BANKRUPTCY PREDICTION SYSTEM

Haoyuan Zhu, Akshay Rita, Priyanka Mulagandla

Abstract

Bankruptcy is a legal status of a person or organization that cannot repay debts to creditors. For many corporations, assessing the credit of investment targets and the possibility of bankruptcy is a vital issue before investment. Data mining and machine learning techniques have been applied to solve the bankruptcy prediction and credit scoring problems. We aim at analyzing the possibility of organizations going bankrupt from a given dataset. We also aim to find out the major attributes that contribute towards an organization going bankrupt.

The dataset contains financial ratios and a bankruptcy indicator for approximately 7000 Taiwanese companies. The data was taken from the Taiwan Economic Journal for the years 1999–2009. As per the report, company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange. Company industries include manufacturing, shipping, tourism, retail, and others, but not financial.

The dataset has 95 feature columns which are numerical financial ratios and one binary categorical column (Liability-Assets Flag) with values 1 (bankrupted) and 0 (failed to go bankrupt). Our goal is to design a Bankruptcy prediction system that gives the best accuracy using the top 10 correlated features of the data

Introduction

Bankruptcy is a legal status of a person or organization that cannot repay debts to creditors. In most jurisdictions, bankruptcy is imposed by a court order, often initiated by the debtor. Bankruptcy offers an individual or business a chance to start fresh by forgiving debts that cannot be paid, while offering creditors a chance to obtain some measure of repayment based on the individual's or business' assets available for liquidation. Upon the successful completion of bankruptcy proceedings, the debtor is relieved of the debt obligations incurred prior to filing for bankruptcy.

Bankruptcy prediction is the art of predicting bankruptcy and various measures of financial distress of public firms. It is a vast area of finance and accounting research and is important due in part to the relevance for creditors and investors in evaluating the likelihood that a firm may go bankrupt. The history of bankruptcy prediction includes application of numerous statistical tools which gradually became available, and involves deepening appreciation of various pitfalls in early analyses. The area is well-suited for testing of increasingly sophisticated, data-intensive forecasting approaches. The business problem we are trying to solve is to help bank lenders by creating a predictive model for which companies will go bankrupt. The greatest risk and cost to the lender would be lended to a company that eventually goes bankrupt and thus losing all principle and interest that company would owe.

AI methods have been developed for a long time in the bankruptcy prediction field. They are used to build models to evaluate whether organisations face financial distress. The financial variables such as financial ratios, extracted from public financial statements, contain a large amount of information about a company’s financial status which may be a factor to cause bankruptcy [1]. Those financial data and other related information from the organisation operational details can be used to establish an effective model.

Along with the development of AI and database technology, data mining techniques are gradually applied in various domains to predict bankruptcy. They can be used as “early warning systems” which are very useful to managers, authorities, Banks etc. that can take actions to prevent business failures. They are very useful when making decisions about a merger of a distressed firm, liquidation or reorganization and associated costs. These systems can help decision makers of financial institutions to evaluate and select firms to collaborate with or to invest in. Such decisions have to take into account the opportunity cost and the risk of failures [2].

Literature review

In 1966, Beaver focused on a single ratio and used various ratios to measure a company's capabilities. His study provides investors with 30 ratios that illustrate key relationships [4]. Altman's Z-score model, which was proposed in 1968, was first used for bankruptcy prediction; later, five variables were selected to generate the Z-score model [3]. Ohlson (1980) used the logit model to select data from 1970 to 1976 to analyze the four factors that affect enterprise bankruptcy. Since then, an increasing number of models have been used for bankruptcy prediction for decades [5].

Balcaen and Ooghe (2006) summarised the development of different models and found that univariate analysis does not require complex statistical knowledge because it requires strict prerequisites, and the risk index model sets a number of different weighting ratios to calculate the score [6].

Taffler and Agarwal (2007) tested the predictive ability by Z-scores and found that the Z-score model was a good choice in most cases; however, it was necessary to develop new models based on specific situations [7]. Some statistical models can then be used for bankruptcy prediction, such as univariate analysis and MDA multi-classification analysis [8].

Currently, with the wide application of machine learning models in various fields, [9–12]. The machine learning model plays an important role in bankruptcy classification. However, most machine learning algorithms for classification assume that the instances of each class are equal.

In addition to linear models, neural network-based methods have also been used to predict the bankruptcy of different firms [13, 14]. Durica(2019) used a decision tree to predict the financial situation of Polish companies; precisely, a classification and regression tree and a chi-square automatic interaction detector were used to obtain valid results [15].

There has been considerable research on the choice of financial indicators as well as the inclusion of other indicators, such as financial structure-related indicators [16]; deep learning of image data has also been used to predict bankruptcy [17]. The combination of financial ratios FRS and corporate governance metrics CGI is used in 2016 to improve the predictive performance of SVM, KNN, NB, CART, and MLP [18].

Bankruptcy prediction models are more and more combining with a machine learning model based on the other fields, such as using the financial ratio to improve the prediction ability [18]. However, few studies focus on the combination of the undersampling method.

Data source

The data was taken from the Taiwan Economic Journal for the years 1999–2009. Per the report, Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange. Company industries include manufacturing, shipping, tourism, retail, and others, but not financial. It contains 6819 enterprises, 96 attributes, two categories. Fig. 1 below illustrates the huge imbalance with 96.774% non-bankruptcy enterprises and 3.226% bankruptcy enterprises. Bankruptcy and non-bankruptcy firms are marked as '1' and '0' separately. The data source is from https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction

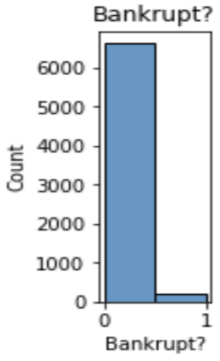


Fig 1. Class distribution. As can be seen from the graph, there is a significant imbalance in the data, with 97% majority and 3.226% minority.

Data preprocessing

Checking for missing values before the study is necessary to ensure the reliability of the data. There are no missing values. In the below Heatmap, all values of the feature Net Income Flag is 1. Therefore, we drop the column Net Income Flag.

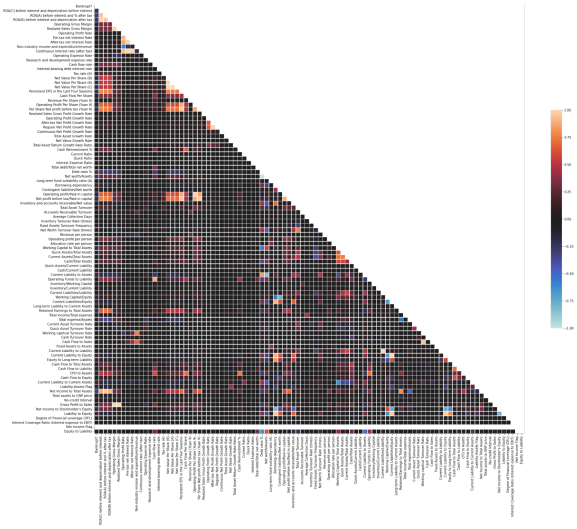


Fig 2. Heatmap

We use Box Plots to remove outliers.A Box Plot is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile and maximum. We used box plots to understand data distribution and identify outliers. We identify the actual indexes of the outlying observations using Tukey's box plot method. Then remove elements that are outside the inner and outer fence.

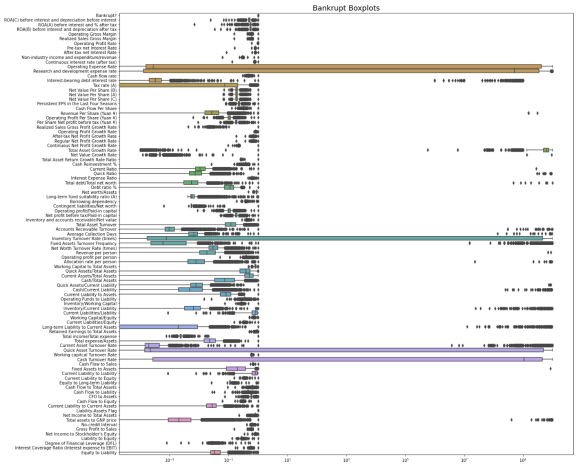


Fig 3. Bankrupt Boxplots

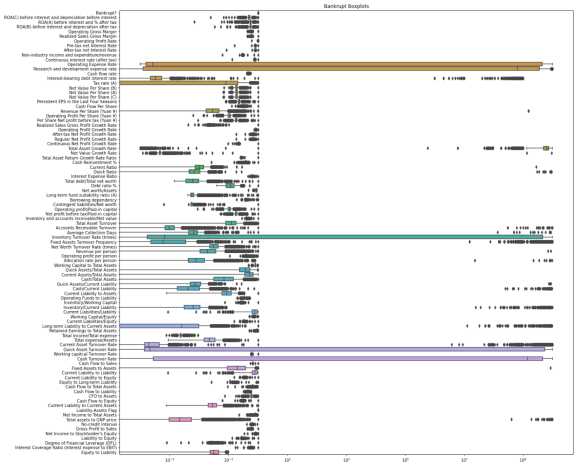


Fig 4. After removing outliers

Data Modeling

Logistic Regression:

Logistic regression is a linear model for classification. It is also known as logit regression, maximum-entropy classification (MaxEnt) or the log linear classifier.

In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.

$$LogitClassifierCostFn(x) = ||w||_1 + C\sum_{i=1}^{n}\left(loglog(expexp\left(-y\left(X_i^T w + c\right)\right) + 1)\right)$$

Random Oversampling and Undersampling:

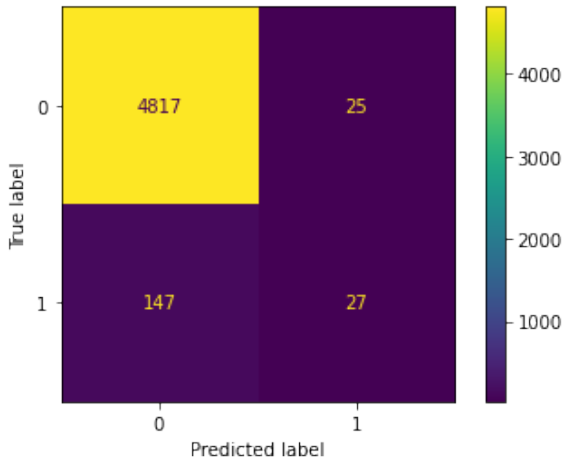
This bias in the training dataset can influence many machine learning algorithms, leading some to ignore the minority class entirely.

One approach to addressing the problem of class imbalance is to randomly resample the training dataset.

Experimental Results

Logistic Regression:

	precision	recall	f1-score	support
0	0.97	0.99	0.98	4842
1	0.52	0.16	0.24	174
accuracy			0.97	5016
macro avg	0.74	0.58	0.61	5016
weighted avg	0.95	0.97	0.96	5016



Thus, we are unable to correctly predict many of the companies. This may be due to the imbalanced data set.

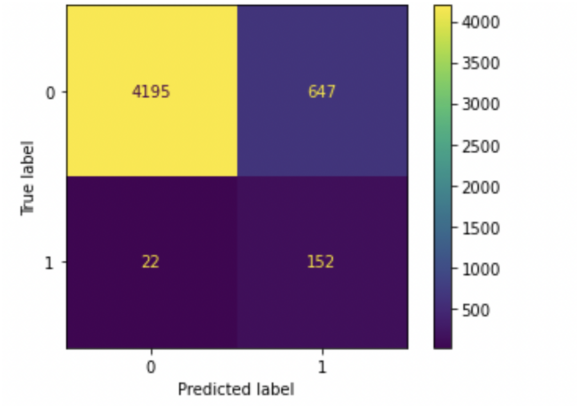
Random Oversampling and Undersampling:

```
[43] recall_score_best_over = recall_score(y_train, best_over_pipeline.predict(X_train_top_10))
     recall_score_best_over

     0.867816091954023

[44] precision_score_best_over = precision_score(y_train, best_over_pipeline.predict(X_train_top_10)
     precision_score_best_over

     0.18969849246231155
```

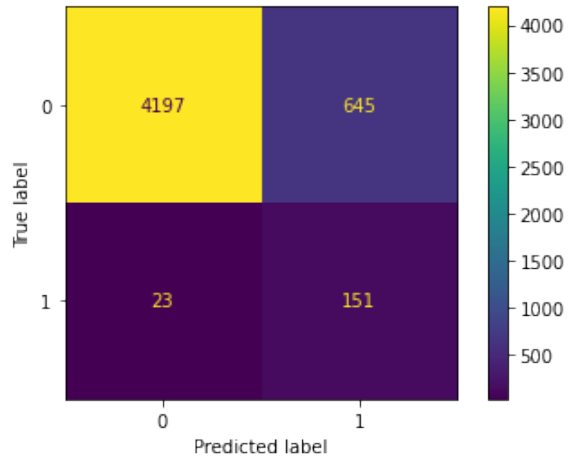



```
[45] precision_score_best_under = precision_score(y_train, best_under_pipeline.predict(X_train_top_1
     precision_score_best_under

     0.1902377972465582

[46] recall_score_best_under = recall_score(y_train, best_under_pipeline.predict(X_train_top_10))
     recall_score_best_under

     0.8735632183908046
```



We can see that our over and under sampling methods yielded higher precision scores than our basic model.

Conclusion

In our study, two models were used to analyze the Taiwanese data. First, we analysed the data using the Logistic Regression model. We are unable to correctly predict many of the companies. This may be due to the imbalanced data set. We used oversampling and undersampling methods to solve this problem. We find that our over and under sampling methods yielded higher precision scores than our basic model, Logistic Regression.

Future Work

So far we have gathered synthetic features. It is possible to generate more complex features and also reduce the dimensionality of the features. We can choose more models and compare their performances

References

[1] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, S. Wu, Credit rating analysis with support vector machines and neural networks: a market comparative study, Decision Support Systems 37 (2004) 543–558.

[2] A.I. Dimitras, S.H. Zanakis, C. Zopounidis, A survey of business failures with an emphasis on prediction methods and industrial applications, European Journal of Operational Research 90 (1996) 487–513.

[3] Altman EI. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance. (1968) 23(4):589–609.

[4] W.H. Beaver, Financial ratios as predictors of failure, Journal of Accounting Research 4 (1966) 4–71.

[5] Ohlson JA. Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research. (1980) 18(1):109.

[6] Balcaen S, Ooghe H. 35 Years of Studies on Business Failure: An Overview of the Classic Statis-

tical Methodologies and Their Related Problems. The British Accounting Review. (2006) 38(1):63–93.

[7] Agarwal V, Taffler RJ. Twenty-five Years of the Taffler Z-score Model: Does It Really Have Predictive Ability? Accounting and Business Research. (2007) 37(4):285–300.

[8] Alaminos D, del Castillo A, Fernández MÁ. A Global Model for Bankruptcy Prediction. PLOS ONE. (2016) 11(11):e0166693.

[9] Srivastava D, Kohli R, Gupta S. Implementation and Statistical Comparison of Different Edge Detection Techniques. In: Bhatia SK, Mishra KK, Tiwari S, Singh VK, editors. Advances in Computer and Computational Sciences. Advances in Intelligent Systems and Computing. Singapore: Springer; (2017) p 211–228.

[10] Javed AR, Beg MO, Asim M, Baker T, Al-Bayatti AH. AlphaLogger: Detecting Motion-Based Side-Channel Attack Using Smartphone Keystrokes. Journal of Ambient Intelligence and Humanized Computing. (2020) p. 1–14.

[11] Abbas S, Jalil Z, Javed AR, Batool I, Khan MZ, Noorwali A, et al. BCD-WERT: A Novel Approach for Breast Cancer Detection Using Whale Optimization Based Efficient Features and Extremely Randomized Tree Algorithm. PeerJ Computer Science. (2021)7:e390. pmid:33817036

[13] Chandra K, Kapoor G, Kohli R, Gupta A. Improving Software Quality Using Machine Learning. In: 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH); (2016) p. 115–118.

[13] Lahmiri S, Bekiros S. Can Machine Learning Approaches Predict Corporate Bankruptcy? Evidence from a Qualitative Experimental Design. Quantitative Finance. (2019) 19(9):1569–1577.

[14] Pozorska J, Scherer M. Company Bankruptcy Prediction with Neural Networks. In: Rutkowski L, Scherer R, Korytkowski M, Pedrycz W, Tadeusiewicz R, Zurada JM, editors. Artificial Intelligence and Soft Computing. vol. 10841. Cham: Springer International Publishing; (2018) 183–189.

[15] Durica M, Frnda J, Svabova L. Decision Tree Based Model of Business Failure Prediction for Polish Companies. Oeconomia Copernicana. (2019) 10(3):453–469.

[16] Veganzones D, Séverin E. An Investigation of Bankruptcy Prediction in Imbalanced Datasets. Decision Support Systems. (2018) 112:111–124.

[17] Hosaka T. Bankruptcy Prediction Using Imaged Financial Ratios and Convolutional Neural Networks. Expert Systems with Applications. (2019) 117:287–299.

[18] Liang D, Lu CC, Tsai CF, Shih GA. Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study. European Journal of Operational Research. ((2016) 252(2):561–572