Machine learning for Bankruptcy Prediction

Abhay DA
SRN : PES1UG19CS011
Computer Science and Engineering
PES University
Bangalore, India
abhayda2001@gmail.com

Akshay SP
SRN : PES1UG19CS046
Computer Science and Engineering
PES University
Bangalore, India
Akshaysp2090@gmail.com

Akash S
SRN : PES1UG19CS042
Computer Science and Engineering
PES University
Bangalore, India
akashpari2013@gmail.com

Akshay SD
SRN : PES1UG19CS045
Computer Science and Engineering
PES University
Bangalore, India
akshaysdoddamani@gmail.com

***Abstract*—Bankruptcy prediction constitutes an important area of research. In recent years artificial intelligence and machine learning methods have achieved promising results in corporate bankruptcy prediction settings. In this report we aim to analyse how well various machine learning models fare in predicting bankruptcy of an organisation. We will compare various models and rank them based on their performance in the test data.**

***Index Terms*—Financial Distress prediction, Machine learning, Bankruptcy prediction**

I. Introduction

Bankruptcy prediction is the problem of detecting financial distress in businesses which will lead to eventual bankruptcy. Bankruptcy prediction has been studied since at least 1930s.The aim of bankruptcy prediction is to help the enterprise stakeholders to get the comprehensive information of the enterprise. Therefore, establishing a reliable enterprise failure prediction model is critical to the company.

Inaccuracy in bankruptcy forecasting can negatively impact finance and lead to a devastating blow to business owners, partners, society, and the entire national economy. Therefore, the company's internal management, the audit, and public authorities are interested in bankruptcy prediction because it affects their decision-making. Therefore, improving the bankruptcy prediction ability seems particularly important. Based on this fact above, an increasing number of scholars use models to identify bankrupt enterprises, thereby reducing the risk to investors. From the research of Altman (1974), Beaver (1966), and Ohlson (1980), The bankruptcy model has evolved over several decades. Ottman first created a multivariate statistical approach, using financial report data to classify the company.. Ohlson(1980) works for the first time to apply logic regression analysis to this field.

Business stress prediction and bankruptcy prediction have been heated-discussed topics for companies and corporations all over the world for the last few decades.

Traditionally, people heavily rely on some traditional statistical models, the assessment and judgment from relevant experts. However, nowadays, the development of novel financial indexes and the explosive growth in the volume of data have made it much harder to tackle the problem of bankruptcy prediction using those traditional approaches.

At the same time, the techniques in data mining, machine learning and deep learning have been developed and improved at a very astonishing speed. Therefore, The field where data mining algorithms and the prediction of bankruptcy are combined together has drawn more and more attention from researchers and experts of related areas. In fact, the application of these methods can help us make bankruptcy predictions and find out those companies with possibility for bankruptcy, which can prevent some possible bankruptcy, or at least help both companies and stockholders to reduce their economic losses in advance.

II. Related Work / Literature review

A. Historical statistical models

Early studies in this field concerning ratio analysis were mainly univariate studies. They focused mainly on individual ratios and compared the ratios of bankrupt companies with that of successful ones. These univariate studies had important implications for the development of models in the future as they laid the groundwork for multivariate studies.

The first study of forecasting can be written back in the early 20th century by Fitzpatrick [1932].He used an economic index to describe the ability to predict an automated business.Net Worth to Debt and Net Profits to Net Worth were the two most important ratios according to him.

Smith and Winakor [1935] analyzed the estimates of 183 failed firms from various industries. They found that the current Total Asset Rates had declined as the company approached the downturn.

In 1942, Merwin published his study focusing on small producers. He found out three important indicators of business

failure - Net Working Capital in Total Assets, Current Ratio, and Net Worth to Total Debt.

Chudson [1945] studied financial structure patterns in an effort to determine whether there was a “normal” pattern. He reported that there was no ”normal” way of making money in a normal, comprehensive economy.

Jackendoff [1962] compared the rates of profitable and unprofitable firms. These initial studies laid the foundation for subsequent studies.

Beaver took his study further and explored individual predictive skills in distinguishing ineffective and distressed firms. Beaver found that In Income to Total Debt has the highest guessing ability (92 % accuracy one year before failure), followed by Net Income to Sales (91%) and Net Income to Net Worth, Cash Flow to Total Debt, and Cash Flow to Total Assets (each with 90% accuracy).

The first multivariate study was published by Altman [1968]. Altman used several discriminatory analyzes to model five factors to predict the collapse of productive firms. Altman’s Z-score model is explained in detail in the next section.

From Altman’s study, the number and severity of the predictive models have increased dramatically.

Multivariate discriminatory analysis (MDA), logit analysis, case analysis, and neural network have been the main methods used for model development since 1968. Early models were largely constructed using MDA. MDA classifies firms into bankrupt / non-bankrupt based on certain characteristics of the company such as financial ratios / features.

B. Atman’s Z-Score model

Edward Altman developed the Z-point model - in 1968. It is a measure that is used to predict the probability of a business collapse in the next two years. It is considered an effective means of predicting the state of financial stress in an organization.

Based on the value of the Z-score, there are three areas that companies fall into - a safe area, a Gray area, a problem area. Under a safe environment the company is considered financially sound. Below the gray zone there is a good chance that the company will die within the next two years of operation. Under the pressure zone, there is a good chance the company will collapse in the near future. The range of values for each category depends on the type of the company.

The Z score model is not intended to predict when a company will actually file a formal downturn, rather it is a measure of how closely it resembles other companies that have applied for a downturn.

The last work of Altman’s (1968) discriminatory work is as follows:

$$Z = 1.2 * X_1 + 1.4 * X_2 + 3.3 * X_3 + 0.6 * X_4 + 1.0 * X_5$$

Where,
X_1 = Operating Fee / Total Asset
X_2 = Savings / Total Assets
X_3 = Earnings before interest and taxes
X_4 = Fair market value / Book full amount
X_5 = Sale of Goods / Total
Z = Z score

The original Z-Score model was based on the market value of the firm and was therefore only available to companies trading publicly. He came up with two revised works as follows:

For private owned companies:

$$Z^{'} = 0.717*X_1+0.847*X_2+3.107*X_3+0.420*X_4+0.998*X_5$$

Where,
X_1 = Operating Fee / Total Asset
X_2 = Savings / Total Assets
X_3 = Earnings before interest and taxes / Total assets
X_4 = Amount of book balance / Book amount of total credit
X_5 = Sale of Goods / Total
$Z^{'}$ = Z Score

For Non-Manufacturing Companies:

$$Z^{''} = 3.25 + 6.56 * X_1 + 3.26 * X_2 + 6.72 * X_3 + 1.05 * X_4$$

Where,
X_1 = Operating Fee / Total Asset
X_1 = Savings / Total Assets
X_1 = Earnings before interest and taxes / Total assets
X_1 = Amount of book balance / Book amount of total credit
$Z^{''}$ = Z score

C. Machine learning for bankruptcy prediction

Machine learning is a field that focuses on drawing conclusions from large amounts of data. This is done by presenting a model with samples from data so that it can find structures and relationships in the data.This process is generally called training. After training the algorithm should be able to generalise what it has learnt on new, unseen data. The field of machine learning combines computation and statistics.

With the development of various machine learning models , analysing and drawing conclusions from data has become more and more reliable. We will experiment with various machine learning models on our data to see which model gives the best result for our set of attributes.

The machine learning algorithms we are going to be using for prediction are as follows :

- Logistic regression
- Support Vector Machines
- Linear Discriminant Analysis
- Gaussian Naive Bayes
- XGBoost
- Neural Networks
- k-Nearest Neighbors
- Random forest classifier

III. Dataset

The dataset we have chosen is the data of Taiwanese companies from 1999 to 2009 taken from the Taiwan Economic Journal. It consists of 96 attributes containing various economic and financial ratios.

Link to the dataset :

https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction

IV. Problem statement

The major problem with our dataset is the problem of imbalanced data.The majority of the companies, that is 97% of the companies are not bankrupt but only 3% are bankrupt which makes the dataset inherently imbalanced in a way that becomes really hard to solve the this issue without any special methods.

The special methods adopted here are the undersampling methods which are in great use when it comes to condensing the features and taking what is really important and reducing the imbalance in the data. Though the use of oversampling is more popular it becomes a matter of simple analysis to take undersampling into account.

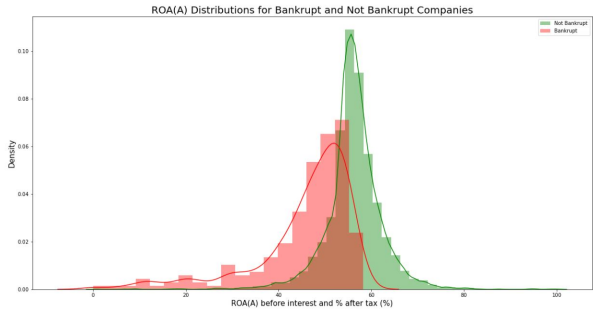
As a primitive measure we preprocess our data using the standard scaling techniques to first normalize the data in a way which becomes imperative for us to use the data keeping into account the equality fator in mind. This measure is really necessary as we don't get the appropriate influencing factors in the next step because of the unequal weights inherited by the features.

Then we use principal component analysis to get the condensed form of the components that are influencing the prediction mainly.

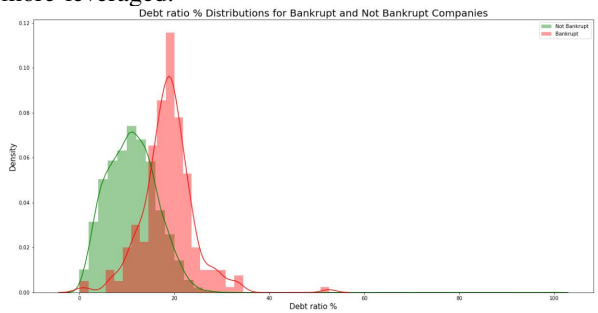
Later, we experiment with various models which gives us the best prediction metrics namely Support Vector Machine (SVM), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (NB), XGBoost (XGB), Neural Networks (NN), k-Nearest Neighbors (KNN), Random Forest (RBF)

V. Exploratory Data Analysis

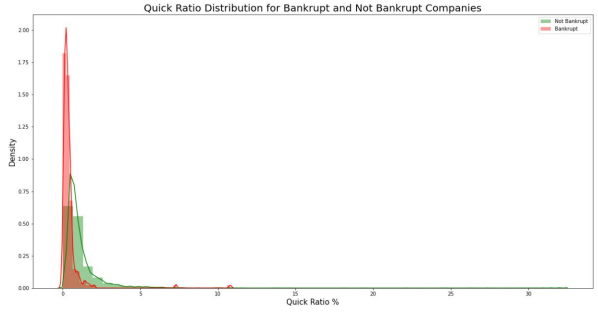
- ROA(A) stands for Return on Assets is a measure of how much profit is generated from its capital. The profitability ratio demonstrates the percentage growth rate in profits that are generated by the assets owned by the company. Return on assets tells investors how efficiently a company generates profit growth from the capital it has been granted, both debt and equity. This metric is used to compare similar companies or to determine how a firm has performed over different periods of time. From the below graph we can clearly see the differences in the distributions for bankrupt companies who have lower and more left skewed ROA(A) relative to not bankrupt companies.



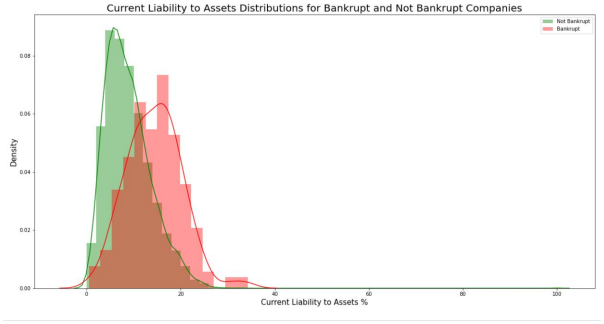
- Debt ratio % is defined as the ratio of total debt to total assets. The debt ratio measures the amount of leverage used by the company in terms of total debt to total assets. The debt ratio also depends upon which the sector the industry or the company belongs to. The higher the ratio the more leveraged the company indicated it has more financial problems.From the above graph we can see that the Bankrupt companies tend to have higher and more right skewed debt ratio % showing that they are typically more leveraged.



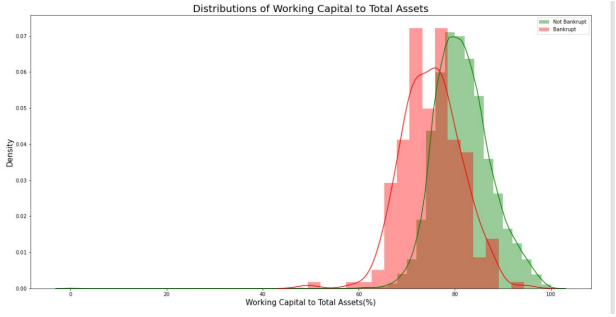
- Quick ratio is a type of liquidity ratio, which measures the ability of the company to use its near cash or quick assets to extinguish or retire its current liabilities immediately without the need to sell its inventory or obtain additional financing. Therefore it is also known as acid test. This is considered more conservative with respect to quick ratio as the latter considers all the assets. The higher the ratio, the better the company's liquidity and financial health and lower the ratio, more likely the company will struggle to pay debts. We found that 75% of companies that went bankrupt had a Quick Ratio less than .5% while only 25% of companies that didn't go bankrupt had quick ratio that low



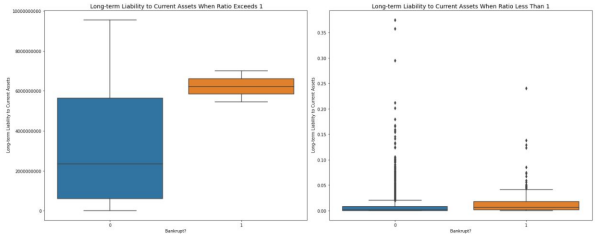
- Current liability to assets is very similar to that of the debt ratio except that current ratio is a liquidity ratio that measures the company's ability to pay short term obligations or those which are due within one year. It compares all the company's assets to its current liabilities. It helps provide an understanding for investors to know about a company's ability to resolve all its short term debt with current assets. The distribution in the distplot below again shows that bankrupt companies will typically have higher Current Liability to Assets relative to companies that didn't go bankrupt.



- Working capital to total assets ratio is meant to indicate how capable a company is of meeting its current financial obligations and is a measure of a company's basic financial solvency. The ratio is defined as current assets to current liabilities and is also known as a measure of liquidity as meaning the company's ability to meet its payment obligations as they fall due. We can see from the distributions that bankrupt companies have smaller ratios of liquid assets to total assets. This may indicate that in time of financial distress, these companies would have more difficult time offloading assets to satisfy their liabilities and avoid bankruptcy.



- Long term liabilities are the financial obligations of a company that are due for more than a year in the future. The long-term debt or liabilities are listed separately to provide a more accurate view of a company's current liquidity and the company's ability to pay current liabilities as they become due. Long-term liabilities are also called long-term debt or non current liabilities. Long Term Liability to Current Assets is also significant for predicting bankruptcy. We can see from the boxplot that bankrupt companies had higher Long-term Liabilities to Current Assets which intuitively makes sense. We saw various companies that had this ratio greater than 1 and less than 1. This relationship holds up for both segments.



VI. Acknowledgements

We would like to acknowledge our Data Analytics Course Professor Dr. Gowri Srinivasa for providing constant guidance during each phase of our project. We would also like to acknowledge our assistant professors who have prepared the course content and also the teaching assistants who have been constantly providing resources to practice the learnt concepts.

References

[1] Sun J, Li H, Huang QH, He KY. Predicting Financial Distress and Corporate Failure: A Review from the State-of-the-Art Definitions, Modeling, Sampling, and Featuring Approaches. Knowledge-Based Systems. 2014;57:41–56.

[2] Shi Y, Li X. An Overview of Bankruptcy Prediction Models for Corporate Firms: A Systematic Literature Review. Intangible Capital. 2019;15(2):114–127.

[3] Beaver WH. Financial Ratios As Predictors of Failure. Journal of Accounting Research. 1966;4:71.

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

[5] Gissel, J. (2007). A Review of Bankruptcy Prediction Studies:1930-Present, Gissel, Don Giacomino, Michael D. Akers. Journal of Financial Education, 33(Winter 2007), 1-42. The author of this document, Jodi L. Gissel, published under th ename Jodi L. Bellovary at the time of publication.