Predicting Bankruptcies

Machine Learning in Econometrics

Hollósi István, Pókász Zoltán, Soós Adrienn, Szabó Gábor

Table of Contents

[**Introduction** 3](#_Toc153935169)

[**Literature Review** 4](#_Toc153935170)

[**About the dataset** 5](#_Toc153935171)

[**Data manipulation** 6](#_Toc153935172)

[**Econometric methodology** 6](#_Toc153935173)

[**Results** 7](#_Toc153935174)

[***Works Cited*** 15](#_Toc153935175)

# **Introduction**

Great importance lies in predicting which companies have higher risk of bankruptcy. This question has been in the center of research for a long time, as the prediction of bankruptcy is relevant from the perspective of financial institutions, investors and the general economy as well. Since machine learning methods started to gain attention as improved analytical tools, research has started to investigate bankruptcies with these techniques.

There has been significant research on the topic of applying machine learning algorithms in predicting financial distress. The literature is conclusive on that machine learning algorithms outperform traditional statistical techniques, but there is still debate on which algorithm performs best on different databases.

The project uses the *US Company Bankruptcy Prediction Dataset* from *Kaggle.* The dataset includes companies that are listed on the New York Stock Exchange and NASDAQ. The dataset contains data from 8262 distinct companies for about 20 years. The dataset contains information about the different attributes of the companies and about whether they have declared bankruptcy in the examined period.

The goal of this project is to examine the probability of bankruptcy of US Companies using Machine Learning methods and therefore contribute to the existing literature.

1 bekezdés a módszertanról

Our results show that the best method from which we compared is the simple regression tree, with 76.30 percent balanced accuracy. With bagging regression tree, random forest and gradient boosting we have better results to sensitivity (in our case the positives are being a non-bankrupt company), but in these models specificity (correct predictions of a bankruptcy) are low. Regarding to the importance of variables the key variables are the changes of market value, retained earnings, and liabilities but because of multicollinearity we can have limited economic conclusions from this finding.

# **Literature Review**

Devi et al (2018) test the methods of artificial neural networks, support vector machines and decision trees and compare these to traditional statistical techniques, such as linear discriminant analysis, multivariate discriminant analysis and logistic regression. Machine learning methods have a higher performance in all four metrics they examined: accuracy, precision, sensitivity and specificity and these effects applied even for smaller datasets. (Devi et al., 2018)

Barbosa et. al. (2017) analyzed over 10 000 firm-year observations from 1985 to 2013 from the North American stock market. They found that machine learning algorithms improved prediction accuracy especially if in addition to using the original Altman’s Z-score variables, they include six complementary financial indicators. These algorithms have approximately 10% better accuracy then traditional tools. They tested different machine learning algorithms and found that bagging, boosting, and random forest models outperform the other techniques. On the other hand they do not prove that SVM leads to higher accuracy rates (Barbosa et. al., 2017).

Wang (2017) has tested three machine learning algorithms on real life data for bankruptcy prediction. She tested the support vector machine, neural network with dropout, autoencoder to determine which has the best accuracy and found that neural network with added layers with dropout perform the best in this regard (Wang, 2017).

Chen (2012) examined decision tree classification methods and artificial network techniques to evaluate their capabilities for predicting financial distress. He found that the support vector machine technique is a better technique for predicting financial difficulties, compared to traditional methods and decision trees (Chen, 2012). This contradicts the previously mentioned Barbosa et. al. (2017) study, that stated that SVM does not lead to higher accuracy. The contradicting results can derive from the fact that machine learning algorithms are a fast improving field, and the newer study had the opportunity to test improved functions.

Zieba et. al. (2016) purpose the approach to use Extreme Gradient Boosting for learning an ensemble of decision trees to predict bankruptcies. They introduce the concept of synthetic features, that is a combination of econometric measures and can be viewed as a single regression model that is developed in an evolutionary manner. They examine this technique on Polish companies to determine financial hardness. The results showed that the boosting model had better performance in predicting bankruptcy. The models that included the synthetic features the authors proposed showed even better performance in accuracy.

There has been significant research recently about introducing machine learning methods to bankruptcy prediction. The literature seems conclusive that machine learning algorithms tend to outperform traditional statistical tools, especially in accuracy. On the other hand, there is still significant debate over which algorithm gives the best performance, as there are no algorithms that consistently outperform others in all types of datasets.

# **About the dataset**

As previously stated, our dataset contains information about 8,262 NYSE and NASDAQ companies for a total of about 78 thousand observations for a time span of 19 years (1999-2018).

There are 18 variables overall, the most important of which is whether a given company had gone bankrupt throughout the 19-year period. After bankruptcy, the given firm is removed from the database. Therefore, it is an unbalanced panel data set, as the available years differ by firm. Overall, seven percent of the firms had gone bankrupt by the end of the sample, the other 93 percent operating in a normal fashion. Bankruptcy is defined as filing bankruptcy according to Chapter 7 or Chapter 11 of the Bankruptcy Code.

Most of the other variables (other than Year) are either items on the companies’ balance sheets or elements of their income statements. Some examples are:

· Current assets

· EBITDA

· Net Sales

· Total Long-term Debt

As such, we will mostly rely on fundamental variables of the firms to predict bankruptcy.

# **Data manipulation**

It should be noted that in the original database, when a company goes bankrupt, its status is retrospectively overwritten as „bankrupt” for the years preceding the bankruptcy too. This would be detrimental to our analysis as we would not be able to exploit the time series nature of the data as well. Accordingly, we modified the data so as to only show a „bankrupt” status the year the company actually goes bankrupt.

Our predictions rely on the three years preceding the bankruptcy as well as the bankruptcy’s year. To be precise, we postulate that the averages of the firms’ characteristics in the last three years leading up to the bankruptcy have significant predictive power when compared to the variable values in the bankruptcy’s year. Therefore, we transform the original variables so as to reflect changes: we compare their previous average values to their values in the given period.

where is calculated by taking the previous three years’ average:

Consequently, if a firm (be it bankrupt or not) has less than four years of observations, the firm is automatically deleted from our dataset. It should be noted that the values are standardized.

Finally, in order to filter outliers, extreme values of the *Change* variable (those above 300 percent, or a 3-time increase compared to the last 3 years) are removed from the dataset. These problems arose especially when looking at the *Inventory* and *Total Long-term Debt* variables, as they often had values of 0. In some cases, this resulted in very high values of *Change* arising from small and somewhat bigger values.

# **Econometric methodology**

Our main goal was to classify which U.S. firms are likely to go bankrupt. Therefore, it was a binary classification problem. Three different models were used for the predictions with a variety of parameters and settings: regression trees (without bagging and with bagging), random forest (with bagging, with the standard variable count) and gradient boosting.

These methods are appropriate when trying to solve classification problems as they are particularly well interpretable graphically in case of a discrete dependent variable such as possible bankruptcy. Diversity of the models is also helped by the fact that random forests do not contain as many variables as decision trees do in most of the cases.

In order to further mitigate overfitting and test which model’s predictions prove to be the most suitable for the task at hand, we split our dataset into train and test samples. We used a 70-30 train-test split: the first 70 percent of observations belong to the training sample, while the remainder constitues the test sample, rather than choosing the 70 and 30 percent randomly. This allows us to exploit the time series structure of the dataset better than as if we had chosen these samples randomly.

The second subjective choice arising when modelling is choosing the appropriate weights. Contrary to equal weighting, we decided to give twice as much weight to the „bankrupt” observations in a somewhat arbitrary fashion. This is due to the scarce number of these observations (only one observation for every bankrupt company, which are themselves also limited in number). We also tried oversampling the „bankrupt” observations, however, this did not yield a significant benefit.

# **Results**

*Table 1: confusion matrix and accuracy of Regression Tree without bagging*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Actual |  |  |  |  |
|  | 0 | 1 |  |  |  |
| Prediction 0 | 8864 | 34 |  | Accuracy | 75.26% |
| 1 | 2909 | 87 |  | Sensitivity | 75.29% |
|  |  |  |  | Specificity | 71.90% |
| *Source: authors’ own calculation* | | |  | Balanced accuracy | 73.60% |

*Table 2: confusion matrix and accuracy of Bagging Regression Tree model*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Actual |  |  |  |  |
|  | 0 | 1 |  |  |  |
| Prediction 0 | 11730 | 110 |  | Accuracy | 98.71% |
| 1 | 43 | 11 |  | Sensitivity | 99.643% |
|  |  |  |  | Specificity | 9.09% |
| *Source: authors’ own calculation* | | |  | Balanced accuracy | 54.36% |

*Table 3: confusion matrix and accuracy of Random Forest*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Actual |  |  |  |  |
|  | 0 | 1 |  |  |  |
| Prediction 0 | 11760 | 112 |  | Accuracy | 98.95% |
| 1 | 13 | 9 |  | Sensitivity | 99.89% |
|  |  |  |  | Specificity | 7.44% |
| *Source: authors’ own calculation* | | |  | Balanced accuracy | 53.66% |

*Table 4: confusion matrix and accuracy of Gradient Boosting Model*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Actual |  |  |  |  |
|  | 0 | 1 |  |  |  |
| Prediction 0 | 11434 | 92 |  | Accuracy | 96.38% |
| 1 | 339 | 29 |  | Sensitivity | 97.12% |
|  |  |  |  | Specificity | 23.97% |
| *Source: authors’ own calculation* | | |  | Balanced accuracy | 60.54% |

We compared the four methods based primarily on the confusion matrices (Table 1-4.). Namely the performance of the test predictions against the observations in the sample with respect to sensitivity, specificity and balanced accuracy. First and foremost, we wanted to maximize accuracy, rather than specificity, that is how we chose the cut-off values. Alternatively, we could have chosen to maximize specificity, which is the ability of the model to correctly predict bankrupt firms; the right choice is a matter of debate.

We find that in the view of sensitivity (true positive rate, where positive is an operating company) the bagging regression tree, the bagging random forest and the gradient boosting overperformed the simple regression tree without bagging. The random forest gave the best result as it predicted the non-failed companies as non-failed correctly in 99.89 percent, while on the contrary, the simple regression tree has only a 75.29 percent true positive ratio. The gradient boosting (97.12 percent) was just slightly overperformed by the random forest and the bagged regression tree (99.64 percent).

However, in the prediction of bankruptcy, there is a crucial role of specificity, since that shows how accurately the method predicts a bankrupt company to be bankrupt (true negative ratio). The regression tree without bagging has a specificity of 71.90 percent, which is better than in case of the other models. In contrast with the sensitivity the random forest and the bagging regression tree did not improve the prediction. They have a specificity of 7.44 and 9.09 percent, respectively. Gradient boosting (23.97 percent) also underperformed the regression tree however overperformed the random forest and the bagging Even when we use the balanced accuracy of the methods to choose, the regression tree is the most useful in this case. This might be due to the over-regularization of ensemble methods in our case. Furthermore, bagging or boosting can potentially introduce bias in our estimation.

Now we turn to examining the importance of the variables (Figures 1-3.). Note that the exact variable name correspondences to balance sheet or financial statement items are found in the Appendix.

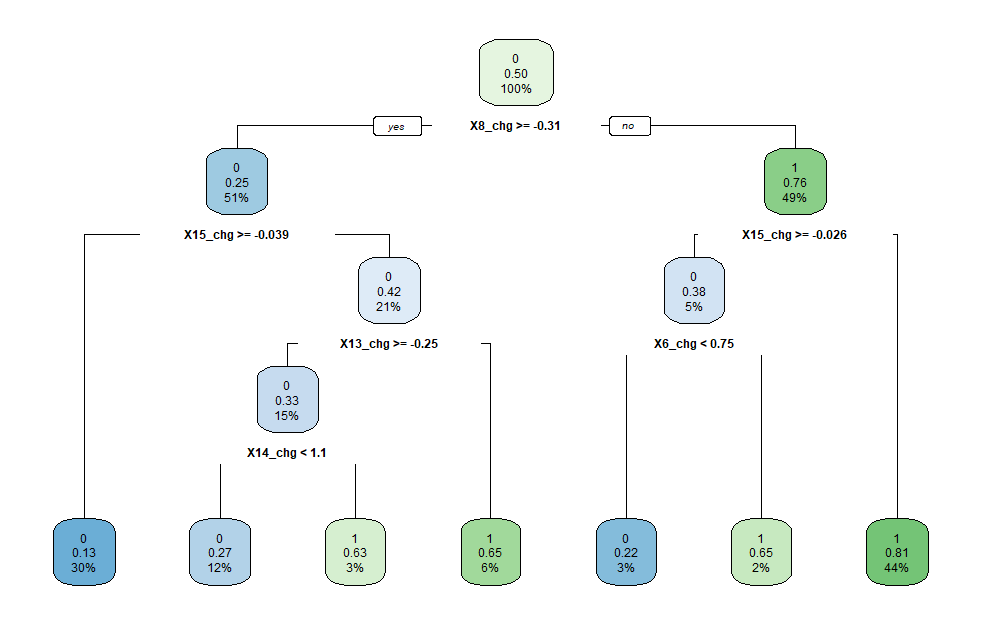
In this table, there are two columns, denoted „Mean Decrease Accuracy” and „Mean Decrease Gini”.

The Mean Decrease Accuracy plot expresses how much accuracy the model loses by excluding each variable. The more the accuracy suffers, the more important the variable is for the successful classification. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. This latter can be read from Fig. 1. as well in case of the bagged Regression Tree.

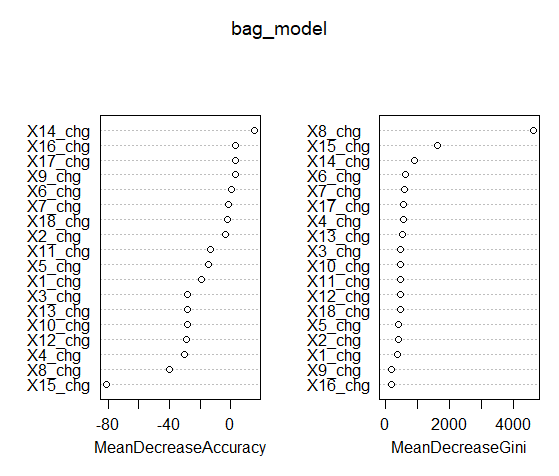
The pattern related to Gini is very similar in all cases (i.e. the same variables are the most important ones), which is consistent with our expectations. Market value change (X8\_chg) seem to be the most important in the regression tree. This suggests that the dataset can be most effectively split into two sub-datasets when looking at the Market value change of the firm. Also Retained earnings change (X15\_chg) is important. We ought to emphasize once again that our variables are changes, therefore not the amount of retained earnings are examined but the growth. It's also intuitive that if a company’s profits decline, that can be a sign of a future bankruptcy. Beside these findings there are lots of variables which could intuitively be more important but they do not affect the model as much (for example the EBIT). This can be explained by the multicollinearity of variables which are given by the nature of the financial statements, but this does not affect the overall accuracy of the models.

In terms of accuracy which is more related to prediction outcomes, X14 (change of Total Liabilities) seems to matter the most in both cases. This is a similar metric to the change in Market value as it signifies the change in the size of a company.

*Figure 1: Representation of the bagged Regression Tree*

*Source: authors’ own calculation*

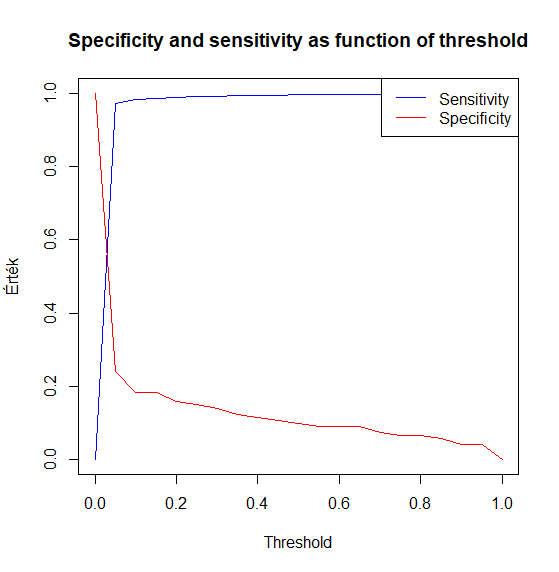
*Figure 2: Feature importance of the Bagging Regression Tree Model*

*  
Source: authors’ own calculation*

*Figure 3: Feature importance of Random Forest*

*  
Source: authors’ own calculation*

*Figure 4: Specificity and Sensitivity as a function of the cut-off value in case of the Boosting model*

  
*Source: authors’ own calculation*

**Summary**

We compared basic machine learning methods to predict a company’s bankruptcy using a 19-year long panel data of more than eight thousand NYSE and NASDAQ companies’ financial reports. After the data manipulation we divided the sample into a training and a test sample and compared a normal regression tree, a regression tree with bagging, a random forest model and a gradient boosting model using confusion matrices. The results showed that in terms of overall accuracy, the random forest model is the most appropriate. However, the most correct classification of bankrupt firms is achieved by a regression tree without bagging, which could indicate potential overfitting in case of the other variables or a subpar choice of the cut-off values in this regard.

**Appendix**: Correspondence of variable names and balance sheet / financial statement items

A képen szöveg, képernyőkép, Betűtípus, szám látható

Automatikusan generált leírás

**Bibliography**

*Barboza, Flavio, et al. “Machine Learning Models and Bankruptcy Prediction.” Expert Systems with Applications, vol. 83, Oct. 2017, pp. 405–417, https://doi.org/10.1016/j.eswa.2017.04.006.*

*Chen, Mu-Yen. “Comparing Traditional Statistics, Decision Tree Classification and Support Vector Machine Techniques for Financial Bankruptcy Prediction.” Intelligent Automation & Soft Computing, vol. 18, no. 1, Jan. 2012, pp. 65–73, https://doi.org/10.1080/10798587.2012.10643227. Accessed 3 Mar. 2019.*

*Devi, S. Sarojini, and Y. Radhika. “A Survey on Machine Learning and Statistical Techniques in Bankruptcy Prediction.” International Journal of Machine Learning and Computing, vol. 8, no. 2, Apr. 2018, pp. 133–139, https://doi.org/10.18178/ijmlc.2018.8.2.676.*

*“US Company Bankruptcy Prediction Dataset.” Www.kaggle.com, www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset/data?fbclid=IwAR00N6E4fd1B\_zSg0fpcFdI-8Kwdrg4ChXS9PwJNMSQ048VHJ3vLI9QlZ5g. Accessed 16 Dec. 2023.*

*Wang, Nanxi. “Bankruptcy Prediction Using Machine Learning.” Journal of Mathematical Finance, vol. 07, no. 04, 2017, pp. 908–918, https://doi.org/10.4236/jmf.2017.74049. Accessed 3 Mar. 2019.*

*Zięba, Maciej, et al. “Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction.” Expert Systems with Applications, vol. 58, 1 Oct. 2016, pp. 93–101, www.sciencedirect.com/science/article/pii/S0957417416301592, https://doi.org/10.1016/j.eswa.2016.04.001.*