# XGBoost and Bagging in Application to Bankruptcy Prediction

Congyang Sun

17210980021

School of Data Science, Fudan University.

## Abstract

Corporate bankruptcy prediction research has been topical now for the better part of four decades. Timely warning of bankruptcy risk makes managers and investors able to do preventative measures. Then the study of bankruptcy provides an early warning signal and detects areas of weaknesses. Accurate bankruptcy prediction usually leads to many benefits such as cost reduction in credit analysis, better monitoring, and an increased debt collection rate. In this paper, I first discover and select the intrinsical and different factors predicting the bankruptcy best per year by XGBoost method. After the above work, these chosen features per year show us changes process in the top k important features on the timeline near bankruptcy. Additionally, the contrast of original XGBoost model and FeaturedBalancedBagging classifier model is given through calculating metric of performance such as precision,recall and fscore. Although some information being lost causes the prediction accuracy decreasing, FeaturedBalancedBagging can let us pay more attention to the most essential things for bankruptcy. Finally, the advice which is based on predicting the probability of bankruptcy analysis though FeaturedBalancedBagging classifier model combining with the features changes process on the timeline near bankruptcy will be prepared for investor to help them do wise decision.

## Introduction

The importance of predicting corporate bankruptcy matter for most academic and business communities in past century. In recent years, the much-publicized collapse of many large global corporations, including Tyco, Vivendi, Royal Ahold, HealthSouth, and in China, Etang, Wanguo Securities and Guangzhou Peugeot, has highlighted the significant economic, social and political costs associated with corporate failure. Therefore, prediction of an enterprise bankruptcy is of great essential in economic decision making.The high social and economic costs as a consequence of corporate bankruptcies have attracted attention of many researchers for better understanding of bankruptcy causes and eventually prediction of business distress.

Foreseeing the financial conditions and its future perspectives of a firm based on

various econometric measures is one of the most important mission concerned by every industry participant like investors shareholders, lawmakers, central banks, auditors and managers. This is a problem that affects the economy on a global scale. Typically, enterprises are quantified by a numerous indicators that describe their business condition that are further used to induce a mathematical model using past observations.

However, the complexity and diversity of the real business world makes such predictions very challenging. For example, the financial indicators describing the business conditions sometimes can not exactly reflect the true operation status of the firm, and as one possible extreme result, it may cause a rare, hard-to-predict bankruptcy with series of serious damages as consequence, which is known as a black swan event. To effectively identify those companies with higher financial risk based on these inaccuracy indicators is one problem which needs to overcome. Moreover, the most concerned issue here associated with bankruptcy prediction is that the historical observations used to train a model are usually influenced by imbalanced data phenomenon, because there are typically much more successful companies than the bankrupted ones. This problem has gained increasingly attention from researchers because it will cause classification models to favour the majority class over the minority class. As a consequent, the trained model tends to predict companies as successful (majority class) even when some of them are distressed firms.

Therefore, how to find features which are most indicative of financial distress and how to improve the accarucy of predicting the bankruptcy under the condition of data imbalance is the problem to be sloved. The solution foreseeing the probability of financial distress provides the approches that show the factors causing bankruptcy in the same industry by analyzing the different data provided and helps investors make decisions to judge the development of company and avoid unnecessary loss.

## Contribution

In this paper, I propose a novel method for bankruptcy prediction and provide advice for investors. There are two essential parts to improve predicting bankruptcy. In the first part of this paper, I focus on discovering and selecting the intrinsical and different factors predicting the bankruptcy best per year by XGBoost method. And in the second part, these chosen features per year show us changes process in the top k important features on the timeline near bankruptcy. Additionally, BalancedBagging classifier with these features, provides the estimate of company bankruptcy. Based on estimate of bankruptcy given by BalancedBagging classifier on the timeline near bankruptcy, the advice shows investor a company bankruptcy probability weighted by the predictor accuracy of the BalancedBagging classifier on the timeline.

## Related Works

In general, there are two research directions that are associated with the bankruptcy prediction. First, the econometric indicators describing the firm's condition are proposed by domain experts and some models combining with these indicators had put forward, ie. the generalized linear models are of special interest because estimated weights of the linear combination of economic indicators in the model can be further used to determine importance of the economic indicators. Second, artificial intelligence and machine learning have become a major research direction in the bankruptcy prediction, such as random forest, svm and neural networks.

One of the first classical studies bankruptcy prediction was the univariate data analysis proposed by Beaver(1966), which was followed by the multi-variate discriminant analysis and regressions (Ohlson, 1980). However, strict assumptions of traditional statistics such as linearity, normality, independence among predictor variables and pre-existing functional form relating to the criterion variable and the predictor variable limit their application in the real-world (Chen and Du, 2009). Since the methods of machine learning have been evolving quickly in recent years, many attempts have been made based on different techniques like support vector machines (SVM) (Shin et al., 2005), neural network (Angelini et al., 2008). The idea of the ensemble learning is to train and combine typically weak classifiers to obtain better predictive performance. First approaches but still very successful were bagging (Breiman, 1996) and boosting (Freund et al., 1996; Friedman, 2001; Zięba et al., 2014). The famous boosting method modified to optimize a Taylor expansion of the loss functions was Extreme Gradient Boosting (Chen and He, 2015a) which can take advantage of different learning paradigm.

## Methods

For bankruptcy prediction problem, it has been discussed and tried to be sovled in both academic research and competition such as Kaggle in recent years. Certainly, with the development of novel technologies, some useful methods are applied in it which include Logistic Regression, Support Vector Machine, Decision Tree and Random Forest, Nerual Network and so on. However, they have not considered time factor and not given a advice to potential customers such as investors although many models continue to improve the metric of performance for predictions. So next I will describe the various parts of the bankruptcy forecast in turn and give an effective advice for investors' reference.

At the beginning of modeling bankruptcy prediction, I first do the necessary data analysis which includes confirming the data content, visualizing the data distribution and cleaning and preprocessing the data. Then I find the big problem presented in Table 1 that the Polish companies dataset is highly skewed. Highly skewed datasets,

where the minority is heavily outnumbered by one or more classes, have proven to be a challenge while at the same time becoming increasingly common. There are some methods to combat imbalanced training data such as Random Over Sampling Technique, Synthetic Minority Oversampling Technique and Adaptive Synthetic and so on. To overcome the skew class problem, I use Synthetic Minority Oversampling Technique (SMOTE) and implement this using imbalanced-learn package of scikit-learn contrib (Lemaître et al., 2017). And to avoid information leak, it is necessary for splitting the dataset in 70% − 30% split prior to adding the generated extra data to the training set.

Table 1: the Polish companies dataset

| | Bankruptcy Compaines Number | NonBankruptcy Companies Number |
|---|---|---|
| 1-year | 271 | 6756 |
| 2-year | 400 | 9773 |
| 3-year | 495 | 10008 |
| 4-year | 515 | 9277 |
| 5-year | 410 | 5500 |

It is necessary to further select the most k-important features because some of the 64 features that are sensitive and insignificant maybe interfere with the results. And feature selection can reduce the number of features and dimensionality so as to make the model generalization ability stronger. Good feature selection can improve the performance of the model, help us understand the characteristics of the data and the underlying structure, which plays an important role in further improving the model and algorithm. There are some feature selection in Scikit-learn such as Removing features with low variance, Univariate feature selection, Lasso and Tree-based feature selection and so on. After comparing about these methods, I choose the effective feature selection based on the importance of xgboost features. The basic principle of feature importance is to calculate the importance of a feature based on the gain of the structure score, which is the sum of the occurrences of it in all trees.

XGBoost(eXtreme Gradient Boosting) is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. Therefore, for feature selection section, first step is finding the best parameters for every XGBoost model because of every year in these five years need a model to forecast bankruptcy. And based on these parameters, top k important features (k is decided by the dataset) are selected based on the importance of xgboost features. Bagging algorithm, also known as bagging algorithm proposed by Leo Breiman in 1994, is a group learning algorithm in the field of machine learning. The Bagging algorithm can be combined with other classification and regression algorithms to improve its accuracy and stability while avoiding over-fitting by reducing the variance of the results. Then, I try to use Balanced Bagging classifier for bankruptcy prediction under given the dataset with the chosen features.

Back to the problem, if there is only one year of data for a company, we can only judge whether is bankrupt or not from the model of the year, and regard the accuracy

of the model Acc_current as the probability of bankruptcy P for company as follows:

$$P = 1.$$

But if there are years of data for a company, it is worthy to consider the impact of previous years on bankruptcy forecasts. So we need to introduce the time factor $\gamma$ which reflects the effectiveness of model on the timeline near bankruptcy. Let us denote by $Acc_i$ the accuracy of the model in the i-th year near the predicted bankruptcy date and by P the probability of predicting future bankruptcy this year. For the weight of the i-th year near the predicted bankruptcy date $W_i$, I can represent it in the following form:

$$W_i = \gamma^i Acc_i,$$

Then I calculate the probability of predicting bankruptcy by weighted summation in the following form:

$$P = \frac{W_{bankruptcy}}{\sum W_j}$$

where $W_{bankruptcy}$ is the sum of the weights of the years in which company bankruptcy is predicated.

**Algorithm 1**

**Input:** *D*: Polish company dataset, *D* tgtcpy: Target company data, *N*: years
*k*: number of features selected, $\gamma$: time factor between [0,1]

**Output:** $M = \{M_1, M_2, \ldots, M_N\}$ : the set of models in the i-th year near the predicted bankruptcy date
*P*: the probability of predicting future bankruptcy this year

**For** n = 1,…,N **Do**
**Select top k important features** based on the importance of xgboost features.
**Train Predicting bankruptcy Model** such as BalancedBagging Classifier
**End**
**Calculating the probability** of predicting bankruptcy by weighted summation

**Return:** $M = \{M_1, M_2, \ldots, M_N\}$ : the set of models in the i-th year near the predicted bankruptcy date
*P*: the probability of predicting future bankruptcy this year

## Experiments

*Dataset*

In this paper, I use the Polish companies bankruptcy dataset. The DataSet hosted on UCI's Machine Learning Repository (Tomczak, 2016) and collected over 2007-2013 describes 64 features and bankruptcy status after 1~5 year of Polish companies. The features considered in the research studies are described in details in Table 2.

Table 2: 64 Features

| ID | Description | ID | Description |
|---|---|---|---|
| X1 | net profit / total assets | X33 | operating expenses / short-term liabilities |
| X2 | total liabilities / total assets | X34 | operating expenses / total liabilities |
| X3 | working capital / total assets | X35 | profit on sales / total assets |
| X4 | current assets / short-term liabilities | X36 | total sales / total assets |
| X5 | [(cash + short-term securities + receiv ables - short-term liabilities) / (operating expenses - depreciation)] * 365, | X37 | (current assets - inventories) / long-term liabilities |
| X6 | retained earnings / total assets | X38 | constant capital / total assets |
| X7 | EBIT / total assets | X39 | profit on sales / sales |
| X8 | book value of equity / total liabilities | X40 | (current assets - inventory - receivables) / short-term liabilities |
| X9 | sales / total assets | X41 | total liabilities / ((profit on operating ac tivities + depreciation) * (12/365)) |
| X10 | equity / total assets | X42 | profit on operating activities / sales |
| X11 | (gross profit + extraordinary items + fi nancial expenses) / total assets | X43 | rotation receivables + inventory turnover in days |
| X12 | gross profit / short-term liabilities | X44 | (receivables * 365) / sales |
| X13 | (gross profit + depreciation) / sales | X45 | net profit / inventory |
| X14 | (gross profit + interest) / total assets | X46 | (current assets - inventory) / short-term liabilities |
| X15 | (total liabilities * 365) / (gross profit + depreciation) | X47 | (inventory * 365) / cost of products sold |
| X16 | (gross profit + depreciation) / total liabil ities | X48 | EBITDA (profit on operating activities - depreciation) / total assets |
| X17 | total assets / total liabilities | X49 | EBITDA (profit on operating activities - depreciation) / sales |
| X18 | gross profit / total assets | X50 | current assets / total liabilities |
| X19 | gross profit / sales | X51 | short-term liabilities / total assets |
| X20 | (inventory * 365) / sales | X52 | (short-term liabilities * 365) / cost of products sold) |
| X21 | sales (n) / sales (n-1) | X53 | equity / fixed assets |
| X22 | profit on operating activities / total assets | X54 | constant capital / fixed assets |
| X23 | net profit / sales | X55 | working capital |
| X24 | gross profit (in 3 years) / total assets | X56 | (sales - cost of products sold) / sales |
| X25 | (equity - share capital) / total assets | X57 | (current assets - inventory - short-term li abilities) / (sales - gross profit - deprecia tion) |
| X26 | (net profit + depreciation) / total liabili ties | X58 | total costs /total sales |
| X27 | profit on operating activities / financial expenses | X59 | long-term liabilities / equity |

| | | | |
|---|---|---|---|
| X28 | working capital / fixed assets | X60 | sales / inventory |
| X29 | logarithm of total assets | X61 | sales / receivables |
| X30 | (total liabilities - cash) / sales | X62 | (short-term liabilities *365) / sales |
| X31 | (gross profit + interest) / sales | X63 | sales / short-term liabilities |
| X32 | (current liabilities * 365) / cost of products sold | X64 | sales / fixed assets |

*Metric of Performance*

As we discussed in Methods section, accuracy is perhaps not a true measure of a model's performance due to the imbalanced nature of training data. we should focus on precision, recall and Fscore that can give more insight into the accuracy of the model than traditional classification accuracy. Certainly, accuracy is divided into sensitivity and specificity and models can be chosen based on the balance thresholds of these values so that Receiver Operating Characteristic (ROC) curve visualize the performance of a binary classifier better (Markham, 2014).

- Precision: A measure of a classifiers exactness.
- Recall: A measure of a classifiers completeness
- F1 Score (or F-score): A weighted average of precision and recall.

*Result*

Choosen features per year selected by XGBoost are presented in Table 3. From Table 3, we can discover that Attr21, Attr34, Attr29, Attr27, Attr24, Attr37, Attr46, Attr15, Attr6 appear in the feature set of each year and specially Attr21, Attr34, Attr29, Attr27, Attr24, Attr46 are in a more important position per year. Therefore, changes process of top k important features features on the timeline near bankruptcy give us basic analysis and feeling for what is connected with bankruptcy prediction.

Table 4 compares the results obtained by Featured-BalancedBagging technique with the XGBoost. It can be observed that some information being lost causes the prediction fscore decreasing in all the datasets but Featured-BalancedBagging can let us pay more attention to the most essential things for bankruptcy.

Table 3: Ranking of features for each of the datasets

| Rank | 1stYear | 2ndYear | 3rdYear | 4thYear | 5thYear |
|---|---|---|---|---|---|
| 1 | Attr21 | Attr27 | Attr5 | Attr24 | Attr27 |
| 2 | Attr34 | Attr29 | Attr24 | Attr46 | Attr29 |
| 3 | Attr29 | Attr46 | Attr6 | Attr29 | Attr6 |
| 4 | Attr27 | Attr34 | Attr46 | Attr27 | Attr34 |
| 5 | Attr24 | Attr6 | Attr27 | Attr34 | Attr37 |
| 6 | Attr37 | Attr21 | Attr29 | Attr21 | Attr21 |
| 7 | Attr46 | Attr9 | Attr34 | Attr6 | Attr46 |
| 8 | Attr5 | Attr24 | Attr9 | Attr37 | Attr9 |
| 9 | Attr15 | Attr58 | Attr59 | Attr61 | Attr35 |
| 10 | Attr9 | Attr15 | Attr15 | Attr39 | Attr15 |
| 11 | Attr6 | Attr37 | Attr58 | Attr60 | Attr24 |

| | | | | | |
|---|---|---|---|---|---|
| 12 | Attr1 | Attr5 | Attr21 | Attr58 | Attr60 |
| 13 | Attr2 | Attr60 | Attr37 | Attr15 | Attr58 |
| 14 | Attr11 | Attr25 | Attr56 | Attr5 | Attr41 |
| 15 | Attr58 | Attr1 | Attr39 | Attr35 | Attr61 |
| 16 | Attr3 | Attr61 | Attr41 | Attr1 | Attr20 |

Table 4:Fscore for each of the datasets

| Fscore | 1stYear | 2ndYear | 3rdYear | 4thYear | 5thYear |
|---|---|---|---|---|---|
| XGBoost | 0.58 | 0.61 | 0.49 | 0.57 | 0.70 |
| Featured-BalancedBagging | 0.52 | 0.53 | 0.41 | 0.48 | 0.60 |

*Example*

And there is an example that show us how to utilize the time factor and model to get the bankruptcy label and probability. A company, in which the first three year near the predicted bankruptcy date predicting 0(Normal), 1(bankruptcy) and 1(bankruptcy) respectively, is created by its data each year being generated from raw dataset randomly. The bankruptcy label per year getted by BalancedBagging claasifier is equal with given label which proves the effectiveness of the model. When time factor $\gamma$ is 0.3, the label of bankrupcty is 0 but the probability is 0.72. It explains although the conclusion that the company will not go bankrupt in the following year is based on the model, the reliability of this conclusion is only 0.72. This is because the company data tends to be bankrupt in the past two years and the company data tends to be normal in the last year. Normally, its potential risks still need to be considered. And when time factor $\gamma$ is more than 0.65, the label of bankrupcty is 1 which indicates that the potential risk of company bankruptcy obtained from the data of the past two years is relatively large. Therefore, the investor can do wise decision based on the above analysis and advice.

## Conclusions

In this paper, a novel method with advice to predict company bankruptcy is provided for investors. The dataset collected over 2007-2013 provides the information about 64 features values and bankruptcy status after different years of different Polish companies so that the impact of company self-factor and economic environment factor on bankruptcy forecast should be ignored. Therefore, there are two parts that one is features selection based on XGBoost method and the other is bankrupcty probability calculation based on Featured-BalancedBagging classifier and time factor. Further, I will develop the method to solve the data imbalance, consider more about features selection and try some other classifier method to predict company bankruptcy better. Certainly, for very highly skewed data, Anomaly detection which detect rare events maybe a good choice for problem similar to bankruptcy prediction. And if considering the time factor, dynamic prediction is also next work.

## References


Angelini, Eliana, di Tollo, Giacomo, and Roli, Andrea. A neural network approach for credit risk evaluation. The quarterly review of economics and finance, 48(4):733–755, 2008.

Beaver, W.H., 1966. Financial ratios as predictors of failure. Journal of accounting research , 71–111.

Breiman, L., 1996. Bagging predictors. Machine learning 24, 123–140.

Chen, Tianqi and He, Tong. Higgs boson discovery with boosted trees. In NIPS 2014 Workshop on High-energy Physics and Machine Learning, pp. 69–80, 2015.

Chen, W and Y. Du, 2009. Using Neural Networks And Data Mining Techniques For The Financial Distress Prediction Model.Expert Systems With Applications, 36: 4075-4086.

Freund, Y., Schapire, R.E., et al., 1996. Experiments with a new boosting algorithm, in: ICML, pp. 148–156. Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics , 1189–1232.

Jason, B., 2015. Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset.URLhttps://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/.

Lemaître, Guillaume, Nogueira, Fernando, and Aridas, Christos K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17):1– 5, 2017. URL http://jmlr.org/papers/v18/16-365.html.

Ohlson, J.A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 18, 109–131.

Shin, Kyung-Shik, Lee, Taik Soo, and Kim, Hyun-jung. An application of support vector machines in bankruptcy prediction model. Expert Systems with Applications, 28 (1):127–135, 2005.

Tomczak, Sebastian. Polish companies bankruptcy data Data Set, 2016. URL https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data.

Zięba, M., Tomczak, J.M., Lubicz, M., Świątek, J., 2014. Boosted svm for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Applied Soft Computing 14, 99–108.

Zikeba, Maciej, Tomczak, Sebastian K, and Tomczak, Jakub M. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications, 2016.