sustainability

MDPI

Article

A Decision Support System for Corporate Tax Arrears Prediction

Õie Renata Siimon and Oliver Lukason *

School of Economics and Business Administration, University of Tartu, 51009 Tartu, Estonia; renata.siimon@ut.ee
* Correspondence: oliver.lukason@ut.ee

Abstract: This paper proposes a decision support system to predict corporate tax arrears by using tax arrears in the preceding 12 months. Despite the economic importance of ensuring tax compliance, studies on predicting corporate tax arrears have so far been scarce and with modest accuracies. Four machine learning methods (decision tree, random forest, k-nearest neighbors and multilayer perceptron) were used for building models with monthly tax arrears and different variables constructed from them. Data consisted of tax arrears of all Estonian SMEs from 2011 to 2018, totaling over two million firm-month observations. The best performing decision support system, yielding 95.3% accuracy, was a hybrid based on the random forest method for observations with previous tax arrears in at least two months and a logical rule for the rest of the observations.

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

Citation: Siimon, Õ.R.; Lukason, O. A Decision Support System for Corporate Tax Arrears Prediction. *Sustainability* **2021**, *13*, 8363. https://doi.org/10.3390/su13158363

Academic Editors: German Gemar and Alberto A. Lopez-Toro

Received: 28 June 2021
Accepted: 23 July 2021
Published: 27 July 2021

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 implement 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. In the machine learning domain, more attention has been ongoingly directed to the detection of tax fraudsters (e.g., [1–4]).

The main drawbacks of current studies (e.g., [5–7]) are that they have mostly concentrated on using financial ratios as predictors of tax arrears, and have only proposed models for predicting tax arrears for the following 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. In addition, they cannot be used if financial reports are unavailable, which is much more likely to happen in the case of financially distressed firms [8,9], which in turn are also more likely to have tax arrears [10]. In addition, the accuracy of models using financial ratios has been moderate. The disadvantage of annual predictions is that they can only be made once a year, and they only predict if a company will have tax arrears any time in the following year, which seems rather vague for practical purposes.

To the best of the authors' knowledge, there are no studies where the behavior of tax arrears in the past has been applied to predict future tax arrears. Thus, this paper intends to fill this research gap by applying monthly time series of corporate tax arrears for predicting tax arrears in the next month. 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 following year, but at any time and in the more immediate future, using the most recent information available. Besides outlining the practical applicability, this paper contributes to the financial management literature by showing whether and how past payment defaults signal future payment defaults.

The aim of this paper is to explore which machine learning methods and types of independent variables work best in predicting companies to have tax arrears in the 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 are considered. One of those types includes statistical measures and counts of events, and the other includes monthly amounts with the aggregation of amounts in earlier months into period means. The machine learning methods used are decision tree (DT), random forest (RF), k-nearest neighbors (KNN) and multilayer perceptron (MLP), which have also been applied in the related area of failure prediction (see, for example, [11–13]). The two areas are related, since defaulting on taxes is a strong sign of financial distress, which in turn might eventually lead to bankruptcy [6].

Data used for this paper 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-13 month period pairs, were collected using the 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 paper 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 in the next month were very unlikely to occur and difficult to predict, were always predicted not to have tax arrears in the next month. Thus, two types of accuracies are reported in this paper: (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 accuracy of type (a) expresses 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 the Keras neural networks library for the MLP model, scikit-learn machine learning libraries for other models and SciPy library for statistical tests.

The rest of the paper is organized as follows. The literature review is provided in Section 2. Section 3 contains a description of the dataset, variables and methodology. The results, accompanied by the discussion, are presented in Section 4, and conclusions in Section 5.

2. Literature Review

The literature focusing on corporate taxes is multifaceted, spanning over finance, accounting, economics, ethics and other domains (see, e.g., [14]). The topic of taxes is also closely interconnected to economic sustainability, as firms with tax arrears are more likely to be insolvent and engage in law violations [15,16]. Bibliometric analyses indicate that economic sustainability remains among the key topics in the literature domain of firms' sustainable development [17,18].

Despite the high economic importance of ensuring tax compliance, studies on predicting corporate tax arrears have so far been scarce. At the same time, there is more abundant recent literature focusing on predicting the tax revenues of governments or tax fraud committed by firms. Nevertheless, these topics are very different, as tax arrears prediction studies usually forecast a binary variable (occurrence of default), the roots of which are in the lack of liquidity [10]. Tax revenue prediction studies, on the contrary, usually forecast continuous variables with predictors portraying growth [19], while tax fraud studies forecast a rare binary event, the roots of which often lie in causes other than poor liquidity [20]. In previous studies on tax arrears prediction, the focus has been on using

companies' financial statements data (in most cases in the form of financial ratios) for tax arrears prediction. Contrary to this paper, all those studies make predictions for the next year, instead of the next month. An overview of previous studies is provided in Table 1.

There have been a handful of studies where the presence of tax arrears has been predicted based on financial ratios. For example, Marghescu et al. [5] 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 [6] 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%. Höglund's [6] dataset was also applied in Abedin et al. [21], with a rich selection of machine learning tools, with accuracy remaining in a comparative range.

Batista et al. [7] 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 the Taxation Effective Rate, which is an indicator associated with tax evasion. The accuracy of all their models was rather similar, with the accuracy level of 72.4% achieved by the discriminant analysis model built for the year 2008 being the best.

The abovementioned four studies reveal that the performance of financial ratios as predictors of tax arrears is rather low. The likely reason for this is that the annual reports 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. The latter argument has been proven by Lukason and Laitinen [22], who indicated that around three quarters of European bankrupt firms witness financial problems portrayed through financial ratios only very shortly before bankruptcy. Namely, when financial problems occur, they might be detectable only through the last annual report, for which, in all jurisdictions, there is a submission time lag set in laws. Thus, the empirically validated theoretical concept by Lukason and Laitinen [22] indicates that payment defaults (including tax arrears) are very likely to occur in circumstances when the available annual report does not signal any financial problems. In addition, it could be assumed that for predicting tax arrears with monthly frequency, as is done in this paper, using solely financial ratios would be even more difficult, because the presence of tax arrears may change monthly, while financial ratios only change annually in the unlisted firm segment.

Another disadvantage of using financial ratios for predicting tax arrears is that the resulting models cannot be used for cases where financial reports are unavailable. This, however, is much more likely to happen in the case of financially distressed firms, which are therefore also more likely to have tax arrears [9,10]. For example, when predicting tax arrears based on financial ratios, Höglund [6] 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. [23], 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 the activity codes of 155 possible activities of ca. 10,000 taxpayers in the Australian tax database, with each activity code being accompanied by

a taxpayer ID and the date and time of the activity. They constructed ca. 16,000 activity sequences from that data with the aim of predicting which sequences lead to social security debts. The accuracy of their best classifier was moderate (76.0%). Additionally, "debt" in their study had a different meaning than "tax arrears" in this paper—instead of corporate 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 that they solved had some similarities to fraud discovery tasks.

Table 1. Previous papers on corporate tax arrears prediction.

Paper	Data	Obs.	Method	Independent Variables	Dependent Variable	Accuracy
Marghescu et al. (2010) [5]	Finland (2004)	328	Logit	Financial ratios	Tax arrears in employer contribution taxes next year (binary)	61.6%
Höglund (2017) [6]	Finland (2012–2014)	768	Linear discriminant analysis (LR)	Financial ratios, bankruptcy risk and payment default risk in industry	Tax arrears next year (binary)	73.8%
Batista et al. (2012) [7]	Portugal (2007–2009)	600	Discriminant analysis and logit	Financial ratios, Taxation Effective Rate	Tax arrears next year (binary)	72.4%
Abedin et al. (2020) [21]	Finland (2012–2014)	768	Logit, LR and 11 machine learning methods	Financial ratios, bankruptcy risk and payment default risk in industry	Tax arrears same year (binary), tax arrears next year (binary)	73.2/71.7%
Su et al. (2018) [24]	China (2015–2016)	120,000	Ensemble model (KNN, MLP and four tree-based models)	17 balance sheet and income statement items, amount of tax arrears, industry, region, taxpayer status, type of registration and accounting system	Tax arrears next year (three classes: tax arrears above or below 5000 RMB, or no tax arrears)	90.6%

Finally, Su et al. [24] built an ensemble classification model composed of k-nearest neighbors, multilayer perceptron and several tree-based algorithms (random forest, extremely randomized trees, gradient tree boosting and XGBoost), using data of 70,000 Chinese companies for training and 50,000 for testing to predict the presence of tax arrears in the next year. To the best of the authors' knowledge, this is the only study where some tax-related variables have been used to predict tax arrears in the future. However, in that study, the majority of independent variables consisted of financial variables originating from balance sheets and income statements. The accuracy of the proposed model was excellent (90.58%). Contrary to this paper, they used classification into three classes (no tax arrears, and tax arrears below or above the threshold of 5000 RMB) instead of binary classification, and 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 [7], 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 [6]. 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. [25], the 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 bankruptcy might be relevant. For example, Lukason and Andresson [10] compared the performance of tax arrears and financial ratios in bankruptcy prediction, and they 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 paper.

In their study, Kubicová and Faltus [26] 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 paper uses previous tax arrears, studies concerning previous payment behavior might also be relevant. In this regard, there have been a few studies where previous payment behavior has been used for predicting company defaults or credit risk. For example, Ciampi et al. [27] 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. [28] used independent variables such as the proportion of invoices paid late among all invoices, sum of days paid before deadline, total debt/total purchases and average amount paid, among other independent variables, for predicting the credit risk that retailers pose for a wholesaler. Finally, Back [29] used, inter alia, independent variables such as numbers of payment disturbances and delays for predicting the financial difficulties of firms. The study by Back [29] also indicates that firms with current payment defaults are likely to have previous payment defaults as well. The latter could be subject to the poor financial management practices inherent to many SMEs [30], while a more general financial explanation could be that these firms are inefficient, struggling with constant liquidity problems and, more broadly, for survival [31,32]. Thus, the findings of previous studies in the financial domain would lend support to the idea proposed in this paper, that the presence or absence of tax arrears in the past could possess value in predicting the same phenomenon in the future.

When choosing the machine learning methods to be used in this paper, methods that have previously been applied in the related area of bankruptcy prediction were considered. According to Veganzones and Severin [33], these 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., [34]) place hazard models and neural networks in separate categories. When comparing recent trends, Veganzones and Severin [33] 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 [35], 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 the predictive performance by combining the models in different ways.

Based on a review by Shi and Li [36] 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 the bankruptcy prediction literature already before 2007, while methods such as random forest, Adaboost, particle swarm optimization, naïve bayes and k-nearest neighbors (KNN) have appeared only after 2007. According to du Jardin [37], 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 [13]. For example, Alaka et al. [11] found that across bankruptcy prediction articles published in 2010–2015, the average accuracy levels 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 the absence of multicollinearity between independent variables [38], and probit assumes cumulative normal distribution [39]. On the other hand, machine learning methods have the advantage that they can deal with non-linear distributions and do not have stringent assumptions on the data [33]. For the reasons above, traditional statistical methods were not used in this paper.

As regards the predictive performance of machine learning methods, there is no consensus on which one of them performs best in bankruptcy prediction [13], since no method performs consistently better than all others across different datasets [40]. Therefore, it was not possible to choose the methods for this paper based on which methods have been established as best-performing in bankruptcy prediction.

The methods chosen for this paper included decision tree (DT), k-nearest neighbors (KNN), multilayer perceptron (MLP) and random forest (RF), where DT and KNN are conventional machine learning methods, MLP is a neural network method and RF is an ensemble method. While DT and neural networks (along with SVM) rank among the three most widely used machine learning methods in bankruptcy prediction [11], RF and KNN have appeared in the literature of this research field only recently [36]. SVM was not used in this paper, since, according to the information in the standard scikit-learn library for SVM, 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 paper can handle well data where classes are not linearly separable, which is also the case with tax arrears.

3. Data, Variables and Methodology

3.1. Data

The dataset used for this paper 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 paper, only the end-of-month figures were used, partly because figures for other dates were not available for all the years. In addition, using only end-of month figures allowed us to disregard cases of less economic importance, where taxes were paid just a few days late. In Estonia, it is very usual that tax arrears lasting a few days occur, which are more subject to technical or negligence causes, rather than portraying a temporary liquidity crisis [10].

In order to increase the homogeneity in the data, only SMEs (by the European Union's definition) that were going concerns and had at least a minimal level of economic activity (being VAT liable with at least 16,000 EUR annual turnover) at the time their tax arrears were recorded. 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. Finally, data on public entities and NGOs were also left out.

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 a 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 were available for each company). The advantage of the moving window approach was that it allowed us to capture the dynamics of tax arrears in the 12 months preceding the prediction while using a large amount of data. Based on Lukason and Andresson [10], it could be assumed that long-horizon prediction (i.e., spanning several years) of future tax arrears with previous tax arrears is not applicable. The latter argument is proven in the empirical section as well.

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. It could also be observed that tax arrears have a tendency to persist. Namely, the larger the number of month-ends with tax arrears during the preceding 12 months, the more likely a company was to have tax arrears also in the next month (see Figures 1 and 2). At the same time, the proportion of companies having a certain number of previous month-ends with tax arrears decreased with each additional month with tax arrears. Thus, on the one extreme, there were companies with no previous tax arrears, making up as much as 76.9% of the observations, for which the probability of having tax arrears in the next month was only 0.8%. On the other extreme, there were companies with tax arrears in all 12 preceding months, which made up only 1.8% of the dataset, but for which the probability of tax arrears in the next month was 93.0%. The rest of the observations lied in between, with the proportion of the respective observations in the dataset decreasing and the probability of tax arrears in the next month increasing with each additional month with previous tax arrears.

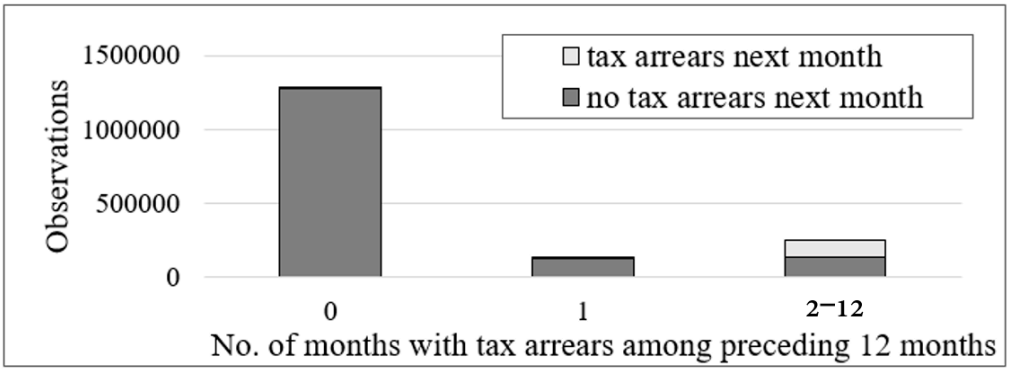

Figure 1. Tax arrears status next month given the number of months with tax arrears (0, 1 or 2–12) among preceding 12 months.

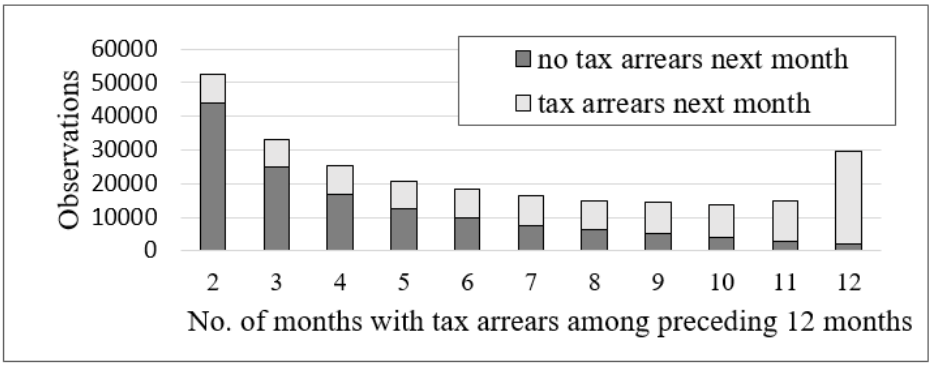

Figure 2. Tax arrears status next month given the number of months with tax arrears (2–12) among preceding 12 months.

While the large proportion of zero values can be explained with companies trying to avoid indebtedness towards tax authorities, 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 [41].

The approach used to reduce data sparsity was to only consider observations with tax arrears in at least any two months during the 12-month 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 in the next month was very low (1.33%), were always predicted not to have tax arrears in the 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. In addition, 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 in the next month if it already had incurred tax arrears at least twice during the 12 preceding months.

Twelve-month periods starting in any month in 2011–2016 (with dependent variable in 2012–2017) were used as the training set and those starting in 2017 (with dependent variable in 2018) were used as the test set. Using different periods for training and testing ensured independence of the training and test data. 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 in the next month, are presented in Table 2.

Table 2. Sizes of parts of the dataset and percentage of observations with tax arrears next month in each.

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 [42]. As shown in Table 2, 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 the training set having tax arrears in the next month. Machine learning models tend to perform better

if they are trained on balanced datasets [42], 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 in the 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 in the next month and "0" for observations without tax arrears in the next month. Due to the low economic significance of tax arrears below 100 EUR, a company was only considered to have tax arrears in the next month if its tax arrears in the next month were at least 100 EUR. 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 a certificate are below 100 EUR.

3.3. Independent Variables

Three types of independent variables were considered in this paper: 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 3). As regards the notation of the months in Table 3 and elsewhere in this paper, 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 a 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 3) 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 variable, 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 paper ("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 [10]. In their research, Lukason and Andresson [10] found that tax arrears were in fact better predictors of bankruptcy than financial ratios.

The motivation for using the M5 type of independent variables was an observation that the Gini importances extracted from a decision tree model of all except the last four 12 monthly tax arrears were very low (below 1%) (see "Gini before aggregation" in Table 4). In essence, 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 us to increase the Gini importances of the resulting independent variables, and was therefore expected to also increase the performance of the models.

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 the 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 4). Decision tree has previously been used as a variable selection method for example by Cho et al. [43], 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 an independent variable selection tool.

Table 3. Initial and final independent variables.

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)

Table 4. Gini importances of monthly tax arrears before and after aggregation.

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

In deciding whether to leave any of the initially selected independent variables out from the final models, their descriptive statistics (Appendix A Table A1), correlation matrix

(Appendix A Table A2) and univariate prediction accuracies (Appendix A Table A3) 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 Appendix A Table A4. 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 Appendix A Table A3), indicating that all independent variables that were initially considered could have been useful predictors of tax arrears in the next month. However, due to high correlations (see Appendix A Table A2), some of the independent variables were left out of the final models (see Table 3). Namely, as regards the STATS variables, "d m in debt" and "d longest" were left out due to high correlations with other counts of event types of variables, "d max" due to high correlations with other statistical measures and "d mean" due to the high correlation with "d med". The independent variables that were left out due to high correlation had lower univariate prediction performance (see Appendix A Table A3) than the independent variables that 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, the distribution properties of the classes were different (see descriptive statistics in Appendix A Table A1), which confirmed that they could be useful predictors of tax arrears in the next month.

In order to ensure the best possible performance of the models, the 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 neighbors. 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 the signed natural logarithm, as defined in [44]:

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

The latter 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 the signed natural logarithm instead of the natural logarithm was that, in the case of M5 and M12 types of independent variables, some values were non-positive.

For MLP, rescaling was done by first using the 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, [45]):

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

where $\overline{x}$ is the mean and s is the standard deviation of the independent variable in the training set. In the case of M5 and M12 types 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.

The Anderson–Darling test was used to check the normality of distributions of the independent variables. Results showed that none of the independent variables were normally distributed. Then, a two-sample Kolmogorov–Smirnov test 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. The test results showed that all independent variables were significant at the 1% significance level.

3.4. Methodology

As provided in Section 3.1, the approach used in this paper 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 in the next month. This was justified because, among observations with previous tax arrears in zero or one months, the probability of tax arrears in the 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 (see also Figures 1 and 2).

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 in the 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 us 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 paper—decision tree (DT), random forest (RF), k-nearest neighbors (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 paper, the criterion used for choosing the best split was *Gini impurity*, which is calculated as [45]:

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

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 Appendix A Table A5. 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.

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 the 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 Appendix A Table A5.

K-nearest neighbors 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 paper are given in Appendix A Table A5. The distance measure used in all KNN models was *Euclidean distance* [45]:

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

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 a 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 are handled in a number of patches. A loss function is used in optimizing the parameter values.

The MLP models used in this paper had three hidden layers, with 4, 4 and 2 neurons, respectively. The parameter values of the models are given in Appendix A Table A5. All MLP models used the Adam optimizer and binary cross-entropy as loss measure. In the hidden layers, the *ReLu* (rectified linear unit) activation function, and in the output layer, the *sigmoid* activation function, were used. The formulas of these functions, as provided in Keras documentation, are:

$$ReLu(x) = \max(x, 0) \quad (5)$$

$$sigmoid(x) = \frac{1}{1+e^{-x}} \quad (6)$$

Using the *sigmoid* activation function meant that the output of the network was given in the form of probabilities. The probabilities were then converted into dependent variables 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 the 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, is given in Table 5. The description of the independent variables is provided in Table 3.

Table 5. Model names and independent variables in each model.

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

The performance of the models was measured based on accuracy. Accuracy is calculated as a ratio of correct predictions over all predictions [46]. Additionally, the misclassification rates were calculated, showing the ratio of falsely classified observations among observations which actually had tax arrears in the next month (type I error), and among observations which actually did not have tax arrears in the next month (type II error).

Cross-validation (1:10) on the 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 the training set and tested on the test set.

The Python machine learning libraries used for building the models were DecisionTreeClassifier, RandomForestClassifier and KNeighborsClassifier in scikit-learn. For building MLP modes, Keras neural networks library was used.

4. Results and Discussion

4.1. Results

The predictive performance of the models is provided in Table 6, both for the training and test set. Following the usual practice, in the following text, the test set results are used. The best-performing machine learning model was random forest trained on monthly tax arrears with aggregation of earlier periods into period means (M5_RF), whose prediction accuracy was 84.46%. In general, the prediction accuracy of all models was in a similar range (between 83.72% and 84.46%), with the best model only slightly outperforming the second best model (M5_DT), whose accuracy was 84.41%. Thus, different machine learning methods are almost equally successful in solving the relevant classification task. The performance of the best model in classifying observations with and without tax arrears in the next month was almost equal, with the respective misclassification rates being 15.79% (type I error) and 15.28% (type II error).

Table 6. Performance of models.

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**

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 variable were also the only ones that outperformed models trained on the 12 monthly amounts without aggregation (M12) with all machine learning methods.

It is interesting to note that the accuracy of the best machine learning model (84.5%) (see Table 6) was only slightly higher than the accuracy of the univariate random forest model, where tax arrears in month 12 were used as the single independent variable (80.4%) (see Appendix A Table A3). This shows that tax arrears in month 12 have the predominant importance in predicting tax arrears in the next month, since adding other independent variables only increased the 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 in the next month, was 98.88% (see Table 6). Results of this approach in Table 6 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 in the next month.

The accuracy of the final model (see Table 6), 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 [46], accuracy might not be the best performance measure in case the dataset is imbalanced, which the test set as a whole indeed was. Namely, the accuracy of the final model was heavily buffed by the overwhelming number of observations that were correctly predicted not to have tax arrears in the next month. In order to obtain a more realistic picture of the model's performance, a closer look can be taken at the misclassification rates of the final model as presented in Table 6. This shows that the model falsely classified 19.8% of the observations with tax arrears in the next month as not having them in the next month (type I error), and 2.5% observations without tax arrears in the next month as having them in the next month (type II error). Therefore, the final model was better at classifying observations without tax arrears in the next month. Indeed, the latter approach would be beneficial for the tax authority, as "good" firms will not be falsely targeted and administrative resources will not be spent on dealing with them, while in turn "bad" firms are targeted with high accuracy, leading to effective usage of public resources and saving the state budget from a bulk of unpaid taxes.

4.2. Discussion

The accuracy of the final model presented in this paper (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 [6] for predicting the tax arrears of Finnish firms, 72.4% in the model by Batista et al. (2012) for predicting the tax arrears of Portuguese real estate agencies and 61.6% in the model by Marghescu et al. [5] for predicting the specific case of arrears in employer contribution taxes in Finnish firms. The performance of the final model also considerably outperformed the pattern mining models presented in the study by Zhao et al. [23], where activities in the taxpayer database were used as independent variables for predicting future social security debts (where the accuracy of the best model was 76.0%).

Therefore, this paper shows that using monthly tax arrears and monthly predictions enables the prediction of 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 the accuracy in previous studies is that financial ratios are not the most appropriate predictors of tax arrears. First, this is because they become available with a considerable time lag, after the tax payment irregularities have already been going on for some time. Secondly, this is because they can result from temporary difficulties that are overcome during the financial year and are therefore never reflected in financial statements. In addition, besides being a sign of financial distress, tax arrears might also reflect certain behavioral 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-term 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. [24], which, to the best of the authors' knowledge is the only study where previous tax variables have been used for predicting tax arrears in the future (accuracy of 90.58%). Contrary to this paper, in that study, the prediction was annual instead of monthly, and tax variables 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 paper.

The overall accuracy of the final model was also higher than the model in Lukason and Andresson [10], where tax arrears in 12 months (using independent variables corresponding

to "d max", "d med", "d m in debt" and "d longest" in this paper) were used for bankruptcy prediction (accuracy 89.5%). However, the model in this paper had a 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 a 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 paper 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 the 12 months preceding bankruptcy are more pronounced and easier to detect than the ones preceding any particular subsequent month with tax arrears. Moreover, companies that will end in bankruptcy often have been de facto insolvent already for 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 paper was random forest, which falls into the category of ensemble methods. Therefore, the results in this paper are in line with the observation made by Domingos [35] 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. [13], where it outperformed all other methods, which included SVM, neural networks, logit, MDA, bagging and boosting. Concerning tax fraud detection, random forest has also been found to have superior performance (e.g., [3]).

The results (see Table 6) show that all machine learning models were characterized by underfitting, with accuracy on the training set being, on average, 3.0 percentage points lower than on the test set. Underfitting occurs when a model is not complex enough to fully represent the underlying relationship between the independent variables and the dependent variable [41]. The main reasons that 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 the tax arrears of 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 in the next month. Instead, there could be other factors besides previous tax arrears history underlying the corporate tax payment behavior.

Besides providing a decision support system of how to predict tax arrears, this paper also provides an important contribution to the financial management literature. First, it clearly shows that the financial decline process of SMEs can be rapid, as tax arrears from further periods are not effective predictors of today's defaults, and therefore, financial problems can emerge suddenly for otherwise healthy firms. This finding contradicts the lengthy failure processes documented in Hambrick and D'Aveni [47] and D'Aveni [31] for large firms, while lending support to the sudden demise postulated by Lukason and Laitinen [22]. Second, the assumption that previous defaults can predict new defaults (e.g., [29]) is valid with reservations. Namely, the given postulate holds in a very short time horizon, i.e., when an SME starts struggling with a liquidity crisis, as indicated by tax arrears in multiple month-ends, then it is very likely that this situation will not resolve on its own and will evolve into a longer series of tax arrears. Third, derived from the latter argument, multiple consecutive months of tax arrears can point to the fact that an SME is likely to enter the "death struggle phase", as indicated by Hambrick and D'Aveni [47], and may thus be the first signal of serious or even "fatal" financial difficulties.

5. Conclusions

The aim of this paper was to explore which machine learning methods and types of independent variables are most useful in predicting companies to have tax arrears in the

next month, given the time series of their tax arrears in the preceding 12 months. The data were the 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 paper 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 in the 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 in the next month was very low (1.33%) 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 neighbors (KNN) and multilayer perceptron (MLP). With each of these 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 final decision support system consists of two parts. First, for firms with at least two months of tax arrears during a twelve-month period, the best method 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 the Gini importances of the 12 monthly amounts. Second, observations with less than two months with previous tax arrears were all simply predicted not to have tax arrears in the next month, which yielded 98.88% accuracy. Thus, 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 in the next month, with the percentage of false classifications among observations without tax arrears in the next month being only 2.5% (type II error), while among observations with tax arrears in the next month, it was 19.8% (type I error).

This paper 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 the more immediate future, using the most recent information available.

This paper 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. In addition, it is important to note that the very low type II error will result in only a small number of good firms being groundlessly targeted, thus ensuring that administrative resources are well employed to deal with likely debtors. The main limitation of the paper is that tax systems, the collection of taxes and punitive measures for not paying taxes differ among countries; thus, the results might not be fully applicable in other environments. Nevertheless, since Estonia holds 12th place in the World Bank's [48] ranking on the ease of paying taxes, and according to the last wave of the World Value Survey [49], in Estonia, the propensity to cheat on taxes is lower than the world's average, the results could be reasonably transferable to many countries. Therefore, a practical guideline to tax authorities free of the latter limitation would be that: (a) occasional tax arrears lasting for a short period of time are usually not a sign of increased risk; (b) several consecutive months of tax arrears demand the relevant authority's attention, while the exact response depends on the specific country's circumstances; (c) when predicting tax arrears, different machine learning tools seem not to have remarkable advantages over each other; (d) concerning variables, a simple approach of accounting for the presence of

tax arrears in each of the most recent months and a more consolidative approach for more further months could be the best choice.

For future research, we suggest implementing additional variables from different domains to even enhance the already high prediction accuracy of the current decision support system. For instance, we believe that real-time information possessed by tax authorities in different countries, e.g., about firms' cooperation partners and members of management, structure of paid taxes or even transactional data from bank accounts, could be put into use for solving the respective classification task. Second, future research could explore the 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. Third, it might also be useful to develop models that predict the probability of tax arrears in the next month. Such models could be then developed further to predict the amount of tax arrears in the next month for cases where the probability of tax arrears in the next month exceeds a certain threshold. Finally, future research could explore the possibilities for predicting the occurrence of tax arrears for different numbers of months ahead, instead of predicting it only for the next month.

Author Contributions: Conceptualization, Õ.R.S. and O.L.; data curation, O.L.; methodology, Õ.R.S.; validation Õ.R.S.; formal analysis, Õ.R.S.; investigation, Õ.R.S. and O.L.; writing—original draft preparation, Õ.R.S.; writing—review and editing, Õ.R.S. and O.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Estonian Research Council's grant PRG791, 'Innovation Complementarities and Productivity Growth', and the University of Tartu's Ernst Jaakson Commemorative Scholarship.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The paper is a substantially extended version of a thesis authored by the first and supervised by the second author.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Descriptive statistics of independent variables.

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	1,426,784	20,624	5566	741	0	1,426,784	26,413	818	0	0	1,404,432	11,863
	d mean	3771	574	0	1,385,285	20,179	6138	1302	0	1,385,285	26,213	1395	234	0	1,355,879	10,721
	d max	8466	2236	1	1,508,237	31,687	11,980	3677	1	1,508,237	38,105	4939	1253	1	1,426,827	23,013
	d std	2420	688	0	707,599	8988	3199	1075	0	497,611	9415	1638	400	0	707,599	8467
Amounts with aggregation	months 1–5	3364	348	0	1,483,347	20,159	4992	716	0	1,483,347	24,273	1730	164	0	1,421,714	14,749
	months 6–9	3862	419	0	1,426,034	21,517	6309	1124	0	1,383,971	27,564	1406	122	0	1,426,034	12,372
	month 10	4202	153	0	1,441,174	23,523	7355	1304	0	1,441,174	31,105	1038	0	0	1,404,432	10,844
	month 11	4357	177	0	1,470,732	24,072	7855	1532	0	1,470,732	32,319	845	0	0	1,404,432	9384
	month 12	4424	164	0	1,508,237	24,322	8251	1704	0	1,508,237	33,184	583	0	0	1,404,432	7109
Amounts without aggregation (for month 10–12 the same as for with aggregation)	month 1	3142	0	0	1,502,340	20,284	4557	236	0	1,502,340	23,755	1723	0	0	1,426,827	15,937
	month 2	3274	0	0	1,502,340	20,779	4753	310	0	1,502,340	24,411	1789	0	0	1,426,827	16,206
	month 3	3364	0	0	1,502,340	20,962	4969	388	0	1,502,340	24,925	1754	0	0	1,426,827	15,864
	month 4	3469	5	0	1,499,340	21,383	5209	473	0	1,499,340	25,414	1723	0	0	1,426,827	16,180
	month 5	3570	22	0	1,496,340	21,538	5472	556	0	1,496,340	26,057	1661	0	0	1,426,827	15,516
	month 6	3674	42	0	1,476,340	21,717	5770	663	0	1,476,340	26,690	1569	0	0	1,426,827	14,871
	month 7	3791	67	0	1,426,827	22,011	6104	778	0	1,426,740	27,562	1470	0	0	1,426,827	14,053
	month 8	3924	97	0	1,426,827	22,547	6477	927	0	1,426,827	28,756	1362	0	0	1,426,827	13,247
	month 9	4058	124	0	1,426,827	23,058	6883	1091	0	1,426,827	29,949	1223	0	0	1,423,743	12,204

Table A2. Correlation matrix of independent variables.

	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

Table A3. Univariate prediction accuracies of independent variables.

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 aggregation	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

Table A4. Parameters used in univariate models.

Method	Parameter	Value
DT	Minimum number of samples in leaf	1000
	Maximum depth of the tree	4
RF	Minimum number of samples in leaf	100
	Maximum depth of the tree	4
KNN	Number of neighbors	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

Table A5. Parameters used in multivariate models.

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 neighbors	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

References

1. Matos, T.; Macedo, J.A.; Lettich, F.; Monteiro, J.M.; Renso, C.; Perego, R.; Nardini, F.M. Leveraging feature selection to detect potential tax fraudsters. *Expert Syst. Appl.* **2020**, *145*, 113128. [CrossRef]
2. Kleanthous, C.; Chatzis, S. Gated mixture variational autoencoders for value added tax audit case selection. *Knowl.-Based Syst.* **2020**, *188*, 105048. [CrossRef]
3. Didimo, W.; Grilli, L.; Liotta, G.; Menconi, L.; Montecchiani, F.; Pagliuca, D. Combining network visualization and data mining for tax risk assessment. *IEEE Access* **2020**, *8*, 16073–16086. [CrossRef]
4. Vanhoeyveld, J.; Martens, D.; Peeters, B. Value-added tax fraud detection with scalable anomaly detection techniques. *Appl. Soft Comput.* **2020**, *86*, 105895. [CrossRef]
5. Marghescu, D.; Kallio, M.; Back, B. Using financial ratios to select companies for tax auditing: A preliminary study. In *Lytras Organizational, Business, and Technological Aspects of the Knowledge Society. WSKS 2010. Communications in Computer and Information Science*; Lytras, M.D., Ordonez de Pablos, P., Ziderman, A., Roulstone, A., Maurer, H., Imber, J.B., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 112. [CrossRef]
6. Höglund, H. Tax payment default prediction using genetic algorithm-based variable selection. *Expert Syst. Appl.* **2017**, *88*, 368–375. [CrossRef]
7. Batista, J.; Cerqueira, A.; Brandão, E.F.M. Modeling Corporate Tax Risk: Evidence from Portugal. 2012. Available online: https://ssrn.com/abstract=2179068 (accessed on 25 July 2021).
8. Altman, E.I.; Sabato, G.; Wilson, N. The value of non-financial information in SME risk management. *J. Credit Risk* **2010**, *6*, 95–127. [CrossRef]
9. Lukason, O.; Camacho-Miñano, M.M. Bankruptcy risk, its financial determinants and reporting delays: Do managers have anything to hide? *Risks* **2019**, *7*, 77. [CrossRef]
10. Lukason, O.; Andresson, A. Tax Arrears Versus Financial Ratios in Bankruptcy Prediction. *J. Risk Financ. Manag.* **2019**, *12*, 187. [CrossRef]
11. Alaka, H.A.; Oyedele, L.O.; Owolabi, H.A.; Kumar, V.; Ajayi, S.O.; Akinade, O.; Bilal, M. Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Syst. Appl.* **2018**, *94*, 164–184. [CrossRef]
12. Kumar, P.R.; Ravi, V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques—A review. *Eur. J. Oper. Res.* **2007**, *180*, 1–28. [CrossRef]
13. Barboza, F.; Kimura, H.; Altman, E. Machine learning models and bankruptcy prediction. *Expert Syst. Appl.* **2017**, *83*, 405–417. [CrossRef]
14. Hanlon, M.; Heitzman, S. A review of tax research. *J. Account. Econ.* **2010**, *50*, 127–178. [CrossRef]
15. Luo, Z.; Hsu, P.; Xu, N. SME default prediction framework with the effective use of external public credit data. *Sustainability* **2020**, *12*, 7575. [CrossRef]
16. Lukason, O.; Camacho-Miñano, M.M. What Best Explains Reporting Delays? A sme population level study of different factors. *Sustainability* **2021**, *13*, 4663. [CrossRef]
17. Bartolacci, F.; Caputo, A.; Soverchia, M. Sustainability and financial performance of small and medium sized enterprises: A bibliometric and systematic literature review. *Bus. Strat. Environ.* **2019**, *29*, 1297–1309. [CrossRef]
18. Meseguer-Sánchez, V.; Gálvez-Sánchez, F.J.; López-Martínez, G.; Molina-Moreno, V. Corporate social responsibility and sustainability. A bibliometric analysis of their interrelations. *Sustainability* **2021**, *13*, 1636. [CrossRef]
19. Buettner, T.; Kauder, B. Revenue forecasting practices: Differences across countries and consequences for forecasting performance. *Fisc. Stud.* **2010**, *31*, 313–340. [CrossRef]
20. Wang, F.; Xu, S.; Sun, J.; Cullinan, C.P. Corporate tax avoidance: A literature review and research agenda. *J. Econ. Surv.* **2019**, *34*, 793–811. [CrossRef]
21. Abedin, M.Z.; Chi, G.; Uddin, M.M.; Satu, S.; Khan, I.; Hajek, P. Tax default prediction using feature transformation-based machine learning. *IEEE Access* **2021**, *9*, 19864–19881. [CrossRef]
22. Lukason, O.; Laitinen, E.K. Firm failure processes and components of failure risk: An analysis of European bankrupt firms. *J. Bus. Res.* **2019**, *98*, 380–390. [CrossRef]
23. Zhao, Y.; Zhang, H.; Wu, S.; Pei, J.; Cao, L.; Zhang, C.; Bohlscheid, H. Debt Detection in Social Security by Sequence Classification Using Both Positive and Negative Patterns. In *Transactions on Petri Nets and Other Models of Concurrency XV*; Springer Science and Business Media LLC: Amsterdam, The Netherlands, 2009; Volume 5782, pp. 648–663.
24. Su, A.; He, Z.; Su, J.; Zhou, Y.; Fan, Y.; Kong, Y. Detection of tax arrears based on ensemble learning model. In Proceedings of the 2018 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Chengdu, China, 15–18 July 2018; pp. 270–274.
25. Kukalová, G.; Moravec, L.; Filipová, D.B.; Bařtipán, M. Success rate of tax arrears recovery: Czech Republic case study. In Proceedings of the International Scientific Conference Hradec Economic Days 2020, Online, 25–26 March 2021; University of Hradec Kralove: Hradec Králové, Czech Republic, 2020.
26. Kubicová, J.; Faltus, S. Tax Debt as an Indicator of Companies' Default: The Case of Slovakia. *J. Appl. Econ. Bus.* **2014**, *2*, 59–74. [CrossRef]

27. Ciampi, F.; Cillo, V.; Fiano, F. Combining Kohonen maps and prior payment behavior for small enterprise default prediction. *Small Bus. Econ.* **2018**, *54*, 1007–1039. [CrossRef]
28. Karan, M.B.; Ulucan, A.; Kaya, M. Credit risk estimation using payment history data: A comparative study of Turkish retail stores. *Cent. Eur. J. Oper. Res.* **2012**, *21*, 479–494. [CrossRef]
29. Back, P. Explaining financial difficulties based on previous payment behavior, management background variables and financial ratios. *Eur. Account. Rev.* **2005**, *14*, 839–868. [CrossRef]
30. Peel, M.J.; Wilson, N.; Howorth, C. Late Payment and credit management in the small firm sector: Some empirical evidence. *Int. Small Bus. J. Res. Entrep.* **2000**, *18*, 17–37. [CrossRef]
31. D'Aveni, R.A. The Aftermath of organizational decline: A longitudinal study of the strategic and managerial characteristics of declining firms. *Acad. Manag. J.* **1989**, *32*, 577–605. [CrossRef]
32. Laitinen, E.K. Financial ratios and different failure processes. *J. Bus. Financ. Account.* **1991**, *18*, 649–673. [CrossRef]
33. Veganzones, D.; Severin, E. Corporate failure prediction models in the twenty-first century: A review. *Eur. Bus. Rev.* **2021**, *33*, 204–226. [CrossRef]
34. Jayasekera, R. Prediction of company failure: Past, present and promising directions for the future. *Int. Rev. Financ. Anal.* **2018**, *55*, 196–208. [CrossRef]
35. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55*, 78–87. [CrossRef]
36. Shi, Y.; Li, X. An overview of bankruptcy prediction models for corporate firms: A Systematic literature review. *Intang. Cap.* **2019**, *15*, 114–127. [CrossRef]
37. Du Jardin, P. Dynamics of firm financial evolution and bankruptcy prediction. *Expert Syst. Appl.* **2017**, *75*, 25–43. [CrossRef]
38. Sun, J.; Li, H.; Huang, Q.-H.; He, K.-Y. Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowl. Based Syst.* **2014**, *57*, 41–56. [CrossRef]
39. Balcaen, S.; Ooghe, H. 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. *Br. Account. Rev.* **2006**, *38*, 63–93. [CrossRef]
40. Chen, N.; Ribeiro, B.; Chen, A. Financial credit risk assessment: A recent review. *Artif. Intell. Rev.* **2015**, *45*, 1–23. [CrossRef]
41. Kelleher, J.D.; Mac Namee, B.; D'Arcy, A. *Fundamentals of Machine Learning for Predictive Data Analytics. Algorithms, Worked Examples, and Case Studies*; MIT Press: Cambridge, MA, USA; London, UK, 2015.
42. Sun, J.; Shang, Z.; Li, H. Imbalance-oriented SVM methods for financial distress prediction: A comparative study among the new SB-SVM-ensemble method and traditional methods. *J. Oper. Res. Soc.* **2014**, *65*, 1905–1919. [CrossRef]
43. Cho, S.; Hong, H.; Ha, B.-C. 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 Syst. Appl.* **2010**, *37*, 3482–3488. [CrossRef]
44. Alessandretti, L.; Baronchelli, A.; He, Y.-H. Machine Learning Meets Number Theory: The Data Science of Birch-Swinnerton-Dyer. *arXiv* **2019**, arXiv:1911.02008v1. Available online: https://arxiv.org/pdf/1911.02008.pdf (accessed on 25 July 2021).
45. Flach, P. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*; Cambridge University Press: Cambridge, UK, 2012.
46. Chawla, N.V. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer: Berlin/Heidelberg, Germany, 2009.
47. Hambrick, D.C.; D'Aveni, R.A. Large Corporate Failures as Downward Spirals. *Adm. Sci. Q.* **1988**, *33*, 1–23. [CrossRef]
48. World Bank. Doing Business Database for Paying Taxes (2019 Benchmark of Rankings). 2019. Available online: https://www.doingbusiness.org/en/rankings (accessed on 15 July 2021).
49. World Values Survey. Wave 7 (2017–2020) (Q180. Justifiable: Cheating on Taxes). 2020. Available online: https://www.worldvaluessurvey.org/WVSOnline.jsp (accessed on 15 July 2021).