Predicting Corporate Bankruptcy in Poland – A Comparison of Naïve Bayes and Random Forest Algorithms

Marc Kendal and Ianne Weber

Description and Motivation of the Problem

We hope to contribute to the existing Bankruptcy Prediction literature, within Management Science, that focuses on trying to predict which companies are likely to become bankrupt in the future and which will survive. Many papers have been written, applying different machine learning algorithms to this problem including boosting [1],[5]. Echoing a study that was carried out with American companies, we utilise a recently created dataset of Polish companies [2]. The data set listed on the UCI Machine Learning Data Set Repository contains 5 years of data. For purposes of simplicity, we focus on the third-year data set, which includes 10,503 instances of company financials, containing 64 attributes (financial ratios) and a binary classification column to describe the company status, bankrupt or not bankrupt. The binary labels were measured for two years from the initial company measurement of its financial attributes over the period of 2000-2013. We compared and contrasted the performance of two algorithms, Naïve Bayes and Random Forest on this classification task.

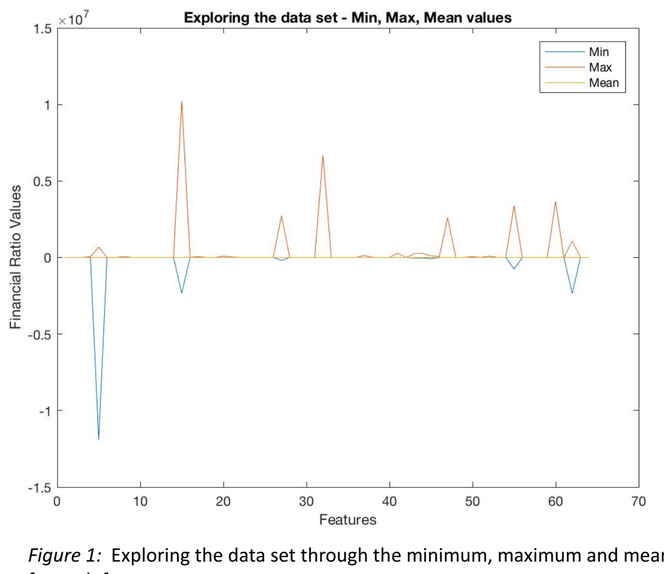

Figure 1: Exploring the data set through the minimum, maximum and mean for each feature.

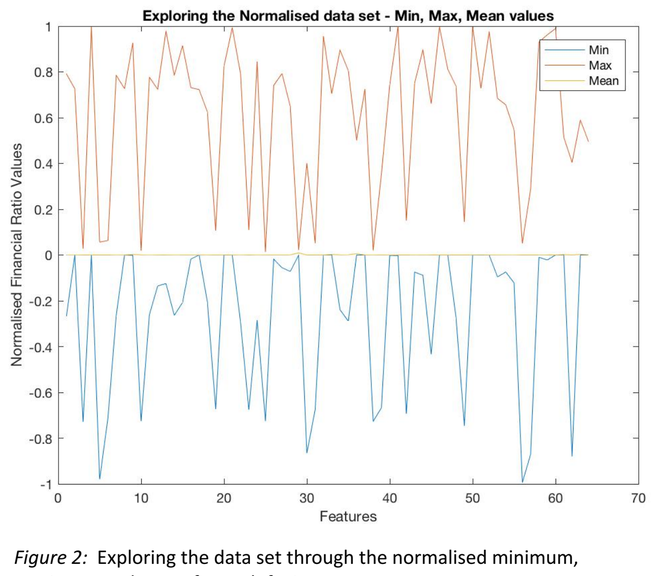

Figure 2: Exploring the data set through the normalised minimum, maximum and mean for each feature.

Initial Analysis of Data set

Each year has a different amount of bankrupt/not bankrupt companies in Poland's manufacturing sector. In our selected dataset, Year 3, there are 495 (4.7%) bankrupt companies and 10,008 (95.3%) not bankrupt companies, for a total of 10,503 evaluated companies. We performed exploratory data analysis on the 64 attributes. We looked at the minimum, maximum and mean for each feature, as shown in figure 1. We also normalised the features in order to better understand the smallest ratios in the data set, as shown in figure 2.

Hypothesis Statement

In the most recent bankruptcy prediction studies [4], Random Forest algorithms and its variant ensemble methods, outperform most machine learning algorithms for bankruptcy prediction, including Naïve Bayes. Therefore we predict that Random Forest will be a better classifier for bankruptcy than Naïve Bayes. A recent paper by Barboza et al. 2017 compares the performance of machine learning algorithms in the classification of bankruptcy of US companies. We will see if we get similar conclusions in the study of Polish companies.

Training and Evaluation Methodology

We used a holdout method of 70% for training data (7,352 data points) and 30% for testing data (3,151 data points). The evaluation method we chose was to find the accuracy and errors obtained for the training and testing data. For the Naïve Bayes (NB) we looked at the test error rate and classification accuracy, confusion matrix, AUC value and ROC curve. For the Random Forest (RF), we used the out-of-bag error for the training set to find the optimal number of trees. We also looked at the confusion matrix to see the amount of correct bankrupt/ not bankrupt predictions the models calculated, AUC value and the ROC curve the testing data.

Naïve Bayes

The Naïve Bayes classifier algorithm is based on Bayes' theorem with an assumption of independence for each attribute. It assumes that each feature in a class is unrelated to another feature. NB is considered to be a probabilistic method.

Pros: NB is often used as a classifier due to its simplicity and its strong independence assumption[3]. NB Classifiers can be used on small data sets.

Cons: When trying to use NB for many features, there can be issues with extensive computation time, overfitting, and confusion created from redundant or unrelated variables [4].

Choice of Parameters: We ran the Naïve Bayes algorithm and tested normal and kernel distributions as well uniform and empirical priors.

Experimental Results: Our results varied, as shown in figure 3. The main results concluded that NB with a kernel distribution, paired with an assumed uniform prior provides the best results.

Naïve Bayes Parameter Extensions and Accuracy Measures:

Distribution Parameters	Assumed Priors	Test Error (%)	AUC	Classification Accuracy of Bankrupt Companies (%)
Normal	Empirical	89.371	0.500	93.9
Kernel	Empirical	16.136	0.579	29
Kernel	Uniform	41.739	0.583	42.7

Figure 3: Parameter testing for Naïve Bayes.

Random Forest

The Random Forest classifier is a probabilistic method based on decision tree models. RF chooses the attributes which provide the highest information gain and takes the majority vote of multiple decision trees. Similar to bootstrapping in its methodology, however, RF avoids the correlation pitfall through randomly selecting subsets of characteristics from each node of the tree.

Pros: The RF classifier handles missing values, noise and outliers very well, and does not typically overfit the data. The RF classifier is very accurate for prediction [4].

Cons: RF are less interpretable and can be considered a black box algorithm.

Choice of Parameters: We tested the number of trees in the forest to see the optimal amount of trees, as shown in figure 4. In addition, we looked 70/30 and 90/10 holdouts for the training and testing data to see which validation method worked best for our data.

Experimental Results: The optimal number of trees were 50 trees for the training and testing data. After 50 trees, the errors increased and were consistently higher thereafter. We also saw that 70/30 holdout provided the best results. When applied to the testing data, the classification error was even lower than the training data, as shown in figure 5.

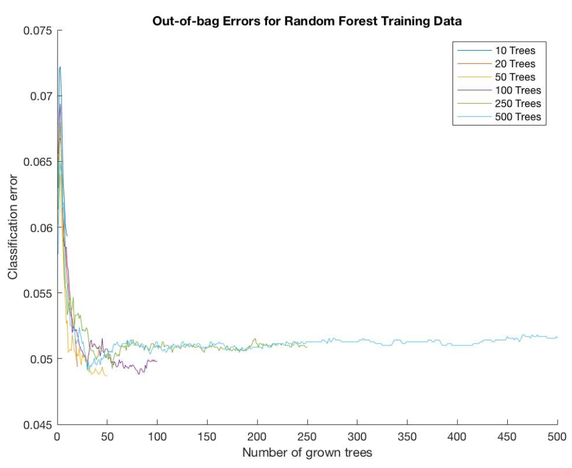

Figure 4: Parameter testing for number of trees to include in the RF.

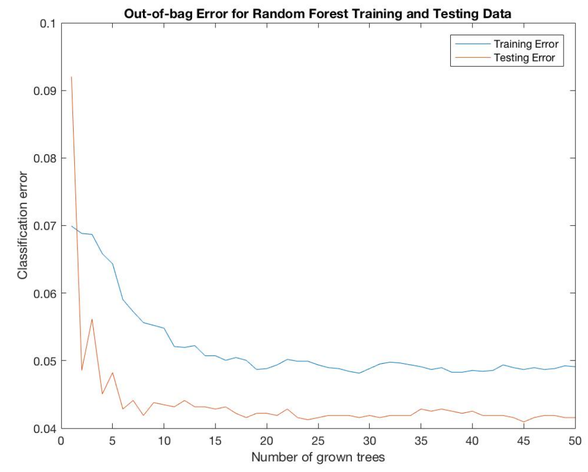

Figure 5: RF Classification error for training and testing data

Analysis and Critical Evaluation of Results

The accuracy results from the Naïve Bayes algorithms are varied. Whilst assuming a normal distribution parameter, the classification accuracy of bankruptcy was high, but the AUC was poor and there was a high test error. However utilising a kernel distribution, AUC accuracy increased and the test error decreased but the the classification accuracy of bankruptcy was significantly decreased. Changing the prior distribution to uniform increased the AUC value slightly and increased the classification accuracy of bankrupt companies whereas the test error increased. Since bankruptcy class 1 classification is what we were interested in the most, we suggest the kernel distribution – empirical model gives the best results.

When looking at the results from the RF, the accuracy for the testing set is 95.7% and the ROC curve has an AUC value of 84.6%. It would seem that the model is good at predicting bankrupt/ not bankrupt companies. Though when you look deeper at the results, the confusion matrix in figure 7 shows a different story. You can see that the model does an great job at predicting not bankrupt companies (99.7%), however, when it comes to predicting bankrupt companies, it is far from ideal with an accuracy of 6.9%. Since there are so many more not bankrupt companies, the accuracy gets skewed towards showing the accuracy of the not bankrupt. Ultimately, a model that is good at predicting not bankruptcy is not our goal. Our objective is to predict companies that will become bankrupt.

While both NB and RF did not perform well for our data set, we would say that the Random Forest algorithm would be slightly better to utilise in this scenario, due to its higher accuracy of not bankrupt classification and stronger AUC indicator. The ROC curve shows this in figure 8.

Figure 6: Confusion Matrix for Naïve Bayes

Figure 7: Confusion Matrix for Random Forest

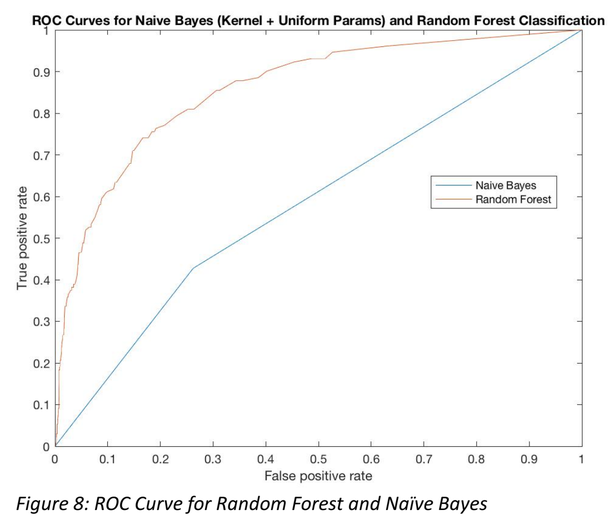

Figure 8: ROC Curve for Random Forest and Naïve Bayes

Lessons Learned and Future Work

While a model can obtain great accuracy measures such as a high AUC and low Test Error, it is important to keep one's objectives and original motivations in mind. Specifically, in our case, we were interested to accurately predict which companies would become bankrupt within the time frame recorded in our data set. In contrast, predicting which companies would survive was of less importance. We did not find these models to perform well for our problem.

For future work, we suggest researchers utilise feature selection[6], as Naïve Bayes assumes feature independence, which is not expected to be true in a financial ratios data set. For the Random Forest, the feature importance errors show that some features are consistently more or less important for the model, as shown in figure 9.

Additionally, the initial data set is partitioned into five smaller data sets, each corresponding to the length of time to bankruptcy for those companies that did go bankrupt. It would be interesting to apply these models and others such as an SVM on each data set to see which model was more accurate, and for which data sets. Lastly, Markov Models could be used understand and model this data over time.

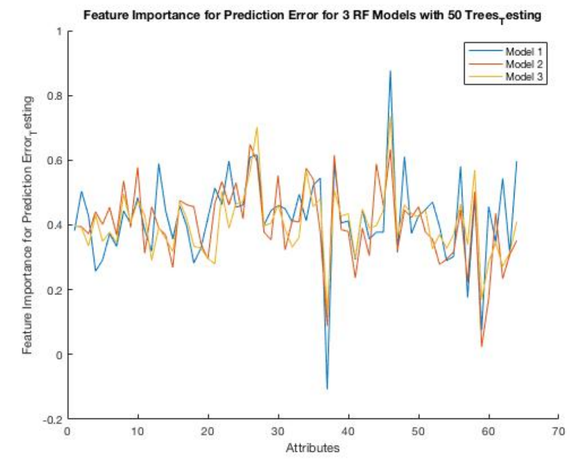

Figure 9: Feature Importance for Prediction Error in Random Forest

References:

[1] M. Zięba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Syst. Appl.*, vol. 58, pp. 93–101, Oct. 2016.
[2] S. Bakar and M. Z. A. Hamid, "Validating a Bankruptcy Prediction by Using Naïve Bayesian Network Model: A case from Malaysian Firms."
[3] L. Sun and P. P. Shenoy, "Using Bayesian networks for bankruptcy prediction: Some methodological issues," *Eur. J. Oper. Res.*, vol. 180, no. 2, pp. 738–753, Jul. 2007.
[4] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," *Expert Syst. Appl.*, vol. 83, pp. 405–417, Apr. 2017.
[5] G. Wang, J. Ma, and S. Yang, "An improved boosting based on feature selection for corporate bankruptcy prediction," *Expert Syst. Appl.*, vol. 41, no. 5, pp. 2353–2361, Apr. 2014.
[6] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, no. 1–4, pp. 131–156, 1997.