Polish companies bankruptcy prediction: a comparison between selected machine learning (and deep learning) algorithms

Joanna Wyrobek[1]

Keywords: bankruptcy prediction, deep learning, machine learning, corporate finance
JEL classification: G33

1. Introduction

Currently, all of us are witnessing an unprecedented growth of the information generated worldwide and on the internet which can be collected, processed and used for prediction purposes. This led to the creation or reinvention of many big data prediction methods among which there are machine learning and deep learning algorithms. The term 'machine learning' can be explained as the algorithms which parse data, learn from it and make a determination (make decisions) or prediction about various processes which are described by this data[2]. Deep learning in practical terms is considered to be a subset of machine learning algorithms, but 'classical' machine learning algorithms are considered to require some guidance during training (if an ML algorithm returns inaccurate prediction it requires a human to introduce adjustments), whereas 'classical' deep learning algorithms can determine by themselves whether the prediction they produced is accurate or not. Deep learning algorithms usually are based on artificial neural networks which allow to structure algorithms in layers and make the information processing to resemble human-like intelligence.

Development of machine learning methods (including neural networks) led to the creation of new job positions in banks, various financial institutions, companies and ministries. They always made use of statisticians, but new developments in the field of data processing created a demand for data scientists, who not only are capable of doing the statistical analysis but also of using programming languages to create semi-automatic and automatic tools for data recovery and processing. Today, data science (and deep learning) algorithms not only are capable of doing risk management, online trading, insurance underwriting, but also of bankruptcy prediction or speech and text recognition which are now popular research areas in all sorts of financial institutions[3]. As far as the application of machine learning to bankruptcy prediction is concerned, there exist specialized companies which concentrate solely on providing accurate insolvency risk information about various businesses and physical persons to their clients who consist of: banks, loan companies, insurers and commercial factoring and commercial companies. These specialized bankruptcy prediction companies constantly look for better classification algorithms because even the tiniest improvement in the accuracy of their classification algorithms produces significant cost savings of their clients (bad debts reductions).

What is interesting, the majority of the analyzed studies dedicated to the analysis of the accuracy of the various machine learning algorithms used relatively modest data samples. This led to the research hypothesis of this paper, which tests the usefulness of various machine learning methods using a relatively large and representative set of Polish

[1]Joanna Wyrobek, Cracow University of Economics, School of Finance and Law, Department of Corporate Finance, Cracow University of Economics, ul. Rakowicka 27, 31-510 Kraków, Poland, email: wyrobekj@uek.krakow.pl.
[2] Lewis N.D., *Machine Learning Made Easy with R: An Intuitive Step by Step Blueprint for Beginners*, CreateSpace Independent Publishing Platform, New York, 2017.
[3] Brynjolfsson E., Mcafee A., *The Business of Artificial Intelligence*, "Harvard Business Review", July 2017, pp. 5.

commercial companies. The paper tests usefulness of the following methods: discriminant analysis (DA), logit (L), support vector machines (SVM), random forest (RF), gradient boosting decision trees (GB), neural network with 1 hidden layer (NN), convolutional neural network (CNN), naïve Bayes (NB). The research hypothesis tested in the paper [H1] states that if one has access to the big sample of companies, the most efficient algorithms (first choice) in the bankruptcy prediction are neural networks, gradient boosting decision trees and random forest algorithms.

2. Literature Review

2.1. Review of international publications

Table 2-1 and Table 2-2 summarize the final accuracy of the models presented in research papers dedicated to bankruptcy prediction. The accuracy of the models was estimated in various ways. Sometimes it was based on the separate validation sample (in some papers the researchers firstly trained the model on the separate training sample, then tested the model on a separate test sample, and finally, they validated the model on the separate validation sample). In other papers, the researchers only tested the model on the test sample (without the validation sample), and in other papers the authors used the cross-validation technique (the model was trained several times, each time another part of data was used as a training sample, and another part of data was used as test sample – in the end, the researchers took the average accuracy of all generated models).

The summary of the publications presented in Tables 2-1 and 2-2 show separately the accuracy of non-ensemble models and ensemble models. Table 2-1 presents the accuracy of non-ensemble models.

Table 2-1. Summary of accuracy single ML methods

Base Classifiers	Aaverage accuracy	No of publications
BN - Bayesian network	86,4	3
CART - classification and regression trees	76,79	4
DA – discriminant analysis	76,1	17
DT – decision trees	75,5	18
GP – Gaussian processes classification	85,67	2
LR – logistic regression	74,78	31
NB - naive Bayes	95,25	3
NN – neural network	76,67	25
Probit	85,48	3
SVM – support vector machines	80,15	27

Source: own research, publications used for presented results are marked with the following abbreviations [BN, CART, DA, DT, GP, LR, NB, NN, PROBIT, SVM] in the literature section of this paper.

As can be seen from Table 2-1, the highest accuracy was found for naïve Bayes algorithms. However, there were only 3 papers which used this method, so it is uncertain whether these results are reliable and generally representative for insolvency detection. A similar situation applies to the Bayesian network (BN), Gaussian processes (GP) and probit. The algorithms which had a relatively good in analyzed publications were as following: neural networks (NN) with the accuracy equal to 76,67%, the logistic regression (LR) with the accuracy equal to 74,78%, decision trees (DT) with the accuracy equal to 75,5% and support vector machines (SVM) with the accuracy equal to 80,15%.

From the practical point of view, it would seem acceptable to use algorithms with the prediction accuracy exceeding 90%, so only in 10 out of 100 cases one would make a mistake. Therefore, the most promising algorithm seemed to be naïve Bayes.

Table 2-2 presents the average accuracy of ensemble methods, which combine basic algorithms into more advanced methods. The most popular and considered as the first-choice methods are the random forest and gradient boosting decision trees, which combine decision trees with majority voting technique and (gradient boosting) creating the next decision tree to correct the mistakes of the previous decision tree. There are dozens, if not hundreds of such ensemble algorithms, but in the paper, there are presented only the most popular algorithms.

Table 2-2. Summary of ensemble ML algorithms accuracy

Technique	Base classifiers	Average accuracy(%)	No of publications
AdaBoost	(A)DT	87,92	5
AdaBoost	(A)NN	84,67	4
Bagging	(B)DT	75,63	2
Backpropagation	(Ba)NN	82,83	2
Bagging	(Bb)NN	77,65	7
Bagging	(Ba)SVM	76,69	4
Gradient Boosting	(Gb)DT	74,62	2
Boosting	(Bo)NN	76,52	3
Boosting	(Bo)SVM	75,63	3
Random Forest	(RF)DT	93,12	3
Random Subspace	(RS)LR	83,71	2
Random Subspace	(RS)NN	80,8	2
Random Subspace	(RS)SVM	87,95	2

Source: own research (symbols used in parentheses were added to identify the literature sources of the following table), publications used for presented results are marked with the following abbreviations [(A)DT, (A)NN, B(DT), (Ba)NN, (Bb)NN, (Ba)SVM, (Gb)DT, (Bo)NN, (Bo)SVM, (RF)DT, (RS)LR, (RS)NN, (RS)SVM] in the literature section of this paper.

Table 2-2 shows the accuracy of ensemble methods which were found more than once in analyzed papers. As can be seen in Table 2-2, the most efficient algorithm proved to be random forest decision trees algorithm. The second best was an algorithm based on SVM, which applied random subspaces applied to SVM. Very close behind was the adaptive boosting algorithm (AdaBoost) applied to decision trees. All three algorithms were very close to 90% accuracy. Underperforming when used individually, neural networks proved to be in the fourth place when they became combined with the AdaBoost algorithm (accuracy 85%).

2.2. Review of publications concerning bankruptcy prediction in Poland

2.2.1. Pawełek and Grochowina paper

One of the Polish publications in which the authors applied machine learning algorithms to bankruptcy prediction is the paper written by Pawełek and Grochowina[4]. In the paper, the researchers took 7223 Polish manufacturing companies which included 42 bankrupt companies and 7181 active companies. Data sample was for years 2013 – 2015 (it came from the Emis database).

The authors used the decision trees (CART) as a basic learning method and then combined it with: boosting (learning of distribution from weak classifiers (weak learners are restricted, smaller models) and adding them to a final strong classifier), bagging (bootstrap aggregation), random subspaces (this method reduces the correlation between variables by training a model based on random subsets of the dataset) and random forest (majority voting ensemble of decision trees). Aggregation of the base models was

[4] Pawełek B., Grochowina D. (2017), Podejście wielomodelowe w prognozowaniu zagrożenia przedsiębiorstw upadłością w Polsce, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 468, p. 171 – 179.

conducted based on the majority voting technique described by Gatnar[5]. Estimation and testing were carried out in the following way: dataset was divided into 10 balanced and 10 unbalanced subsets. Each subset was divided into 9 training sets and 1 testing set (each time different observations belonged to the training and testing sets). Final models were based on aggregation of partial models trained on subsets. Authors did not use any validation sample, but the testing sample was for every subsample different. Results were presented in the form of total classification error of the final models. Table 2-3 shows a small fraction of the results, only for the models with prediction period equal to one year before the bankruptcy and for the ensemble consisting of 100 base models (but this number of base models produced the best validation results).

In Table 2-3 ensemble methods: random subspaces and random forest are presented in 2 columns (in total, in 4 columns). These methods were applied in two ways. In the first approach (symbol (1)), models were simply trained on 16 financial ratios. In the second approach, models were trained on 8 randomly chosen ratios (symbol (2)).

Table 2-3. General error rates of models comprising 100 base models [in %] one year prior to bankruptcy (second approach)

Error type	Type of subset	*CART*	*Bagging*	*Boosting*	Random Subspaces 1	Random Subspaces 2	*Random Forest 1*	*Random Forest 2*
Total error	balanced	35.5	29.5	34.1	31.9	28.4	29.2	27.7
	unbalanced	27.9	21.8	23.2	22.3	21.6	22.5	21.6
1st type error	balanced	34.8	29.1	36.2	31.3	27.9	30.3	28.6
	unbalanced	57.7	61.1	56.6	57.4	61	57.1	56.7
2nd type error	balanced	36.1	29.8	32	32.6	28.9	28.1	26.9
	unbalanced	17.9	8.7	12.1	10.6	8.5	10.9	9.9

source: Pawełek and Grochowina (2017), p. 175.

As can be seen in Table 2-3, the most accurate non-ensemble algorithm turned to be random forest and random subspaces methods for randomly selected ratios. For the balanced sample, the best algorithm turned to be random forest also for 8 randomly selected financial ratios (random selection was carried out separately for every base model).

According to the researchers, their research confirmed dependency between a model accuracy and the number of base models (positive relationship) and the most recommended and promising methods turned to be (for unbalanced sample and random selection of financial ratios for base models) Random Subspaces, Random Forest, and bagging.

2.2.2. Pociecha, Pawełek, Baryła and Augustyn book

Another important publication which analyzes the accuracy of various machine learning techniques in bankruptcy prediction is the book written by Pociecha, Pawełek, Baryła, and Augustyn[6]. Researchers used data for 7147 active companies and for 182 bankrupt companies. Observations were for years 2005 – 2009. Researchers used two approaches to testing of their models: 70%/30% and 60%/40%. They did not use the validation sample.

---

[55] Gatnar E., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa, 2008, str. 63.

[6] Pociecha J., Pawełek B., Baryła M, Augustyn S., Statystyczne metody prognozowania bankructwa w zmieniającej się koniunkturze gospodarczej, Wydawnictwo UEK, 2014, Kraków.

Researchers used the following algorithms: logit, discriminant analysis, decision trees (CART –Classification and Regression Trees), and neural networks (only 3-layer networks: an input layer, one hidden layer, and one output layer). To teach neural network researchers used the BFGS method (Broyden-Fletcher-Goldfarb-Shanno). Researchers tested many different activation functions: linear function, logistic function, hyperbolic tangent, an exponential function and, softmax function. Table 2-4 shows the best models trained on the balanced sample.

Table 2-4. Ranking of the best models in Pociecha et al. (2014) research for a balanced sample one year before bankruptcy (abbreviations of variables used in each model were omitted)

Ranking	Model Type	Training:Test Proportion	Test sample		
			Error Type 1	Error Type 2	Accuracy
1	NN (Equity + Long-Term Liabilities) / Total Assets; (Operating Profit + Depreciation) / Total Assets; Operating Costs / Short-Term Liabilities)	6:4	95.83	83.33	89.58
2	DA (Equity + Long-Term Liabilities) / Total Assets; (Operating Profit + Depreciation) / Total Assets; Operating Costs / Short-Term Liabilities)	6:4	95.83	75	85.42
3	CART (Net Profit (Loss) * 100 / Total Assets)	6:4	95.83	70.83	83.33

Source: Pociecha et al. (2014, p. 109), NN – neural network, DA – discriminant analysis, CART, CART – decision trees.

It can be seen in Table 2-4, the most efficient proved to be the 3-layer neural network with proportions between training and test samples as 60% to 40%. The total error rate was 89,58%. The second best model was the discriminant analysis model and the third best – a single decision tree. Yet, only the neural network accuracy was close to 10% error rate in classification.

Table 2-5 shows the results for models trained on the unbalanced sample. The most accurate method proved to be (again) neural network but based on different ratios than in the case of the balanced sample. Only this model total error rate was below 10% (4,47%). The second best model was also a neural network, but based on a different combination of financial ratios and its accuracy was 91,67%. Finally, the third best model was a simple decision tree with total accuracy equal to 89,58%.

Table 2-5. Ranking of the best models in Pociecha et al. (2014) research for the unbalanced sample (random sampling) one year before bankruptcy

Ranking	Model Type	Training:Test Proportion	Test sample		
			Error Type 1	Error Type 2	Accuracy
1	NN((Current Assets - Inventories - Short-Term Receivables) / Current Liabilities; (Operating Profit + Depreciation) / Total Assets)	6:4	100.00	91.67	95.83
2	NN((Operating Profit + Depreciation) / Total Assets; Operating Profit / Total Assets; 2 * Net Sales Revenues / (Short-Term Receivables (t) + Short-Term Receivables (t-1)); Operating Costs / Short-Term Liabilities)	7:3	100.00	83.33	91.67
3	CART(Net Profit (Loss) * 100 / Total Assets)	6:4	100.00	79.17	89.58

Source: Pociecha et al. (2014, p. 112), NN – neural network, DA – discriminant analysis, CART, CART – decision trees.

2.2.4. Korol and Prusak book

The Korol and Prusak book[7] is another publication dedicated to bankruptcy detection of Polish companies. Researchers used data for 180 manufacturing companies which were published in Monitor Polski B in years 1998 – 2001. Dataset was divided into a training set, which consisted of 39 bankrupt companies and 39 active companies and the test sample, which included 39 bankrupt companies and 39 active companies. Researchers used three different approaches: K1 – model trained on all financial ratios, K2 – model trained on financial ratios selected based on variance matrix, K3 – researchers selected the ratios for training arbitrarily. Researchers used the neural network with 1 hidden layer and experimented with different numbers of neurons looking for the most efficient combination (which produced the highest accuracy of the model).

Researchers tested different structure of the training sample: either 1:1 proportion between bankrupt and active companies, 3:1 proportion (3 active companies for each 1 bankrupt company), finally 10:1 proportion (10 active companies for each 1 bankrupt company). Selected results of the estimated models are presented in Table 2-6, 2-7 and 2-8.

Table 2-6. The accuracy of Korol and Prusak models, K1 approach

proportions training sample	Hidden neurons no	proportions test sample								
		`1:1 Error 1	`1:1 Error 2	`1:1 Accuracy	`3:1 Error 1	`3:1 Error 2	`3:1 Accuracy	`10:1 Error 1	`10:1 Error 2	`10:1 Accuracy
1:1	28	0	2,25	98,72	0	2,56	98,08	0	2,56	97,67
3:1	10	15,38	0	92,31	0	0	100	0	0	100
10:1	10	41,02	0	79,49	7,69	0	98,08	0	0	100

Source: Korol and Prusak, 2015, p. 161 – 162.

Table 2-7. Accuracy of Korol and Prusak models, K2 approach

proportions training sample	Hidden neurons no	proportions test sample								
		`1:1 Error 1	`1:1 Error 2	`1:1 Accuracy	`3:1 Error 1	`3:1 Error 2	`3:1 Accuracy	`10:1 Error 1	`10:1 Error 2	`10:1 Accuracy
1:1	28	0	5,12	97,44	0	5,12	96,16	0	5,12	95,35
1:1	56	15,38	0	92,31	0	0	100	0	7,69	100
3:1	10	28,2	0	95,9	0	0	100	0	0	100
10:1	28	12,82	12,82	87,18	7,69	12,82	88,47	0	12,82	88,38
10:1	10	30,76	2,56	83,34	15,38	2,56	94,24	0	2,56	97,68

Source: Korol and Prusak, 2015, p. 161 – 162.

[7] Korol T., Prusak B.,Upadłość przedsiębiorstw, CeDeWu, Warszawa, 2014.

Table 2-8. Accuracy of Korol and Prusak models, K3 approach

proportions training sample	hidden neurons no	proportions test sample								
		`1:1 Error 1	`1:1 Error 2	`1:1 Accuracy	`3:1 Error 1	`3:1 Error 2	`3:1 Accuracy	`10:1 Error 1	`10:1 Error 2	`10:1 Accuracy
`1:1	`8:4:2	15,38	0	92,31	0	0	100	0	0	100
`3:1	`8:4:2	20,51	0	89,75	7,69	0	98,08	0	0	100
`10:1	`8:4:2	25,64	0	87,18	7,69	0	98,08	0	0	100

Source: Korol and Prusak, 2015, p. 170.

What can be seen from the results presented above is that the accuracy of the network was deteriorating along with the increasing imbalance of the training sample. For a balanced training sample, the model's accuracy was very high for all approaches: K1, K2, and K3. All accuracies were above 90%; some even reached 100%. For unbalanced samples, the classification accuracy was deteriorating and for the training sample proportions 10:1 the accuracy was below 90%. Researchers did not use validation samples.

3. Research method presentation

Table A1 in the Appendix 1 shows definitions of financial ratios used in the training of the models. For the algorithms like logit where one has to avoid multicollinearity (logit, SVM with linear kernel, naïve Bayes without the PCA application) we used only variables which were not highly correlated with each other (correlation had to be below 0.3). For models which are capable of creating their own ratios (algorithms based on the decision trees, neural networks) we used also nominal data from financial statements (but after the normalization). The limit of the variables (features) left in the final model training was 8 (the rest of the variables were removed beginning from the least important one).

Dataset chosen for the research included years 2008 – 2018 and 152 340 companies (with double-sided accounting). Data were downloaded from the database Orbis which belongs to Bureau Van Dijk company. Data included the balance sheet (statement of financial position) and income statement elements. Data was tested whether the sum of assets was equal to the sum of equity and liabilities. Moreover, any suspicious or error records or columns were also removed. Finally, a panel of 1526 bankrupt companies and 1561 active companies was selected from the sample. Active companies were randomly chosen, but we tried to match the type of economic activity between each bankrupt company and a 'matching' active company. As a status change date, we assumed the year when the company had a negative equity for the first time. We assumed that the insolvency application must have been filed one year earlier and that the model should be able to know about it one year later.

The data sample was then divided into a training set including the companies data for years 2008-2013 (1411 active and 1376 bankrupt companies from which 10% of the sample was used during each training as a testing set) and a testing (evaluation) set including the companies data for years 2014-2018 (150 active and 150 bankrupt companies). Companies selected for every dataset were different between the sets (companies in the validation set did not belong to the training set). After validity checking and processing, the collected data was normalized.

For processing, we took advantage of the skleran Python library. Training data sample was divided into 10 parts and for each iteration we used 9 parts for training and 1 part for testing (cross-validation).

As it was explained earlier, the following models were trained (all with skleran library): discriminant analysis (DA), logit model (L), support vector machines (SVM), Random Forest (RF, 100 trees), Gradient Boosting Decision Trees (GB), Neural Network with 1 hidden layer (NN, the number of input neurons and in the hidden layer was equal to the number of variables), Convolutional Neural Network (CNN) and Naïve Bayes (NB)

The code was written in Jupyter Notebook (Python version 3.6).

For gradient boosting decision trees model, the number of estimators was assumed to be 100, the function which measured the quality of split was mean squared error with improvement score by Friedman, the learning rate was 10% and the improvement of the loss function was calculated as the deviation between the value for the out-of-the-bag samples and the value for the previous iteration. Maximum depth of the individual regression estimators was set to 3.

For the NN model, we assumed 100 trees. The activation function used for RF was Scaled Exponential Linear Unit (we also tested ReLU). As initializer was used LeCun uninform initializer.

CNN was also based on SeLU function and LeCun uinformed initializer and the sigmoid activation function. The loss function was a binary cross-entropy. Random forest assumed 100 decision trees, nodes were split based on the Gini criteria.

4. Results and Discussion

Table 4-1 shows the results of the model training and validation. As can be seen, the accuracies for every training subset were similar, at the bottom, there are calculated average accuracies for every method.

Table 4-1. Comparison of accuracy

training subset	Logit	DA	SVM	RF	GB	NN	CNN	NB
1	84,2	82,4	66,9	94,4	95,0	79,9	79,8	69,4
2	82,7	83,2	68,4	94,5	94,5	79,1	79,9	67,6
3	83,8	82,8	66,4	94,5	95,2	81,6	78,2	70,7
4	84,2	82,4	64,5	94,9	94,5	79,6	77,5	71,0
5	84,6	82,9	65,4	94,1	94,2	76,8	77,1	69,8
6	82,9	82,9	67,5	93,9	93,6	77,5	78,2	68,0
7	84,0	83,2	65,8	93,8	94,7	80,1	77,8	70,0
8	84,3	81,8	65,3	94,3	94,3	80,7	79,3	72,1
9	83,8	83,4	64,2	93,9	94,5	77,7	78,9	68,0
10	83,3	82,7	69,2	94,3	94,5	82,6	78,5	68,4
average	**83,8**	**82,8**	**66,3**	**94,3**	**94,5**	**79,6**	**78,5**	**69,5**
validation sample accuracy	**81,2**	**80,5**	**64,8**	**93,7**	**93,8**	**78,1**	**76,3**	**67,1**
error 1	11,7	11,7	28,9	0	0	14,9	11,7	25,7
error 2	25,9	27,4	41,6	12,7	12,4	28,9	35,7	40,2

Source: own calculations.

As can be seen in Table 4-1, the most accurate method proved to be the Gradient boosting decision trees algorithm. The accuracy for the validation sample was 93,8% which means that for 100 companies the model would make a mistake for 7 companies. The second best was random forest decision trees with the accuracy for the validation sample 93,7%. In terms of testing 100 companies against the bankruptcy risk, this model would be also wrong for 7 companies. The rest of the models did not achieve the accuracy levels which would seem useful for the practical application, although we did not use macroeconomic data, which could significantly improve the quality of the classifications (this is one of the suggestions for further research).

5. Conclusions

Presented research was designed to observe the behavior of various machine learning methods when they were trained on a representative sample of the companies. The best and most efficient algorithms proved to be the GBDT and RFDT. However, low accuracy of other analyzed machine learning algorithms must be treated with caution, because there exist many variants of these methods which potentially can significantly change their accuracy.

Based on the empirical evidence, there seems that the hypothesis H1 must be partially rejected (*if one has access to the big sample of companies, the most efficient algorithms (first choice) in the bankruptcy prediction are Neural Networks, Gradient Boosting Decision Trees and Random Forest algorithms*). Neural network with one hidden layer did not prove to be sufficiently accurate to confirm this hypothesis, however, this could be the problem of insufficient tuning of the model.

A significant advantage of the decision tree based algorithms may be due to their technical properties. GBDT and RFDT algorithms do not require any removal of the outliers, they are also believed to be resistant to multicollinearity because both methods put outliers and different variables into different leaves of the decision trees and they do not influence the prediction process. Both methods also handle automatically missing values, so they can learn from the cases which were useless for classical methods. Both algorithms deal well with the lack of the normal distribution of the independent variables. Random Forest algorithm is based on the majority voting, so it uses a technique quite popular in the financial analysis where one often uses multiple bankruptcy prediction models (from the corporate finance literature) and estimates solvency of the company based on multiple models of different authors taken together.

Yet another advantage of both methods was that they used the nominal data taken from the financial statements (after normalization) and there was no need to calculate any financial ratios for them. The algorithms managed to train themselves truly automatically from the given data. Both algorithms are also capable of detecting non-linear feature interactions and adjust to them. Some feature may achieve very high and very low values for insolvent companies and the algorithms are capable of recognizing this behavior and adjusting to it.

6. Practical implications and discussion

For a long time bankruptcy prediction was based on the models which were provided by researchers in the form of a linear equation with given coefficients and formulae of financial ratios to use as variables in such a formula. The most famous Altman model survived for more than 50 years and it is still being used by the practitioners. However, due to the significant differences in the accuracy, companies all over the world switch to machine learning algorithms which usually takes the form of server access purchase which offers a commercial credit scoring/bankruptcy prediction system. They prefer to pay a monthly fee and participate in the system which trains itself from the big data than to risk mistakes and use free non-commercial linear bankruptcy prediction models. With the differences in the accuracy presented in the paper between the DA and LOGIT algorithms compared to RFDT and GBDT, this decision pays for itself.

The second important implication for the economic practice is that if only one has access to big enough sample of financial data, training of their own model is not very difficult, because the RFDT and GBDT work very well almost out of the box. One does not have to make decisions which data should be removed as outliers, and this not only facilitates training of the model but also leaves potentially valuable information in the sample. Decision Trees based models may not work very well with time, and due to their non-linear nature may not allow a significant extrapolation, but if they are fed regularly with new information, they can learn from that and maintain their high accuracy. They are rather short-term solutions than long-term prediction models, but they are relatively easy to train and tune-up. Adjustments to the neural network concerning the number of neurons, the number of hidden layers, activation functions, can take a very long time, as well as the training of the neural network.

Another observation from the paper is that the ensemble methods generally perform better than non-ensemble algorithms. On one hand, it is understandable and natural, but there seems to be the price to be paid for it. Decision tree-based algorithm and neural networks operate like black boxes and it is challenging to understand how exactly the models work. This is a significant disadvantage of these methods, they cannot be shared between interested parties as easily as logit or discriminant analysis models.

7. Literature

Journals:

Aktan S., *Application of machine learning algorithms for business failure prediction*, "Investment Management and Financial Innovations", 2011, volume 8, issue 2, p. 52-65 [SVM, DT, BN, NB].

Alaminos D., del Castillo A., Fernandez M., *A global model for bankruptcy prediction*, "PLoS One", 2016, volume 11, issue 11, doi: 10.1371/journal.pone.0166693 [CART, DT, BN, SVM, NB].

Alfaro, E., Garcia, N., Games, M., Elizondo, D., *Bankruptcy forecasting: an empirical comparison of Ada Boost and neural networks*, "Decision Support Systems", 2008, volume 45, pp. 110-122 [DA, DT, NN, (A)DT].

Altinirmak S., Karamasa C., *Comparison of Machine Learning Techniques for Analyzing Banks' Financial Distress*, "Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi Cilt", 2016, volume 19, issue 36, Aralık, pp. 291 – 303 [SVM].

Anandarajan M., Lee P., Anandarajan A., *Bankruptcy Prediction of Financially Stressed Firms: An Examination of the Predictive Accuracy of Artificial Neural Networks*, "International Journal of Intelligent Systems in Accounting, Finance & Management", 2001, volume 10, pp. 69–81 [DA, NN, (Ba)NN].

Arieshanti I., Purwananto Y., Ramadhani A., Nuha M.U., Ulinnuha N., *Comparative Study of Bankruptcy Prediction Models*, "Telkomnika", 2013, volume 11, number 3, September, pp. 591 – 596 [SVM].

Barboza F., Kimura H., Altman E., *Machine Learning Models and Bankruptcy Prediction*, "Journal of Expert Systems with Applications: An International Journal", 2017, volume 83, issue C, October, pp. 405 – 417 [DA, LR, NN, SVM, (A)NN, (Rf)DT, (Bg)NN].

Blanco-Oliver A., Irimia-Dieguez A., Oliver-Alfonso M., Wilson N., *Improving Bankruptcy Prediction in Micro-Entities by Using Nonlinear Effects and Non-Financial Variables*, "Czech Journal of Economics and Finance", 2015, volume 65, number 2, pp. 144 – 166 [LR].

Bryniolfsson E., Mcafee A., *The Business of Artificial Intelligence*, "Harvard Business Review", July 2017, pp. 1-8.
Chaudhuri A., De K., *Fuzzy Support Vector Machine for bankruptcy prediction*, "Applied Soft Computing", 2011, volume 11, issue 2, pp. 2472 – 2486 [LR, NB, SVM].
Chen M.-Y., *Comparing Traditional Statistics, Decision Tree Classification And Support Vector Machine Techniques For Financial Bankruptcy Prediction*, "Intelligent Automation & Soft Computing", 2012, volume 18, issue 1, pp. 65-73 [DT, CART, SVM].
Cho S., Hong H., Ha B.C., *A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction*, "Expert Systems with Applications", 2010, volume 37 pp. 3482–3488 [DT, LR, NN].
Cho, S., Kim, J., Bae, J. K., *An integrative model with subject weight based on neural network learning for bankruptcy prediction*, "Expert Systems with Applications", 2009, volume 36, pp. 403-410 [DA, DT, LR, NN, SVM].
Chuang, C. L., *Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction*. "Information Sciences", 2013, volume 236, pp. 174-185 [LR].
Dellepiane U., Marcantonio M., Laghi E., Renzi S., *Bankruptcy Prediction Using Support Vector Machines and Feature Selection During the Recent Financial Crisis*, "International Journal of Economics and Finance", 2015, volume 7, number 8, pp. 182 – 195 [SVM].
Ecer F., *Comparing the bank failure prediction performance of neural networks and support vector machines: the Turkish case*, "Ekonomska Istraživanja-Economic Research", 2013, volume 26, issue 3, pp. 81-98 [SVM].
Edrogan B.E., *Prediction of bankruptcy using support vector machines: an application to bank bankruptcy*, "Journal of Statistical Computation and Simulation", 2013, volume 83, issue 8, pp. 1543 – 1555 [SVM].
Fedorova, E., Gilenko, E., Dovzhenko, S., *Bankruptcy prediction for Russian companies: application of combined classifiers*, "Expert Systems with Applications", 2013, volume 40, pp. 7285-7293 [DA, (A)NN].
Geng, R., Bose, I., Chen, X., *Prediction of financial distress: an empirical study of listed Chinese companies using data mining*, "European Journal of Operational Research", 2015, volume 241, pp. 236-247 [DT,NN, SVM].
Ghodselahi A., Amirmadhi A., *Application of Artificial Intelligence Techniques for Credit Risk Evaluation*, "International Journal of Modelling and Optimization", 2011, volume 1, number 3, August, pp. 243 – 249 [DA, DT, LR, NN, SVM, (B(DT), (Bg)NN, (Gb)DT, (Bo)NN, (Bo)SVM].
Hauser R., Booth D., *Predicting Bankruptcy with Robust Logistic Regression*, "Journal of Data Science", 2011, volume 9, pp. 565 – 584 [LOGIT].
Heo, J., Yang, J. Y., *AdaBoost based bankruptcy forecasting of Korean construction companies*, "Applied Soft Computing", 2014, volume 24, pp. 494-499 [DT, NN, SVM, (A)DT].
Hu Y-C., Tseng F-M., *Functional-link net with a fuzzy integral for bankruptcy prediction*, "Neurocomputing", 2007, volume 70, pp. 2959–2968. [DA, LR, PROBIT].
Jardin P., *Bankruptcy prediction models: How to choose the most relevant variables?*, [in:] "Munich Personal RePEc Archive", 2009, MPRA Paper, number 44380 [DA, DT, CART, LR, (Bo)NN, (Bg)NN, NN, (RS)LR].
Jayanthi J., Suresh J.K., Vaishnavi J., *Bankruptcy Prediction using SVM and Hybrid SVM Survey*, "International Journal of Computer Applications", 2011, volume 34, number 7, November, pp. 39-45 [SVM].
Kasgari A.A., Salehnezhad S.H., Ebadi F., *The Bankruptcy Prediction by Neural Networks and Logistic Regression*, "International Journal of Academic Research in Accounting, Finance and Management Sciences", 2013, volume 3, number 4, October 2013, pp. 146–152 [LR, NN].
Kim M.J., Kang D.K., *Ensemble with neural networks for bankruptcy prediction*, "Expert Systems with Applications", 2010, volume 37, pp. 3373 – 3379 [DT, NN, SVM, (B)DT, (Bo)NN, (Gb)NN, (Bo)SVM].

Kim M-J., Kang D.-K, *Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction*, "Expert Systems with Applications", 2012, volume 39, pp. 9308–9314 [NN, (A)NN, (Bg)NN].
Kim, S. Y., Upneja, A., *Predicting restaurant financial distress using decision tree and Ada Boosted decision tree models*, "Economic Modelling", 2014, volume 36, pp. 354-362 [DT, (A)DT].
Ko L.-J., Blocher E., Lin P., *Prediction of Corporate Financial Distress: An Application of the Composite Rule Induction System*, "The International Journal of Digital Accounting Research", 2001, volume 1, number 1, pp. 69-85 [LR, NN].
Krichene A., *Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank*, "Journal of Economics, Finance and Administrative Science", 2017, volume 22, issue 42, pp.3-24 [NB].
Laitinen E., Laitinen T., *Bankruptcy prediction: Application of the Taylor's expansion in logistic regression*, "International Review of Financial Analysis", 2000, volume 9, issue 4, winter, pp. 327 – 349 [LOGIT].
Li, H., Lee, Y. C., Zhou, Y. C., Sun, J., *The random subspace binary logit (RSBL) model for bankruptcy prediction*, "Knowledge-Based Systems", 2011, volume 24, pp. 1380-1388 [DA, LR, PROBIT, (RS)LR].
Li, H., Sun, J., *Majority voting combination of multiple case-based reasoning for financial distress prediction*, "Expert Systems with Applications", 2009, volume 36, pp. 4363-4373 [DA, LR].
Li, H., Sun, J., *Business failure prediction using hybrid2 case-based reasoning H2CBR*, "Computers and Operations Research", 2010, volume 37, pp. 137-151 [DA, LR].
Li, H., Sun, J., *Principal component case-based reasoning ensemble for business failure prediction*, "Information and Management", 2011, volume 48, pp. 220-227 [DA, LR].
Liao, J. J., Shih, C. H., Chen, T. F., Hsu, M. F., *An ensemble-based model for two-class imbalanced financial problem*, "Economic Modelling", 2014, volume 37, pp. 175-183 [DA, DT, (Rf)DT, LR, BN].
Marques, A. I., Garcia, V., Sanchez, J. S., *Exploring the behaviour of base classifiers in credit scoring ensembles*, "Expert Systems with Applications", 2012, volume 39, pp. 10244-10250 [DT, LR, NN, SVM, (Rs)SVM, (A)DT, (Bg)NN].
Min J., Jeong C., *A binary classification method for bankruptcy prediction*, "Expert Systems with Applications", 2009, volume 36, pp. 5256–5263 [DA, DT, LR, NN].
Min J., Lee Y.C., *Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters*, "Expert Systems with Applications", 2005, volume 28, issue 4, pp. 603 – 614 [DA, LR, NN, SVM].
Mirzaei M., Ramakrishnan S., Bekri M., *Corporate Default Prediction with Industry Effects: Evidence from Emerging Markets*, "International Journal of Economics and Financial Issues", 2016, volume 6, issue 53, pp. 161-169 [DT, CART, LR].
Nagaraj K., Sridhar A., *A predictive system for detection of bankruptcy using machine learning techniques*, "International Journal of Data Mining & Knowledge Management Process (IJDKP)", 2015, Vol.5, No.1, January, p. 29 – 40 [LR, NN, SVM, BN, (Rf)DT].
Nanni L., Lumini A., *An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring*, "Expert Systems with Applications", 2009, volume 36, pp. 3028–3033 [NN, SVM, (Bg)SVM, (Bo)NN, (Bo)SVM, (Rs)SVM].
Pawełek B., Grochowina D., *Podejście wielomodelowe w prognozowaniu zagrożenia przedsiębiorstw upadłością w Polsce*, „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu", 2017, number 468, p. 171 – 179.
Ramakrishnan S., Mirzaei M., Naveed M., *Corporate Bankruptcy Prediction: A Case of Emerging Economies*, "International Journal of Sciences Basic and Applied Research", 2015, volume 19, number 1, pp 177-187 [DT, NN, SVM].
Shin K.S., Lee T.S., Kim H.J., *An application of support vector machines in bankruptcy prediction model*, "Expert Systems with Applications", 2005, volume 28, issue 1, January, pp. 127 – 135 [SVM].

Shin K-S., Lee T.S., Kim H.J., *An application of support vector machines in bankruptcy prediction model*, "Expert Systems with Applications", 2005, volume 28, pp. 127 – 135 [SVM].
Sun, J., Li, H., *Financial distress prediction based on serial combination of multiple classifiers*, "Expert Systems with Applications", 2009, volume 36, pp. 8659-8666 [LR, NN, SVM, DA, DT].
Sun, J., Li, H., *Financial distress prediction using support vector machines: ensemble vs. individual*, "Applied Soft Computing", 2012, volume 12, pp. 2254-2265 [SVM].
Sun, J., Jia, M. Y., Li, H., *AdaBoost ensemble for financial distress prediction: an empirical comparison with data from Chinese listed companies*, "Expert Systems with Applications", 2011, volume 38, pp. 9305-9312 [DT, SVM, (A)DT].
Tsai C.-F., Cheng K.-C., *Simple instance selection for bankruptcy prediction*, "Knowledge-Based Systems", 2012, volume 27, pp. 333–342 [DT, LR].
Tsai, C. F., *Combining cluster analysis with classifier ensembles to predict financial distress*, "Information Fusion", 2014, volume 16, pp. 46-58 [DT, LR, NN].
Tseng F.M., Hu Y.C., *Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks*, "Expert Systems with Applications", 2010, volume 37, pp. 1846–1853 [LOGIT, (Ba)NN].
West, D., Dellana, S., Qian, J., *Neural network ensemble strategies for financial decision application*, "Computers and Operations Research", 2005, volume 32, pp. 2543-2559 [NN, (Bg)NN, (A)NN].
Xiao, Z., Yang, X. , Pang, Y., Dang, X., *The prediction for listed companies' financial distress by using multiple prediction methods with rough set and Dempster-Shafer evidence theory*, "Knowledge-Based Systems", volume 202, issue 26, pp. 196-206 [LR, NN, SVM].
Zhou L., Lai K.K., Yen J., *Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimization*, "International Journal of Systems Science", 2014, volume 45, issue 3, pp. 241 – 253 [SVM].

Books:
Gatnar E., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa, 2008.
Ghatak A., *Machine learning with R*, Springer, New York, 2017.
Korol T., Prusak B., *Upadłość przedsiębiorstw*, CeDeWu, Warszawa, 2014.
Lee J., Jang D., Park S., *Deep Learning-Based Corporate Performance Prediction Model Considering Technical Capability*, Sustainability, 9, doi:10.3390/su9060899, 2017 [DT, CART, LR, NN, GP].
Lewis N.D., *Deep Learning Made Easy - A gentle introduction for data science*, Create Space Independent Publishing Platform, New York, 2016.
Lewis N.D., *Machine Learning Made Easy with R: An Intuitive Step by Step Blueprint for Beginners*, CreateSpace Independent Publishing Platform, New York, 2017.
Pena T., Martinez S., Abudu B., *Bankruptcy Prediction: A Comparison of Some Statistical and Machine Learning Techniques*, Banco de Mexico Working Papers, no 18, 2009 [DA, LR, GP, PROBIT, SVM, BN].
Pociecha J., Pawełek B., Baryła M., Augustyn S., *Statystyczne metody prognozowania bankructwa w zmieniającej się koniunkturze gospodarczej*, Wydawnictwo UEK, Kraków, 2014.

Appendix 1:

Table A1. Construction of Financial ratios included in training of the models

Symbol	Definition
RSHF	(Profit before tax / Shareholders funds) * 100
RCEM	(Profit before tax + Interest paid) / (Shareholders funds + Non current liabilities) * 100
RTAS	(Profit before tax / Total assets) * 100

Symbol	Definition
ROE	(Net income / Shareholder funds) * 100
ROCE	(Net income + Interest paid) / (Shareholder funds + Non current liabilities) * 100
ROA	(Net income / Total Assets) * 100
PRMA	(Profit before tax / Operating revenue) * 100
GRMA	(Gross profit / Operating revenue) * 100
ETMA	(EBITDA / Operating revenue) * 100
EBMA	(EBIT / Operating revenue) * 100
NAT	Operating revenue / (Shareholders funds + Non current liabilities)
IC	Operating profit / Interest paid
STOT	Operating revenue / Stocks
COLL	(Debtors / Operating revenue) * 360
CRPE	(Creditors / Operating revenue) * 360
RDOP	(Research & development / Operating revenue) * 100
CURR	Current assets / Current liabilities
LIQR	(Current assets - Stocks) / Current liabilities
SHLQ	Shareholders funds / Non current liabilities
SOLR	(Shareholders funds / Total assets) * 100
SOLL	(Shareholders funds / (Non current liabilities + Current liabilities)) * 100
GEAR	((Non current liabilities + Loans) / Shareholders funds) * 100
NatWym	Short-Term Investments / Trade Liabilities
ZadlAkt	(Total Assets - Shareholder's Funds)/Total Assets
ZadlKapWl	(Total Assets - Shareholder's Funds) / Equity
ZadlDlug	Long-Term Liabilities/Total Assets
PozKoszOp	Operating Costs / Operating Revenue

Source: own calculations.