BALANCED BAGGING WITH EXPECTATION MAXIMIZATION IMPUTATION IN BANKRUPTCY PREDICTION – APPLICATION ON ROMANIAN COMPANIES

Claudiu CLEMENT[1]

Alexandru Ioan Cuza University of Iasi

Abstract

Bankruptcy prediction models are widely used by lending institutions, policy makers or investors. Despite the large volume of international research, limited studies have addressed the particularities of Romanian companies. Balanced Bagging is an Ensemble Method that uses a voting mechanism for a classification task. Expectation Maximization Imputation helps replacing the missing data. In this study we report a promising accuracy performance of 90.03% for the model of Balanced Bagging with Expectation Maximization Imputation on a dataset of more than 20,000 Romanian companies.

Keywords: *bankruptcy, machine learning, classification*

JEL classification: *C58, G33, M10*

1. Introduction

Bankruptcy, *failure* or *default* all refer to a difficult financial position a company is facing, Levratto (2013) describing a defaulted company as "unable to be profitable or whose capital does not produce value". The effect of large bankruptcies can be devastating to economies at large (Alaka et al., 2018). Bankruptcy prediction on the other hand is of great importance to multiple parties such as lending institutions, policy makers or investors in helping predict the risk of bankruptcy and allow companies time to reorganize.

As the benefit of predicting bankruptcy is significant, during the last 50 years there were a few studies developing or testing methodologies on this

[1] *Ph.D. Student, Doctoral School of Economics and Business Administration, "Alexandru Ioan Cuza" University of Iasi, e-mail: c.clement@gausian.com*

topic. In fact, methods of predicting bankruptcy have been implemented since the end of the 19th century. The focus has been around two methodological approaches: (1) bankruptcy prediction models (BPM) using statistical models (i.e. Multiple Discriminant Analysis (Altman, 1968)); (2) BPM using intelligent techniques (Neural Networks, Ensemble Methods, Support Vector Machine etc.).

In fact, Artificial Neural Networks, Support Vector Machine and Ensemble models are the ones that shown to be the most accurate in BPM.

With a constant goal of achieving good performance, recent studies focused on feature selection techniques as well (such as PCA) (Alaka et al., 2018).

In this paper, we introduce a new approach for BPM using an ensemble method Balanced Bagging and testing it on an empirical study using Mean data imputation and Expectation Maximization to impute for missing data. Additionally, the model is tested on a dataset about Romanian companies. Because the split between bankrupt and non-bankrupt companies was not exactly 50-50, this study applied SMOTE (Synthetic Minority Oversampling Technique) to equally balance the dataset. To compare our proposed method, we also employed Logistic Regression and Decision Trees models.

The rest of the paper is organized as follows. In section 2, we present previous work and opportunity for our study. In section 3, we detail the methodology used and rationale of using it. In section 4, we explain the empirical study and show the results. In section 5, we summarize, and we present the conclusion.

2. Literature Review

Considering that the methodology and the various measures proposed for bankruptcy prediction have a quite long existence (first major study being published in FitzPatrick (1932), a good amount of studies have been published covering both statistical and intelligent techniques. We propose a division of the historical evolution of BPM in three parts. First, from 1930 to 1970, this initial stage has seen focus from researchers mostly on Multiple Discriminant Analysis (FitzPatrick (1932), Smith & Winakor (1935), Beaver (1967), Altman (1968)) as it did not require significant computing power. Second, the 1980s constitute the period when the logistic regression started to win more ground in the studies of Ohlson (1980) and Zmijewsk, (1984). Decision trees also started to pick up in this time frame (Quinlan (1986) or Breiman et al.

(1984)). Finally, from 1990s to present, BMP research started to get more and more attention and due to access to computing power, the usage of intelligent methods started. Methods such as Neural Networks, Support Vector Machine and Ensemble Methods were used extensively. Some of the studies include Martín-del-Brío & Serrano-Cinca (1993), Min & Lee (2005), Zoričák et al. (2020), Tsai et al. (2014), Kim et al. (2021) or M. Smith & Alvarez (2021).

On the Romanian context, Brîndescu-Olariu (2016) notes that although there are several studies on Romanian BPMs these either used isolated samples or had deficiencies in their statistical methodology hence not being useful for general applicability. With this in mind, our approach of testing Balanced Bagging on a Romanian dataset is novel.

3. Methodology

The classification task of bankruptcy prediction has received great attention in the last more than 30 years, consequently the amount of studies employed is so large that is virtually impossible to test them all in one study (Alaka et al., 2018). In this study we tested Balanced Bagging and compared its results with two of the most used methods in the literature, Logistic Regression and Decision Trees.

Logistic Regression (Berkson, 1944) uses a logistic sigmoid function to model the dependent variable. This method is well known to be effective in classification tasks such as the one of bankruptcy prediction.

As a simple, explainable, and very powerful method, we decided to also test Decision Trees. This model proved to be reliable in both classification and regression tasks. In our study we employed an optimized version of Classification and Regression Tree (CART), a model proposed by Breiman et al. (1984).

Bagging is the short version of Bootstrap Aggregation that was proposed by Breiman (1996) and was designed to provide more stability and better accuracy while helping to avoid overfitting. In Bagging most often Decision Trees are used for classification tasks however we used Random Forests with "Entropy" as the determinant for how the decision trees split data. As Decision Trees are known for being sensitive to the training set by balancing the dataset before training the prediction accuracy is improved (Breiman, 1996). This balancing prevents the algorithm from focusing on the majority class which would decrease accuracy and tend to overfit. We used 5 estimators (number of trees in the random forest) in our study.

Kang (2013) noted that almost all datasets have partly missing data and that this can reduce statistical power and results in biased estimates thus drawing the wrong conclusions. Gold & Bentler (2000) noted that "*Incomplete data may be the only certainty in empirical research.".* To address the issue of missing data we employed two imputation methods: Mean and Expectation Maximization.

Mean Imputation is a quite straight forward approach where for all variables the mean is calculated and where there is missing data for one variable that field is replaced by the mean of the variable.

While the Mean Imputation method is fairly intuitive, we employed a more advanced method: Expectation Maximization Imputation (EM). Proposed by Jamshidian & Bentler (1999), this method obtains the maximum likelihood estimates of missing data by cycling iteratively through two steps (E) step and (M) step. In the (E) step based on the observed data the log-likelihood is calculated and then passed into the (M) step where it is maximized to obtain the parameter estimates. In our model we set the loop number to 50 thus the method cycles 50 times to find each parameter estimate.

To deal with the class imbalance we used a common approach of over-resampling strategy SMOTE (Synthetic Minority Over-Sampling Technique). Introduced by Chawla et al. (2002) the idea is to introduce synthetic examples by joining any/all the *k* minority class nearest neighbour (Zhou, 2013). Only the minority class (bankrupt companies) was over-sampled.

4. Empirical study

To develop this research, we collected financial statement data for a total of more than 20,000 Romanian companies. The process of selecting the companies consisted of randomly selecting 29,298 bankrupt companies from the database of bankrupt companies in Romania (Buletinul Procedurilor de Insolventa) and 15,000 healthy companies from the database of the Romanian Ministry of Finance. Due to totally incomplete financial reports, we excluded several companies (3,523 bankrupt and 9,225 non-bankrupt) hence the final database is presented in Table 1 below.

Table 1: Final database

Years before bankruptcy	Number of healthy companies	Number of bankrupt companies	Total	Percentage of minority class
1-Year	13067	8052	21119	38.12%
2-Years	11962	9079	21041	43.14%
3-Years	10746	9805	20551	47.71%

We did not take into account industries or regions in Romania.

A total of 23 features included in this study are presented in Table 2 below. Features X1 to X13 are included in the financial statements for Romanian companies while X14 to X23 are financial ratios calculated and suggested by the study of Zięba et al. (2016).

Table 2: Set of features considered for classification

ID	Description	ID	Description
X1	Revenue	**X13**	Expenses
X2	Net Profit	**X14**	Net profit / Total Assets
X3	Number of Employees	**X15**	Liabilities / Total Assets
X4	Fixed Assets	**X16**	Stockholder's Equity / Total Assets
X5	Current Assets	**X17**	Expenses / Revenue
X6	Inventory	**X18**	Liabilities / Stockholder's Equity
X7	Cash on Hand	**X19**	Revenue / Inventory
X8	Account Receivables	**X20**	Expenses / Liabilities
X9	Shareholder's Equity	**X21**	Log(Total Assets)
X10	Social Capital	**X22**	Net profit / Revenue
X11	Liabilities	**X23**	Revenue / Total Assets
X12	Total Revenue		

The complete dataset includes financial statements from the years 2016-2019. In 2019 the bankrupt companies were declared bankrupt and thus we split the dataset in 3:

- Year1 – financial statements from 2018, one year before bankruptcy
- Year2 – financial statements from 2017, two years before bankruptcy
- Year3 – financial statements from 2016, two years before bankruptcy

The dependent variable was a binary variable (Y) coded 0 for healthy companies and 1 for bankrupt companies.

The goal of the study was to identify what is the predictive power of Balanced Boosting and compare its performance with Logistic Regression and Decision Trees.

To show the results we utilized the accuracy metric, precision, recall and F1-Score. To test the proposed imputation method, we also show the differences for every model.

For computing the models, we utilized Python libraries: sklearn, numpy, pandas, seaborn, matplotlib, fancyimpute and imblearn.

Table 3: Experimental results

Model	Accuracy (%)	Recall	Precision	F1 Score
Mean_year1_Logistic Regression	78.6447	[0.709, 0.864]	[0.839, 0.749]	[0.768, 0.802]
Mean_year2_Logistic Regression	80.3252	[0.903, 0.704]	[0.753, 0.88]	[0.821, 0.781]
Mean_year3_Logistic Regression	74.8977	[0.75, 0.748]	[0.749, 0.752]	[0.748, 0.749]
EM_year1_Logistic Regression	82.7925	[0.846, 0.81]	[0.817, 0.84]	[0.831, 0.825]
EM_year2_Logistic Regression	76.0659	[0.931, 0.591]	[0.695, 0.895]	[0.795, 0.711]
EM_year3_Logistic Regression	72.7667	[0.8, 0.656]	[0.699, 0.767]	[0.746, 0.707]
Mean_year1_Decision Tree	85.3371	[0.847, 0.86]	[0.858, 0.849]	[0.852, 0.854]
Mean_year2_Decision Tree	78.8372	[0.783, 0.794]	[0.792, 0.785]	[0.787, 0.789]
Mean_year3_Decision Tree	74.1997	[0.733, 0.751]	[0.747, 0.738]	[0.74, 0.744]
EM_year1_Decision Tree	84.8511	[0.843, 0.854]	[0.853, 0.844]	[0.848, 0.849]
EM_year2_Decision Tree	78.8539	[0.783, 0.794]	[0.792, 0.785]	[0.787, 0.79]
EM_year3_Decision Tree	73.5344	[0.735, 0.736]	[0.736, 0.735]	[0.735, 0.736]
Mean_year1_Balanced Bagging	89.9824	[0.923, 0.877]	[0.882, 0.919]	[0.902, 0.897]
Mean_year2_Balanced Bagging	85.4372	[0.89, 0.819]	[0.831, 0.881]	[0.859, 0.849]

Mean_year3_Balanced Bagging	81.4489	[0.855, 0.774]	[0.791, 0.842]	[0.822, 0.807]
EM_year1_Balanced Bagging	90.036	[0.922, 0.879]	[0.884, 0.918]	[0.902, 0.898]
EM_year2_Balanced Bagging	85.3954	[0.891, 0.817]	[0.83, 0.882]	[0.859, 0.848]
EM_year3_Balanced Bagging	81.3559	[0.854, 0.773]	[0.79, 0.842]	[0.821, 0.806]

Results of the experiments are presented in Table 3. For each of the considered method we present the four performance metrics and highlighting in grey the best model.

As we can see in Table 3, the best model is Balanced Bagging with Expectation Maximization with an accuracy of 90.03% while the second best is the same algorithm but with Mean Imputation (89.98%). Moving from 3 years before bankruptcy to 1 year both Balanced Boosting and Decision Trees are improving the accuracy performance showing that the models are able to better identify the differences between bankrupt and non-bankrupt companies. However, Logistic Regression shows a better performance 2 years before bankruptcy than it does for 1 year before bankruptcy. With an accuracy metric of 80.32% for the 2-year-before financial statements, the prediction power of Logistic Regression is still lower than that of Decision Trees and Balanced Bagging.

Previous studies' results for the same methods (Table 4) show that the accuracy achieved for our model on the Romanian dataset is in line with the literature and that financial statements data and financial ratios can help in prediction of bankruptcy for Romanian companies.

Table 4: Past studies results on Logistic Regression and Decision Tree

Previous Study	**Method**	**Accuracy (%)**
Tseng & Hu (2010)	Logistic Regression	86.25
Cho et al. (2010)	Logistic Regression	72.2
(Divsalar & Roodsaz, 2011)	Logistic Regression	76.47
(Zhou, 2013)	Logistic Regression	73.99
	Decision Tree	50.67
(Tsai et al., 2014)	Logistic Regression	87.28
	Decision Tree	86.83
(Wang et al., 2014)	Logistic Regression	73.90
	Decision Tree	75.99

5. Conclusions

This paper investigates the effect of two imputation techniques on the performance of two commonly used models in BPM and a novel approach of Balanced Bagging. Each imputation-model pair is tested on a dataset of more than 20,000 companies with financial statements 1, 2 and 3 years before bankruptcy. The experimental results show that the Expectation Maximization Imputation method employed with the Balanced Bagging model provide the best results. The achieved accuracy metric of 90.03% 1 year before bankruptcy is showing to be better than that of Logistic Regression and Decision Trees, opening the field for more research on the Romanian companies.

While the methods employed show good results on the Romanian dataset, further research should test other models such as Xtreme Gradient Boosting, Support Vector Machine or Neural Networks methods. Moreover, other Imputation Methods, Over-Sampling Techniques and Feature Engineering should be employed to have an overall picture of the best model on the Romanian data particularities.

6. References

- Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., & Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, *94*, 164–184. https://doi.org/10.1016/j.eswa.2017.10.040
- Altman, E. I. (1968). Financial ratios, Discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *XXIII*.
- Beaver, W. H. (1967). Financial Ratios as Predictors. *Journal of Accounting Research*, *4*(1966), 71–111. https://doi.org/https://doi.org/10.2307/2490171
- Berkson, J. (1944). Application to the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, *39*(227), 357. https://doi.org/10.2307/2280041
- Breiman, L. (1996). Bagging predictors. *Machine Learning 1996 24:2*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. *Classification and Regression Trees*, 1–358.

https://doi.org/10.1201/9781315139470/CLASSIFICATION-REGRESSION-TREES-LEO-BREIMAN-JEROME-FRIEDMAN-RICHARD-OLSHEN-CHARLES-STONE

- Brîndescu -Olariu, D. (2016). MULTIVARIATE MODEL FOR CORPORATE BANKRUPTCY PREDICTION IN ROMANIA. *Network Intelligence Studies*, *IV*(1).
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/JAIR.953
- Cho, S., Hong, H., & Ha, B. C. (2010). 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*, *37*(4), 3482–3488. https://doi.org/10.1016/j.eswa.2009.10.040
- Divsalar, M., & Roodsaz, H. (2011). *A Robust Data-Mining Approach to Bankruptcy Prediction*. *523*(March 2011), 504–523.
- FitzPatrick, P. J. (1932). *A comparison of the ratios of successful industrial enterprises with those of failed companies*.
- Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A Monte carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*, *7*(3), 319–355. https://doi.org/10.1207/S15328007SEM0703_1
- Jamshidian, M., & Bentler, P. M. (1999). ML estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and Behavioral Statistics*, *24*(1), 21–41. https://doi.org/10.3102/10769986024001021
- Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402. https://doi.org/10.4097/KJAE.2013.64.5.402
- Kim, H., Cho, H., & Ryu, D. (2021). Corporate Bankruptcy Prediction Using Machine Learning Methodologies with a Focus on Sequential Data. *Computational Economics*, *0123456789*. https://doi.org/10.1007/s10614-021-10126-5
- Levratto, N. (2013). From failure to corporate bankruptcy: a review. *Journal of Innovation and Entrepreneurship 2013 2:1*, *2*(1), 1–15.

https://doi.org/10.1186/2192-5372-2-20

- Martín-del-Brío, B., & Serrano-Cinca, C. (1993). Self-organizing neural networks for the analysis and representation of data: Some financial cases. *Neural Computing & Applications*, *1*(3), 193–206. https://doi.org/10.1007/BF01414948
- Min, J. H., & Lee, Y.-C. (2005). *Business Failure Prediction With Support Vector Machines And Neural Networks: A Comparative Study* *.
- Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, *18*(1), 109. https://doi.org/10.2307/2490395
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/bf00116251
- Smith, M., & Alvarez, F. (2021). Predicting Firm-Level Bankruptcy in the Spanish Economy Using Extreme Gradient Boosting. In *Computational Economics* (Vol. 2). Springer US. https://doi.org/10.1007/s10614-020-10078-2
- Smith, R. F., & Winakor, A. H. (1935). *Changes in the financial structure of unsuccessful industrial corporations, by Raymond F. Smith ... and Arthur H. Winakor ...* University of Illinois.
- Tsai, C. F., Hsu, Y. F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing Journal*, *24*, 977–984. https://doi.org/10.1016/j.asoc.2014.08.047
- Tseng, F.-M., & Hu, Y.-C. (2010). Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks. *Expert Systems with Applications*, *37*(3), 1846–1853. https://doi.org/10.1016/J.ESWA.2009.07.081
- Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, *41*(5), 2353–2361. https://doi.org/10.1016/j.eswa.2013.09.033
- Zhou, L. (2013). Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems*, *41*, 16–25. https://doi.org/10.1016/j.knosys.2012.12.007
- Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble

boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, *58*, 93–101. https://doi.org/10.1016/j.eswa.2016.04.001

- Zmijewski, M. E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, *22*, 59. https://doi.org/10.2307/2490859
- Zoričák, M., Gnip, P., Drotár, P., & Gazda, V. (2020). Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets. *Economic Modelling*, *84*, 165–176. https://doi.org/10.1016/j.econmod.2019.04.003