable (https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier.feature_importances_).

[13] Built with minimum of 1000 samples in leaf and otherwise with default parameter values.

Table 3. Gini importances of monthly tax arrears before and after aggregation (composed by author)

Month	Gini before aggregation	Gini after aggregation
1	0.004	0.012
2	0.002	
3	0.002	
4	0.007	
5	0.003	
6	0.002	0.019
7	0.001	
8	0.003	
9	0.014	
10	0.022	0.024
11	0.108	0.110
12	0.833	0.835

In order to decide which months to aggregate, all possible combinations for aggregating the first ten monthly amounts into period means were explored, with the restriction that as a result of the aggregation, all Gini importances were to be above 1%. The best possible choice for aggregation, chosen based on accuracy of the resulting decision tree model (81.18%), was to aggregate months 1–5 and 6–9 into period means, and not to aggregate the last three months (see 'Gini after aggregation' in Table 3). Decision tree has previously been used as variable selection method for example by Cho *et al.* (2010), who also used it as a preliminary technique to select independent variables that were subsequently used for building models with other machine learning methods. In their study, decision tree outperformed stepwise logistic regression as independent variable selection tool.

In deciding whether to leave any of the initially selected independent variables out from the final models, their descriptive statistics (Annex 1), correlation matrix (Annex 2) and univariate prediction accuracies (Annex 3) were considered. The latter were obtained by training univariate models for each independent variable with all four machine learning methods that were later also used for training the multivariate models. The parameters used in the univariate models that differ from the default parameters are given in Annex 4.1. Univariate prediction accuracies and correlations were not calculated for the

M12 type of independent variables, since leaving out any of the 12 monthly amounts would have jeopardized the integrity of the time series.

All univariate prediction accuracies were satisfactory (above 60%) (see Annex 3), indicating that all independent variables that were initially considered could have been useful predictors of tax arrears next month. However, due to high correlations (see Annex 2), some of the independent variables were left out of the final models (see Table 2). Namely, as regards the STATS variables, 'd m in debt' and 'd longest' were left out due to high correlations with other counts of events type of variables, 'd max' due to high correlations with other statistical measures, and 'd mean' due to high correlation with 'd med'. The independent variables that were left out due to high correlation had lower univariate prediction performance (see Annex 3) than the independent variables they correlated with. Since the M5 type of variables essentially constituted a time series, and autocorrelation is a typical property of time series data, none of the M5 type of variables were excluded due to high correlation. For all independent variables included in the final models, distribution properties of the classes were different (see descriptive statistics in Annex 1), which confirmed that they could be useful predictors of tax arrears next month.

In order to ensure best possible performance of the models, independent variables used in KNN and MLP models needed to be on similar scales. For KNN, this requirement was due to distance calculations performed in order to find the closest neighbours. For MLP, rescaling was necessary because having independent variables on different scales makes it harder for the algorithm to learn appropriate weights.

The rescaling method used for KNN was signed natural logarithm, as defined in (Alessandretti *et al.*, 2019):

$$sLog(x) = \begin{cases} sign(x)\, log(x), x \neq 0 \\ 0 \quad , x = 0 \ , \end{cases} \tag{1}$$

which was applied to all independent variables that were expressed in euros (i.e. all except counts of events, where the variances were small). The reason for using signed natural logarithm instead of natural logarithm was that in case of M5 and M12 type of independent variables, some values were non-positive.

For MLP, rescaling was done by first using signed natural logarithm in the same way as for KNN, and then applying a widely used standardization method that consists in subtracting the mean and dividing by standard deviation and is sometimes called z-score (see, for example, Flach (2012)):

$$z_i = \frac{x_i - \bar{x}}{s}, \quad (2)$$

where $\bar{x}$ is the mean and s is the standard deviation of the independent variable in the training set. In case of M5 and M12 type of independent variables, the mean and standard deviation of all variables belonging to the respective type of independent variables were used instead of standardizing each variable separately.

Anderson-Darling test[14] was used to check normality of distributions of the independent variables. Results showed that none of the independent variables were normally distributed. Then, two-sample Kolmogorov–Smirnov test[15] was used to check the statistical significance of the independent variables in discriminating between the classes. The advantage of this test is that, unlike many other statistical tests, it does not require data to be normally distributed. In the bankruptcy prediction literature, sometimes tests having normality assumption, like t-test, are used to test the significance of financial ratio variables (Alaka *et al.*, 2018) despite that they are well known for not having normal distribution (Balcaen and Ooghe, 2006). Given that violating assumptions of a test makes its results questionable, two-sample Kolmogorov-Smirnov test was deemed better suited in this case. The test results showed that all independent variables were significant at 1% significance level.

3.4. Methodology

As provided in section 3.1, the approach used in this thesis for reducing the overwhelming proportion of zero values among the monthly tax arrears was to build machine learning models only for observations which had tax arrears in at least two among any of the 12 preceding months, while the rest of the observations were always predicted not to have tax arrears next month. This was justified because among observations with previous tax

[14] Documentation of the test in SciPy library: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.anderson.html#scipy.stats.anderson.
[15] Documentation of the test in SciPy library: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html.

arrears in zero or one months, the probability of tax arrears next month was very low (0.77% and 6.69%, respectively), and due to all or nearly all monthly figures being zero, they were difficult to predict.

In the final model, predictions made by the best performing machine learning model were combined with predictions made for observations with less than two months with tax arrears. More specifically, for each test set observation, prediction in the final model was made either according to the best machine learning model if there were at least two, or predicted not to have tax arrears next month if there were less than two months with tax arrears among any of the 12 months preceding the prediction. This way, predictions were obtained for all test data, not depending on the number of months with previous tax arrears, which allowed to make the results comparable to previous studies.

For building the machine learning models, four widely used classification methods, which have also worked well in the related area of bankruptcy prediction, were used in this thesis – decision tree (DT), random forest (RF), k-nearest neighbours (KNN) and multilayer perceptron (MLP).

Decision tree is a classification method where decision rules are learnt from the values of the independent variables. The trained model can be represented as a binary tree structure, where classification is based on the predominant class in the leaf node at the end of the decision path. In the learning process, at each iteration the best split is chosen. In this thesis, the criterion used for choosing the best split was Gini impurity, which is calculated as (Flach, 2012):

$$Gini\ impurity\ = 2 \times p \times (1 - p), \tag{3}$$

where p is the proportion of observations belonging to one class among total observations in the node.

Other parameter values of the DT models are provided in Annex 4.2. All three parameters are criteria for stopping the recursive splitting process of nodes (i.e. for pruning the tree). Setting stopping criteria helped to avoid overfitting. The DT algorithm used in scikit-learn is an optimized version of the CART algorithm[16].

[16] Decision trees section in the User Guide of scikit-learn: https://scikit-learn.org/stable/modules/tree.html#tree.

Random forest is a classification method that consists in building a certain number of DT models, each time using only a randomly chosen part of the training data. Classification is then performed using the averaged results of all DT models. Since RF combines results of a number of models, it is considered an ensemble model. Similar to DT models, Gini impurity was also used in RF models as the criterion for choosing the best split. Other parameter values are provided in Annex 4.2.

K-nearest neighbours is a classification method that maps training set observations in the multi-dimensional space and makes the prediction for each test set observation based on k training set observations that are closest to it. The values of k used for the models in this thesis are given in Annex 4.2. The distance measure used in all KNN models was Euclidean distance (Flach, 2012):

$$Euclidean\ distance\ = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}\ , \tag{4}$$

where d is the number of independent variables, and x_i and y_i are the values of the i-th independent variable in training set observations x and y.

Multilayer perceptron is a neural network method. The network consists of an input layer, a number of hidden layers and an output layer. Each layer contains certain number of neurons. The learning process is split into several epochs where the network parameters (weights of the edges between neurons in each layer and the bias term of each neuron) are learnt using back-propagation mechanism. Within each epoch, data is handled in a number of patches. A loss function is used in optimizing the parameter values. In each neuron, dot product of the input vector of values from the previous layer and the weights vector being learnt in that neuron is calculated and the bias term is added to the result. The result then passes an activation function and is thereafter passed on to the neurons in the next layer, until the output layer gives out the result.

The MLP models used in this thesis had three hidden layers, with 4, 4, and 2 neurons, respectively. The parameter values of the models are given in Annex 4.2. All MLP models used Adam optimizer, which, according to Keras documentation[17], is a stochastic gradient descent based optimizer that uses adaptive estimation of first and second order moments.

[17] Adam optimizer in Keras: https://keras.io/api/optimizers/adam/.

The loss measure used in the models was binary cross-entropy, defined as (Ho and Wookey, 2020):

$$Binary\,cross\text{-}entropy = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \times \log\left(h_\theta(x_i)\right)\right) + (1 - y_i) \times \log\left(1 - h_\theta(x_i)\right)], \quad (5)$$

where N is the number of training set observations, y_i is the value of the dependent variable of training set observation i, x_i is input for training set observation i, and h_θ is model with neural network weights θ.

In the hidden layers, ReLu (rectified linear unit) activation function, and in the output layer, sigmoid activation function were used. The formulas of those functions, as provided in Keras documentation[18], are:

$$ReLu(x) = \max(x, 0), \quad (6)$$

$$sigmoid(x) = \frac{1}{1+e^{-x}}\,. \quad (7)$$

Using sigmoid activation function meant that the output of the network was given in the form of probabilities. The probabilities were then converted into dependent variable with value '1' (tax arrears next month) if they exceeded 0.5, and with value '0' (no tax arrears next month) otherwise.

In order to compare performance of different machine leaning methods on different types of independent variables, separate models were built for each of the three types of independent variables described in section 3.3 using each of the four machine learning methods. The encoding of the model names, as well as independent variables included in each model are given in Table 4. The description of the independent variables was provided in Table 2.

The performance of the models was measured based on accuracy. Accuracy is calculated as percentage of correct predictions among all predictions (Chawla 2009). Also, the misclassification rates were calculated, showing the percentage of falsely classified observations among observations which actually had tax arrears next month (type I error), and among observations which actually did not have tax arrears next month (type II error).

[18] Layer activation functions in Keras: https://keras.io/api/layers/activations/.

Table 4. Model names and independent variables in each model (composed by author)

Type of independent variables	Independent variables	Method			
		DT	**RF**	**KNN**	**MLP**
STATS	d first, d last, d med, d std	STATS_DT	STATS_RF	STATS_KNN	STATS_MLP
M5	months 1–5, months 6–9, month 10, month 11, month 12	M5_DT	M5_RF	M5_KNN	M5_MLP
M12	month 1, ..., month 12	M12_DT	M12_RF	M12_KNN	M12_MLP

Cross-validation (1:10) on training set was used for choosing the best parameters for each of the models. The criteria for choosing the model with the best parameters were the arithmetic means of accuracies on cross-validation test sets, as well as minimum overfitting and underfitting. Then models with the best parameters were trained on training set and tested on test set.

The Python machine learning libraries used for building the models were DecisionTreeClassifier[19], RandomForestClassifier[20] and KNeighborsClassifier[21] in scikit-learn[22]. For building MLP modes, Keras[23] neural networks library was used. In Keras, MLP models fall into the category of Sequential model[24].

4. Results and discussion

4.1. Results

The predictive performance of the models is provided in Table 5. The final results are the results on test set. For comparison, results on training set have also been provided.

[19] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.
[20] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.
[21] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html.
[22] Each library's webpage contains a reference to the respective chapter in scikit-learn's User Guide where the method and its implementation are explained in more detail.
[23] Keras API: https://keras.io/api/.
[24] Sequential model in Keras: https://keras.io/guides/sequential_model/.

Table 5. Performance of models (composed by author)

Months with tax arrears	Model	Test set			Training set		
		Accuracy	Type I error	Type II error	Accuracy	Type I error	Type II error
2 or more	STATS_DT	0.8386	0.1487	0.1745	0.8068	0.1988	0.1876
	STATS_RF	0.8392	0.1591	0.1626	0.8077	0.2129	0.1718
	STATS_KNN	0.8372	0.1566	0.1692	0.8087	0.2083	0.1743
	STATS_MLP	0.8377	0.1612	0.1634	0.8054	0.2162	0.1731
	M5_DT	0.8441	0.1607	0.1509	0.8128	0.2173	0.1572
	M5_RF	**0.8446**	**0.1579**	**0.1528**	**0.8133**	**0.2132**	**0.1602**
	M5_KNN	0.8422	0.1504	0.1655	0.8154	0.1993	0.1699
	M5_MLP	0.8424	0.1652	0.1498	0.8115	0.2205	0.1565
	M12_DT	0.8422	0.1586	0.1571	0.8107	0.2109	0.1678
	M12_RF	0.8417	0.1589	0.1577	0.8145	0.2130	0.1580
	M12_KNN	0.8403	0.1556	0.1638	0.8145	0.2057	0.1653
	M12_MLP	0.8422	0.1583	0.1573	0.8127	0.2148	0.1599
0	**Predict no tax arrears next month**	0.9935	1.0000	0.0000	0.9923	1.0000	0.0000
1		0.9265	1.0000	0.0000	0.9331	1.0000	0.0000
0–1		**0.9888**	**1.0000**	**0.0000**	**0.9663**	**1.0000**	**0.0000**
0–12	**FINAL MODEL**	**0.9528**	**0.1982**	**0.0255**	**0.9430**	**0.2476**	**0.0255**

*The models used for the final model are marked in bold.

The best performing machine learning model was random forest trained on monthly tax arrears with aggregation of earlier periods into period means (M5_RF), which prediction accuracy was 84.46%. In general, prediction accuracies of all models were in a similar range (between 83.72% and 84.46%), with the best model only slightly outperforming the second best model (M5_DT), which accuracy was 84.41%. The performance of the best model in classifying observations with and without tax arrears next month was almost equal, with the respective misclassification rates being 15.79% (type I error) and 15.28% (type II error).

As regards types of independent variables, models trained on monthly amounts with aggregation of months 1–5 and 6–9 into period means (M5) performed best with all machine learning methods. Models with this type of independent variables were also the only ones that outperformed models trained on the 12 monthly amounts without aggregation (M12) with all machine learning methods. As regards machine learning methods, random forest could be considered the best, as it outperformed other methods

in case of two types of independent variables (M5 and STATS), while MLP performed best only with one type of independent variables (M12).

The results (see Table 5) show that all machine learning models suffered from underfitting, with accuracies on training set being on average 3.0 percentage points lower than on test set. Underfitting occurs when a model is not complex enough to adequately represent the underlying relationship between the independent variables and the dependent variable (Kelleher *et al*., 2015), which may lead to poor generalization (Hastie *et al*., 2017). The main reasons why underfitting may happen are that either the amount of training data is insufficient or that the independent variables used are insufficient to fully describe the phenomenon that is being predicted. In this case, the former could not have been an issue, as tax arrears of practically the entire population of Estonian SMEs over six years were used for training the models. Therefore, it could only be assumed that historical tax arrears time series alone do not fully explain whether a company will incur tax arrears next month. Instead, there must be other factors besides previous tax arrears history underlying the corporate tax payment behaviour.

It is interesting to note that accuracy of the best machine learning model (84.5%) (see Table 5) was only slightly higher than accuracy of the univariate random forest model where tax arrears in month 12 were used as the single independent variable (80.4%) (see Annex 3). This shows that tax arrears in month 12 have predominant importance in predicting tax arrears next month, since adding other independent variables only increased accuracy to a limited extent.

The accuracy of the approach applied for observations with less than two months with previous tax arrears, which consisted in simply predicting them all not to have tax arrears next month, was 98.88% (see Table 5). Results of this approach in Table 5 are also presented separately for observations with zero and one months with tax arrears, with the respective accuracies being 99.35% and 92.65%. It must be noted, though, that accuracy in this case simply corresponded to the percentage of observations without tax arrears next month.

The accuracy of the final model (see Table 5), where predictions made by the best performing machine learning model (M5_RF) were combined with predictions made for

observations with less than two months with previous tax arrears, was 95.28%, which can be considered excellent. However, as shown by Chawla (2009), accuracy might not be the best performance measure in case the dataset is imbalanced, which the test set as a whole indeed was. Namely, accuracy of the final model was heavily buffed by the overwhelming number of observations that were correctly predicted not to have tax arrears next month. In order to get a more realistic picture of the model's performance, a closer look can be taken at misclassification rates of the final model as presented in Table 5. This shows that the model falsely classified 19.8% of the observations with tax arrears next month as not having them next month (type I error), and 2.5% observations without tax arrears next month as having them next month (type II error). Therefore, the final model was better at classifying observations without tax arrears next month.

4.2. Discussion

The accuracy of the final model presented in this thesis (95.28%) was considerably higher than the accuracies of models in previous studies where financial ratios were used for predicting tax arrears: 73.8% in the study by Höglund (2017) for predicting tax arrears of Finnish firms, 72.4% in the model by Batista *et al*. (2012) for predicting tax arrears of Portuguese real estate agencies, and 61.6% in the model by Marghescu *et al.* (2010) for predicting the specific case of arrears in employer contribution taxes in Finnish firms. Performance of the final model also considerably outperformed the pattern mining models presented in the study by Zhao *et al.* (2009), where activities in taxpayer database were used as independent variables for predicting future social security debts (where accuracy of the best model was 76.0%).

Therefore, this thesis shows that using monthly tax arrears and monthly predictions enables to predict future tax arrears with remarkably higher accuracy than using financial ratios and annual predictions. While having predictions for a more immediate future than a year would seem more useful for practical purposes, the question of whether it would be easier or harder to predict tax arrears for more than one month ahead using monthly tax arrears still remains to be explored. However, very likely the main reason for the accuracy of the final model being considerably higher than accuracies in previous studies is that financial ratios are not the most appropriate predictors of tax arrears. First, because

they become available with a considerable time lag after the tax payment irregularities have already been going on for some time. Secondly, because they can result from temporary difficulties that are overcome in course of the financial year and therefore never reflected in financial statements. Also, besides being a sign of financial distress, tax arrears might also reflect certain behavioural aspects of a company's financial management. For example, it is possible that some companies otherwise in good financial standing might use tax arrears as a form of short-time credit.

The performance of the final model was also slightly better than the already quite high performance of the ensemble model by Su *et al.* (2018), which to the best knowledge of the author is the only study where previous tax arrears have been used for predicting tax arrears in the future (accuracy of 90.58%). Contrary to this thesis, in that study the prediction was annual instead of monthly, and tax arrears were used only as one among many independent variables. Since also classification into three classes (no tax arrears, or tax arrears below or above certain threshold) was used in that study, its accuracy is not directly comparable to the accuracy in this thesis.

The overall accuracy of the final model was also higher than the model in (Lukason and Andresson, 2019), where tax arrears in 12 months (using independent variables corresponding to 'd max', 'd med', 'd m in debt' and 'd longest' in this thesis), were used for bankruptcy prediction (accuracy 89.5%). However, the model in this thesis had lower type II error rate (i.e. it was better at classifying observations that will not have tax arrears than the model in their study was at classifying companies that will not go bankrupt), but higher type I error rate (i.e. it was worse at classifying observations that will have tax arrears than the model in their study was at classifying companies that will go bankrupt). A possible explanation for the higher type I error rate in this thesis is that prior to bankruptcy, the indebtedness of a company, including towards the government, has grown more severe and therefore the tax arrears patterns in 12 months preceding bankruptcy are more pronounced and easier to detect than the ones preceding any particular next month with tax arrears. Also, companies that will end in bankruptcy often have been *de facto* insolvent already quite some time before the bankruptcy proceedings are finally launched. This makes predicting bankruptcies easier, since insolvency is likely to be also reflected in the financial ratios. Having tax arrears, on the contrary, is often a

temporary situation that is more easily reversible, which makes predicting monthly tax arrears more difficult.

The best performing machine learning method in this thesis was random forest, which falls into the category of ensemble methods. Therefore results in this thesis are in line with the observation made by Domingos (2012) that ensemble models often provide better results than single models. An example from bankruptcy prediction where also random forest was found to be the best performing method is a study by Barboza *et al.* (2017), where it outperformed all other methods, which included SVM, neural networks, logit, MDA, bagging and boosting.

5. Conclusions

The aim of this thesis was to explore which machine learning methods and types of independent variables are most useful in predicting companies to have tax arrears next month, given the time series of their tax arrears in the preceding 12 months. The data were monthly tax arrears of Estonian SMEs in 2011–2018.

A specific characteristic of tax arrears is that they are rare events, showing that companies usually pay their taxes in time. Since learning from mostly zero-valued data is a difficult task for machine learning models, the approach in this thesis was to build those models only for observations which had tax arrears in at least two among any of the 12 preceding months, while the rest of the observations were always predicted not to have tax arrears next month. The approach was justified because among observations with previous tax arrears in less than two months (85% of the data), the probability of tax arrears next month was very low (1.12%) and, due to all or nearly all monthly figures being zero, they were difficult to predict. This approach succeeded in reducing zero values in the dataset from 91.82% to 48.84% and resulted in a nearly balanced dataset.

The machine learning methods used were decision tree (DT), random forest (RF), k-nearest neighbours (KNN) and multilayer perceptron (MLP). With each of those methods, models were built using three alternative types of independent variables: 12 monthly amounts of tax arrears, statistical measures and counts of events, and monthly amounts with aggregation of months 1–5 and 6–9 into period means.

The best performing model was random forest trained on monthly tax arrears with aggregation of months 1–5 and 6–9 into period means (accuracy 84.46%), where the months to aggregate were chosen based on Gini importances of the 12 monthly amounts. The usefulness of such aggregation approach was further confirmed by all machine learning methods performing best with this type of independent variable. In general, prediction accuracies of all models were in a similar range (between 83.72% and 84.46%).

Results showed that accuracy of the best model was only slightly higher than accuracy of the univariate random forest model where tax arrears in month 12 were used as the single independent variable (80.42%). This indicated that tax arrears in month 12 have predominant importance in predicting tax arrears next month, since adding other independent variables only increased accuracy to a limited extent.

In the final model, predictions made for observations with less than two months with previous tax arrears, which were all simply predicted not to have tax arrears next month, were added to the predictions made by the best machine learning model. This way, predictions were obtained for all test data with any number of months with previous tax arrears, in order to make results comparable to previous studies. The accuracy of the final model was 95.28%, which could be considered excellent. The model was better at correctly predicting a company not to have tax arrears next month, with the percentage of false classifications among observations without tax arrears next month being only 2.5% (type II error), while among observations with tax arrears next month it was 19.8% (type I error). Although from practical perspective, identifying companies which will have tax arrears next month seems more important, the performance of the final model in identifying them can still be considered quite good.

This thesis represents the first attempt to predict corporate tax arrears based on the historical monthly time series of previous tax arrears. While there have been a handful of studies where tax arrears have been predicted based on financial ratios or annual tax arrears among other independent variables, using data with monthly instead of annual frequency has much higher practical value. This is because in carrying out their daily activities, tax authorities would greatly benefit from being able to detect companies likely to incur tax arrears not only once a year and not only for the entire next year, but at any time and for a more immediate future, using the most recent information available.

This thesis has high practical value, since the proposed approach could enable tax authorities to better target their tax audits to companies that are likely to default on their corporate tax obligations, and better focus preventive measures aimed at ensuring the timely payment of taxes.

6. Future research

This thesis sets ground for future research in the field of corporate tax prediction, especially for studies where monthly tax arrears data are used among the independent variables. A major challenge for future research would be to find suitable additional independent variables that could improve classification performance, given that underfitting of the models in this thesis indicated that historical tax arrears alone do not fully explain whether a company will incur tax arrears next month.

For example, in their model for predicting tax arrears next year, Su *et al.* (2018) used, in addition to tax arrears in the previous year, 17 financial statements items as well as dummies for industry, region, registration type, type of accounting system and taxpayer status as independent variables. It could be interesting to explore if any of such variables would improve the performance of models built on monthly data. Also, variables related to a company's management, which have already been proven to improve accuracy of bankruptcy prediction models (see, for example, Ciampi (2015) and Back (2005)), might also prove useful in tax arrears prediction.

The approach used in this thesis was to gather all available 12 months periods of all companies into a single dataset, without distinguishing between the companies nor the years. The disadvantages of this approach are that models built on such dataset can only discover patterns that are generally applicable to all companies, and that multiannual dynamics are not taken into account. Therefore, future research could explore possibilities for building multiannual models, or separate models for each company, or combining patterns discovered for each company with general patterns applicable to all companies.

The practical value of the proposed models could be further enhanced if they were used as an input to developing models where in addition to tax arrears, also tax payments are taken into account. For practical purposes, it might also be interesting to develop models

that instead of predicting whether there will be tax arrears next month, predict the probability of tax arrears next month. Such models could be then developed further to predict the amount of tax arrears next month for cases where the probability of tax arrears next month exceeds certain threshold. Also, future research could explore possibilities for predicting the occurrence of tax arrears different numbers of months ahead, instead of predicting it only for next month.

References

1. Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications,* 94, 164–184. doi: 10.1016/j.eswa.2017.10.040.
2. Alessandretti, L., Baronchelli, A., He, Y.-H. (2019). Machine Learning meets Number Theory: The Data Science of Birch-Swinnerton-Dyer. arXiv:1911.02008v1 [math.NT], https://arxiv.org/pdf/1911.02008.pdf.
3. 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. doi: 10.1016/j.bar.2005.09.001.
4. Back, P. (2005). Explaining financial difficulties based on previous payment behavior, management background variables and financial ratios. *European Accounting Review,* 14(4), 839–868. doi: 10.1080/09638180500141339.
5. Barboza, F., Kimura, H., Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications,* 83, 405–417. doi: 10.1016/j.eswa.2017.04.006.
6. Batista, J., Cerqueira, A., Brandão, E. F. M. (2012). Modeling Corporate Tax Risk: Evidence from Portugal. Available at SSRN: https://ssrn.com/abstract=2179068. doi: 10.2139/ssrn.2179068.
7. Chawla, N. V. (2009). Data Mining for Imbalanced Datasets: an Overview. *In: Maimon O., Rokach L. (eds) "Data Mining and Knowledge Discovery Handbook"*. Springer, Boston, MA. Chapter 40, 853–867.
8. Chen, N., Ribeiro, B., Chen, A. (2016). Financial credit risk assessment: a recent review. *Artificial Intelligence Review,* 45, 1–23. doi: 10.1007/s10462-015-9434-x.
9. Cho, S., Hong, H., Ha, B.-C. (2010). A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction. *Expert Systems with Applications,* 37(4), 3482–3488. doi: 10.1016/j.eswa.2009.10.040.
10. Ciampi, F. (2015). Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms. *Journal of Business Research,* 68(5), 1012–1025. doi: 10.1016/j.jbusres.2014.10.003.

11. Ciampi, F., Cillo, V., Fiano, F. (2020). Combining Kohonen maps and prior payment behavior for small enterprise default prediction. *Small Business Economics, 54,* 1007–1039. doi: 10.1007/s11187-018-0117-2.
12. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM,* 55(10), 78–87. doi: 10.1145/2347736.2347755.
13. du Jardin, P. (2017). Dynamics of firm financial evolution and bankruptcy prediction. *Expert Systems with Applications,* 75, 25-43. doi: 10.1016/j.eswa.2017.01.016.
14. Flach, P. (2012). Machine Learning: The Art and Science of Algorithms that Make Sense of Data. New York: Cambridge University Press.
15. Hastie, T., Tibshirani, R., Friedman, J. (2017). The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer series in Statistics. Second edition. Springer.
16. Ho, Y., Wookey, S. (2020). The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. *IEEE Access,* 8, 4806–4813. doi: 10.1109/ACCESS.2019.2962617.
17. Höglund, H. (2017). Tax payment default prediction using genetic algorithm-based variable selection. *Expert Systems With Applications,* 88, 368–375. doi: 10.1016/j.eswa.2017.07.027.
18. Jayasekera, R. (2018). Prediction of company failure: Past, present and promising directions for the future. *International Review of Financial Analysis,* 55, 196–208. doi: 10.1016/j.irfa.2017.08.009.
19. Karan, M. B., Ulucan, A., Kaya, M. (2013). Credit risk estimation using payment history data: a comparative study of Turkish retail stores. *Central European Journal of Operations Research,* 21, 479–494. doi: 10.1007/s10100-012-0242-y.
20. Kelleher, J. D., Mac Namee, B., D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics. Algorithms, Worked Examples, and Case Studies. Cambridge, Massachusetts, London, England: The MIT Press.
21. Kubicová, J., Faltus, S. (2014). Tax Debt as an Indicator of Companies' Default: the Case of Slovakia. *Journal of Applied Economics and Business,* 2(4), 59–74. doi: 10.2139/ssrn.2543257.
22. Kukalová, G., Moravec, L., Bína Filipová, D., Bařtipán, M. (2020). Success Rate of Tax Arrears Recovery: Czech Republic Case Study. *HED - Hradec Economic Days*

2020, Konference Hradec Economic Days 2020. doi: 10.36689/uhk/hed/2020-01-047.

23. Lukason, O., Andresson, A. (2019). Tax Arrears Versus Financial Ratios in Bankruptcy Prediction. *Journal of Risk and Financial Management,* 12(4), 187, 1–13. doi: 10.3390/jrfm12040187.
24. Marghescu D., Kallio M., Back B. (2010). Using Financial Ratios to Select Companies for Tax Auditing: A Preliminary Study. *In: Lytras M.D., Ordonez de Pablos P., Ziderman A., Roulstone A., Maurer H., Imber J.B. (eds) Organizational, Business, and Technological Aspects of the Knowledge Society. WSKS 2010. Communications in Computer and Information Science,* vol 112. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-16324-1_45.
25. Ravi Kumar, P., Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research,* 180(1), 1–28. doi: 10.1016/j.ejor.2006.08.043.
26. Shi, Y., Li, X. (2019). An overview of bankruptcy prediction models for corporate firms: A systematic literature review. *Intangible Capital,* 15(2), 114–127. doi: 10.3926/ic.1354.
27. Su, A., He, Z., Su, J., Zhou, Y., Fan, Y., Kong, Y. (2018). Detection of Tax Arrears Based on Ensemble Leaering Model. *2018 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Chengdu,* 270–274. doi: 10.1109/ICWAPR.2018.8521362.
28. Sun, J., Li, H., Huang, Q.-H., He, K.-Y. (2014a). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems,* 57, 41–56. doi: 10.1016/j.knosys.2013.12.006.
29. Sun, J., Shang, Z., Li, H. (2014b). Imbalance-oriented SVM methods for financial distress prediction: a comparative study among the new SB-SVM-ensemble method and traditional methods. *Journal of the Operational Research Society,* 65, 1905–1919. doi: 10.1057/jors.2013.117.
30. Veganzones, D., Severin, E. (2020). Corporate failure prediction models in the twenty-first century: a review. *European Business Review* (ahead-of-print). doi: 10.1108/EBR-12-2018-0209.

31. Zhao Y., Zhang, H., Wu, S., Pei, J., Cao, L., Zhang, C., Bohlscheid, H. (2009). Debt Detection in Social Security by Sequence Classification Using Both Positive and Negative Patterns. *In: Buntine W., Grobelnik M., Mladenić D., Shawe-Taylor J. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2009. Lecture Notes in Computer Science*, 5782, 648–663. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-04174-7_42.

Annex 1. Descriptive statistics of independent variables (composed by author)

Type	Independent variable	All					Tax arrears next month					No tax arrears next month				
		Mean	Median	Min	Max	St.dev.	Mean	Median	Min	Max	St.dev.	Mean	Median	Min	Max	St.dev.
Counts of events and statistical measures	d first	3.2	1.0	0.0	12.0	4.3	5.6	4.0	0.0	12.0	4.6	0.8	0.0	0.0	12.0	2.1
	d last	1.7	0.0	0.0	10.0	2.6	0.4	0.0	0.0	10.0	1.3	3.0	2.0	0.0	10.0	2.9
	d m in debt	6.2	6.0	2.0	12.0	3.6	8.1	9.0	2.0	12.0	3.4	4.4	3.0	2.0	12.0	2.6
	d longest	5.0	4.0	1.0	12.0	3.7	6.7	6.0	1.0	12.0	3.9	3.3	2.0	1.0	12.0	2.6
	d med	3196	2	0	1426784	20624	5566	741	0	1426784	26413	818	0	0	1404432	11863
	d mean	3771	574	0	1385285	20179	6138	1302	0	1385285	26213	1395	234	0	1355879	10721
	d max	8466	2236	1	1508237	31687	11980	3677	1	1508237	38105	4939	1253	1	1426827	23013
	d std	2420	688	0	707599	8988	3199	1075	0	497611	9415	1638	400	0	707599	8467
Amounts with aggregation	months 1–5	3364	348	0	1483347	20159	4992	716	0	1483347	24273	1730	164	0	1421714	14749
	months 6–9	3862	419	0	1426034	21517	6309	1124	0	1383971	27564	1406	122	0	1426034	12372
	month 10	4202	153	0	1441174	23523	7355	1304	0	1441174	31105	1038	0	0	1404432	10844
	month 11	4357	177	0	1470732	24072	7855	1532	0	1470732	32319	845	0	0	1404432	9384
	month 12	4424	164	0	1508237	24322	8251	1704	0	1508237	33184	583	0	0	1404432	7109
Amounts without aggregation*	month 1	3142	0	0	1502340	20284	4557	236	0	1502340	23755	1723	0	0	1426827	15937
	month 2	3274	0	0	1502340	20779	4753	310	0	1502340	24411	1789	0	0	1426827	16206
	month 3	3364	0	0	1502340	20962	4969	388	0	1502340	24925	1754	0	0	1426827	15864
	month 4	3469	5	0	1499340	21383	5209	473	0	1499340	25414	1723	0	0	1426827	16180
	month 5	3570	22	0	1496340	21538	5472	556	0	1496340	26057	1661	0	0	1426827	15516
	month 6	3674	42	0	1476340	21717	5770	663	0	1476340	26690	1569	0	0	1426827	14871
	month 7	3791	67	0	1426827	22011	6104	778	0	1426740	27562	1470	0	0	1426827	14053
	month 8	3924	97	0	1426827	22547	6477	927	0	1426827	28756	1362	0	0	1426827	13247
	month 9	4058	124	0	1426827	23058	6883	1091	0	1426827	29949	1223	0	0	1423743	12204

* The descriptive statistics for the last three months are the same as under "Amounts with aggregation" and have not been repeated.

Annex 2. Correlation matrix of independent variables (composed by author)

	d m in debt	**d longest**	**d first**	**d last**	**d mean**	**d med**	**d max**	**d std**	**months 1-5**	**months 6-9**	**month 10**	**month 11**	**month 12**
d m in debt	1.000	0.929	0.784	-0.502	0.196	0.205	0.158	0.106	0.181	0.195	0.178	0.169	0.165
d longest	0.929	1.000	0.841	-0.392	0.219	0.226	0.180	0.122	0.202	0.221	0.199	0.189	0.183
d first	0.784	0.841	1.000	-0.492	0.213	0.218	0.172	0.106	0.173	0.214	0.218	0.218	0.220
d last	-0.502	-0.392	-0.492	1.000	-0.084	-0.087	-0.072	-0.052	-0.044	-0.090	-0.107	-0.114	-0.120
d mean	0.196	0.219	0.213	-0.084	1.000	0.974	0.896	0.628	0.941	0.973	0.909	0.885	0.858
d med	0.205	0.226	0.218	-0.087	0.974	1.000	0.809	0.522	0.922	0.967	0.863	0.828	0.801
d max	0.158	0.180	0.172	-0.072	0.896	0.809	1.000	0.878	0.829	0.855	0.835	0.836	0.825
d std	0.106	0.122	0.106	-0.052	0.628	0.522	0.878	1.000	0.580	0.595	0.592	0.595	0.576
months 1–5	0.181	0.202	0.173	-0.044	0.941	0.922	0.829	0.580	1.000	0.874	0.749	0.717	0.694
months 6–9	0.195	0.221	0.214	-0.090	0.973	0.967	0.855	0.595	0.874	1.000	0.900	0.851	0.817
month 10	0.178	0.199	0.218	-0.107	0.909	0.863	0.835	0.592	0.749	0.900	1.000	0.928	0.879
month 11	0.169	0.189	0.218	-0.114	0.885	0.828	0.836	0.595	0.717	0.851	0.928	1.000	0.936
month 12	0.165	0.183	0.220	-0.120	0.858	0.801	0.825	0.576	0.694	0.817	0.879	0.936	1.000

Annex 3. Univariate prediction accuracies of independent variables (composed by author)

Type	Independent variable	Decision tree	Random forest	KNN	MLP	Avg. accuracy
Counts of events and statistical measures	d first	0.7872	0.7872	0.7845	0.7889	0.7773
	d last	0.7845	0.7845	0.7845	0.7985	0.7698
	d m in debt	0.7203	0.7203	0.7144	0.6931	0.7028
	d longest	0.7012	0.7028	0.6918	0.6806	0.6843
	d med	0.7291	0.7289	0.7288	0.7077	0.7131
	d mean	0.6906	0.6912	0.6910	0.7162	0.6922
	d max	0.6464	0.6462	0.6597	0.6777	0.6546
	d std	0.6362	0.6361	0.6366	0.6671	0.6428
Amounts with aggre-gation	months 1-5	0.6214	0.6214	0.5979	0.5986	0.6068
	months 6-9	0.6797	0.6797	0.6752	0.6817	0.6706
	month 10	0.7269	0.7269	0.7053	0.7255	0.7090
	month 11	0.7632	0.7632	0.7623	0.7689	0.7573
	month 12	0.8042	0.8042	0.8031	0.8176	0.7965

Annex 4.1. Parameters used in univariate models (composed by author)

Method	Parameter	Value
DT	Minimum number of samples in leaf	1,000
	Maximum depth of the tree	4
RF	Minimum number of samples in leaf	100
	Maximum depth of the tree	4
KNN	Number of neighbours	61
	Distance measure	Euclidean distance
MLP	No. of neurons in hidden layers	3, 2
	Activation function in hidden layers	ReLu
	Activation function in output layer	Sigmoid
	Loss function	Binary cross-entropy
	Optimizer	Adam
	Learning rate	0.001
	Number of epochs	15
	Batch size	100
	Validation split	0.05

Annex 4.2. Parameters used in multivariate models (composed by author)

Method	Parameter	Type pf independent variables		
		STATS	M5	M12
DT	Minimum No. of samples in leaf	250	300	500
	Maximum depth of the tree	6	7	7
	Minimum impurity decrease	0.0005	0.00002	0.00002
RF	Minimum No. of samples in leaf	70	25	80
	Maximum depth of the tree	5	6	6
	Minimum impurity decrease	0.00002	0.00003	0.00001
	Estimators (i.e. number of trees)	200	150	200
KNN	Number of neighbours	175	105	101
	Distance measure	Euclidean distance	Euclidean distance	Euclidean distance
MLP	No. of neurons in hidden layers	4, 4, 2	4, 4, 2	4, 4, 2
	Activation function in hidden layers	ReLu	ReLu	ReLu
	Activation function in output layer	Sigmoid	Sigmoid	Sigmoid
	Loss function	Binary cross-entropy	Binary cross-entropy	Binary cross-entropy
	Optimizer	Adam	Adam	Adam
	Learning rate	0.0002	0.0002	0.0002
	Number of epochs	30	30	30
	Batch size	70	70	70
	Validation split	0.05	0.05	0.05

Non-exclusive licence to reproduce thesis and make thesis public

I, Õie Renata Siimon,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

"Corporate Tax Arrears Prediction Based on Monthly Time Series of Tax Arrears",

supervised by Oliver Lukason.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Õie Renata Siimon
11.08.2020