UNIVERSITY OF TARTU
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
Institute of Computer Science
Software Engineering Curriculum

Taavi Ilves

Impact of Board Dynamics in Corporate Bankruptcy Prediction: Application of Temporal Snapshots of Networks of Board Members and Companies

Master's thesis (30ECP)

Supervisor: Peep Küngas, PhD

Tartu 2014

Impact of Board Dynamics in Corporate Bankruptcy Prediction: Application of Temporal Snapshots of Networks of Board Members and Companies

Abstract:

Corporate bankruptcy affects significantly a variety of stakeholders, such as investors, creditors, competitors, employees, and is therefore an event, in which there is a serious economic interest to predict it well ahead. Although this topic is widely studied, typically annual financial data is used to make predictions. However, due to significant delay in publication of such data, the predictions are often outdated. At the same time, changes in board membership of companies are made public with significantly shorter delay. This thesis investigates whether usage of network metrics of networks of board members and companies will positively impact accuracy and timeliness of bankruptcy prediction. More specifically, the thesis reveals that network metrics, especially PageRank, degree and eccentricity, indeed improve bankruptcy prediction models. Furthermore, by using random forest learning method and network metrics, the author was able to construct a classification model that was capable of predicting bankruptcy up to nine months in advance.

Keywords:

Bankruptcy prediction, Machine learning, Graph metrics, Graph evolution.

Juhatuse liikmete ja firmade võrgu meetrikate mõju firmade pankrottide ennustamisel

Lühikokkuvõte:

Firma pankrott mõjutab erinevaid ettevõttega seotud huvigruppe, näiteks investoreid, võlausaldajad, konkurente, töötajad, ja seetõttu on pankroti ennustamise vastu tõsine majanduslik huvi. Kuigi seda probleemi on juba laialdaselt uuritud, on enamasti ennustuste tegemiseks kasutatud ettevõtete varasemaid finantsandmeid. Kuna majandusaasta aruanded koostatakse ja avalikustatakse alles peale majandusaasta lõppu, ei ole ennustused enam ajakohased. Samal ajal avalikustatakse juhatuse liikmete muudatused ilma erilise viivituseta. Antud töö uurib, kas juhatuse liikmete ja firmade graafi võrgumeetrikad mõjutavad ennustuste täpsust ning seeläbi muudaks ennustused ajakohasemaks. Töös tehtud eksperimentide tulemused näitavad, et võrgumeetrikad, eriti *PageRank*, *degree* ja *eccentricity*, suurendavad mudelite täpsust. Parimaks mudeliks osutus otsustuspuul põhinev *random forests*, mis suutis pankrotti klassifitseerida kuni üheksa kuud ette.

Märksõnad:

Pankroti ennustamine, Masinõppe, Graafi meetrikad, Graafi evolutsioon.

Abbreviations

ROC	receiver operating characteristic
AUC	area under the ROC curve
WACC	weighted accuracy
PPV	positive predictive value
SMOTE	Synthetic minority oversampling technique
ENN	Wilson's edited nearest neighbor rule
RF	random forest
DT	decision tree
BDT	boosted decision tree
NB	naive Bayes
SVM	support vector machine
GLM	generalized linear model
ANN	artificial neural networks
CBR	case-based reasoning
logit	logistic regression

Table of Contents

1. INTRODUCTION

Bankruptcy prediction of companies, especially banks, has been a well-researched area since the late 1960s [1]. It is an important problem since it can have a high influence on business decisions and profitability. In fact, the forecast of bankruptcies is necessary for different types of public and commercial organizations as a failed business can cause failures of other companies and affect the rest of the financial system, and a shifting bankruptcy rate can indicate changes in the economic environment.

Previously, the most common approach has been using companies' financial data to forecast bankruptcy. The techniques to predict bankruptcy used in the past are divided into two extensive categories: linear and non-linear methods. For example, the linear techniques that have been used include linear discriminant analysis [1], multivariate discriminant analysis [2] and logistic regression (logit) [3]. There are some limiting presumptions when using linear statistical methods such as the linearity, normality and independence amongst predictor or input variables. Statistical methods can have problems with effectiveness and validity, because violation of these presumptions for independent variables frequently occurs with business data [4].

Some of the more advanced non-linear techniques previously used are different artificial neural network (ANN) architectures [4, 5], decision trees (DT) [6], random forests (RF) [7], case-based reasoning (CBR) [8] and support vector machine (SVM) [4].

However, using financial data has some limitations. Firstly, the financial reports are usually published after the end of the fiscal year, and only listed companies publish some of their economic data quarterly. For instance, in Estonia all companies have to publish their annual financial statements within six months of the end of a fiscal year, with some exceptions where the period can be extended even further. A study shows that over 80% companies soon to be bankrupted do not submit their last financial report at all [9]. This

leads to that the predictions are not made with up-to-date data and many companies have to be discarded from training set or making predictions as there is no data available.

In recent years, graph-based analysis of different systems has become more popular. Many real-life systems can be represented as a graph, for example web pages, social relations, topology, and in our case, board members network. For example, graph analysis and metrics have been used to find patterns and evolution rules [10, 11], detect and classify trends [12] and find influential nodes from the graph [13]. Therefore it is worth to study the effect of the board members graph metrics on the bankruptcy prediction.

This study compares different machine learning methods by comparing their key classification performance metrics to determine if using graph metrics can improve the accuracy of bankruptcy prediction, and thanks to this, make the predictions more up to date. The compared models include random forests (RF), support vector machine, naive Bayes (NB), decision tree, artificial neural networks, generalized boosted regression model (GBM) and generalized linear model (GLM). The models were trained and compared with three features sets a) using only company tax debt and their classification of economic activities, b) board member graph metrics and c) both feature sets together.

The main argument to use graph metrics is that the changes in the company board member network are visible and available almost immediately so the input data is more up-to-date than financial data, which makes the predictions more accurate and practical [14]. As well, tax debt information is published and available without significant delay. Furthermore, some companies do not publish their financial year reports at all or go over the deadline. So using board member network and tax debt data increases the number of companies one can use for training the models and making predictions.

The rest of the paper is organized as follows. Chapter 2 provides an overview of related works in the field of bankruptcy prediction and graph evolution. Chapter 3 gives a brief review of the tools, techniques, and performance metrics used in this research. Chapter 4 presents the descriptions and analyses of the datasets and its features. In chapter 5, the results of the experiments are presented, and practical application based on the findings and analyses is proposed. Finally, Chapter 6 draws some conclusions about the proposed models and outlines future work.

2. RELATED WORK

2.1 Bankruptcy Prediction

One of the first studies about bankruptcy prediction was conducted by William H. Beaver [15] in 1966. He used 158 samples of failed and non-failed firms to investigate the suitability of 14 financial ratios. The best performing two financial ratios were working capital/debt ratio and net income/total assets ratio. In our study we can't use financial ratios, because this data is not available to us. Only some of the companies make their financial statements public, but we want to apply our model to as many companies as possible.

Beaver's research was followed by Altman, [1] who proposed a model based on the multiple discriminant analysis to classify the companies into known categories. He concluded that bankruptcy can be explained by using a combination of five (selected from the original list of 22) financial ratios. The classification of Altman's model had a predictive score of 96% for a prediction one year before the bankruptcy. In our case his approach is also only practical when predicting bankruptcy of larger companies as we do not have financial data available for smaller companies.

In 1999 Clive Lennox [3] studied the causes of bankruptcy using samples of 949 United Kingdom companies. He argued that the most important indicators of bankruptcy are profitability, leverage, cash-flow, company size, industry sector, and the economic cycle. He compared linear and non-linear probit, logit, and discriminant analysis (DA) models and discovered that probit and logit performed better than DA.

In their work Park and Han [8] used case-based reasoning using feature weights and concluded that this approach will perform very well when modeling bankruptcy prediction. They compared different models and the best model - AHP (analytical hierarchy process) CBR - yielded average accuracy of 83%. They used different company financial ratios to

construct the models. Similarly to their approach, we used k-fold cross- validation, however; we did not compare the CBR approach with other techniques as this model was not available in the tool set we used.

Shin *et al.* [4] investigated the effect of applying support vector machines to the bankruptcy prediction problem. They showed using financial ratios that the proposed classifier of SVM approach outperforms back-propagation neural network (BPN) to the problem of corporate bankruptcy prediction. Their results demonstrated that the accuracy and generalization performance of SVM is better than BPN as the training set size gets smaller. Similarly to us, the authors performed their studies on an unbalanced dataset where the number of positive samples was about % of the whole dataset. Our experiments revealed that SVM was the next best approach after random forests.

In his work, Wo-Chiang Lee [6] compared CART (classification and regression trees), C5.0 (decision tree based algorithm) and genetic programming decision tree (GP) classifiers with logit model and NN model. He used samples of Taiwan-listed electronic firms and to predict bankruptcy he also used company financial ratios. He showed that GP decision tree outperformed other models. Under the 0.5 cutoff value, the training sample overall percentage of correct of the logit model scored 81%, and test sample yielded 69.23%. The NN model scored 96% and 73.85% respectively. GP decision tree scored 100% and 92.91% subsequently. Likewise to us, they concluded that decision tree based models outperformed other approaches and used k-fold cross-validation.

Another popular widely used machine learning method for forecasting bankruptcy is artificial neural networks. Jardin [5] compared the accuracy of different classification methods, including discriminant analysis, logistic regression and artificial neural networks to predict bankruptcy. He concluded that neural networks performed best with appropriate variable selection techniques such as zero-order, first-order or out-of-sample error technique. In contrast to their findings, we found that ANN approach performed worse, however; Jardin used financial data, but we used graph metrics and tax debt features.

In 2011 Kartasheva and Traskin [7] used random forests classification model to predict the insolvency of insurers. They used the model to order companies according to their probability to default. They were able to show that RF delivers a higher quality of prediction compared to logistic regression. They also concluded that RF can be used on unbalanced

dataset as their dataset had about 1% of bankruptcy cases versus 99% normal companies. We had the same problem where positive samples made up under 1% of the dataset, but RF was able still to outperform other techniques.

Random forests together with logistic regression were also used in research conducted by Creamerk [16] to predict corporate bankruptcy. He used a 10-fold cross-validation approach to study Latin American depository receipts and banks. He used RF and logit to rank the most important variables that affect the performance of the companies. He concluded that random forests with logit can be used to forecast the performance and rank the variables. He also highlighted that the use of machine learning methods in finance requires time-series cross-sectional data in order to calculate meaningful results.

Some less commonly applied methods to forecast bankruptcies are naive Bayesian and generally boosted models. Sarkar *et al.* [17] used and compared different models including naive Bayesian to build an automated system to assess bankruptcy risks. They pointed out that the naive Bayesian model performed the best. They only used financial ratios, but suggested that other features should be also used.

Philosophov *et al.* [18] studied bankruptcy prediction using multi-period formulation instead of the common one-period forecast solution. They used financial ratios and the company's tax debt repayment schedule data to train the models. Similarly to us, they trained and compared Bayesian rules and Altman Z-score approaches using different feature sets, i.e. financial ratios with and without schedule information. This approach is much the same as ours where we compare different techniques and feature sets.

2.2 Graph Evolution and Metrics

Leskovec *et al.* [10] studied the properties of the evolution of real graphs using different types of graphs, for example, citation and affiliation graphs that are similar to board member network. They were able to show that there are patterns in dynamic networks. For example, they found that most of the graphs they studied have out-degrees that grow over time and the diameter gradually decreases as the network grows. They also proposed a Forest Fire model based on only two parameters that can capture graph patterns. This study illustrates that there are patterns in graph development, and it is possible to find and design them.

Huang *et al.* [19] studied a services network system graph. They investigated the static and dynamic version of the graph using different graph metrics such as degree centrality and edge weight centrality. The static graph was similarly to us divided into several snapshots. They concluded that it is possible to develop a rigorous theoretical framework to trace and predict the evolution of the service ecosystem.

Berlingerio *et al.* [20] followed a frequent pattern-mining approach and defined the relative time patterns and introduced the problem of extracting graph evolution rules. They implemented an effective solution to mine the patterns, and extensively tested it on four large real-world networks. In conclusion, they showed that graph evolution rules characterize different types of networks.

In their work Charanpa and Clmenon [13] constructed co-authorship graph from the database of authors and their publications that are very similar to a board member graph. Then they used the same graph metrics like us - PageRank, betweenness and closeness to obtain a ranked list of experts of a chosen topic. They were also able to demonstrate that, using graph metrics, it is possible to find influential nodes from the graph. Unlike our study, they only used a static graph, not a dynamic time graph.

Lambiotte and Ausloos [12] analyzed a bipartite network of people's music-listening habits where the nodes represented music groups and listeners using correlation matrices as random walks exploration. They pointed out that their described methods can be used to detect and classify trends in the graph. Similarly to our approach they used both bipartite version of the graph and also projected the graph to analyze the unipartite version of the music groups.

Another great example of dynamic social network analyses is the study conducted by Iba *et al.* [11] where they studied editing patterns of Wikipedia contributors. Using their own developed tool they converted the edit flow of contributors into a temporal social network by constructing snapshots of the graph. They studied 2580 featured articles of the English Wikipedia and found different editing patterns and identified the most valuable contributors. Unlike our study they only utilized degree out of a variety of metrics.

Nicosia *et al.* [21] have studied graph metrics of temporal networks. They described how to represent and construct dynamic time-varying graphs and observed different metrics such as betweenness, closeness and the spectral centrality which, first two we have also used

in our study. They argued that the structural properties of a complex network usually reveal important information about its dynamics and function.

Similar research was conducted by Bilgin *et al.* [22] where they reviewed contributions to literature of dynamic graph evolution. They pointed out following graph metrics many of which were coincident with metrics we used in our study - degree, clustering coefficient, eccentricity and closeness. Finally, they pointed out that real dynamic graphs have patterns and laws and knowing them can help finding anomalies and construct models to synthetically generate such graphs.

3. BACKGROUND

3.1 Overview of R

To process and analyze the data also train and test the models we used software named R. R is a programming language and environment mainly for statistical computing and graphics [23]. R provides a wide variety of statistical (time-series analysis, classification, clustering, linear and nonlinear modeling, classical statistical tests, etc.) and graphical techniques, and is highly extensible.

R functionality is extendable by different packages. For graph construction, analyzes and graph metrics computation we used $igraph$ package. This library is used for building and manipulating both undirected and directed graphs [24]. It can also calculate different graph metrics and can be used for various graph manipulation operations.

For model training and tuning we chose the package named $caret$. $Caret$ is short of classification and regression training and the package includes set of functions that attempt to unify the process for creating predictive models [25]. For example, the package contains tools for:

- Data splitting.
- Pre-processing.
- Feature selection.
- Model tuning using resampling.
- Variable importance estimation.

We decided to use this package as it includes over 40 different predictive models and supports diverse training methods especially 10-fold validation that is recommended for datasets with a small number of positive or negative samples [8, 4, 26].

3.2 Networks

3.2.1 Network of Board Members

In this section we detail our study on transforming the raw network data and representing it as both static and dynamic time graph. The network is presented as an $undirected\ time\ graph\ G = (V, E)$, i.e., each node $v \in V$ has an associated interval $D(v)$ on the time axis (called the $duration$ of v) and each edge $e \in E$ is a triple (u, v, t) where u and v are nodes in V and t is a point in time in the interval $D(u) \bigcap D(v)$ [27]. In particular, for anytime t, there is a natural graph G_t that comprises all the nodes and edges that have arrived up until a time t; here we assume that the end points of an edge always arrive during or before the edge itself [27]. We chose time interval of one month to construct set of snapshots from the graph and we used these snapshots to analyze the graph.

The dataset consists of all Estonian companies and its co-executives. The relationship between a firm and a co-executive has the relation start and end date. The changes of board members in the dataset are available from the beginning of 2012 and end with the march of 2014. Previous board member change dates are unknown and not available. When the start date is not known then it is represented as $NULL$, and if the end date is not set (this means that, this person is still a board member of the company), then it is represented as NA. The dataset also includes a separate list of companies who have gone bankrupt with the date of bankruptcy.

This structure described previously leads directly to a bipartite network for the whole system. Namely, it is a graph composed by two kinds of nodes, i.e., persons, called board members or co-executives, and companies. Bipartite graph representation allows connections only between two nodes in different sets. Please see Figure 3.1 for an example of one company that has two board members. The network is presented as a graph with edges running between a company i and a person u, if u is a board member of i.

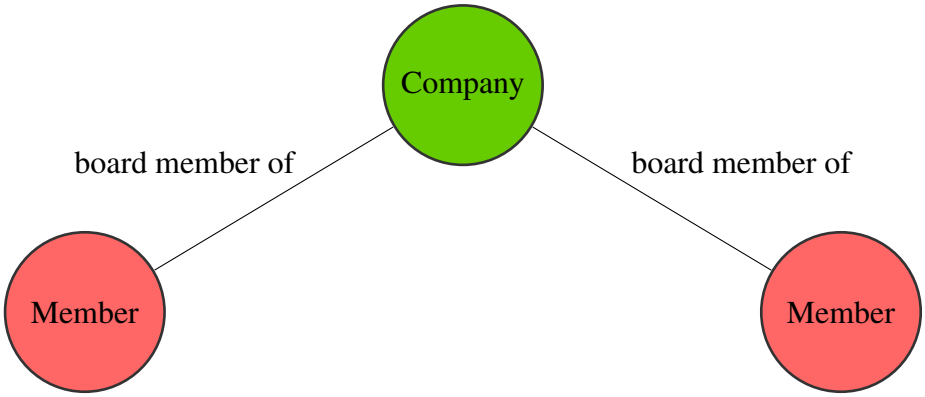


Figure 3.1 – *Example of bipartite network of one company with two board members.*

For confidentiality reasons all companies and board members are identified by a unique number - this will not affect the presentation of our results.

It is possible to make a static or a dynamic graph by using the available date attributes. To construct time-graphs from the dataset available we first transformed unknown ($NULL$ and NA) membership start dates and end dates. So for an unknown start date we chose a long time in the past, for example - $1900/01/01$. For NA end date, we chose a time in the future, for example, $2020/01/01$. The next step was transforming the edge-list into a graph. That task was done by using R and $igraph$ package. Also, as the graph was bipartite, $type$ parameter had to be assigned to each vertex to distinguish company and board member nodes.

3.2.2 Graph Construction and Types

As described in previous chapters, the bipartite graph was constructed from our dataset whose nodes are divided up into two sets X and Y, and only connections between two nodes in different sets are allowed [28]. However, bipartite graph can also be compressed into two one-mode networks. The method is known as network projection (PR) [12]. This means that the newly projected graph contains nodes only from either of the two sets, and two X (or, alternatively, Y) nodes are connected only if when they have at least one common neighboring Y (or, alternatively, X) node [28].

In this study, we used two graph types, the bipartite graph and the company graph projection. The board member projection was not used, because only company node metrics are meaningful when predicting bankruptcy. The bipartite graph gave us the company's relation

to its board members and the projection represents relations between the company nodes.

To clean the dataset, isolated firms were removed from the graph. A company was defined as isolated when it had zero degrees in the company node projection during the whole dataset lifetime, because the company was not related to any other firm during the whole dataset lifetime. This removal was done because the nodes did not have any changes, are isolated, and they do not influence the global metrics.

3.3 Models

In this chapter, we are describing briefly the different models we are going to use and compare the performance. Our choice of the models is mainly based on previous studies where the models were used for bankruptcy prediction and have achieved best results. Another criterion was that the model is available using R software and $caret$ package.

3.3.1 Logistic Regression

Logistic regression (also known as logit regression) is a technique that is used for predicting the probability of occurrence of an event. It is trained by fitting one or more predictor variables (features) to a logistic curve. The predictor variables can be either numerical or categorical (factors) [29, 30].

The formula of the model can be written as [29]

$$f(z) = \frac{1}{1 + e^{-z}}, z = x_0 + \beta_1 x_1 + \beta_2 x_2 ... + \beta_n x_n \tag{3.1}$$

where $x_1, x_2, ..., x_n$ are predictor variables for the regression. The classification output is a type A if $f(z) \geq 0$, otherwise a type B [29]. In our case, the predictor variables or features would be the time-graph metrics, debt information and field of activity. The output A would mean bankruptcy and B means no bankruptcy.

Logistic regression models are heavily used in different fields. For example, it can be used to predict the like-hood of a home-owner defaulting on a mortgage [30].

3.3.2 Support Vector Machines

Support vector machines (SVMs, also known as support vector networks) is a supervised learning method that can be used for classification and regression. SVM uses statistical learning to solve classification and regression problems. SVMs analyze data and try to recognize patterns in it. The model is trained by using the training set that has two types of examples (two classes classification). The training algorithm works by building a model that assigns new examples into one category or the other. The model can be viewed as points in space, placed so that the categories are divided by a clear gap that is as wide as possible (see Figure 3.2 for an example) [29, 31].

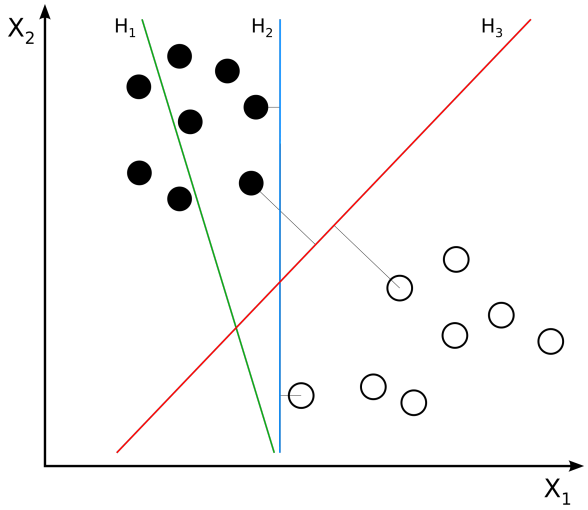


Figure 3.2 – *Example of SVM model graphical representation. H_1 does not separate the classes. H_2 does, but only with a small margin. H_3 separates them with the maximum margin [31].*

3.3.3 Naive Bayes Classifier

Naive Bayes classifier based on Bayesian theorem is simple probabilistic classifier that has strong independence assumptions. Even though the method is simple or naive, it can often perform better than more sophisticated classification models. Naive Bayes classifier works by assuming that the presence or absence of a particular predictor or feature is unrelated to the presence or absence of any other feature. The advantage of this classifier is that it only requires a small amount of training data to estimate the parameters that are necessary for classification [32].

3.3.4 Artificial Neural Networks

Artificial neural networks are computational models that learn from examples and are inspired by central nervous systems. The model is represented as a system of interconnected ”neurons” that can compute values from inputs by feeding information through the network. One of the reasons why this technique is widely used is that they can represent non-linear relationships. Neural networks also require much training data and training cycles [33, 34].

In addition, their ability to represent non-linear relationships makes them well-suited to modelling the frequently non-linear relationship between the likelihood of bankruptcy and commonly used variables (i.e. financial ratios) [5].

3.3.5 Decision Tree Based Models

Decision trees work using recursive partitioning theory and use measures as entropy to create decision trees from a data set. They produce a human readable flowchart-like structure in which internal node represents a test on an attribute and each branch represents an outcome of a test and each leaf node represents class label (decision). The path from a root to leaf represents classification rules. Some disadvantages are overfitting, and they require many data samples to make reliable predictions [34].

Random forests are an ensemble learning method which means that it is utilizing a number models to obtain better predictive performance. RF is used for both classification and regression. The idea is to construct a number of decision trees at the training time and output the class that is the mode of the classes outputted by individual trees. This method has very good accuracy among other training methods and can handle thousands of input variables without variable deletion and give estimates of what features are important in the classification [35].

We compared decision tree, boosted tree model (BTD) and random forests techniques to train the decision tree based models.

3.4 Performance Measurements

Many previous studies we reviewed in Chapter 2 used classification accuracy as the performance metric of a model. As our dataset consists of less than 1% of minority class and 99% of majority class, it is considered imbalanced. Therefore accuracy is not the best measurement to evaluate a model's performance, because a trivial classifier can achieve high accuracy score by ranking all the cases as the majority class.

In our experiments, we used performance metrics like true negative rate, true positive rate, weighted accuracy, G-mean, precision, recall, and F-score (also known as F1 score or F-measure). According to [36], these metrics are widely used to compare classifier performance. Performance metrics are functions of the confusion matrix, also known as contingency table or an error matrix, as shown in Table 3.1. The rows of the matrix represent actual classes, and the columns are represent predicted classes [36]. Based on Table 3.1 the performance metrics can be defined as:

	Actual Positive class	Actual Negative class
Predicted Positive Class	TP (True Positive)	FP (False Positive)
Predicted Negative Class	FN (False Negative)	TN (True Negative)

Table 3.1 – *Confusion matrix.*

$$
\begin{aligned}
\text{True Negative Rate } (Acc^{-}) &= \frac{TN}{TN+FP} \\
\text{True Positive Rate } (Acc^{+}) &= \frac{TP}{TP+FN} \\
\text{G-mean} &= (Acc^{-} \times Acc^{+})^{1/2} \\
\text{Weighted Accuracy (WACC)} &= \beta Acc^{+} + (1-\beta)Acc^{-} \\
\text{Precision (PPV)} &= \frac{TP}{TP+FP} \\
\text{Recall} &= \frac{TP}{TP+FN} = Acc^{+} \\
\text{F-score (F1)} &= \frac{2 \times Precision \times Recall}{Precision + Recall}
\end{aligned}
$$

There is always a trade-off between true positive rate and true negative rate. Same applies for recall and precision [36]. In our case, the rare class is of more interest than the majority class. For instance, in a credit risk application that uses bankruptcy risk as one of the inputs, it is more useful to have a model that results high prediction accuracy over the rare class (Acc^{+}), while the majority class (Acc^{-}) maintains reasonably good score because, one is more interested in finding the companies that can go bankruptcy in the near future to minimize the risk. Sometimes weighted accuracy is used in similar situations. Weights are fine-tuned to fit the application. In our study, we used equivalent weights for both the true positive rate and true negative rate; i.e., β equals 0.5 suggested by [36]. Another metric that has been used for one-side sampling in imbalanced training is Geometric Mean (G-mean) [37]. Precision, recall and F-score are very common performance measure in the information retrieval area [36].

We also used the ROC curve, more precisely the area under the ROC curve (AUC) to assess and compared the performance of a model. The ROC curve is presented usually as a graphical diagram to illustrate the trade-off between the false negative and false positive rates at every possible cut off point. This measure is commonly used by the machine learning community for model comparison [38]. However, using AUC alone to evaluate the model's performance is not good enough as recent studies have found that AUC can be quite noisy as a classification measure [39].

We compared different models by using a ten-fold cross validation where 70% of the data was used to train the model and 30% of the data to test the model. This is important because many models can over fit and thus give very good performance results, so we need to test the model by using the samples that the model has not seen yet.

3.5 Sampling

This section describes how we sampled the data using Synthetic Minority Over-sampling Technique (SMOTE) and Wilsons Edited Nearest Neighbor Rule (ENN).

As our dataset is heavily unbalanced, we needed to use some method to balance the dataset because some models underperform or cannot correctly classify minority class when using unbalanced datasets. Previous studies have noted that a balanced data set provides

better overall classification performance compared to an imbalanced data set [40].

SMOTE is an oversampling method. It works as by creating new minority class examples by interpolating between other minority class examples that lie together. By doing that the over fitting problem is avoided and forces the decision boundaries for the rare class to spread further into the majority class space [41].

Even though oversampling balances the class distributions, other problems still are there. Sometimes class clusters are not well-defined since some majority class samples are invading the rare class space. The opposite can also be true, since interpolating rare class examples can expand the minority class clusters [42]. In order to generate better-defined class clusters, it is reasonable to use ENN. ENN is expected to provide an in depth data cleaning. It is used to remove examples from both classes. Thus, any example that is misclassified by its three nearest neighbors is removed from the training set [41].

We used R package named $unbalanced$ to do the sampling. Firstly, we used $ubSMOTE$ method with parameters: $perc.over = 300$, $k = 5$ and $perc.under = 155$. These internal parameters help to obtain desired class distribution. We added or removed examples until a balanced distribution was reached. This decision is motivated by the results presented in [43], in which it is shown that when AUC is used as a performance measure, the best class distribution for learning tends to be near the balanced class distribution and by the fact that balanced dataset provides better classification results [40]. Running $ubSMOTE$ method yielded 3243 samples. After $ubSMOTE$ method we ran $ubENN$ method that removed 268 samples. At the end we had 2975 samples with 50.42% belonging to the positive class.

4. DATASETS

4.1 Bankruptcy Statistics

The full static bipartite graph consists of 151622 company nodes and 149656 member nodes so total of 301278 nodes and the total edge count is 231403. Table 4.1 contains the number of bankrupted companies in each year. The isolated and non isolated companies are shown separately (read about the isolated nodes from the previous chapter), because only non-isolated companies were used in this study. The number of businesses bankrupted in 2012 is low because the dataset includes bankruptcy records starting from mid 2012.

Year	Bankruptcy (not isolated)	Bankruptcy (isolated)
2012	100	17
2013	294	79
2014	7	3
Total:	401	99

Table 4.1 – *The number of annual bankruptcies of isolated and non-isolated companies.*

4.2 Graph Metrics

This section contains descriptions and observations of different graph metrics that are usually used when analysing graphs and their evolution. Each subsection describes one metric and its measure over snapshots is presented by figures. The figure show first, second and third quartile of the measured data points, but the outliers are removed. As noted before only the companies that are not isolated, i.e., removed were used in when making the observations.

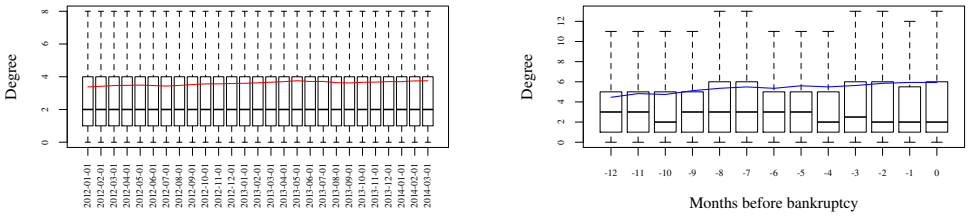
4.2.1 Degree

The degree or valency of a vertex of a graph is the number of edges incident to the vertex, with loops counted twice [44]. When considering the bipartite version of the graph then the degree of a company node represents the number of board members the company has at some point. However, when looking the degree size of the graph projection it shows how many other companies are related to the company through its board members.

When looking bipartite graph in Figure 4.1 we can see that the average degree of normal companies is decreasing very slowly. The average number of board members of the company has decreased from 1.5 to about 1.45. The average number of board members of bankrupted companies is decreasing slightly faster.

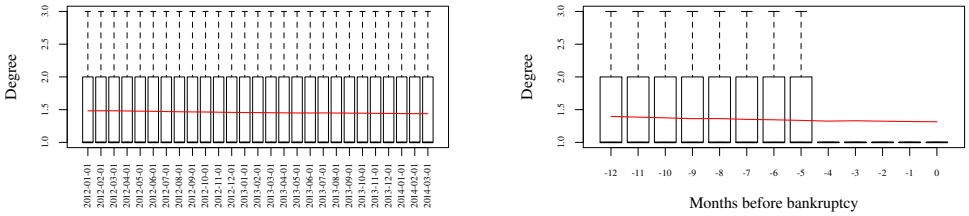
When looking graph projection in Figure 4.1 we can observe that the average degree of both plots is growing slowly but before bankruptcy it is growing more aggressively. The increasing average for both cases shows that the node neighborhood is growing, and this usually means that previous board members are starting new companies or members that already are connected with other companies are joined with the company.

In Estonia, there are professional companies you can hire that deal with businesses that are about to go bankruptcy. Usually someone is assigned to a board so that person can make legal actions to assist the bankruptcy process. The aggressive growth on the average degree in the company projection plot could imply that the companies are joined with bigger clusters that can be professional businesses that deal with bankruptcy.



(a) *Normal companies (projection)*

(b) *Before bankruptcy (projection)*



(c) *Normal companies (bipartite)*

(d) *Before bankruptcy (bipartite)*

Figure 4.1 – *Measured average, mean, upper and lower quartile of degree.*

4.2.2 Closeness Centrality

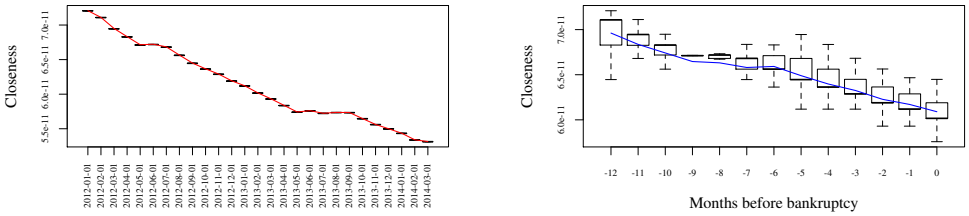
Closeness centrality measures how many steps are required to access every other vertex from a given vertex or how close all other vertices are to the current one [24]. Closeness can be expressed as a measure of how long it would take to spread information from a single node to all other nodes sequentially. According to $igraph$ package documentation, the closeness centrality of a vertex is defined by the inverse of the average length of the shortest paths to/from all the other vertices in the graph and it can be expressed as:

$$C(u) = \frac{1}{\sum\limits_{v \in V \setminus u} d(u, v)} \tag{4.1}$$

If there is no path between vertex u and v then the total number of vertices is used in the formula instead of the path length [24].

In Figure 4.2 we can see that the average closeness is decreasing slowly. The graphs of bipartite and projection were the same, so we only included the projected version of the graph. This is probably affected by new companies that enter the graph over time. We can

also observe that both plots follow a similar trend. This is due that closeness is global metric because it considers all other nodes represented in the graph.



(a) *Normal companies (projection)* **(b)** *Before bankruptcy (projection)*

Figure 4.2 – *Measured average, mean, upper and lower quantile of closeness.*

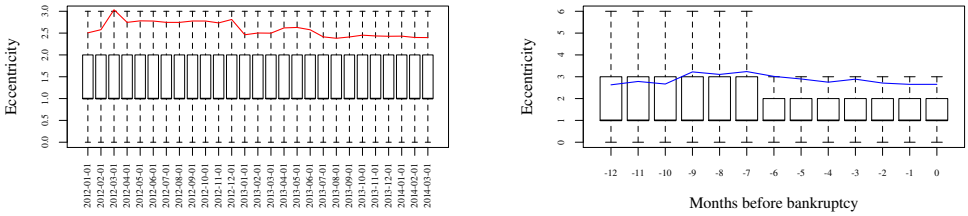
4.2.3 Eccentricity

Eccentricity is the shortest path distance from the farthest other node in the graph. The smallest eccentricity in a graph is called its radius and the maximum measure is the graph diameter [24]. The eccentricity of a node v can be expressed as:

$$ecc(u) = max\{d(u, v)|v \in V)\} \tag{4.2}$$

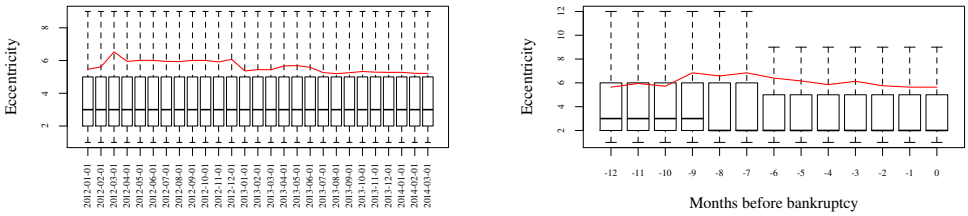
where $d(u, v)$ is the distance between vertexes v and u.

In Figure 4.3 we can see that the quartiles of both graph types of normal companies stay the same only the average eccentricity is moving a little bit. However, when looking companies before the bankruptcy we can see that six months before bankruptcy the quartiles have changed and the mean eccentricity starts decreasing.



(a) *Normal companies (projection)*

(b) *Before bankruptcy (projection)*



(c) *Normal companies (bipartite)*

(d) *Before bankruptcy (bipartite)*

Figure 4.3 – *Measured average, mean, upper and lower quartile of eccentricity.*

4.2.4 Betweenness Centrality

Betweenness of a vertex is roughly defined by the number of geodesics (shortest paths) going through that vertex. It is a measure of a node's centrality in the network [24]. The betweenness centrality of a node v can be expressed as:

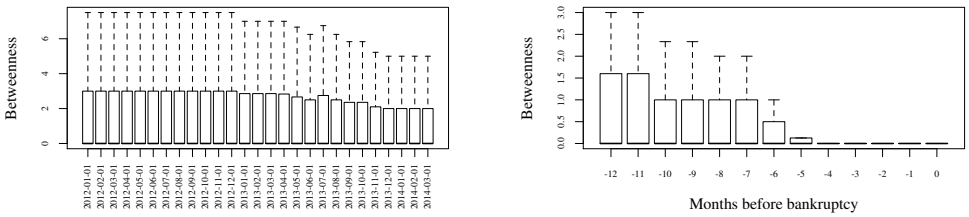
$$B(u) = \sum_{s \neq u \neq v} \frac{\sigma_{st}(u)}{\sigma_{st}} \tag{4.3}$$

where σ_{st} is the total number of shortest paths from s to node t and $\sigma_{st}(u)$ is the number of these paths that go through u [45].

The betweenness centrality of a single node reflects the amount of control that this particular node exerts over the interactions of other nodes in the network [46]. This measure favours nodes that join communities (dense subnetworks), rather than nodes that lie inside a community.

In Figure 4.4 one can see that we only included the bipartite graph as the projected version resulted a graph that measures were zero. Also, the average line is missing because

it is higher than the quantiles. We can see that the betweenness is decreasing in both cases, however, before bankruptcy it is decreasing faster and the value of the measure is lower than normal companies have.



(a) *Normal companies (bipartite)* **(b)** *Before bankruptcy (bipartite)*

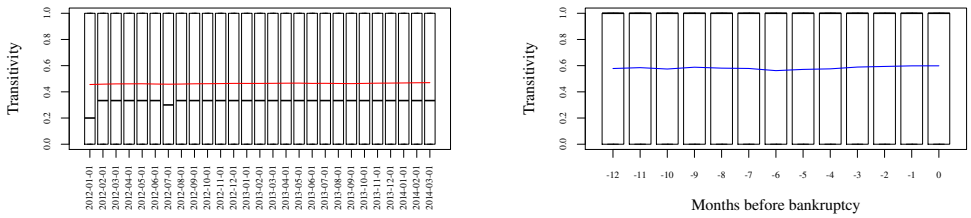
Figure 4.4 – *Measured average, mean, upper and lower quartile of betweenness.*

4.2.5 Transitivity

Transitivity (clustering coefficient) measures the probability that the adjacent vertices of a vertex are connected [47]. There are two types of transitivity measures - global and local. The global measure is simply the ratio of triangles and connected triples in the graph. The local measure is the ratio of triangles connected to the vertex and triples centered on the vertex [47].

We used only local measure because the global metric would not make sense as it would increase the chance of overfitting as the metric does not represent the individual vertex closely. The global version was designed to give an overall indication of the clustering in the network, whereas the local gives an indication of the embeddedness of single nodes [47].

In Figure 4.5 we can see that the transitivity measure is almost the same throughout the whole time span, however, the average transitivity is about 0.2 points higher than the normal companies have, and it is increasing slightly before bankruptcy. We did not include bipartite version as its calculated transitivity was always zero.



(a) *Normal companies (projection)* **(b)** *Before bankruptcy (projection)*

Figure 4.5 – *Measured average, mean, upper and lower quartile of transitivity.*

4.2.6 PageRank

PageRank is one of the most known scoring functions of networks in general and of the World Wide Web graph in particular. For example, it is used by Google Search Engine to rank websites in their search engine results [48]. Today the idea of PageRank has been applied to many other types of graph other than Web pages.

In the sense of our graph, the PageRank is a link analysis algorithm, and it assigns a numerical weighting to each company. The weight calculated shows the importance of a single company relative to the whole graph. This means that PageRank is global metric as it takes account all nodes in the graph and could mean that this metric can be a very good feature to predict bankruptcy.

PageRank works by counting the number links (edges) to a vertex to determine a rough estimate of how important the vertex or in our case a company is. The underlying assumption is that more important vertexes are likely to have more links from other vertexes. A vertex that is linked to by many pages with high PageRank receives a high rank itself. The metric is expressed as a numeric value between 0 and 1 [24].

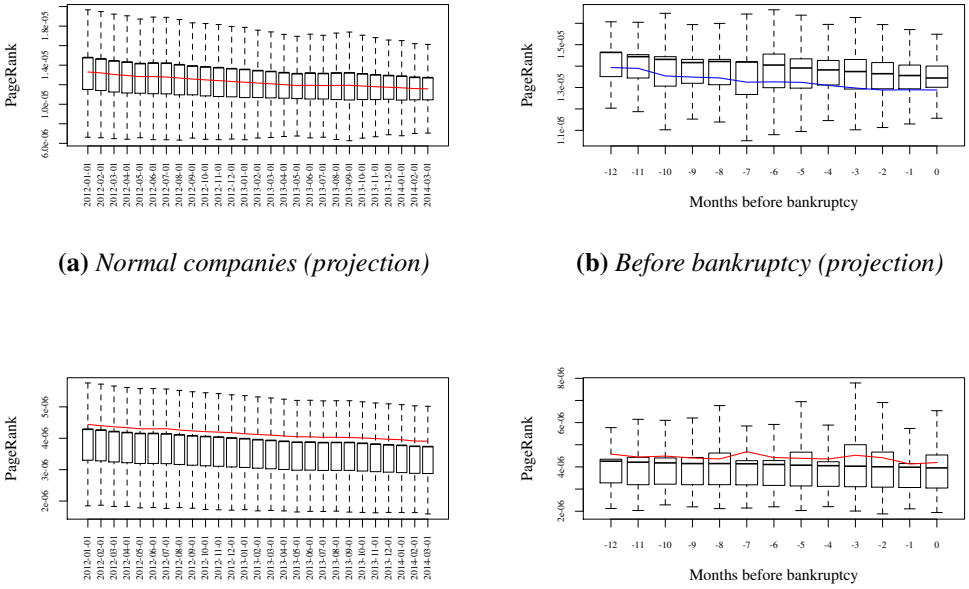
We are using $igraph$ package method named $pagerank$ to calculate PageRank for each vertex. The computational process is iterative. The PageRank of a given vertex can be expressed as [49]:

$$P(v) = (1 - d) + d(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) \tag{4.4}$$

PageRanks form a probability distribution over all graph vertexes, so the sum of all

values will be one [49].

In Figure 4.6 we can see that both graph versions of normal companies, the PageRank measure is decreasing, this is due to that the score is global and when new companies are added the average measure is decreasing. However, before bankruptcy the projection version is not decreasing as fast as normal companies. The bipartite version graph before bankruptcy is almost the same throughout 12 months despite that normal companies average measure and quartiles are decreasing.



(a) *Normal companies (projection)*

(b) *Before bankruptcy (projection)*

(c) *Normal companies (bipartite)*

(d) *Before bankruptcy (bipartite)*

Figure 4.6 – *Measured average, mean, upper and lower quartile of PageRank.*

4.3 Tax Liabilities

The dataset of company tax liabilities consisted of company identification number, record time-stamp, tax type, debt sum, disputed and postponed debt sum. The number of different tax types was about 70, but we used six most common types and also summed up all owned liabilities and represented it as total tax debt. We also created two binary variables that indicate if the company has debt or not, and if the debt is postponed.

In Figure 4.7 we can see that 88% of bankrupted companies have had tax debt and only

38% of regular firms during 2012-01-01 to 2014-03-01 period. Based on this fact, we can say that tax debt is common among bankrupted companies and can be a good indicator to predict bankruptcy.

In the following subsections, we describe and investigate the chosen tax types and variables created from the dataset.

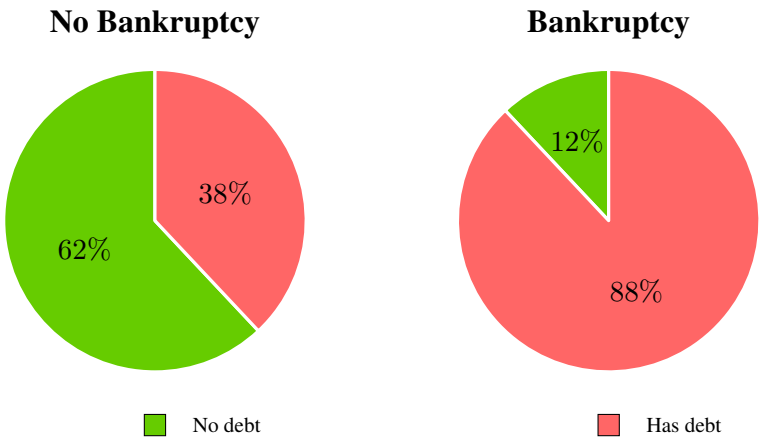
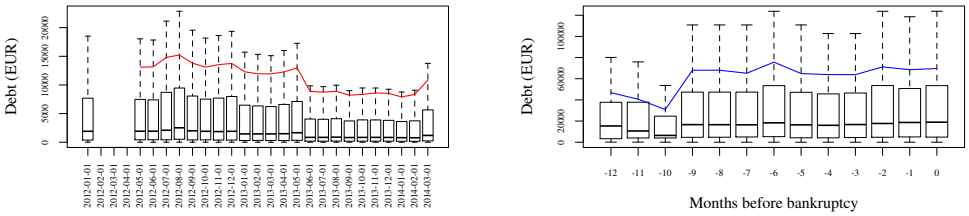


Figure 4.7 – *Percentage of companies that have or had tax liabilities between 2012-01-01 and 2014-03-01.*

4.3.1 Value Added Tax Debt

Value added tax (VAT) is a form of a consumption tax. In Estonia, all companies that are VAT obligatory have to register its collected VAT by the 20th day of the next month.

In Figure 4.8 we can see that normal companies' average and mean VAT debt is decreasing slowly, however, bankrupted companies' debt is gradually rising. We can also observe that the average and mean debt of normal companies is noticeably lower. This suggests that companies that are soon to be declaring bankruptcy do not pay VAT debt and the debt is increasing.



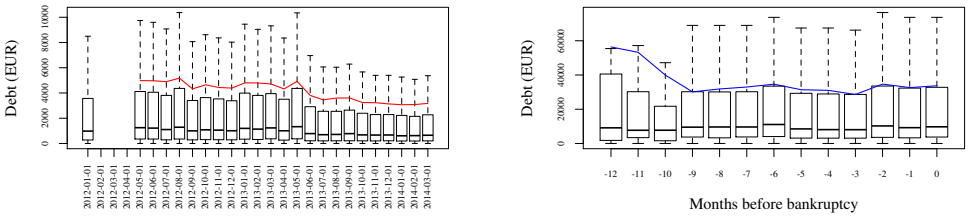
(a) *Normal companies* **(b)** *Before bankruptcy*

Figure 4.8 – *Value added tax debt measured average, mean, upper and lower quartile.*

4.3.2 Social Tax Debt

Social tax also known as the Federal Insurance Contributions Act tax (FICA) is paid by all companies that pay salaries to their employees. FICA is paid on almost all payments made to employees except some special cases that are regulated by the Estonian law.

When examining Figure 4.9 we can observe that bankrupted firms have higher mean and average FICA debt. Also the average and mean debt of non bankrupted companies' is decreasing significantly, however, before the bankruptcy the debt is steady, but it is much higher than normal companies have.



(a) *Normal companies* **(b)** *Before bankruptcy*

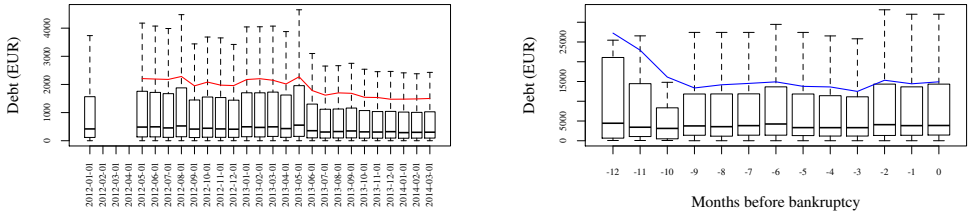
Figure 4.9 – *Social income tax debt measured average, mean, upper and lower quartile.*

4.3.3 Personal Taxes Debt

All residents in Estonia have to pay the personal income tax on their worldwide income. This tax is usually detained by the employee when the salary is paid. Employees also have to pay some percentage of their gross salary as unemployment insurance tax and funded

pension payment that is also paid by the employer. As these three tax types are usually paid from the gross salary, therefore, we described them together but provided only personal income tax figure as the trends are almost identical, only values are different, depending on the percentage of the salary.

In Figure 4.10 we can see that it acts similarly to social tax debt (Figure 4.9). This is expected, because both taxes are related with the salary and are paid to the government when wages are paid.



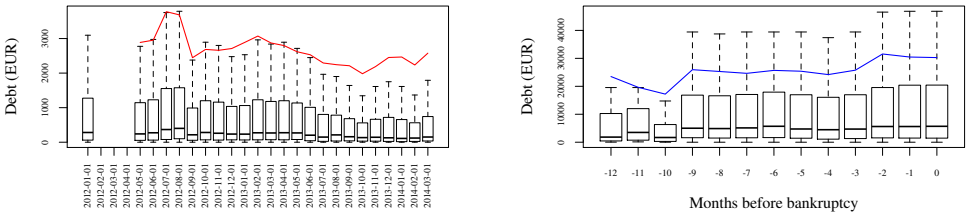
(a) *Normal companies* **(b)** *Before bankruptcy*

Figure 4.10 – *Personal income tax debt measured average, mean, upper and lower quartile.*

4.3.4 Tax Debt Interest

When companies fail to pay the tax on time, they have to pay interest of the remaining debt. A high interest debt indicates that the company has not paid the debt in time and has long lasting financial difficulties.

When examining Figure 4.11 we can see that failed companies' average and mean debt is growing, and the normal companies' debt is decreasing. Also, the average and mean interest debt is higher. We can see very clear increasing debt before bankruptcy. This indicates that companies that are due to bankruptcy do not pay a tax deb interests.



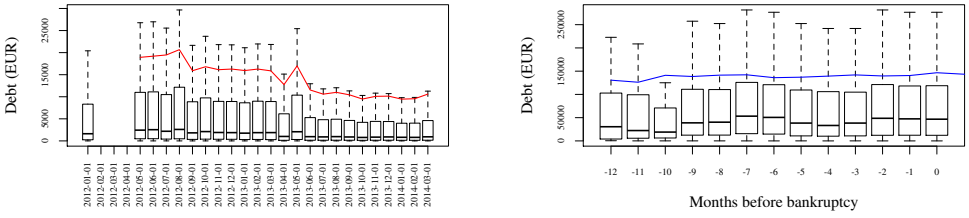
(a) *Normal companies* **(b)** *Before bankruptcy*

Figure 4.11 – *Tax debt interest measured average, mean, upper and lower quartile.*

4.3.5 Total Tax Debt

This variable is calculated by adding all the company different tax liabilities at a certain month. It is done because there are almost 70 different tax debt types available in our dataset and using all of them would not be practical. Most of them are uncommon, and they would result in high number new variables that would slow down the process of training the model and may cause overfitting.

When observing Figure 4.12 we can see that failed companies' average and mean total debt is increasing slowly, however, normal companies have a clear decreasing trend. Also, the average and mean debt is much lower than bankrupted companies have. We can also see that the debt is going up and down, but before the bankruptcy it is clearly higher than 12 months before bankruptcy.



(a) *Normal companies* **(b)** *Before bankruptcy*

Figure 4.12 – *Total tax debt measured average, mean, upper and lower quartile.*

4.3.6 Postponed Debt

Postponed debt shows the amount of the debt that the company has postponed by submitting an application to the tax office, and the tax office has granted the postponement. This can indicate if the company has motivation to deal with the debt or not, and this knowledge can be used to predict the bankruptcy.

When investigating the dataset we discovered that under 5% of the failed companies have postponed their tax debts, so we only included a binary feature (value 1 or 0) that indicates if the company has any of its debts postponed.

4.4 Field of Activity

As the probability of bankruptcy probably vary across the field of activities, it can be used to better predict bankruptcy [3]. In Estonia, the Estonian Classification of Economic Activities (EMTAK) is used to indicate the field of activity of the company. EMTAK is the national version of the international harmonized NACE classification [50].

EMTAK is the basis for determining the field of activities that in turn is an important source of statistics for various categories. Division into fields of activities improves also the international comparability within a category [50].

The dataset has a company main field of activity code (a letter from A to U). However, not all the companies have the activity code available in our dataset. The distribution of normal and bankrupted companies across the different field of activities can be seen in Table 4.2. We can observe that the highest number of failed companies operate in the filed of construction; the next one is a wholesale and retail trade and finally manufacturing.

4.5 Annual Report

The probability of bankruptcy can also be influenced by the fact if the company has submitted its annual report on time. In Estonia, all companies have to submit their previous year annual report at least six months after the end of the financial year.

The dataset does not have data for all the companies, but the difference is clear between normal and failed companies. When observing Figure 4.13 we can see that most (87%)

Code	Description	Normal	Failed
G	Wholesale and retail trade; Repair of motor vehicles and motorcycles	8158	71
F	Construction	4023	101
M	Professional, scientific and technical activities	3731	13
K	Financial and insurance activities	3584	17
C	Manufacturing	3307	61
L	Real estate activities	2439	17
N	Administrative and support service activities	2004	10
H	Transportation and storage	1903	30
J	Information and communication	1383	8
A	Agriculture, forestry and fishing	1133	9
I	Accommodation and food service activities	865	9
S	Other service activities	774	2
R	Arts, entertainment and recreation	493	1
Q	Human health and social work activities	335	1
P	Education	330	1
E	Water supply; sewerage, waste management, remediation activities	169	3
D	Electricity, gas, steam and air conditioning supply	124	0
O	Public administration and defense; compulsory social security	44	2
B	Mining and quarrying	4	0
T	Activities of households as employers; undifferentiated goods and services producing activities for households for own use	1	0

Table 4.2 – *Normal and failed companies known number per field of activity.*

failed companies did not submit their last annual report on time or did not do it at all. This can be a good indicator that companies going soon bankrupt avoid submitting the annual report.

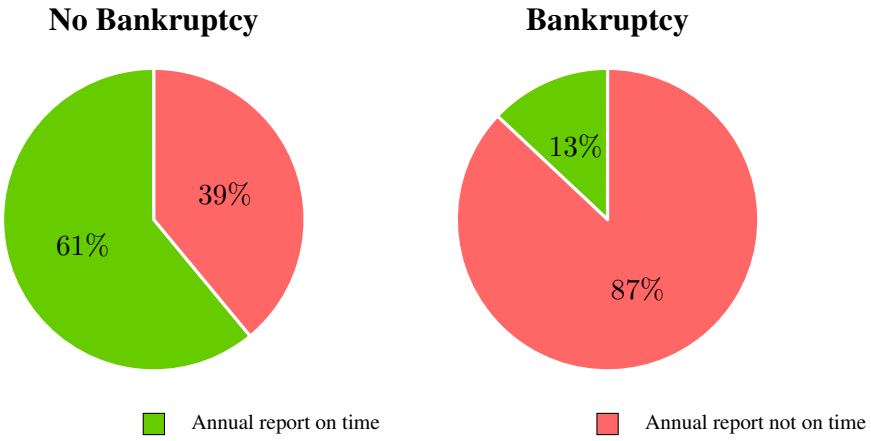


Figure 4.13 – *Percentage of companies that have submitted annual report on time or not.*

5. EXPERIMENTAL RESULTS AND ANALYSES

In this chapter we present the experiments and the results obtained by training and comparing different machine learning techniques described previously using graph metrics, tax debt, the field of activity and annual report data. Our workflow was as follows:

Step one - Choosing predictor variables. The first step was to choose the variables to use to train the models.

Step two - Training the models using the variables selected in step one a) using only graph metrics, b) using only tax debt and the field of activity data, c) using both data combined.

Step three - Choosing the model by comparing key performance metrics to determine the best performing model and validate the hypothesis that the graph metrics improve bankruptcy prediction models.

Step four - Determining the optimal feature set size in months by investigating how large in months should the dataset optimally be to make accurate enough predictions.

Step five - Determining the prediction period by studying how many months in advance the model can accurately classify bankruptcy. Similar approach was used by Philosophov *et al*. [18].

Step six - Proposing practical use. Finally, we propose practical use for our proposed models and present some real predictions using the models we have previously trained.

5.1 Feature Selection

Based on graph metrics and tax debt data observations in Chapter 4 it was optimal to use six months of data, i.e. five months before bankruptcy together with the month after the bankruptcy was declared to train and compare the models. Especially there were clear changes in eccentricity, betweenness and degree bipartite graph metrics six months before bankruptcy. In addition to the observation's analyses, using a larger number of monthly data would not be practical as the training would take much more time and computation power. As well as the main purpose of the first experiment is to compare the models and after that study the optimal dataset size using the best-performed approach.

The sample group consists of both healthy and bankrupted companies. Some samples were filtered out, firstly companies that were isolated such that during the whole time span (2012-01-01 until 2014-03-01) the company had no relations to another firm (read more in Subsection 3.2.2). These companies were removed due to the fact they were isolated, and no changes happened during that time span and their graph metrics were constant and did not evolve. Also, this reduction is important to reduce the sample size as bigger dataset size results more unbalanced samples. In addition companies that did not have data available for at least six months were removed.

The features used for training the models using the debt, the company activity field, and annual report data are presented in Table 5.1.

1	Failed or normal company	2	Value added tax debt (VAT)
3	Social tax debt (FICA)	4	Unemployment insurance tax debt (FUTA)
5	Detained income tax debt	6	Funded pension debt
7	Tax interests (INTERESTS)	8	Total tax debt (TOTAL)
9	Has Debt	10	Has Postponed Debt
11	Acticity sector (SECTOR)	12	Annual report on time (REPORT)

Table 5.1 – *Features used for training models without formal network metrics.*

The features used for training the models using the graph metrics are presented in the Table 5.2. We decided not to use closeness as the closeness is global metric and when observing closeness Figure 4.2 the data had almost no variation as the mean, average and

upper and lower quartile followed the same decreasing line. Due to that there was danger of overfitting. To address the problem we also conducted some experiments that showed that not using this metric did not degrade performance when predicting bankruptcy in advance.

1	Failed or normal company	2	Degree (PR)
3	Degree (PP)	4	Pagerank (PR)
5	Pagerank (PP)	6	Betweeness (PR)
7	Betweeness (PP)	8	Eccentricity (PR)
9	Eccentricity (PP)	10	Transitivity (PP)

PP - Graph projection metric

PR - Bipartite graph metric

Table 5.2 – *Features used for training models using formal network metrics.*

5.2 Choosing the Model

Due to the small number of bankrupted companies the minority class oversampling was applied to the dataset, 10-fold cross validation was used to train the models. It works by randomly choosing 70% of both failed and healthy companies as a training sample and 30% of samples are used as a test sample. This setup also helps to reduce the effects of any possible over fitting problems, although there is no complete guarantee. In addition, using the test sample group is important as the trained model has not seen those samples yet. The same or similar setup was used in studies [8, 4, 26].

We trained all the models described previously (RF, SM, DT, BDT, NB, GLM and NN) using three different feature sets. Firstly, only tax debt and company activity field and annual report data was used, and the results of the test samples are observed in Table A.1. Then the models were trained by using only graph metrics data, and the results of that experiment are presented in Table A.2. Lastly, all the models were trained using both datasets combined, and the results of that experiment are shown in Table A.3. The results comparing F-Score are seen in Figure 5.1, comparison of recall is seen in Figure 5.2 and precision in Figure 5.3.

In Figure 5.1 we can see that the best performing model was RF. It leads across all three feature sets. Second was SVM and then came BDT and DT. All three RF, BDT and DT are a

decision tree based so in conclusion, the decision trees based approach gives the best results using the proposed features and our dataset.

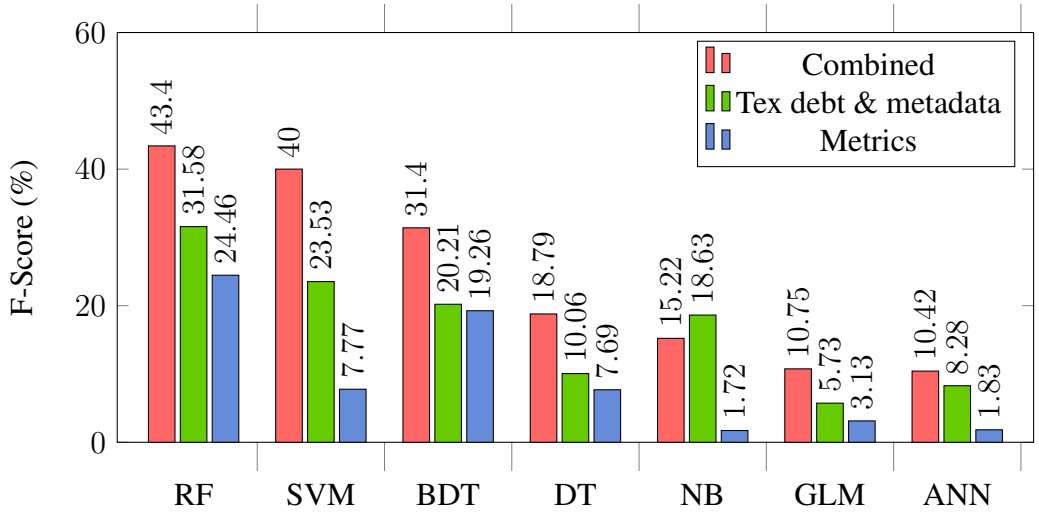


Figure 5.1 – *Comparison of combined, graph metrics and financial models validation F-score results.*

Figure 5.2 and 5.3 indicate that RF did not result best performance metrics, though, one can see that when recall is high then precision is low, because the model classifies positive samples well, but it also produces a high number of false positive classifications. F-score takes both in account and is more suitable when comparing model overall performance.

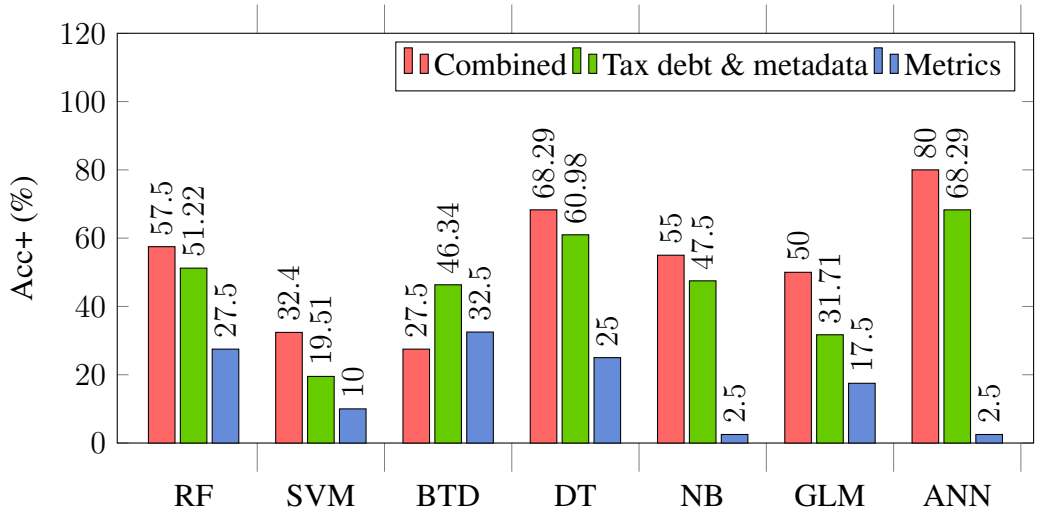


Figure 5.2 – *Comparison of combined, graph metrics and financial models validation recall score results.*

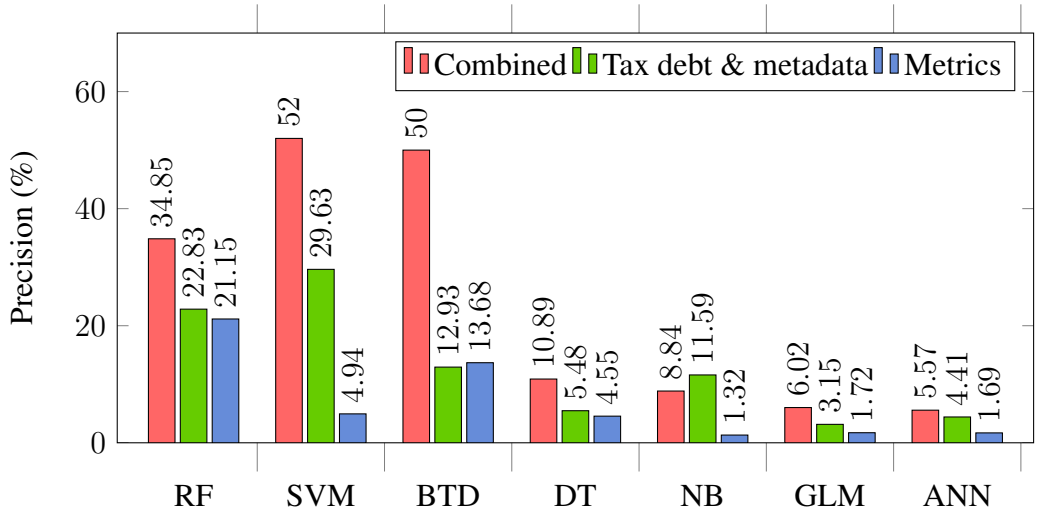


Figure 5.3 – *Comparison of combined, graph metrics and financial models validation precision score results.*

Using different feature sets the best-performed model was RF that yielded the best F-score in each feature set category. Based on these results we decided to use RF approach to conduct next experiments with. Also, the results clearly indicate that using companies formal network graph metrics evolution data improves company bankruptcy classification whatever the technique we used with one exception - NB - that yielded the best results using tax debt data.

5.3 Optimal Feature Set Size in Months

In this section, we investigate how many months of data we optimally need to make accurate predictions. To find that out RF model was trained using increasing dataset size, namely firstly using only one month of data, next two months and up to 12 months of data. This time SMOTE and ENN techniques were not used to sample the dataset instead all the samples were used, because RF can handle unbalanced datasets, and it is recommended to use as many samples as possible. The training method was the same, 10-fold cross validation while 70% of the samples were used as a training set, and 30% were used for testing the model, similarly in the previous experiment.

The results of this experiment are observed in Figure 5.4. We discovered that using 11 months data gives the best results, however, when considering our dataset short time

span size, it is not optimal to use so large number of monthly data in our next experiments. The subsequent best choice is six months and after that three and four months. In the next section, we used four and six months data to try to determine how many months in advance we can accurately classify bankruptcy.

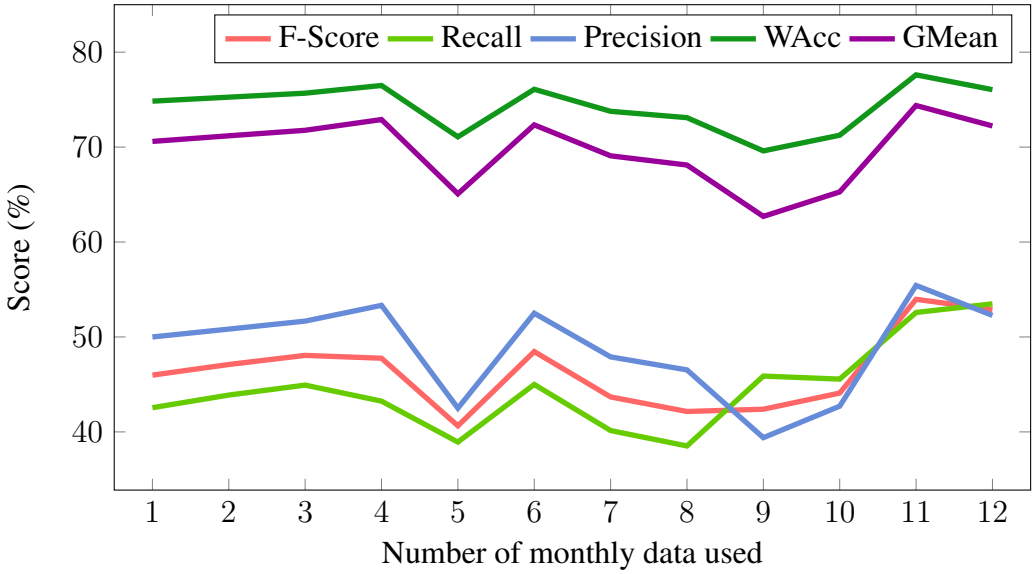


Figure 5.4 – *Key performance metrics comparison using different data set size in months.*

5.4 Prediction Period

Previously the models were trained using monthly data starting from the month when the bankruptcy was recorded and, according to the results, a conclusion was made, that the graph metrics can be used to classify and predict bankruptcy. However, this type of model has no practical purpose as one would want to predict bankruptcy at least some reasonable time in advance to use the probability to make decisions. In this section, we try to determine how long in advance the model can predict the bankruptcy while maintaining reasonable good F-Score.

The same setup as in the previous experiment was used, i.e. no sampling and 70% of samples used as a training set and 30% as a testing set and the training method was 10-fold cross-validation. The results of this experiment are presented in Figure 5.5. We can observe that using our proposed approach the model can predict bankruptcy using six months of data reasonably good nine months in advance before bankruptcy is declared. The models

scored between 40% and 58% F-Score. After eight months, key performance metrics start to decrease. One can also notice that predicting 10 months in advance yields very low F-Score. This is most probably due to that the dataset contains little positive samples for a period that long.

We also tried to use only four months of data to optimize the amount of features needed to make predictions; however, using only four months of data produced worse results than using six.

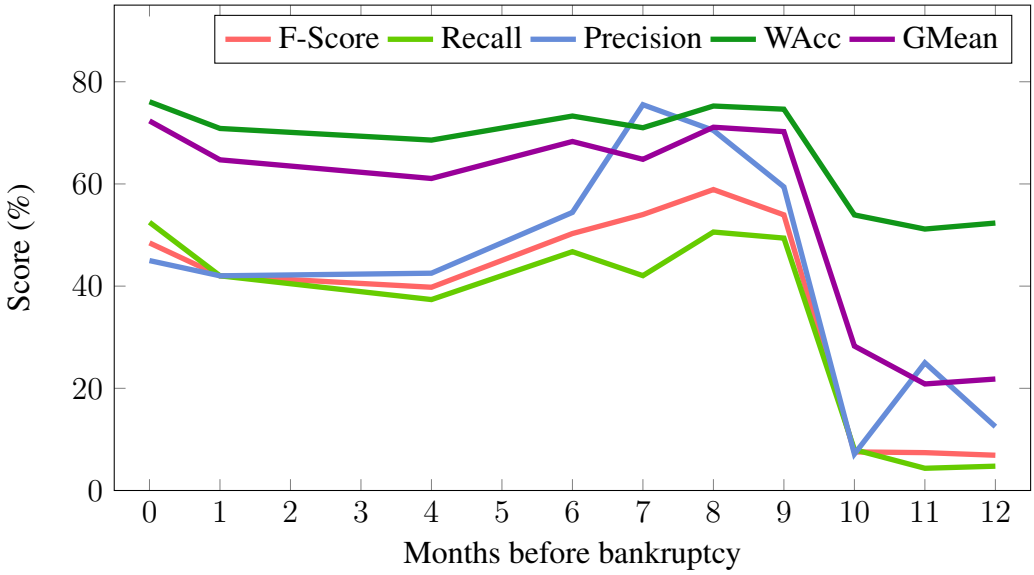


Figure 5.5 – *Performance metrics over time horizons between 0 and 12 months before bankruptcy.*

5.5 Variable Importance Analyze

In this section, we analyze the variables that were used to classify bankruptcy. The main motivation behind this analysis is to study what variables influence bankruptcy as the model produced by best-performed technique, random forests, is not easily interpretable, unlike individual classification tree, we must use other techniques to evaluate the variable importance.

Firstly, we describe which variables we found significant based on features observations in Chapter 4. We concluded that an average bankrupted business has the following description. The last annual report is not submitted on time; the business field of activity is construction, wholesale, manufacturing or transportation; firm has tax debts, and the total

tax debt is over 20 000 EUR and is growing; tax debt interest is also increasing and is at least or more than 1000 EUR. Finally, the company is connected to three or more companies through its board members.

Secondly, we used DT to construct a classification tree. The decision tree graphical representation without attribute tests is seen in Figure 5.6; for full textual representation with attribute tests see Appendix B. To train the model, all positive and negative samples were used. The training method was 10-fold cross-validation and all the features described in Table 5.2 (graph metrics, blue nodes) and Table 5.1 (tax and meta data, red nodes) were used. Each node represents feature and path to the child node represents a decision. The leaf nodes show how many samples do follow that path and their distribution. N being normal company and Y being bankrupted sample. We can see that most of the variables presented in the tree are tax debt related, but only eccentricity and PageRank are from graph metrics set.

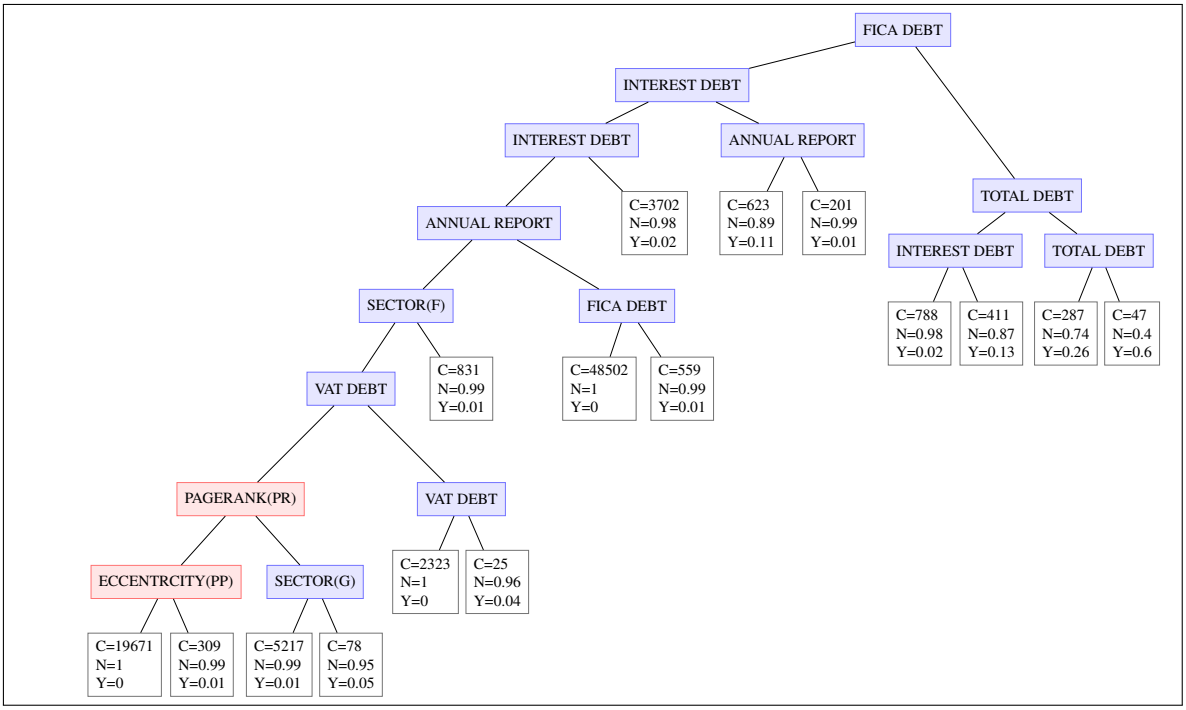


Figure 5.6 – *Simplified version (without attribute tests) of bankruptcy classification decision tree.*

Finally, we used random forests variable importance measures to assess variables importance. The measure is computed from permuting OOB data: For each tree, the prediction error on the out-of-bag portion of the data is recorded. Then the same is done after permuting each predictor variable. The difference between the two are then averaged over all trees,

and normalized by standard deviation of the differences [51]. Same setup, as with decision tree training, was used. The results are seen in Figure 5.7. Other remaining less important variables are not shown in the figure. We can clearly see that in general tax related variables are more important than graph metrics; however, both PageRank versions are considered very important and are placed second and third.

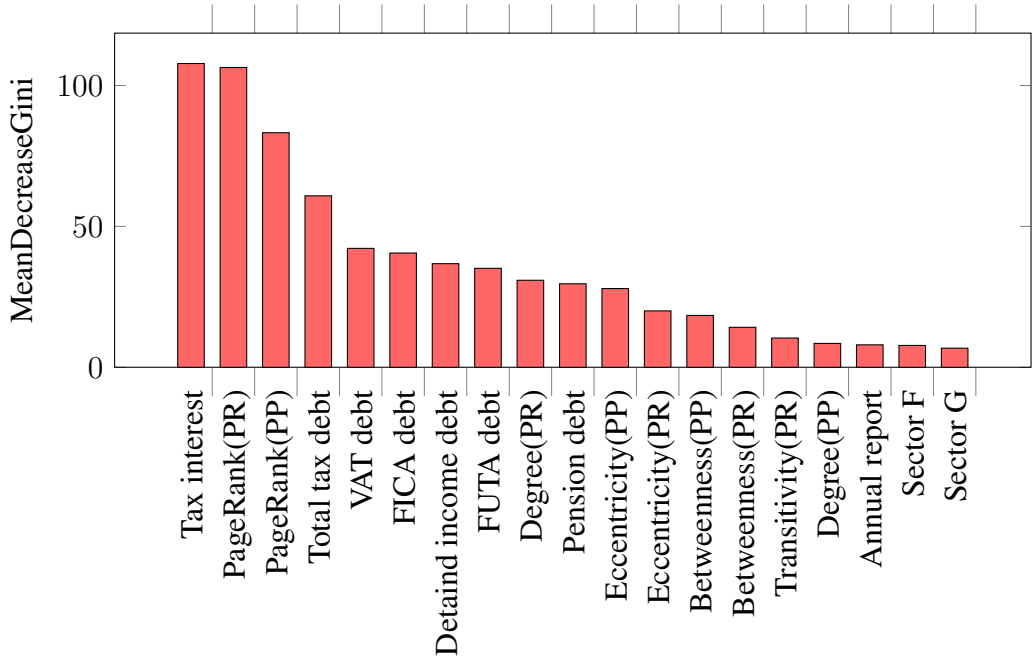


Figure 5.7 – *Random forest variable importance measures.*

In conclusion, we can argue that according to random forest importance measure and classification decision tree graph metrics especially PageRank and Eccentricity are important when classifying bankruptcy. All three approaches revealed that tax interest debt is important as bankrupted companies have accumulated high debt. It was clearly seen on Figure 4.11, appeared several time on the decision tree and was most important based by RF importance plot. Also, activity fields like construction (code F) and wholesale (code G) are relevant. However, we note that current analyzes do not consider the evolution of the metrics as only one month of data was used when training the models.

5.6 Practical Usage

Based on previous experiments that concluded that it is optimal to use a set of at least six months of data, and the fact that the proposed RF model could classify bankruptcy reasonably a good nine months in advance we proposed an algorithm to utilize these models in a practical application. The proposed method computes the chosen company's most probable month to go bankrupt by comparing different month classification results together with the probability that bankruptcy can happen during that month or not. The procedure can be described by following pseudo code:

$Month \leftarrow 1$
$Data \leftarrow getCompanyData()$
$ProbabilityByMonth \leftarrow []$
while $Model \leftarrow getTrainedModel(Month)$ **do**
  $ProbabilityByMonth[Month] \leftarrow predict(Model, Data)$
  $Month++$
end while
$MostProbableMonth \leftarrow key(max(ProbabilityByMonth))$
$Probability \leftarrow ProbabilityByMonth[MostProbableMonth]$

Where $getCompanyData()$ returns company's formal network metrics, debt and metadata from previous 6 months, $getTrainedModel()$ returns corresponding trained month based on its first argument. $ProbabilityByMonth$ is a list that holds the company's computed bankruptcy probability for each month. In the end, the algorithm finds the highest probability from the $ProbabilityByMonth$ list and its corresponding month.

This algorithm can be appropriately illustrated by one real example. Using the method we computed the following bankruptcy probabilities: 1) 0.16 2) 0.244 3) 0.226 4) 0.198 5) 0.116 6) 0.706 7) 0.264 8) 0.1 9) 0.032. Based on the classification results, we can clearly see that this company most probably will go bankrupt during the sixth month. Assuming that the bankruptcy will occur during the next nine months the probabilities of each month can be represented as valid probability distribution. It is found by adding all probabilities and dividing each probability by the previously found sum (see Figure 5.8). Now we can say that the probability of the bankruptcy happening during the sixth month is 35%.

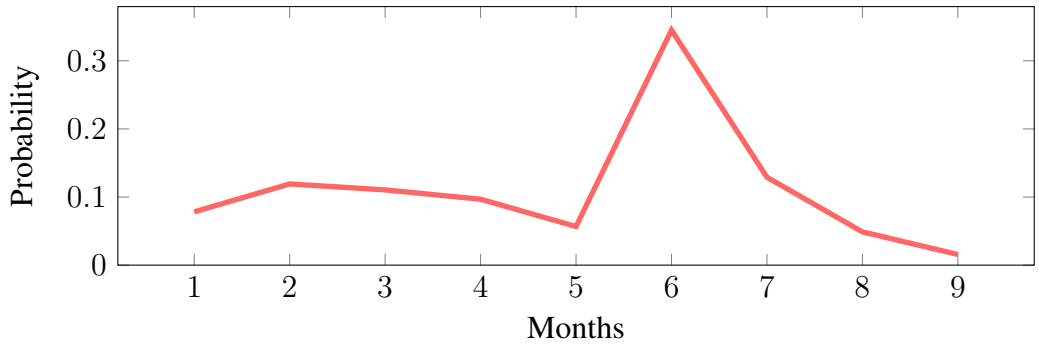


Figure 5.8 – *Bankruptcy probability distribution of a sample company.*

Finally, we applied this procedure to all companies that are not bankrupted yet and forecast how many companies will default in the next nine months. The results are plotted in Figure 5.9. On the plot, we can observe how many companies have previously gone bankrupt (black line), and the number of companies the models classified that can go bankrupt in the next months. The high number of bankruptcies at the beginning of 2013 is caused by the data set quality as in some instances the exact bankruptcy date was not available at the time the data was recorded. The red line represents a confident level of 50%. This means it represents the number of companies that have the bankruptcy probability of 50% or higher. The dark gray area represents a confidence interval between 40% and 60% and light gray 25% to 75%. Every company is plotted once i.e.; the most probable month is shown.

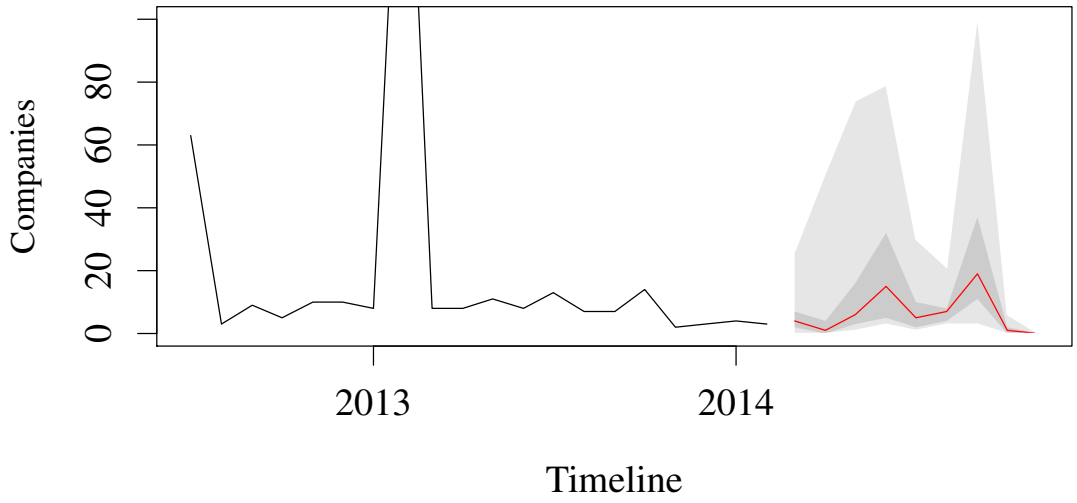


Figure 5.9 – *Predicted number of bankruptcies in the future based on most probable bankruptcy month. Red line is confidence level 50%, dark gray 60% to 40% and light gray 75% to 25%.*

5.7 Threats to Validity

Firstly, this study is based on a relatively small sample of 500 Estonian companies that became bankrupt within an arguably short time interval from 2012 through the beginning of 2014. A more comprehensive study, including a larger samples of firms and covering bankruptcies of latter years, is desirable for a better understanding of how graph evolution metrics affect the bankruptcy prediction. Also, the results of this study could be affected by the unbalanced dataset, but there are not many solutions to this threat other than using a larger number of samples, as bankruptcy is a rare event.

Another potential threat is the quality of the data. There are a few periods where data was missing from the dataset due to reasons beyond our reach. For example, tax debt dataset did not have information about three months including 2012 February to April and some bankruptcy dates in the middle of 2012 and at the beginning of 2013 are possibly marked with a later date than the bankruptcy really happened.

This study did not cover all possible machine learning techniques, only those that were previously mostly used and available in the chosen R software and tool set. Therefore some approaches could perform better using our dataset and chosen features. Due to the limited time, we also did not try to optimize the models using different model-specific techniques and therefore some models could perform better than indicated in this study. For example artificial neural networks have historically performed reasonably well when using financial ratios to predict bankruptcy however in our case NN performed the worst when using tax and graph metrics data.

There are some more graph metrics that we did not use as their computation was not possible using the R software and tool set we used. We also decided not to use closeness graph metric as it is global metric and based on observations and experiments caused worse performance when predicting bankruptcy in advance. Therefore more analyses should be performed using other graph metrics and the effect of global metrics.

Finally, due to the previously described potential threats one should not rely only on computed bankruptcy probability using the proposed procedure in the previous section when making business decisions because a greater number of positive samples are needed to confirm the suitability of the model for practical use. However, it should be safe to use the

probability as an indicator of possible risk or use it as one component in credit-risk computation for a company.

6. CONCLUSION AND FUTURE WORK

Predicting bankruptcy is an important problem since it can have a significant impact on business decisions and profitability and is therefore a well-researched topic. However, usually only company financial ratios have been used to make predictions, but as fiscal reports are published long after the end of the financial year the predictions are not up to date.

This study researched if using the company board member formal network graph metrics evolution can improve bankruptcy prediction accuracy. Unlike financial data, the company's board member changes are available without notable delay and using board member graph metrics to predict bankruptcy has not been studied widely yet.

We were able to demonstrate that indeed using graph metrics, especially PageRank, degree and eccentricity, can improve classification models key performance metrics. We trained and compared different machine learning techniques and concluded that the best-performing approach using our dataset and chosen features was decision tree based random forests. Then we determined that the optimal size of the feature set is six months. Finally, we studied that our proposed model can optimally predict bankruptcy nine months in advance using our dataset. We also proposed practical application based on our findings and presented algorithm that can compute the probability if the company would go bankrupt or not for each month for the next nine months. In addition, we calculated how many companies could go bankrupt in the next following nine months.

This study has the following limitations that need further research. Firstly, the dataset used in this study had data from January 2012 to February 2014 along with only about 500 samples of Estonian bankrupted companies. To verify the accuracy of the model proposed

one should use larger dataset with a higher positive sample number to verify the findings. Also, it is recommended to investigate if using global graph metrics such as PageRank and especially closeness that was not used in this study are justified and are not causing over fitting the models and improve the prediction accuracy.

An interesting direction of further research should consist in involving widely used financial data and the company's information, like the number of employees, location and evaluating their contribution on improving the efficiency of the proposed model ability to classify a bankruptcy. In addition, it would be interesting if the random forest model could be tuned to perform even better.

Acknowledgments. I would like to thank my supervisor, Dr. Peep Küngas, for his continuous support for this research. I would also like to thank the Estonian Information Technology Foundation and Skype Technologies OÜ for giving me a scholarship and supporting my master studies.

Juhatuse liikmete ja firmade võrgu meetrikate mõju firmade pankrottide ennustamisel

Magistritöö (30 eap)

Taavi Ilves

Resümee

Pankroti ennustamine on oluline probleem, kuna firmade pankrotid võivad mõjutada ettevõtete ja asutuste kasumlikkust ning pankroti potensiaal on aluseks ärilistele otsustele. Tänu selle tähtsusele on antud probleemi varemgi uuritud, kuid enamasti on panrkoti ennustamiseks kasutatud ettevõtete varasemaid finantsandmeid, aga kuna finantsandmed avaldatakse peale finantsaasta lõppu, ei ole sel moel tehtud ennustused enam ajakohased.

Käesoleva töö eesmärk oli uurida kas ettevõtte juhatuse liikmete graafi meetrikate muutumise dünaamika võiks aidata pankroti ennustamise täpsust parandada, kuna juhatuse liikmete muudatused on kättesaadavad ilma viivituseta ning tänu sellele on ennustused ajakohasemad ja juhatuse liikmete dünaamika mõju pankroti ennustamiseks on varasemalt vähe uuritud.

Eksperimentide tegemisel kasutasime Eesti ettevõtete juhatuse liikmete andmeid alates 2012 jaanuarist kuni 2014 märtsini. Töös võrdlesime erinevaid masinõppe meetodeid (otsustuspuu, *random forests*, tehislikud närvivõrgud, tugivektor-masinad, naiivne Bayes) kasutades iga ühe puhul 3 andmekogu - juhatuse liikmete graafi meetrikad, firma maksuvõlad ning firmat iseloomustavad andmed (tegevusvaldkond, aasta aruande esitamine) ja mõlemad kombineeritult. Seejärel võrdlesime mudeleid klassifitseerimise meetrikate abil ja jõudsime järeldusele, et graafi meetrikad, eelkõige *PageRank*, *degree* ja *eccentricity*, parandavad ennustuste täpsust ning parimaks meetodiks osutus otsustuspuudel põhinev *random forests*.

Seejärel uurisime kui suurt andmekogu on optimaalne kasutada ning leidsime, et selleks on 6 kuu andmed. Samuti katsetasime, kui palju on võimalik pankroti ette ennustada nii, et mudel oleks veel effektiivne ja leidsime, et antud andmetega saab pankroti ette ennustada kuni 9 kuud. Lõppuks pakkusime välja meie poolt koostatud mudelile praktilise lahenduse ning arvutasime välja järgneva 9 kuu potensiaalsed pankroti minevad firmad.

Bibliography

[1] Altman, Edward I. ”Financial ratios, discriminant analysis and the prediction of corporate bankruptcy.” The journal of finance 23.4 (1968): 589-609.

[2] Canbas, Serpil, Altan Cabuk, and Suleyman Bilgin Kilic. ”Prediction of commercial bank failure via multivariate statistical analysis of financial structures: the Turkish case.” European Journal of Operational Research 166.2 (2005): 528-546.

[3] Lennox, Clive. ”Identifying failing companies: a re-evaluation of the logit, probit and DA approaches.” Journal of Economics and Business 51.4 (1999): 347-364.

[4] Shin, Kyung-Shik, Taik Soo Lee, and Hyun-jung Kim. ”An application of support vector machines in bankruptcy prediction model.” Expert Systems with Applications 28.1 (2005): 127-135.

[5] du Jardin, Philippe. ”Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy.” Neurocomputing 73.10 (2010): 2047-2060.

[6] Lee, Woh-Chiang. ”Genetic Programming Decision Tree for Bankruptcy Prediction.” JCIS. 2006.

[7] Kartasheva, Anastasia V., and Mikhail Traskin. ”Insurers Insolvency Prediction using Random Forest Classification.” (2011).

[8] Park, Cheol-Soo, and Ingoo Han. ”A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction.” Expert Systems with Applications 23.3 (2002): 255-264.

[9] Krediidiinfo AS. ”PANKROTID EESTIS 2013.” Paneeluuring.

[10] Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos. ”Graphs over time: densification laws, shrinking diameters and possible explanations.” Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005.

[11] Iba, Takashi, et al. ”Analyzing the creative editing behavior of Wikipedia editors: through dynamic social network analysis.” Procedia-Social and Behavioral Sciences 2.4 (2010): 6441-6456.

[12] Lambiotte, R., and M. Ausloos. ”Uncovering collective listening habits and music genres in bipartite networks.” Physical Review E 72.6 (2005): 066107.

[13] Dhanjal, Charanpal, and Stphan Clmenon. ”Learning Reputation in an Authorship Network.” arXiv preprint arXiv:1311.6334 (2013).

[14] Chava, Sudheer, and Robert A. Jarrow. ”Bankruptcy prediction with industry effects.” Review of Finance 8.4 (2004): 537-569.

[15] Beaver, William H. ”Financial ratios as predictors of failure.” Journal of accounting research (1966): 71-111.

[16] Creamerk, Germn. ”Using Random Forests and Logistic Regression for Performance Prediction of Latin American ADRS and Banks.” Journal of CENTRUM Cathedra 2.1 (2009): 24-36.

[17] Sarkar, Sumit, and Ram S. Sriram. ”Bayesian models for early warning of bank failures.” Management Science 47.11 (2001): 1457-1475.

[18] Philosophov, Leonid V., Jonathan A. Batten, and Vladimir L. Philosophov. ”Predicting the event and time horizon of bankruptcy using financial ratios and the maturity schedule of long-term debt.” Mathematics and Financial Economics 1.3-4 (2008): 181-212.

[19] Huang, Keman, Yushun Fan, and Wei Tan. ”An Empirical Study of Programmable Web: A Network Analysis on a Service-Mashup System.” Web Services (ICWS), 2012 IEEE 19th International Conference on. IEEE, 2012.

[20] Berlingerio, Michele, et al. ”Mining graph evolution rules.” Machine learning and knowledge discovery in databases. Springer Berlin Heidelberg, 2009. 115-130.

[21] Nicosia, Vincenzo, et al. ”Graph metrics for temporal networks.” Temporal Networks. Springer Berlin Heidelberg, 2013. 15-40.

[22] Bilgin, Cemal Cagatay, and Blent Yener. ”Dynamic network evolution: Models, clustering, anomaly detection.” IEEE Networks (2006).

[23] ”Introduction to R.” R-project. 4. April. 2014. http://www.r-project.org/

[24] ”Introduction.” The igraph library. The igraph Project. 28. Feb. 2014. http://igraph.sourceforge.net/introduction.html

[25] ”Introduction.” The caret Package. 4. April. 2014. http://caret.r-forge.r-project.org/

[26] Lee, TsunSiou, and YinHua Yeh. ”Corporate governance and financial distress: evidence from Taiwan.” Corporate Governance: An International Review 12.3 (2004): 378-388.

[27] Kumar, Ravi, et al. ”On the bursty evolution of blogspace.” World Wide Web 8.2 (2005): 159-178.

[28] ”Bipartite network projection.” Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 31 July 2013. Web. 28 Feb. 2014.

[29] Huang, Hsinyu Liu Shianchang. ”Integrating GA with boosting methods for financial distress predictions.” Journal of Quality Vol 17.2 (2010): 131.

[30] ”Logistic regression.” Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 11 January 2014. Web. 26 Jan. 2014.

[31] ”Support vector machine” Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 19 January 2014. Web. 26 Jan. 2014.

[32] ”Naive Bayes classifier” Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 1 December 2013. Web. 27 Jan. 2014.

[33] ”Artificial neural network” Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 24 January 2014. Web. 27 Jan. 2014.

[34] Ravi Kumar, P., and Vadlamani Ravi. ”Bankruptcy prediction in banks and firms via statistical and intelligent techniques a review.” European Journal of Operational Research 180.1 (2007): 1-28.

[35] ”Random forrest” Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 23 January 2014. Web. 27 Jan. 2014.

[36] Chen, Chao, Andy Liaw, and Leo Breiman. ”Using random forest to learn imbalanced data.” University of California, Berkeley (2004).

[37] Kubat, Miroslav, and Stan Matwin. ”Addressing the curse of imbalanced training sets: one-sided selection.” ICML. Vol. 97. 1997.

[38] ”Receiver operating characteristic” Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 28 January 2014. Web. 28 Jan. 2014.

[39] Hanczar, Blaise, et al. ”Small-sample precision of ROC-related estimates.” Bioinformatics 26.6 (2010): 822-830.

[40] He, Haibo, and Edwardo A. Garcia. ”Learning from imbalanced data.” Knowledge and Data Engineering, IEEE Transactions on 21.9 (2009): 1263-1284.

[41] Batista, Gustavo EAPA, Ronaldo C. Prati, and Maria Carolina Monard. ”A study of the behavior of several methods for balancing machine learning training data.” ACM Sigkdd Explorations Newsletter 6.1 (2004): 20-29.

[42] Suman, Sanjeev, Kamlesh Laddhad, and Unmesh Deshmukh. ”Methods for Handling Highly Skewed Datasets.” Part I-October 3 (2005).

[43] Weiss, Gary M., and Foster J. Provost. ”Learning when training data are costly: The effect of class distribution on tree induction.” J. Artif. Intell. Res.(JAIR) 19 (2003): 315-354.

[44] ”Degree (graph theory).” Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 5 January 2014. Web. 19 Jan. 2014.

[45] ”Betweenness centrality.” Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 6 February 2014. Web. 2 March. 2014.

[46] Yoon, Jeongah, Anselm Blumer, and Kyongbum Lee. ”An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality.” Bioinformatics 22.24 (2006): 3106-3108.

[47] ”Clustering coefficient.” Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 11 December 2013. Web. 2 March. 2014.

[48] ”PageRank” Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 28 January 2014. Web. 28 Jan. 2014.

[49] Brin, Sergey, and Lawrence Page. ”The anatomy of a large-scale hypertextual Web search engine.” Computer networks and ISDN systems 30.1 (1998): 107-117.

[50] ”EMTAK fields of activities.” E-Business Register. Centre of Registers and Information Systems. 12. May. 2014. http://www.rik.ee/en/e-business-registry/emtak-fields-activities

[51] ”Breiman and Cutlers random forests for classification and regression.” Package randomForest. 4. May. 2014. http://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Appendices

A. Model Comparison Results

A.1 Tax Debt and Sector Data Results

Model	AUC	Cut*	Acc^{+}	Acc^{-}	PPV	F1	G-mean	WACC
RF	95.3	0.9	51.22	99.15	22.83	31.58	71.26	75.18
BDT	96.6	0.9	46.34	98.46	12.93	20.21	67.55	72.40
DT	79.6	0.8	60.98	94.82	5.48	10.06	76.04	77.90
NB	97.3	0.9	47.50	98.26	11.59	18.63	68.32	72.88
GLM	97.8	0.1	31.71	95.19	3.15	5.73	54.94	63.45
ANN	62.3	0.8	68.29	92.70	4.41	8.28	79.57	80.50
SVM	94.8	0.9	19.51	99.77	29.630	23.53	44.12	59.64

*The cut off value was choosen to maximize the F1 score and weighted acuraccy measure.

Table A.1 – *Results of financial features using SMOTE+ENN sampling*

A.2 Graph Metrics Data Results

Model	AUC	Cut*	Acc^{+}	Acc^{-}	PPV	F1	G-mean	WACC
RF	95.3	0.9	27.50	99.48	21.15	23.91	52.30	63.49
BDT	96.6	0.9	32.50	98.96	13.68	19.26	56.71	65.73
DT	79.6	0.9	25.00	97.34	4.55	7.69	49.33	61.17
NB	97.3	0.9	2.50	99.05	1.32	1.72	15.74	50.77
GLM	97.8	0.7	17.50	94.91	1.72	3.13	40.75	56.21
ANN	62.3	0.9	2.50	99.26	1.69	2.02	15.75	50.88
SVM	94.8	0.9	10.00	99.02	4.94	6.61	31.47	54.51

*The cut off value was choosen to maximize the F1 score and weighted acuraccy measure.

Table A.2 – *Results of graph metric features using SMOTE+ENN sampling*

A.3 Combined Data Results

Model	AUC	Cut*	Acc^{+}	Acc^{-}	PPV	F1	G-mean	WACC
RF	95.3	0.9	57.50	99.45	34.85	43.40	75.62	78.48
BDT	96.6	0.9	27.50	99.86	50.00	35.48	52.40	63.68
DT	79.6	0.8	68.29	97.25	10.89	18.79	81.49	82.77
NB	97.3	0.9	55.00	97.12	8.84	15.22	73.09	76.06
GLM	97.8	0.1	50.00	96.04	6.02	10.75	96.04	69.30
ANN	62.3	0.9	80.0	93.12	5.57	10.42	86.31	86.56
SVM	94.8	0.9	32.40	99.85	52.00	99.85	56.97	66.17

*The cut off value was choosen to maximize the F1 score and weighted acuraccy measure.

Table A.3 – *Results of combined features using SMOTE+ENN sampling*

B. Decision Tree Textual Representation

```
1  1) FICA <= 5454.06; criterion = 1, statistic = 5096.897
2    2) INTEREST <= 832; criterion = 1, statistic = 789.959
3      3) INTEREST <= 0; criterion = 1, statistic = 314.436
4        4) REPORT <= 0; criterion = 1, statistic = 90.754
5          5) SECTOR(F) <= 0; criterion = 1, statistic = 24.949
6            6) VAT <= 0; criterion = 1, statistic = 19.771
7              7) PAGERANK(PR) <= 1.335303e-05; criterion = 0.997, statistic = 17.946
8                8) ECCENTRICITY(PP) <= 37; criterion = 1, statistic = 40.613
9                  9)*  weights = 19671
10               8) ECCENTRICITY(PP) > 37
11                 10)*  weights = 309
12             7) PAGERANK(PR) > 1.335303e-05
13               11) SECTOR(H) <= 0; criterion = 0.995, statistic = 14.813
14                 12)*  weights = 5217
15               11) SECTOR(GH) > 0
16                 13)*  weights = 78
17           6) VAT > 0
18             14) VAT <= 239849.9; criterion = 1, statistic = 33.969
19               15)*  weights = 2323
20             14) VAT > 239849.9
21               16)*  weights = 25
22         5) SECTOR(F) > 0
23           17)*  weights = 831
24       4) REPORT > 0
25         18) FICA <= 1271.14; criterion = 1, statistic = 72.588
26           19)*  weights = 48502
27         18) FICA > 1271.14
28           20)*  weights = 559
29     3) INTEREST > 0
30       21)*  weights = 3702
31   2) INTEREST > 832
32     22) REPORT <= 0; criterion = 0.999, statistic = 18.214
33       23)*  weights = 623
34     22) REPORT > 0
35       24)*  weights = 201
36 1) FICA > 5454.06
37   25) TOTAL <= 71300.79; criterion = 1, statistic = 132.25
38     26) INTEREST <= 1269.9; criterion = 1, statistic = 39.903
39       27)*  weights = 788
40     26) INTEREST > 1269.9
41       28)*  weights = 411
42   25) TOTAL > 71300.79
43     29) TOTAL <= 353993.8; criterion = 0.997, statistic = 15.392
44       30)*  weights = 287
45     29) TOTAL > 353993.8
46       31)*  weights = 47
```

C. License

Non-exclusive licence to reproduce thesis and make thesis public

I, Taavi Ilves (date of birth: 24.03.1989),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

   of my thesis
   Impact of Board Dynamics in Corporate Bankruptcy Prediction: Application of Temporal Snapshots of Networks of Board Members and Companies,
   supervised by Peep Küngas,

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 26.05.2014