APRIL 2023

Project Progress Report

BANKRUPTCY PREDICTION

FEDERAL RESERVE NOTE
THE UNITED STATES OF AMERICA
THIS NOTE IS LEGAL TENDER FOR ALL DEBTS, PUBLIC AND PRIVATE
B 51637442 D
WASHINGTON, D.C.
SERIES 2017
ONE DOLLAR

FA-582 -Group : Tashveen Kaur, Abhinav Ganguly, Clinton Nwokike

INTRODUCTION

Background of the project

One of the primary objectives of credit risk assessment is predicting business insolvency. Particularly since the financial crisis of 2007–2008, it has elevated in importance for most financial institutions, professionals, and scholars.

Bankruptcy or corporate failure can hurt both the individual company and the global economy. Business practitioners, investors, governments, and academic academics have long sought techniques to identify the risk of business failure in order to reduce the economic losses associated with bankruptcies.

Research Question: Can we accuarately predict bankruptcy using Machine Learning?

Overview of the project

The project aims to analyze financial market data by applying data collection, data preparation, feature extraction, data cleaning, analytical processing and algorithms. Data clustering, data classification, outlier analysis, and data mining techniques may be implemented. The data can be obtained from the financial markets (stocks, financial statements, etc.) or from text data sources (twitter, news, etc.). You may implement new methods and argue the advantage of them over traditional methods

Significance of the project

Forecasting insolvency is an important task for many financial institutions. The purpose is to forecast the likelihood of a corporation going bankrupt. Effective prediction models are required by financial institutions in order to make suitable lending decisions.

Several models for predicting bankruptcy were able to be developed thanks to recent developments in machine learning (ML).

A recent glance at these bankruptcy activities is the bankruptcy of Silicon Valley Bank recently and Signature Banks. These recent bank failures led us to solve the issue that was in the mind of various investors about the prediction of these banks and companies and to create a model that could accurately predict the failure.

• The DATA

Description of the data

The data set we have chosen was collected from the Taiwan Economic Journal from 1999 to 2009. The data contains Bankrupt companies that were identified based on the business regulations of the Taiwan Stock Exchange.

Features of the data

- The data runs across different industries (electronic manufacturing, retail, shipping, tourism...) Each industry has a sufficient amount of companies in similar size in order to do the comparison
- There are 95 features (X1-X95, business regulations of Taiwan Stock Exchange) and 1 label (bankrupt or not)

Source of the data

Deron Liang and Chih-Fong Tsai, deronliang '@' gmail.com; cftsai '@' mgt.ncu.edu.tw, National Central University, Taiwan
The data was obtained from UCI Machine Learning Repository:
https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

In our data as the table below would show the data set contains various ratios like (ROA, Gross Margin, Operating Profit, etc.) of said banks and other quantitative metrics like (Net Income or Equity to Liability) that are used as variables.

Variable Name	Description	Data Type
Bankrupt.	Whether the company had gone bankrupt, in 0 and 1	Factor
ROA.C..before.interest.and.depreciation.before.interest	Return on assets, calculated as earnings before interest and taxes divided by total assets.	Numeric
ROA.A..before.interest.and...after.tax	Return on assets, calculated as earnings before interest divided by total assets, then multiplied by (1 - tax rate)	Numeric
ROA.B..before.interest.and.depreciation.after.tax	Return on assets, calculated as earnings before interest and taxes divided by total assets	Numeric
Operating.Gross.Margin	Gross margin, calculated as gross profit divided by revenue, showing the percentage of revenue left after deducting cost of goods sold	Numeric
Realized.Sales.Gross.Margin	Gross margin on actual sales, calculated as gross profit on actual sales revenue minus cost of goods sold, divided by actual sales revenue.	Numeric
Operating.Profit.Rate	Operating profit divided by revenue.	Numeric
Pre.tax.net.Interest.Rate	Pre-tax net profit divided by revenue.	Numeric
After.tax.net.Interest.Rate	After-tax net profit divided by revenue.	Numeric
Non.industry.income.and.expenditure.revenue	The proportion of non-operating income and expenses to revenue.	Numeric
Continuous.interest.rate..after.tax.	After-tax continuous profit rate.	Numeric
Operating.Expense.Rate	Operating expenses divided by revenue.	Numeric
Research.and.development.expense.rate	Research and development expenses divided by revenue.	Numeric
Cash.flow.rate	Cash flow from operating activities divided by revenue.	Numeric
Interest.bearing.debt.interest.rate	Interest expense on interest-bearing debt divided by total interest-bearing debt.	Numeric
Tax.rate..A.	Tax rate, calculated as taxes paid divided by revenue.	Numeric

Net.Value.Per.Share..B.	**Net value per share, calculated as equity attributable to shareholders of the parent company divided by total shares outstanding.**	**Numeric**
Net.Value.Per.Share..A.	**Net asset value per share, calculated as total assets minus total liabilities divided by total shares outstanding.**	**Numeric**
Net.Value.Per.Share..C.	**Net asset value per share (C), which represents the net asset value (remaining value after subtracting liabilities) divided by the total number of outstanding shares, is an indicator for evaluating a company's value.**	**Numeric**
Persistent.EPS.in.the.Last.Four.Seasons	**Earnings per share (EPS) for the last four quarters, which is the net profit per share that the company has continuously earned in the past four quarters and used to evaluate the company's profit capabilities during the economic cycle.**	**Numeric**
Cash.Flow.Per.Share	**Cash flow per share, which is the total amount of cash inflows and outflows divided by total outstanding shares, used to evaluate a company's cash flow situation.**	**Numeric**
Revenue.Per.Share..Yuan...	**Revenue per share (in Chinese yuan), which is the company's revenue divided by total outstanding shares, used to evaluate the company's revenue generating capabilities.**	**Numeric**
Operating.Profit.Per.Share..Yuan...	**Operating profit per share (in Chinese yuan), which is the profit obtained by the company after deducting operating expenses from operating revenue, divided by total outstanding shares, used to evaluate the company's operating profit.**	**Numeric**
Per.Share.Net.profit.before.tax..Yuan...	**Pre-tax net profit per share (in Chinese yuan), which is the company's net profit before deducting various expenses and taxes, divided by total outstanding shares.**	**Numeric**
Realized.Sales.Gross.Profit.Growth.Rate	**Percentage change in gross profit from sales that has been realized over a specific period.**	**Numeric**
Operating.Profit.Growth.Rate	**Percentage change in operating profit over a specific period.**	**Numeric**
After.tax.Net.Profit.Growth.Rate	**Percentage change in net profit after taxes over a specific period.**	**Numeric**
Regular.Net.Profit.Growth.Rate	**Percentage change in regular net profit over a specific period.**	**Numeric**
Continuous.Net.Profit.Growth.Rate	**Percentage change in net profit that occurs continuously over a specific period.**	**Numeric**
Total.Asset.Growth.Rate	**Percentage change in total assets over a specific period.**	**Numeric**
Net.Value.Growth.Rate	**Percentage change in net value over a specific period.**	**Numeric**

Total.Asset.Return.Growth.Rate.Ratio	Ratio of percentage change in total assets to percentage change in return over a specific period.	Numeric
Cash.Reinvestment..	The amount of cash reinvested into the business for future growth and expansio	Numeric
Current.Ratio	Ratio of current assets to current liabilities, used to assess a company's short-term liquidity and ability to pay its obligation	Numeric
Quick.Ratio	Ratio of quick assets (current assets minus inventory) to current liabilities, used to assess a company's immediate liquidity without relying on inventory.	Numeric
Interest.Expense.Ratio	Ratio of interest expense to operating income, used to assess a company's ability to cover its interest obligations.	Numeric
Total.debt.Total.net.worth	Ratio of total debt to total net worth, used to assess a company's leverage and financial risk.	Numeric
Debt.ratio..	Ratio of total debt to total assets, used to assess a company's solvency and risk of default.	Numeric
Net.worth.Assets	Ratio of net worth to total assets, used to assess a company's financial health and stability.	Numeric
Long.term.fund.suitability.ratio..A.	Ratio of long-term funds to fixed assets, used to assess a company's ability to finance its fixed assets with long-term funds.	Numeric
Borrowing.dependency	The extent to which a company relies on borrowing to finance its operations and investments.	Numeric
Contingent.liabilities.Net.worth	Ratio of contingent liabilities to net worth, used to assess a company's exposure to potential liabilities in relation to its net worth.	Numeric
Operating.profit.Paid.in.capital	Amount of paid-in capital generated from operating profit.	Numeric
Net.profit.before.tax.Paid.in.capital	Amount of paid-in capital generated from net profit before taxes.	Numeric
Inventory.and.accounts.receivable.Net.value	Net value of inventory and accounts receivable, used to assess a company's liquidity and working capital.	Numeric
Total.Asset.Turnover	atio of net sales to average total assets, used to assess a company's efficiency in generating sales from its assets.	Numeric

Accounts.Receivable.Turnover	**Ratio of net sales to average accounts receivable, used to assess a company's efficiency in collecting receivables.**	**Numeric**
Average.Collection.Days	**Number of days, on average, it takes for a company to collect its accounts receivable.**	**Numeric**
Inventory.Turnover.Rate..times.	**Ratio of cost of goods sold to average inventory, used to assess a company's efficiency in managing its inventory.**	**Numeric**
Fixed.Assets.Turnover.Frequency	**Ratio of net sales to average fixed assets, used to assess a company's efficiency in generating sales from its fixed assets.**	**Numeric**
Net.Worth.Turnover.Rate..times.	**Ratio of net sales to average net worth, used to assess a company's efficiency in generating sales from its net worth.**	**Numeric**
Revenue.per.person	**Average revenue generated per employee, indicating workforce efficiency**	**Numeric**
Operating.profit.per.person	**Average operating profit generated per employee, indicating operational efficiency**	**Numeric**
Allocation.rate.per.person	**Average allocation rate per employee, ratio of allocated costs to total costs, assessing cost allocation efficiency**	**Numeric**
Working.Capital.to.Total.Assets	**Ratio of working capital (current assets - current liabilities) to total assets, indicating liquidity and financial health**	**Numeric**
Quick.Assets.Total.Assets	**Ratio of quick assets (current assets - inventory) to total assets, measuring liquidity and ability to cover liabilities quickly**	**Numeric**
Current.Assets.Total.Assets	**Ratio of current assets to total assets, measuring liquidity and short-term solvency**	**Numeric**
Cash.Total.Assets	**Ratio of cash and cash equivalents to total assets, indicating liquidity and financial flexibility**	**Numeric**
Quick.Assets.Current.Liability	**Ratio of quick assets to current liabilities, measuring ability to cover short-term obligations**	**Numeric**
Cash.Current.Liability	**Ratio of cash and cash equivalents to current liabilities, indicating ability to cover short-term obligations with cash**	**Numeric**
Current.Liability.to.Assets	**Ratio of current liabilities to total assets, measuring short-term solvency and risk**	**Numeric**
Operating.Funds.to.Liability	**Ratio of operating funds (operating income + depreciation) to total liabilities, indicating ability to cover liabilities with operating income**	**Numeric**

Inventory.Working.Capital	**Ratio of inventory to working capital, measuring efficiency of inventory management**	**Numeric**
Inventory.Current.Liability	**Ratio of inventory to current liabilities, assessing ability to cover short-term obligations with inventory**	**Numeric**
Current.Liabilities.Liability	**Ratio of current liabilities to total liabilities, measuring short-term solvency and risk**	**Numeric**
Working.Capital.Equity	**Ratio of working capital to shareholders' equity, indicating financial health and liquidity**	**Numeric**
Current.Liabilities.Equity	**Ratio of current liabilities to shareholders' equity, assessing short-term solvency and ris**	**Numeric**
Long.term.Liability.to.Current.Assets	**Ratio of long-term liabilities to current assets, indicating ability to cover short-term obligations with long-term funds**	**Numeric**
Retained.Earnings.to.Total.Assets	**Ratio of retained earnings to total assets, measuring profitability and reinvestment in the business**	**Numeric**
Total.income.Total.expense	**Ratio of total income to total expense, assessing profitability and efficiency of expense management**	**Numeric**
Total.expense.Assets	**Ratio of total expenses to total assets, indicating cost efficiency and expense management**	**Numeric**
Current.Asset.Turnover.Rate	**Ratio of net sales to average current assets, measuring how efficiently current assets are used to generate sales**	**Numeric**
Quick.Asset.Turnover.Rate	**Ratio of net sales to average quick assets, indicating how efficiently quick assets are used to generate sales**	**Numeric**
Working.capitcal.Turnover.Rate	**Ratio of net sales to average working capital, measuring how efficiently working capital is used to generate sales**	**Numeric**
Cash.Turnover.Rate	**Ratio of net sales to average cash and cash equivalents, indicating how efficiently cash is used to generate sales**	**Numeric**
Cash.Flow.to.Sales	**atio of cash flow from operations to net sales, measuring cash generation relative to sales**	**Numeric**
Fixed.Assets.to.Assets	**Ratio of fixed assets to total assets, indicating proportion of assets invested in fixed assets**	**Numeric**

Current.Liability.to.Liability	**Ratio of current liabilities to total liabilities, measuring short-term solvency and risk**	**Numeric**
Current.Liability.to.Equity	**Ratio of current liabilities to shareholders' equity, assessing risk and leverage**	**Numeric**
Equity.to.Long.term.Liability	**Ratio of shareholders' equity to long-term liabilities, indicating leverage and financial stability**	**Numeric**
Cash.Flow.to.Total.Assets	**Ratio of cash flow from operations to total assets, measuring cash generation relative to total assets**	**Numeric**
Cash.Flow.to.Liability	**Ratio of cash flow from operations to total liabilities, assessing cash generation relative to liabilities**	**Numeric**
CFO.to.Assets	**Ratio of cash flow from operations to total assets, measuring cash generation relative to total assets**	**Numeric**
Cash.Flow.to.Equity	**Ratio of cash flow from operations to shareholders' equity, indicating cash generation relative to equity**	**Numeric**
Current.Liability.to.Current.Assets	**Ratio of current liabilities to current assets, measuring short-term solvency and liquidity**	**Numeric**
Liability.Assets.Flag	**Indicator flagging potential financial distress or risk based on the ratio of total liabilities to total assets**	**Numeric**
Net.Income.to.Total.Assets	**Ratio of net income to total assets, indicating profitability and return on assets**	**Numeric**
Total.assets.to.GNP.price	**Ratio of total assets to gross national product (GNP) price, assessing relative size and significance of assets**	**Numeric**
No.credit.Interval	**Indicator of time period without credit, measuring liquidity and ability to operate without additional credit**	**Numeric**
Gross.Profit.to.Sales	**Ratio of gross profit to net sales, indicating profitability and gross margin**	**Numeric**
Net.Income.to.Stockholder.s.Equity	**Ratio of net income to shareholders' equity, indicating profitability and return on equity**	**Numeric**
Liability.to.Equity	**Ratio of total liabilities to shareholders' equity, assessing leverage and financial risk**	**Numeric**
Degree.of.Financial.Leverage..DFL.	**Measure of financial risk, indicating the sensitivity of earnings to changes in operating income**	**Numeric**
Interest.Coverage.Ratio..Interest.expense.to.EBIT.	**Measure of ability to cover interest expenses with earnings before interest and taxes (EBIT)**	**Numeric**
Net.Income.Flag	**Indicator flagging potential financial distress or risk based on net income**	**integer**
Equity.to.Liability	**Ratio of shareholders' equity to total liabilities, measuring leverage and financial stability**	**Numeric**

PROJECT PROGRESS

Since the previous submission of our Project Proposal, we have come a long way into our project where we delved deeper into the Exploratory Data Analysis with our financial dataset.
We began by clearing the R environment and installing and loading the necessary packages for data manipulation and visualization, such as ggplot2, skimr, GGally, car, nortest, and moments. Then we set the working directory to the location where the data file was stored and we read in the data.
We then proceeded to examine the data by printing its summary statistics, displaying the first and last few rows, and showing the names of the columns.

We then checked for missing values in our dataset using the is.na() function which returns a logical vector of the same length as the input vector, indicating which elements of the vector are missing values. We used sum() function to count the number of missing values in our dataset by summing up the logical vector returned by is.na(). This gave us the total number of missing values in the data set.

Post this, we used the skim function from the skimr package to generate a more detailed summary of the data. We then created histograms for all the columns using a for loop and the hist function to visualize the distribution of data in each column. These functions helped us in understanding the overall structure of the data and the distributions of the variables. This would help us in understanding and selecting appropriate statistical analyses and models for our project.

PROJECT PROGRESS

We then used visualizations to explore the data for our project. We created histograms of each column in the dataset. Then we created a scatterplot of two specific columns in the dataset, Accounts Receivable Turnover and Quick Ratio, using the ggplot2 package in R. The visualizations allowed us to gain insights into the distribution and relationships of the variables in our dataset.

We then performed some statistical analysis to investigate the numerical variables in the dataset. First, we identified the continuous numerical variables and created histograms for each of them. Then, we visualized the summary statistics for all variables using the skim() function. Next, we created histograms for all features except for the "Bankrupt." column using the ggplot2 package in R. After that, we ran normality tests for each variable with a sample size between 3 and 5000. Following this, we calculated the skewness and kurtosis for each variable. This statistical analysis helped us to explore the distribution and properties of the numerical variables in the dataset, which is crucial for further data modeling and analysis.

Lastly, we performed logistic regression on our dataset to predict bankruptcy using selected features. We first set out the bankrupt column as our target variable in the dataset and extracted the features. Then we removed Net.Income.Flag variable as it had the same value in each row.

PROJECT PROGRESS

Then we went on to standardize all the features and performed logistic regression on our dataset using the glm function in R. We printed the summary of our model, and the overall Pseudo R-squared and its p-value were also calculated. We generated the predicted probabilities, and a plot of predicted probabilities over time was also created. We then generated predicted values and created a confusion matrix to evaluate the performance of our classification model. Finally, the overall fraction of correct predictions was calculated, which gave us an estimate of the model's accuracy.

NEXT STPES

- The "Can we accurately predict bankruptcy using Machine Learning?" project is progressing well, and the next steps involve performing further analysis to predict bankruptcy using machine learning techniques. We plan to use KNN, QDA, and LDA analysis to predict bankruptcy based on significant variables that were tested by our logistic regression. We will use different combinations of predictors, including possible transformations and interactions, for each of the methods. We will report the variables, method, and associated confusion matrix that provides the best results on the held-out data.

- Furthermore, we will compare the difference in prediction and the level of accuracy of the models. This will allow us to assess the effectiveness of each method and determine which technique works best for predicting bankruptcy.

- In addition to these techniques, we also plan to perform different machine learning techniques like "DecisionTreeClassifier," "RandomForestClassifier," "XGBClassifier," and "CatBoostClassifier." These models will be trained and tested to determine their rate of accuracy and to compare them with the previous techniques used.

- Overall, our aim is to find the most effective machine learning technique to accurately predict bankruptcy. The results obtained will provide insights into how machine learning can be utilized to predict bankruptcy and assist businesses in identifying potential financial risks.

ISSUES ENCOUNTERED

- We encountered some issues during the EDA part as the data we selected was very diverse and needed to be standardized and normalised so we needed to handle that in our EDA.
- The other issue we encountered was during regression, where we needed to calculate the p-value and check whether the data was statistically significant.

REFERENCES

Liang, D., Lu, C.-C., Tsai, C.-F., and Shih, G.-A. (2016) Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study. European Journal of Operational Research, vol. 252, no. 2, pp. 561-572. https://www.sciencedirect.com/science/article/pii/S0377221716000412