Bankruptcy Predictive Analysis

Kalyani Nikure, Bhagyashree Uttareshwar, Manasa Pendyala
Sai Kiran Reddy Lingampally, Akshay Ragi
Department of Computing and Engineering
University of Missouri-Kansas City
Kansas City, Missouri, USA
{kmn6bg, buwfc, mmvmq, slpgk, arrmw}@umsystem.edu

ABSTRACT

Corporate bankruptcy is one of the major drivers of credit risk and is receiving primary attention from lenders and investors. The financial loss caused by corporate bankruptcy cannot be overstated. However, bankruptcy forecasting has been a topic of interest for over a century and is still one of the most talked-about topics in economics. The primary goal is to develop a strong, accurate, and successful assessment model for corporate bankruptcy prediction.

The results of a mid-term study of the Project Bankruptcy Predictive Analysis are summarized in this paper. Statistical hypothesis testing, statistical modeling, Logistic Regression, Random Forest Classification, and Gradient Boosting have all been used to evaluate various approaches in this domain. Some methods, such as Principal Component Analysis, are used to minimize the size of datasets and improve their interpretability.

The initial findings for the dataset used are discussed and presented in this paper. Although the methods used are insufficient to produce the best results, we have listed a few additional techniques that should be reviewed in order to improve accuracy and efficiency. To better form our research in this field, we will work on k-fold cross validation techniques and also explore more relevant works.

1 Introduction

Bankruptcy or business failure can adversely affect the organization and the global economy. Business practitioners, investors, governments, and academic researchers have long studied ways to identify the potential for business failure in reducing the financial loss caused by bankruptcy.

Bankruptcy forecasting is a method of assessing bankruptcy that is a different indicator of the financial difficulties of government agencies. This is a great area to study finance and accounting. The value of the field is somewhat due to the leniency of investors and investors in determining the risk of a company going bankrupt. The quantity of analysis is also a feature of the availability of data, various accounting ratios that may suggest danger can be measured for public entities that went bankrupt or did not, and numerous other possible explanatory variables are also available.[2] The field is therefore well-suited for evaluating increasingly complex, data-intensive approaches to forecasting. In short, bankruptcy assessment is a very important task for affiliated financial institutions.

The objective of predicting financial distress is the development of an assessment model that allows to predict financial condition of an organization by combining different financial measures.[1] Effective prediction of bankruptcy is crucial for making reasonable lending decisions by financial institutions. In order to determine the creditworthiness of prospective borrowers, many potential lenders use credit scoring models to help lenders select the factors that allow them to distinguish between good and bad credit risks, such as default or bankruptcy.

The first part of this paper is a review of similar studies (Ȣ2) in the field of bankruptcy prediction. The methodology for implementing predictive models (Ȣ3) is then discussed, as well as the representation of preliminary experimental findings (Ȣ4) by presenting details about the dataset used. Finally, we examine the contributions of the authors and draw a conclusion.

KEYWORDS

Bankruptcy Prediction, Gradient Boosting, Logistic Regression, Random Forest Classifier, Principal Component Analysis

2 Related work

Failure, according to Altman and Hitchhikes (2006), is described as a realized rate of return on invested capital that is significantly and consistently lower than current rates on comparable investments that take risk into account. When a company's liabilities exceed its assets, it declares insolvency. It renders a company unable to support its existing obligations, reflecting a liquidity shortage. Default happens when a company fails to meet an obligation, such as repaying a loan or appearing before a lower court.[4]

Using more corporate language, default occurs when a debtor meets a clause of a creditor's agreement, and it can result in legal proceedings. Scholars in the field of business failure prediction have historically focused their research on particular aspects or stages of the business failure process, often based on personal experiences or interests, with little or no regard for theoretical background. It leads to a jumbled analysis of the key loss, as well as fear of business failure.

Other researchers look at early warning signals that can be used to predict a company's likelihood of bankruptcy. The term "early alert" was first used in the military, but it is now commonly used in a variety of fields, including macroeconomics, business management, environmental monitoring, and so on. As a result, early warning of company collapse is a key word in the field of bankruptcy prediction research. It refers not only to the increasing number of papers released, but also to the wide range of models used to predict business failure.

An increasing number of different predictive methods have been used to develop a more reliable bankruptcy prediction model in recent years as computational techniques and information technology have progressed.

3 Methodology

This section provides a detailed view of methodological implementation. We first define techniques used in the project (Ȣ3.1) and introduce their functionality in (Ȣ3.1.1) and subsections. We then discussed Principal Component Analysis approach in (Ȣ3.2).

3.1 Techniques Used

Different models are adopted to accurately predict the bankruptcy. The right choice of models affects the accuracy which can be decided by testing various techniques.[3] We implemented below three models as of now:

3.1.1 Logistic Regression. Logistic Regression is regression analysis to conduct when the dependent variable that is output data is categorical or binary. Logistic regression is a predictive analysis used to describe data and to explain the relationship between one dependent variable and one or more independent variables. Also called a generalized linear model.

3.1.2 Random Forest Classifier. It is a type of ensemble algorithm. Instead of relying on 1 decision tree. Random Forest Classifier combines multiple decision trees from a randomly selected subset of the training set. It then combines votes from different decision trees to decide the final class of the test object.

3.1.3 Gradient Boosting. It is one of the best possible models and a very high-performing model or algorithm. It is robust, very easy to use and it improves the performance of the algorithm by reducing over-fitting. This model also minimizes the prediction error compare to logistic regression and the random forest classifier model.

3.2 Principal Component Analysis

We performed Dimensionality Reduction on dataset using Principal Component Analysis. PCA reduces the dimensionality of large data sets, by transforming a large set of variables into a smaller one. In the dataset, we cannot drop any of the data directly because we are not sure which one is helpful, and they all have same variance. But with the PCA we can predict which of the new features are generated and have high variance. In our project we calculated PCA and found that first 35 components are having higher variance, so we tested accuracy again with the reduced data.

The figure 1 depicts the variance ratios of various principal components in descending order. The higher variance of an attribute refers to the higher contribution to the prediction.[9] This result helped us to analyze the significance of various attributes in forming the prediction decision by considering only the higher variance values.

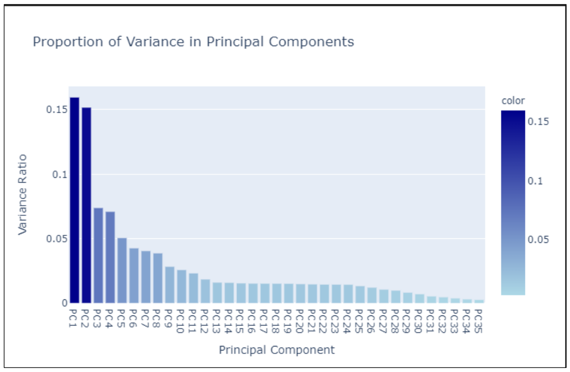

Figure 1: Proportion of Variance in Principle Components

4 Preliminary Results

The experiments were conducted to compare the model performance for the dataset containing financial information of polish companies. We have noted the results to achieve the best prediction accuracy and find the most fitting model for the dataset. The detailed information of results and dataset is discussed in following sections.

4.1 Dataset

To assess the quality of the policy, we collected data on the financial condition of Polish companies. The selection data process depends on the selection of the database, the research period, the number of companies and the number of financial indicators to be analyzed. We used data from the Emerging Markets Information Service (EMIS) [11], a global database of information on emerging markets, for this project. The data collection is about the likelihood of Polish businesses going bankrupt. The bankrupt firms were studied from 2000 to 2012, while the companies that were still active were assessed from 2007 to 2013. In CSV format, the dataset contains 43,400 Instances and 64 Attributes. [5]

4.2 Experiment Setup

The aim of the experiment was to identify the best classification model for each bankruptcy prediction case represented by the training data described in the previous subsection. The Jupyter Notebook tool was used to try and model different techniques using python language. The implementation was done in following phases as discussed:

4.2.1 Pre-Processing. In very first steps before using various techniques, we identified the null or unnamed attributes in the dataset which we cleaned as part of pre-processing activities. Another set of steps involved converting string data into float values, replacing the invalid values with zeros, and also randomly splitting the data into training and testing datasets. Finally, the scaling was also performed on the split datasets.

4.2.2 Modelling. This phase was conducted after importing all the required libraries like `pandas` to perform dataset operations, `plotly.express` to plot graphs and `sklearn` to test and train models. The `sklearn` library functions were also used to implement various models.

4.3 Results

The following classification methods were considered, and the below results were recorded:

4.3.1 Logistic Regression. Logistic regression is a form of regression analysis that involves estimating the parameters of a logistic model (a form of binary regression).

The relationship between one simple binary dependent variable and one (categorical or continuous) independent variable is analyzed using logistic regression analysis. This differs from linear regression analysis, which uses a continuous variable as the dependent variable. We were able to forecast bankruptcy data with a **95.14%** accuracy.

4.3.2 Random Forest Classifier. A random forest is a meta estimator that uses averaging to improve predictive accuracy and control over-fitting by fitting a number of decision tree classifiers on various sub-samples of the dataset. [10] The prediction accuracy of the Random Forest Classifier was **96.14 %**, which was significantly higher than the Logistic Regression.

4.3.3 Gradient Boosting. It's a greedy algorithm that can easily overfit a training dataset. It may benefit from regularization approaches, which restrict different parts of the algorithm and increase the algorithm's efficiency by reducing overfitting.[7] With a **99.94%** accuracy score, this model outperformed the other two versions and proved to be the best of the three currently that are being used.

4.3.4 Dimensionality Reduction. It is the method of reducing the dimensionality of the feature space by acquiring a collection of primary features. We were able to delete unnecessary and irrelevant features without causing significant data loss in this way. We plotted the graph using `plotly.express` library.[8]

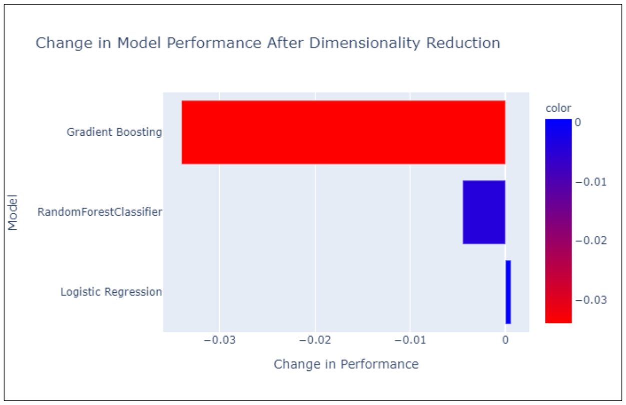

Figure 2: Change in model performance after dimensionality reduction

Figure 2 above shows the performance after dimensionality reduction. Red bar which depicts we have negative performance for Gradient Boosting. Other models have increased in performance after the Principal Component Analysis indicated by the blue bar.

4.4 Future Work

After performing PCA and comparing the results among different models, there was no significant accuracy improvement for the models. Ironically, the performance dropped for the Gradient boosting model which was not satisfactory. In order to achieve the best prediction accuracy, we target to experiment more prediction models.

We anticipate that the K-fold cross-validation test will help us to resample the methods of evaluating machine learning models in a limited data model. We will try to test the F1 score which measures the accuracy of the model in the dataset. The F1-score is a way to combine model accuracy and recall, which is defined as the harmonic mean of the model's precision and recall.

AUTHOR CONTRIBUTIONS

With reference to our Professor Dr. Uddin's review comments on the submitted project proposal, we changed our initial dataset with the new proposed dataset after his approval over email. Bhagyashree actively took up this task and communicated with the Professor.

Bhagyashree being the significant contributor for this project wrote the initial code for two of the techniques. Kalyani added additional Random Forest Classifier model to compare with the existing models already implemented. We maintained the code on GIT Repository [6] in order to have integration platform. Kalyani and Bhagyashree created the Power point document for the mid review presentation in the class.

Kalyani extensively worked on the preparation of the mid report for the project. Bhagyashree helped in the technical writeup of the document. Manasa gathered information about the Related Works and took care of the report part as well. Sai Kiran drafted for Results and Akshay did it for the Dataset section. The team is still working towards further implementation and other related research to achieve best results for the project in a collaborated fashion.

ACKNOWLEDGMENTS

A special thanks to our Professor Dr. Uddin for his valuable guidance and extensively helping us with choosing the right dataset to perform these testing experiments. He also provided a critical feedback to improve on the model's accuracy which would benefit us for the selection of right approach. He also provided many good suggestions, and identified a few research challenges along the way.

REFERENCES

[1] Zieba, M., S. K. Tomczak, and J. K. Tomczak. 2016. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. In *Proceedings of the Expert Systems with Applications Volume 58, 1 October 2016, Pages 58:93–101.* DOI: https://dl.acm.org/doi/10.1016/j.eswa.2016.04.001

[2] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai and Guan-An Shih. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive

study. In *Proceedings of the European Journal of Operational Research Volume 252, Issue 2, 16 July 2016, Pages 561-572.* DOI: https://doi.org/10.1016/j.ejor.2016.01.012

[3] Mu-Yen Chen, 2011. Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. In *Proceedings of the Computers & Mathematics with Applications Volume 62, Issue 12, December 2011, Pages 4514-4524.* DOI: https://doi.org/10.1016/j.camwa.2011.10.030

[4] Shi, Y., & Li, X. (2019). An overview of bankruptcy prediction models for corporate firms: A systematic literature review. *Intangible Capital, Volume 15, Issue 2, April 2019, Pages 114-127.* DOI: https://doi.org/10.3926/ic.1354

[5] The dataset is about bankruptcy prediction of Polish companies. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013. https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

[6] GIT Code repository URL: https://github.com/Big-data-mgmt/Bankruptcy-Prediction-Analysis

[7] XGBoost Documentation: https://xgboost.readthedocs.org/en/latest/

[8] Principal Component Analysis with Python: https://www.geeksforgeeks.org/principal-component-analysis-with-python/

[9] Principal Component Analysis: Your Tutorial and Code: https://towardsdatascience.com/principal-component-analysis-your-tutorial-and-code-9719d3d3f376

[10] Random Forest Classifier: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

[11] Emerging Markets Information Service: https://www.emis.com/