



Business Credit Scoring of Estonian Organizations

Jüri Kuusik^{1,3(✉)} and Peep Kungas^{1,2}

¹ Register OÜ, Tallinn, Estonia

{jyri.kuusik, peep.kungas}@ir.ee

² University of Tartu, Tartu, Estonia

³ STACC, Tartu, Estonia

Abstract. Recent hype in social analytics has modernized personal credit scoring to take advantage of rapidly changing non-financial data. At the same time business credit scoring still relies on financial data and is based on traditional methods. Such approaches, however, have the following limitations. First, financial reports are compiled typically once a year, hence scoring is infrequent. Second, since there is a delay of up to two years in publishing financial reports, scoring is based on outdated data and is not applied to young businesses. Third, quality of manually crafted models, although human-interpretable, is typically inferior to the ones constructed via machine learning.

In this paper we describe an approach for applying extreme gradient boosting with Bayesian hyper-parameter optimization and ensemble learning for business credit scoring with frequently changing/updated data such as debts and network metrics from board membership/ownership networks. We report accuracy of the learned model as high as 99.5%. Additionally we discuss lessons learned and limitations of the approach.

Keywords: Business credit scoring · Machine learning · Boosted decision tree
Hyper-parameter tuning

1 Introduction

Credit scoring is an effective tool to assess credit risks of individuals and businesses. In business credit scoring literature credit risk is often defined as likelihood of default and its numerical representation is called credit score. In this paper, credit score is defined as likelihood of default within the following 12 months such as commonly used in industry [24].

Most often credit scoring is applied in the context of loan applications (in B2C domain) or when setting payment terms of invoices (in B2B domain). Hence the quality of a credit scoring model affects directly the amount of cash companies loose. In the context of business credit scoring, history of credit risk modelling goes back to late 1960s when numerous studies were devoted to model business failure using publicly available data and combining it with statistical classification techniques [27]. Pioneering work and one of the first attempts to perform modern statistical failure analysis was done by Tamari [3]. Since 1990s there has been a steady growth in the

number of articles related to credit risk using artificial intelligence and machine learning methods [22, 23, 26]. A few possible reasons can be associated with the increased interest in credit risk modelling - rapid development of some new data mining techniques, the availability of more open credit datasets, the growth of credit products and credit markets [4]. By following the trends, first attempts have been made in application of machine learning methods (support vector machines [28], neural networks [29] and extreme gradient boosting (XGBoost) [1, 26]). XGBoost [16] has proven its superiority in discriminating negative and positive events wrt other learning methods [1, 26] including very popular logistic regression which is extensively used in credit scoring area [2]. Further details on recent advances in the field can be found from a review compiled by Zięba et al. [26].

The business credit scoring models still rely mainly on financial data having the following drawbacks. First, financial reports are compiled typically once a year leading to infrequent scoring. Second, since there is a delay of up to a year in publishing financial reports, scoring is based on outdated data.

Regarding our contribution and goal of paper we advanced the state of art in business credit scoring by demonstrating that combination of rapidly changing datasets and top machine learning method XGBoost will lead to business credit scoring model with utmost accuracy in practical settings. More specifically, we used features extracted from company debts and network metrics from board membership/ownership networks in addition to features used in business credit scoring with traditional approaches. Based on such features we defined multiple models, which in turn we combined with ensemble learning into a model, which will predict with accuracy as high as 99.5% what will be the likelihood of a default event in a company within the following 12 months. Also, by our knowledge we made a first attempt to model forced deletions of companies together with bankruptcies in one model along with using Bayesian hyper-parameter optimization. Wrt model validation credit decisions made by credit specialists were compared with credit scores of fitted model. Wrt model description approach not implemented in [1, 26] has been used (described in Sect. 6.2).

The paper is organized as follows. Problem description is explicitly formulated in Sect. 2. In Sect. 3 dataset and sampling methodology are described. In Sect. 4 modeling and scoring process is described. Section 5 outlines evaluation of results. In Sect. 6 the learned model is described. Finally, conclusions are summarized together with discussion related to threats to validity and future work in Sect. 7. For more detailed documentation refer please to our full report [25] which includes also information of how to access training and validation data set along with used R functions for training and validation of all models.

2 Problem Description and Preliminaries

In its essence, corporate credit risk evaluation is to make a classification of good and bad companies. Importance of the ability to accurately evaluate credit worthiness of customers cannot be overrated as extending credit to insolvent customers directly results in loss of profit and money. Correct determination whether a customer is solvent or there are some indications of default enables the company to minimize costs and

maximize earnings by choosing reliable customers. Default occurs when a company cannot meet its contractual financial obligations as they come due. More specifically, a company is technically insolvent if it cannot meet its current obligations as they come due, despite the value of its assets exceeding the value of its liabilities. A company is legally insolvent if the value of its assets is less than the value of its liabilities and a company is bankrupt if it is unable to pay its debts and files a bankruptcy petition [5]. In the model proposed by this paper the bad companies (with positive event of risk) are bankruptcies or forced deletions of companies (informal bankruptcy), good companies (with negative event of risk) are companies non-failed and operating companies (an overview of all research done on credit risk models shows that this is appropriate way of defining bad and good companies for credit risk [6]). To be precise, in Estonia there are certain requirements established for companies that need to be met for legally staying as an operating company. Those requirements are set by Commercial Code of Estonia [7]. These requirements also include regulations for equity and submission of annual reports. So, rather than being just a traditional bankruptcy prediction model, this model is predicting the risk of a company either going bankrupt or being deleted due to their failure of meeting specific requirements set by Commercial Code.

3 Dataset Description

In this section we describe data sample and variables used for modeling.

3.1 Sample Used for Modeling

The sample dataset used in the model training consists of bad companies (in total of 17,953 cases) and good companies that have been active in the same period (in total of 190,804 cases). Information regarding bankrupt and forcefully deleted companies was gathered from Business Register of Estonia and the bankruptcies and forceful deletions from the event period 01.01.2015–31.05.2017 were used. Events from earlier period were not included into analysis due to lack of data availability of explanatory variables (variables used in the models). The date of the bankruptcy or deletion is set to the earliest date of bankruptcy announcement available for us. Companies that have a lower credit risk (negative cases) are the companies that have been active in the event period. Companies registered later than 31.05.2015 were removed from the sample as these companies have little or no historical information that could be used in the training phase. The event dates for negative cases were set so that the event dates of positive and negative cases would be similarly distributed allowing us to reduce the effect that the event period (incl. external macro-economic factors) might have on events if the dates for negative cases would have been selected with a different method.

3.2 Variables Used in Modeling

Variables used in the model were selected by the insights and knowledge about insolvent companies - variables that characterize a typical insolvent company according to domain experts. More precisely, variables formatted as *italic* in the Table 1 below

are being used in expert models. Most of the variables have been used in modeling multiple times, i.e. we have used values of these variables at different time points (1–6 months/quarters and 2 years before the event by the results of analysis done in [8]). Set of different time points is used to provide sufficient information for the model to recognize patterns of change in time (e.g. increase or decrease in tax debts, changes in the VAT registration). We used the same 164 variables in each individual model for modeling a positive event happening in each subsequent month of 12-month period and since the computation time for fitting all 12 individual models was acceptable (1 h per model in average) no variable selection was needed.

For fitting individual models following explanatory variables have been used:

Table 1. Variables used for modelling

Category of variable	Granularity	#Variables	Variables
Tax	Month	48	<i>Total tax debt</i> , postponed tax debt, disputed tax debt, tax debt interest, postponed tax debt interest, <i>whether VAT is registered or not</i> , number of days past due date of VAT registration, number of unsubmitted tax declarations
	Quarter	12	<i>Paid state taxes, paid personnel taxes</i>
Annual reports of companies	Year	16	<i>Turnover</i> , balance, floating assets, capital assets, current liabilities, long term liabilities, equity, net income
	Month	6	<i>Number of unsubmitted annual reports</i>
Network data (board members/co-executives/owners and companies)	Month	42	owner degree*, owner PageRank**, board degree, board PageRank, overall degree, overall PageRank, board degree for negative score***
Company related (as of end of historical period****)	N/A	28	Age of company, initial capital of company, main activity of company, type of company
Debt	Month	6	<i>Amount of total debts to suppliers</i>
Derivates	Month	6	<i>Historical company reputation score*****</i>

* Degree is a feature reflecting the information available on the registry card of the company. Degree shows the number of incoming and outgoing relations of the company.

** PageRank is a link analysis algorithm that assigns a numerical weight to each company. The calculated weight shows the importance of each single company relative to the whole graph. PageRank considers all nodes of the graph and is therefore a global metric. PageRank is a probability of ending up at that node after starting from a random node in the network [8] and therefore it is a numerical value between 0 and 1 for each company.

*** Board degree for negative score shows the number of companies with a bad reputation (reputation of the company is a negative number) related to the board members of the respective company.

**** Historical period is the period before event period, last month/quarter/year in that period belongs to event period.

***** Historical score is an internal derivate of Register OÜ which is computed by different business rules considering different features reflecting the “health” of company.

4 Modeling and Scoring

In this section we first describe sampling and data partitioning approach for individual models and ensemble model (using outputs from all individual models as inputs predicting probability that positive event will be happen within the following 12 months). Then the process of model building is elaborated where in total 13 models are learned. After this scoring mechanism is described and finally we described how we use credit score distribution to define credit classes, which are used in credit management and other applications.

4.1 Sampling

Since the number of positive cases is significantly smaller than the number of negative cases, we need to address the imbalance problem in the training set. In literature various techniques have been documented for both resampling and algorithmic ensemble techniques in addressing the imbalance problem [9, 10]. In treating the imbalance problem of our training set, we applied extreme gradient boosting, avoiding resampling techniques which have many undesirable properties. As a result, no sampling methods are applied to both training and test sample. Therefore, the training and evaluation of the model performance is based on the imbalanced data set which is representative of the actual real-life data.

We actually applied previously under-sampling for balancing training data sets for all individual models and it resulted in too high number of false positives (positive event was predicted by model while no such event actually happened). It was due to the essence of under-sampling – it threw out majority of negative cases from training data and some patterns were lost for distinguishing two classes.

We used 50/50 ratio to split overall sample set of cases into training and test sample via random selection without replacement. Then by using the training cases we created in total, 13 training datasets - 12 datasets for fitting of individual models and one dataset for fitting an ensemble model. All training datasets for individual models contained the same list of companies from training sample. Each training dataset contains therefore 104,378 companies. Each training dataset for certain single model contained the fact of event with event rate of 8.6% and explanatory variables from aligned historical period. The training dataset for ensemble model was constructed from test datasets of individual models. It contains predictions of all individual models on their test datasets and true label (whether a positive event has happened or not).

In total, we created two test datasets from the rest of 50% of the initial sample (from which 50% of data has been used for modeling of all individual models) with random selection without replacement with sample rate of 50%. One test data set was reserved for validation of all individual models while second one was reserved for validation of ensemble model.

4.2 Model Building

In total 13 models were fitted - 12 individual models and an ensemble model. We used state of the art classification method of XGBoost both for individual models and

ensemble model. This method has been used in different Kaggle competitions with great success [11] and for creation of credit scoring models both for B2C and B2B domain [1, 26]. For optimizing hyperparameters of the models most promising hyperparameter tuning method was used: model-based (aka Bayesian hyper-parameter) optimization [12]. We compared it also with other optimization methods, but it turned out that above-mentioned method gave the best performance. As implementation, R packages “mlr” and “mlrMBO” were used [12, 13]. In Fig. 1 modeling process of individual models and the ensemble model is summarized. Yellow boxes in Fig. 1 with label “Y” denote months (so-called “target windows” [20], p. 96) where events are found in initial sample and boxes with label “X” denote historical time-periods (so-called “observation windows” [20], p. 96) the values for explanatory variables are taken from.

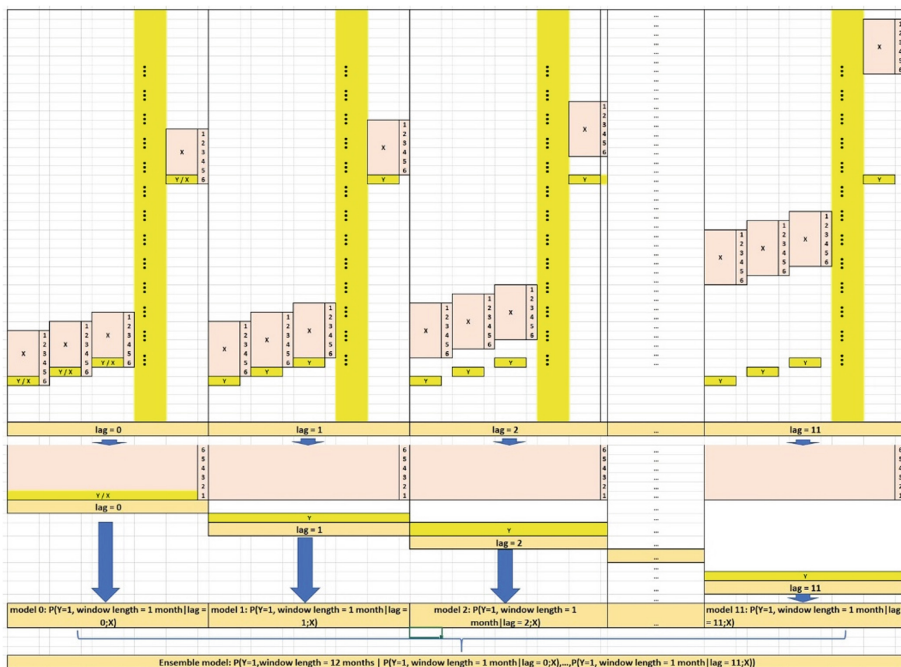


Fig. 1. Modeling scheme of fitting individual models and ensemble model. (Color figure online)

4.3 Scoring

We trained 12 individual models that use data from different time-periods before the event allowing us to predict default of a company at the scoring date and up to 11 months before the potential default event as follows. For “Model 0” we used for training timeseries variables from 0 to 5 months before the event for features with monthly granularity. Using this model with current data allows us to predict if the company becomes insolvent on the current month. In “Model 1” we used for training

the model features from 1 to 6 months prior to the event with monthly granularity. Using this model with current data allows us to predict if the company becomes insolvent on the next month from the current date. Analogously other individual models are trained and used until “Model 11” where training data involves data from 11 to 17 months before the event for time series features with monthly granularity. Using this model with current data allows us to predict if the company becomes insolvent on the 11th month from the current date.

Since we defined credit score as likelihood of default within 12 months we combine the preceding 12 individual models via ensemble model to provide a single credit risk score. Additionally, when scoring we by default assign credit score of 1 to all bankrupted companies and companies in liquidation as of scoring date.

4.4 Credit Risk Classes

In practical decision-making often a credit score as a probability is too abstract for humans to interpret. Therefore in practice often (usually, 5 classes by industry standard [24]) credit classes are defined and used instead of credit scores. Hence we also divided credit scores into 5 classes to be usable in the credit management and other processes. We used Jenks natural breaks optimization [14] to calculate boundaries for credit risk classes by using probabilities of event from scored ensemble model as of 28.11.2017 as input. Allocation points for the risk classes are chosen to be the credit score values which divide the whole spectrum of values of credit scores into clusters where density of credit scores is as high as possible - that should indicate that the distinction between the groups is the best at those points [14]. In Fig. 2 credit score distribution (on left side of figure) and the corresponding division into credit score classes (on right side of figure) are depicted. In the left hand side of Fig. 2 one can see usual distribution of scored probabilities of strong classifier i.e. they are distributed like a “bathtub” - majority of probabilities are near values 0 and 1.

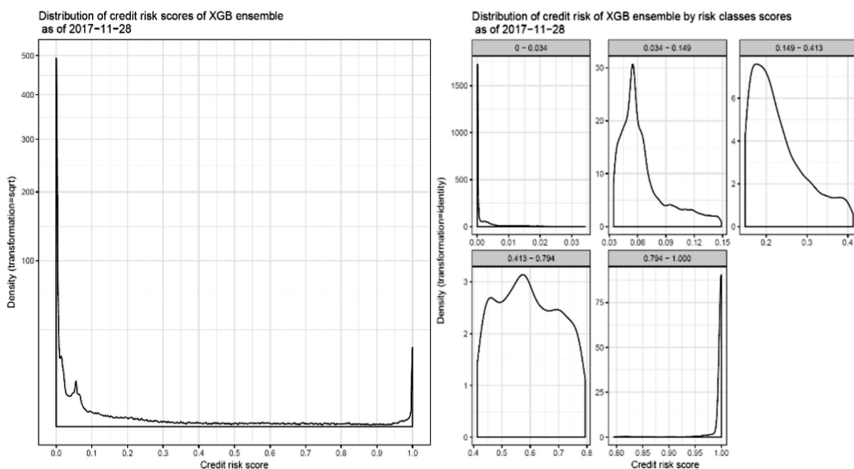


Fig. 2. Distribution of credit scores (left); Distribution of credit scores by risk classes (right).

5 Model Evaluation

In this section we evaluate the learned model from three perspective. First we evaluate model performance wrt test dataset. Then we compare scoring results from model wrt credit specialist decisions for the same cases. Finally, we perform retrospective analysis of the model.

5.1 Performance of the Model on Test Dataset

In Fig. 3 overview of assessment results of ensemble model by test data is compactly presented.

Model	Accuracy	H	Gini	AUC	AUCH	KS	MER	MWL	ER	Accuracy	Accuracy
Ensemble (xgboost)	99.5516	0.9744	0.9982	0.9991	0.9993	0.9826	0.004	0.003	0.004	H	H-measure
Maximum (xgboost)	98.6875	0.9605	0.9969	0.9985	0.9986	0.9736	0.008	0.004	0.013	Gini	Gini coefficient
Maximum (rf)	96.2311	0.9253	0.9922	0.9961	0.9962	0.9512	0.015	0.008	0.038	AUC	Area Under the ROC Curve
Model	Sensitivity	Specificity	Precision	Recall						AUCH	Area Under the convex Hull of the ROC Curve
Ensemble (xgboost)	0.9722	0.9977	0.9761	0.9722						KS	Kolmogorov-Smirnov statistic
Maximum (xgboost)	0.9859	0.9870	0.8781	0.9859						MER	Minimum Error Rate
Maximum (rf)	0.9861	0.9600	0.7015	0.9861						MWL	Minimum cost-Weighted Error Rate
										ER	Error rate
Model	TP	TN	FP	FN	TPR	FPR	FMES	Youden	NRObs	Sensitivity	Sensitivity
Ensemble (xgboost)	4,411	47,545	108	126	0.9722	0.0023	0.9742	0.9700	52,190	Specificity	Specificity
Maximum (xgboost)	4,473	47,032	621	64	0.9859	0.0130	0.9289	0.9729	52,190	Precision	Precision
Maximum (rf)	4,474	45,749	1,904	63	0.9861	0.0400	0.8198	0.9462	52,190	Recall	Recall
										TP	True Positives
										TN	True Negatives
										FP	False Positives
										FN	False Negatives
										TPR	True Positive Rate / Error II
										FPR	False Positive Rate / Error I
										F (FMES)	F measure
										Youden	Youden index
										NRObs	Number of observations

Fig. 3. Performance of ensemble models and reference models.

For comparison purposes, individual models have been fitted and evaluated with two different methods - extreme gradient boosting (“xgboost”) and random forests with under-sampling (“rf”) as reference model which was used in production at the time of experimenting with new approach. We used two heuristics for aggregating results of 12 models - “Maximum” means that final score is calculated by taking maximum of scores by individual models, while “Ensemble” means that ensemble learning was used to combine results of individual models. Experimental results indicate that ensemble model based on extreme gradient boosting has best performance wrt accuracy, AUC and several other metrics. Regarding false positives (FP) and false negatives (FN) we can see that extreme gradient boosting based model is producing more false negatives (positive event was not predicted by model while such event has happened actually) while at the same time less false positives at much higher rate. Further details regarding used performance metrics can be found in [15].

5.2 Credit Management Specialist Decisions vs Model Estimates

For understanding in which extent the learned model reflects credit decisions made by humans, we compared scores of companies with the credit decisions made by credit specialists. The decision dataset was provided by company Kredix OÜ for the period 15.09.2016–14.03.2017. Credit decisions are made in the process of analyzing trade receivables (sales invoices) in the context of Invoice-to-Cash business process [25]. Credit decisions for each debtor may be either “Yes”, “No” or “Wait”. Credit decision “Yes” means that the invoice receiver (debtor) company is eligible to credit, “No” means that it is not advised to extend credit to the debtor, “Wait” means that further data should be collected before making the credit decision (basically “No”). Credit decisions are made by using in addition to company’s public data (plentiful of what we used for model learning) and inside knowledge (e.g. creditor’s notes) also invoice data such as the amount of debt and the number of days payment is due date.

Distribution of credit scores along with bootstrapped confidence intervals for the median by credit decisions “NO”, “WAIT” and “YES” are depicted in Fig. 4. We can see that credit scores (as logarithm of probabilities of event) by model are quite well in accordance with credit decisions of credit specialists, i.e. distributions of credit scores across different type of credit decisions are quite different and aligned wrt credit decisions (the lower the credit score the better the credit decision). In summary, we can conclude that model reflects credit decisions made by humans.

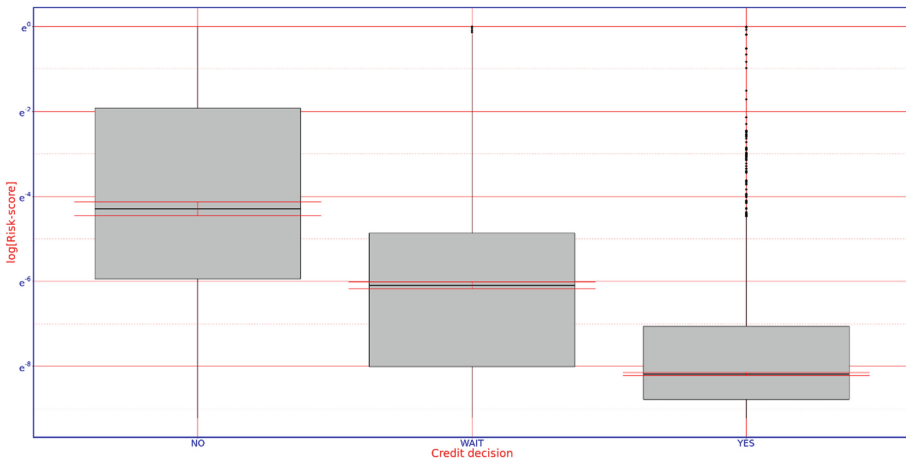


Fig. 4. Comparison of scores of XGboost ensemble model with credit decisions.

5.3 Retrospective Analysis

Here we found retrospective performance of model i.e. knowing actual events during time-period 1.06.2017–31.10.2017 we predicted what would have been the probability of event as of 0, 1, 2, 5, 8 and 11 months before event for each company by scoring them with fitted ensemble model. Sample for validation has been created similarly to

the sampling process used during modeling - i.e. including all positive events which occurred in before mentioned time-period and sampling negative events from the same time-period having distribution of dates as similar as possible for both negative and positive events.

As expected, it follows from Fig. 5 that the model is most accurate as of 0 months before the event (bankruptcies and deletions are evaluated with a much higher credit risk compared to normal companies) and less accurate as of 11 months before the event. The prediction accuracy for several months ahead is not very satisfactory wrt prediction accuracy for few months ahead. It might be influenced by used modeling framework in which case for predicting the positive event for several months ahead the observation window and target window are located quite far away from each other and therefore quite outdated data are used for fitting of individual models. We can also see from figure that predicting deletions is more accurate compared to bankruptcies. This can be explained with the presence of annual reports submission information in the model along with larger proportion of deleted companies in training dataset and the fact that not multinomial but binary classification has been used. Also, we can see that variance of predicted credit scores by confidence intervals of median for bankruptcies is very high wrt deletions.

However, in summary, comparing primarily the distribution of credit scores of good companies wrt to credit scores of bad companies we can conclude that model behaves quite acceptable on data set used for retrospective analysis.

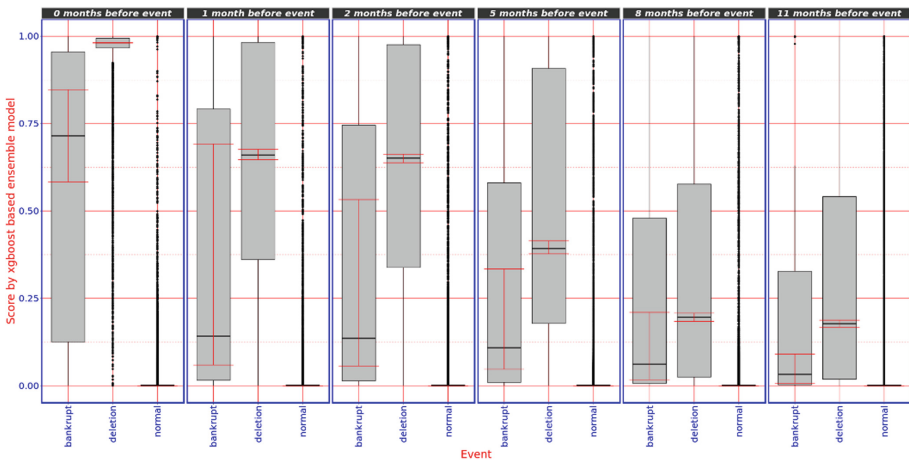


Fig. 5. Credit risk predictions for companies on the data of retrospective analysis.

6 Model Description

In this section description we analyze the key characteristics both of the individual models (taking as an example “Model 0”) and ensemble model. Since “black-box” method is used for fitting the classification model, there are by far lesser options to

describe the resulting individual models compared to logistic regression with its mature framework [2]. However, there are some tools which could help us to understand the individual models. First, we used relative variable importance (RVI) provided by extreme gradient boosting model [16]. Second, we used so-called partial dependence plots (PDP) [17]. Finally, we described the ensemble model wrt risk classes by comparing distributions of explanatory variables in all of them giving some general patterns for distinction of credit classes and gaining a formal description of credit risk scores classes.

6.1 Relative Variable Importance and Partial Dependence Plots

Here the aim is to understand what kind of explanatory variables are playing supreme role in predicting of positive event in individual models. In the current framework all variables are used for fitting individual models. Therefore, no variable importance evaluation was needed during modeling. However, since extreme gradient boosting is a “black-box” method, then for describing this type of model, RVI can be used. RVI shows how important is the variable in contributing to the ability for differentiate two classes. Also, PDP can be used for the similar purpose. PDP are low-dimensional graphical renderings of the prediction function so that the relationship between the outcome and predictors of interest can be more easily understood [17]. In Fig. 6 RVI (on left) and PDP (on right) partial for the “Model 0” and variable “score” (company reputation score) have been produced.

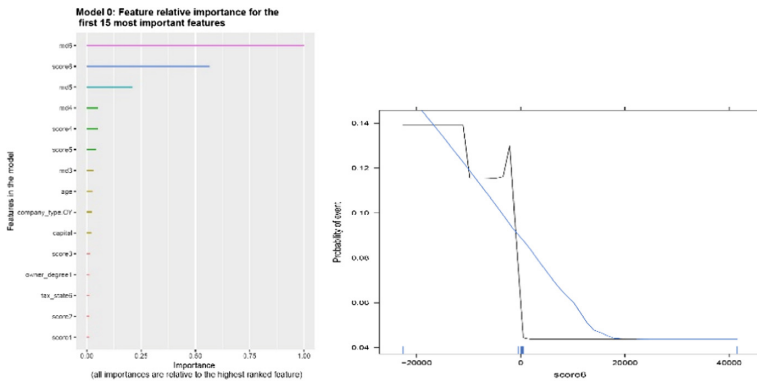


Fig. 6. Plots of relative variable importance (left) and partial dependence (right).

From the RVI we can see that most important variables (considering first 15 most important features in the “Model 0”) by extreme gradient boosting model are number of unsubmitted annual reports (the one as of most recent month is the most important variable in “Model 0”), historical score, company age, company type (OÜ/not OÜ), owner degree and state tax. From the PDP we can see that the higher historical reputation score as of most recent month the lower the probability of event.

6.2 Formal Description of Credit Risk Scores Classes

In this section the goal is to describe ensemble model by exploring and differentiating credit risk classes with help of scoring data. Similar approach is currently being used in SAS® Enterprise Miner™ [18].

Columns of matrix (“1”–“5”) in Fig. 7 denote credit risk classes, rows of matrix present a selection of explanatory variables used in modelling (the most recent time-period has been chosen). In the figure, next to the name of variable in the brackets there are mean values of that variable by credit risk classes in original scale of variable. Cells of matrix represent result of comparison of distributions in given credit class and population for a given variable (the brighter the cell (lower the value in the cell) the lower the values of variable in risk class than in population in average). For example, if credit risk class is “1” then by population we mean all the rest of other risk classes (“2”–“5”). In the last column “Trend” trend for a given variable is presented wrt moving from best credit risk class (“1”) towards worst one (“5”). For instance, we can see that there is a linear trend for variable “Score history” - the lower the values for variable “score history” in average, the worse the credit risk class in terms of credit score.

Risk class	1	2	3	4	5	Scale	Trend
As of 28.11.2017							
Number of companies in risk score class	219,201	17,851	3,310	1,175	8,754		
% of companies in risk score class	88%	7%	1%	0.5%	3.5%		
1. Score history (1257.6) 3(-91.05) 3(-829.41) 4(-2 879.91) 5(-3 755.81)	6 - Higher than in population	5 - A bit higher than in population	3 - A bit lower than in population	2 - Lower than in population	1 - Much lower than in population	identity	↓
2. Number of unsubmitted annual reports (1-0.6) 2(-1.35) 3(-0.43) 4(0.93) 5(6.86)	3 - A bit lower than in population	2 - Lower than in population	5 - A bit higher than in population	6 - Higher than in population	7 - Much higher than in population	identity	↕
3. Number of unsubmitted tax declarations (10.14) 2(0.31) 3(1.26) 4(1.8) 5(0.32)	4 - No significant difference in class and population	4 - No significant difference in class and population	5 - A bit higher than in population	5 - A bit higher than in population	4 - No significant difference in class and population	identity	↔
4. Tax debts (1178.73) 2(1 339.15) 3(37 0095.17) 4(15 329.02) 5(13 084.72)	2 - Lower than in population	5 - A bit higher than in population	5 - A bit higher than in population	6 - Higher than in population	6 - Higher than in population	log	↑
5. Tax debts interest (118.11) 2(149.7) 3(70 139.48) 4(6 485.22) 5(6 680.17)	2 - Lower than in population	5 - A bit higher than in population	6 - Higher than in population	7 - Much higher than in population	7 - Much higher than in population	log	↑
6. Tax debts postponed (198.16) 2(304.42) 3(412.99) 4(232.24) 5(31.57)	3 - A bit lower than in population	5 - A bit higher than in population	4 - No significant difference in class and population	4 - No significant difference in class and population	4 - No significant difference in class and population	log	↔
7. Tax debt disputed (13.06) 2(71.96) 3(1 097.57) 4(280.45) 5(369.17)	3 - A bit lower than in population	4 - No significant difference in class and population	5 - A bit higher than in population	5 - A bit higher than in population	5 - A bit higher than in population	log	↔
8. Paid taxes (personnel) (14 862.45) 2(536.72) 3(507.67) 4(277.17) 5(176.75)	5 - A bit higher than in population	3 - A bit lower than in population	3 - A bit lower than in population	2 - Lower than in population	2 - Lower than in population	log	↓
9. Paid taxes (state) (10 443.37) 2(739.93) 3(893.5) 4(1 504.98) 5(991.35)	5 - A bit higher than in population	4 - No significant difference in class and population	4 - No significant difference in class and population	3 - A bit lower than in population	2 - Lower than in population	log	↓
10. Age of company (13 310.56) 2(17 317.36) 3(1 269.98) 4(2 403.36) 5(4 984.2)	6 - Higher than in population	1 - Much lower than in population	2 - Lower than in population	3 - A bit lower than in population	5 - A bit higher than in population	identity	↕
11. Network centrality - degree (score < 0) (10.33) 2(1.21) 3(5.59) 4(9.94) 5(2.59)	4 - No significant difference in class and population	4 - No significant difference in class and population	5 - A bit higher than in population	6 - Higher than in population	4 - No significant difference in class and population	log	↕
12. Network centrality - board degree (10.99) 2(0.17) 3(0.43) 4(0.91) 5(0.4)	5 - A bit higher than in population	4 - No significant difference in class and population	3 - A bit lower than in population	3 - A bit lower than in population	4 - No significant difference in class and population	log	↓
13. Network centrality - owner pagerank (10) 2(0) 3(0) 4(0) 5(0)	5 - A bit higher than in population	3 - A bit lower than in population	3 - A bit lower than in population	3 - A bit lower than in population	3 - A bit lower than in population	log	↔
14. Network centrality - board pagerank (11.5) 2(1.26) 3(1.28) 4(1.21) 5(0.75)	6 - Higher than in population	2 - Lower than in population	2 - Lower than in population	2 - Lower than in population	3 - A bit lower than in population	log	↓
15. Network centrality - board pagerank (10) 2(0) 3(0) 4(0) 5(0)	4 - No significant difference in class and population	4 - No significant difference in class and population	4 - No significant difference in class and population	3 - A bit lower than in population	3 - A bit lower than in population	log	↔
16. Network centrality - degree of all connections (13.25) 2(1.62) 3(2.08) 4(2.81) 5(2.04)	6 - Higher than in population	2 - Lower than in population	4 - No significant difference in class and population	5 - A bit higher than in population	2 - Lower than in population	log	↕
17. Network centrality - pagerank of all connections (10) 2(0) 3(0) 4(0) 5(0)	5 - A bit higher than in population	3 - A bit lower than in population	3 - A bit lower than in population	4 - No significant difference in class and population	3 - A bit lower than in population	log	↕
18. VAT registration (registered or not) (131.01) 2(43.74) 3(100.88) 4(252.5) 5(123.47)	3 - A bit lower than in population	2 - Lower than in population	4 - No significant difference in class and population	2 - Lower than in population	1 - Much lower than in population	identity	↕
19. VAT registration (days past due date) (131.01) 2(43.74) 3(100.88) 4(252.5) 5(123.47)	4 - No significant difference in class and population	4 - No significant difference in class and population	3 - A bit lower than in population	3 - A bit lower than in population	5 - A bit higher than in population	log	↕
20. Claims (18.65) 2(74.41) 3(290.94) 4(909.05) 5(663.9)	3 - A bit lower than in population	5 - A bit higher than in population	4 - No significant difference in class and population	5 - A bit higher than in population	5 - A bit higher than in population	log	↕
21. Balance (1376.91) 2(32 730) 3(44 084) 4(73 81) 5(113 833)	7 - Much higher than in population	1 - Much lower than in population	1 - Much lower than in population	2 - Lower than in population	1 - Much lower than in population	log (min) + 1	↕

Fig. 7. Description of credit risk classes.

7 Conclusions, Threats to Validity and Future Work

In this paper we described a novel approach for modelling credit risk of companies. We applied extreme gradient boosting with Bayesian hyper-parameter optimization and ensemble learning with rapidly changing data to learn a model, which predicts likelihood of a default event of a business in 12 months. The learned model has accuracy of 99.5%. For confirming that the high accuracy has impact in practical settings, we performed additionally retrospective analysis and a case study where we analyzed credit scoring results by the model wrt credit management specialist decisions. In retrospective analysis we reviewed the credit scores that the model provided for the insolvent companies 0, 1, 2, 5, 8 and 11 months before event. Retrospective analysis proved that the predictive ability of the model is acceptable even 11 months before the event giving confidence that the model is applicable in predicting events up to 12 months ahead.

Validation of the model wrt credit decisions made by domain experts shows that the model performs quite well in terms of making credit decisions. Furthermore, the analysis shows, that our credit risk model is able to capture some cases where a manual credit decision made by a specialist may be questionable. However, it should be noted that there may be deviations both ways - manual assessments may be incorrect in times and credit risk model also has some deviations and exceptions. By now the model is put into practice and is used by Estonian businesses in everyday decision-making via products of Register OÜ and its partners.

However, there are some threats to validity related to the proposed model wrt to bias in sampling, availability of data, modeling process and usability of model. For remedies please refer to future work activities provided at the end of current section.

Regarding sample used in modeling, we have excluded young companies from training sample since they have only few data points available for training. In the credit model it is reflected in young companies having scores from slightly higher risk class compared to average companies, which seems to make sense if we assume that if not much is known about the company then the potential risk is higher as well. Also, we have excluded from negative cases companies for which tax debts were monotonically increasing since this is an early indicator of a potential default event. Anyway, there should be no need for that explicitly once time series derivatives (incl. univariate statistics, linear and non-linear trends, different statistics wrt variations around trends) will be added into modeling process. Positive cases mostly consist of company deletion events due to multiple unsubmitted annual reports. There has been recently significant rise in such events since Estonian Business Register deletes more systematically companies, which have failed to submit their annual reports multiple consecutive years in a row. In the model this means that the number of unsubmitted annual reports is a dominant feature. However, importance of this feature is subject of change if behavior of Estonian Business Register will change.

Regarding modeling process, all missing values for numeric variables were replaced with zeros. This was done primarily due to fact that the same method for handling missing values has been used for reference model. Regarding classification methods “state of the art” algorithm XGBoost has been used alone due to satisfactory

results. Currently, in total 13 models need to be fitted. In case the number of potentially useful explanatory variables and number of observations will grow some more compact and flexible approach for modeling and variable selection is needed. There are also several threats to validity wrt retrospective analysis which are described in according section.

Regarding usage of current model for other than Estonian market there might be present some restrictions - more precisely, some important explanatory variables like tax debts might not be available at all or the collecting process of all data needed for modeling might be too expensive for larger markets.

Currently we use only a limited amount of network features (degree, PageRank [19]) derived from networks of board members/owners and companies. In future we will add support for a wider variety of network types, e.g. networks of companies and locations, and network metrics, e.g. eigenvector centrality, Kleinberg's authority centrality etc. We also see value in adding additional variables derived from time series for improving the model wrt detection of "edge" cases [19–21]. Current model does not consider default events related to disruptions in payments. We have started collecting late payment data and believe that features from this data could be beneficial to the model and everyday credit scoring as well. For improving accuracy of predicting defaults financial ratios have been applied in practice [26]. We are also interested in exploring multinomial discrete hazard survival data mining model which have several good properties wrt current approach [30]. Finally, we have recently initiated a new project, which aims at performing credit scoring by experimenting with plenty of different classification algorithms using freely available web data only. This project represents a shift towards big data scoring, where we intend to explore the enhancements outlined earlier.

References

1. Xia, Y., Liu, C., Li, Y., Liu, N.: A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Syst. Appl.* **78**, 225–241 (2017)
2. Siddiqi, N.: *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards*, 2nd edn. Wiley, Hoboken (2017)
3. Tamari, M.: Financial ratios as a means of forecasting bankruptcy. *Manag. Int. Rev.* **6**(4), 15–21 (1966)
4. Baxter, R.A., Gawler, M., Ang, R.: Predictive model of insolvency risk for Australian corporations. In: *Proceedings of the Sixth Australasian Conference on Data Mining and Analytics*, vol. 70. Australian Computer Society, Inc. (2007)
5. Investopedia: Bankruptcy risk definition. <https://www.investopedia.com/terms/b/bankruptcyrisk.asp>
6. Yu, L., Wang, S., Lai, K.K., Zhou, L.: *Bio-inspired Credit Risk Analysis: Computational Intelligence with Support Vector Machines*. Springer, Heidelberg (2008). <https://doi.org/10.1007/978-3-540-77803-5>
7. Riigi Teataja: Commercial Code. <https://www.riigiteataja.ee/en/eli/504042014002/consolide>
8. Ilves, T.: *Impact of board dynamics in corporate bankruptcy prediction: application of temporal snapshots of networks of board members and companies*. Master thesis. Tartu University (2014)

9. Yap, B.W., Rani, K.A., Rahman, H.A.A., Fong, S., Khairudin, Z., Abdullah, N.N.: An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In: Herawan, T., Deris, M.M., Abawajy, J. (eds.) *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*. LNEE, vol. 285, pp. 13–22. Springer, Singapore (2014). https://doi.org/10.1007/978-981-4585-18-7_2
10. Analytics Vidhya: How to handle Imbalanced Classification Problems in machine learning? (2017). <https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/>
11. Kaggle Inc.: Kaggle competitions. <https://www.kaggle.com/competitions>
12. Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., Lang, M.: mlrMBO: a modular framework for model-based optimization of expensive black-box functions. Cornell University Library (2017)
13. Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., Jones, Z.M.: mlr: machine learning in R. *J. Mach. Learn. Res.* **17**, 1–5 (2016)
14. Jenks, G.F.: The data model concept in statistical mapping. In: *International Yearbook of Cartography*, vol. 7, pp. 186–190 (1967)
15. Hand, D.J., Anagnostopoulos, C.: Measuring classification performance. hmeasure.net
16. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: *KDD 2016 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, pp. 785–794 (2016)
17. Greenwell, B.M.: pdp: an R package for constructing partial dependence plots. *R J.* **9**(1), 421–436 (2017)
18. SAS Institute Inc.: SAS® Enterprise Miner™ 14.3: Reference Help; Chapter 67 Segment Profile Node (2017)
19. Luke, D.A.: *A User’s Guide to Network Analysis in R*. UR. Springer, Cham (2015). <https://doi.org/10.1007/978-3-319-23883-8>
20. Svolba, G.: *Data preparation for analytics using SAS*. SAS Institute Inc. (2006)
21. Афанасьев, В.Н., Юзбашев, М.М.: Анализ временных рядов и прогнозирование. *Финансы и статистика* (2012)
22. Abdou, H., Pointon, J.: Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intell. Syst. Account. Financ. Manag.* **18**(2–3), 59–88 (2011)
23. Hooman, A., Mohana, O., Marthandan, G., Yusoff, W.F.W., Karamizadeh, S.: Statistical and data mining methods in credit scoring. In: *Proceedings of the Asia Pacific Conference on Business and Social Sciences*, Kuala Lumpur (2015)
24. Tarver, E.: Business credit score: everything you should know to build business credit. <https://fitsmallbusiness.com/how-business-credit-scores-work/>
25. Register OÜ: Credit risk prediction for Estonian companies. <https://docs.google.com/document/d/1aG9Y6B8J8Q9Ee75tA6X3SJBByo9QAJCvugYas7Q9p2VE>
26. Zięba, M., Tomczak, S.K., Tomczak, J.M.: Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Syst. Appl.* **58**, 93–101 (2016)
27. Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **23**, 589–609 (1968)
28. Shin, K.S., Lee, T.S., Kim, H.J.: An application of support vector machines in bankruptcy prediction model. *Expert Syst. Appl.* **28**, 127–135 (2005)
29. Geng, R., Bose, I., Chen, X.: Prediction of financial distress: an empirical study of listed Chinese companies using data mining. *Eur. J. Oper. Res.* **241**, 236–247 (2015)
30. Tutz, G., Schmid, M.: *Modeling Discrete Time-to-Event Data*. SSS. Springer, Cham (2016). <https://doi.org/10.1007/978-3-319-28158-2>