Company Bankruptcy Prediction

1. Data overview

Source

The data collected from the Taiwan Economic Journal for the years 1999 to 2009 were found on the Kaggle website. According to the source, company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange. Full list of data features can be found in ANNEX A.

Content

Raw data imported from Kaggle website include a csv table with 96 columns (95 data features and class label) and 6819 rows. Class label indicates whether a company, represented by a single row, is bankrupt (‘1’ label) or not (‘0’ label). Data features are mostly financial/accountant parameters, however some flags are also included.

Exploration plan

Downloaded csv table was turned into Pandas dataframe in order to perform data cleaning, quality checks and feature engineering. Taking into account size of the data, visual inspection of dependencies between all the features was not possible, however other actions were taken to identify redundant columns and rows.

2. Data cleaning and feature engineering

Initial checks

In the first step data quality was checked. Tests revealed that in the dataframe there were no other data types than INT64 and FLOAT64, therefore one-hot encoding of categorical variables with string values was not needed.

Additionally, it was found that there were no missing values or duplicated rows in the dataframe. On the other hand two duplicated columns were identified and removed.

It was also noticed that a big disproportion between output class labels (“Bankruptcy?’) count was present. Only 220 of the rows had label ‘1’, while 6599 times label ‘0’ was assigned.

Statistics

Mean and median of each feature were calculated for 3 different sets:

- For all data
- For rows labeled ‘0’ (‘set 0’)

- For rows labeled ‘1’ (set 1’)

Then, both parameters for set 0 and set 1 were compared and their relative difference (in %) was listed in the new dataframe. It has to be noted that for 30 features discrepancy between both - means and medians - were smaller than 3%. It could suggest that some of those features might not be very useful in company bankruptcy prediction (however not necessarily).

Visual analysis

As mentioned earlier, due to the high number of columns it was computationally expensive to visually check dependencies between all of the features. However, some chosen feature pairs were analyzed in order to determine if any correlation is present. For example, it was found that ‘*Net Value Per Share (A)*’, ‘*Net Value Per Share (B)*’ and ‘*Net Value Per Share (C)*’ are almost linearly dependent, therefore the latter two were removed from the data (Figure 1).

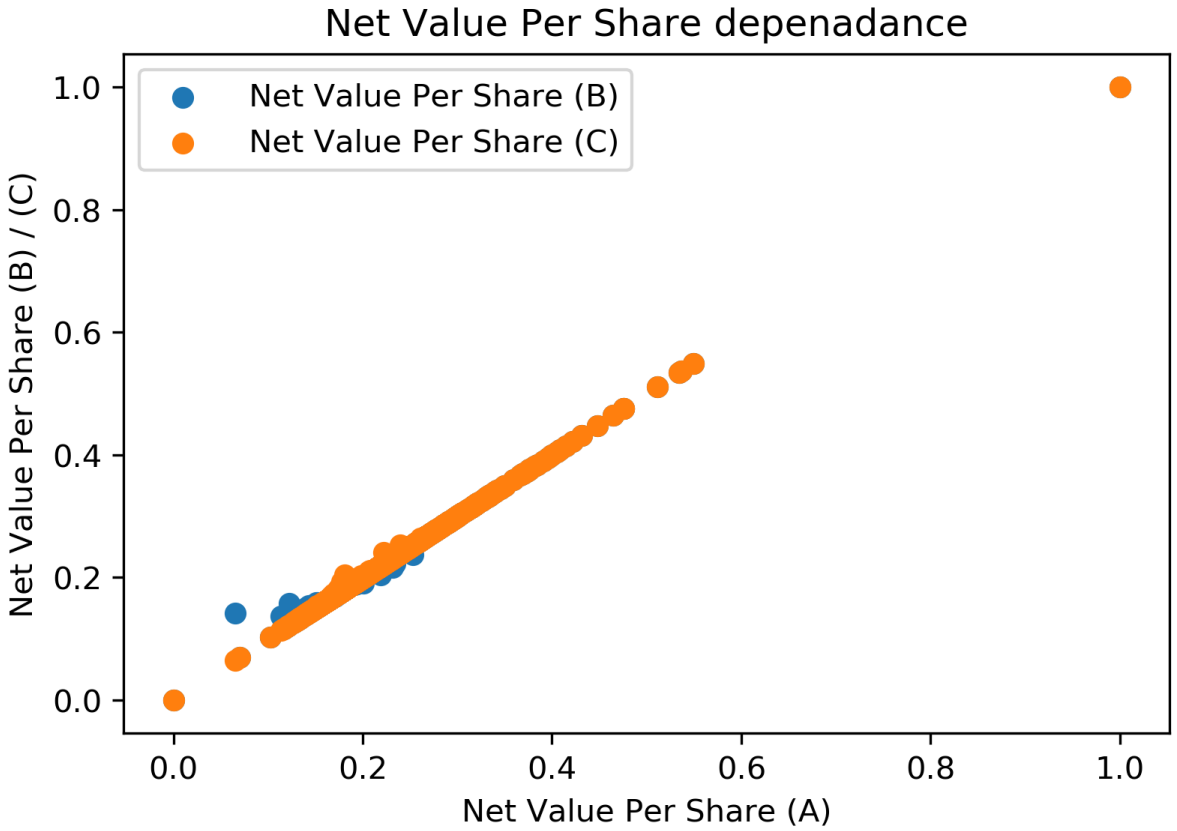

Figure 1. Net Value Per Share Comparison

Similar behaviour was discovered in relations between ‘*ROA(A) before interest and % after tax*’, ‘*ROA(B) before interest and depreciation after tax*' and ‘*ROA(C) before interest and depreciation before interest*' (Figure 2). The latter two were removed from the data.

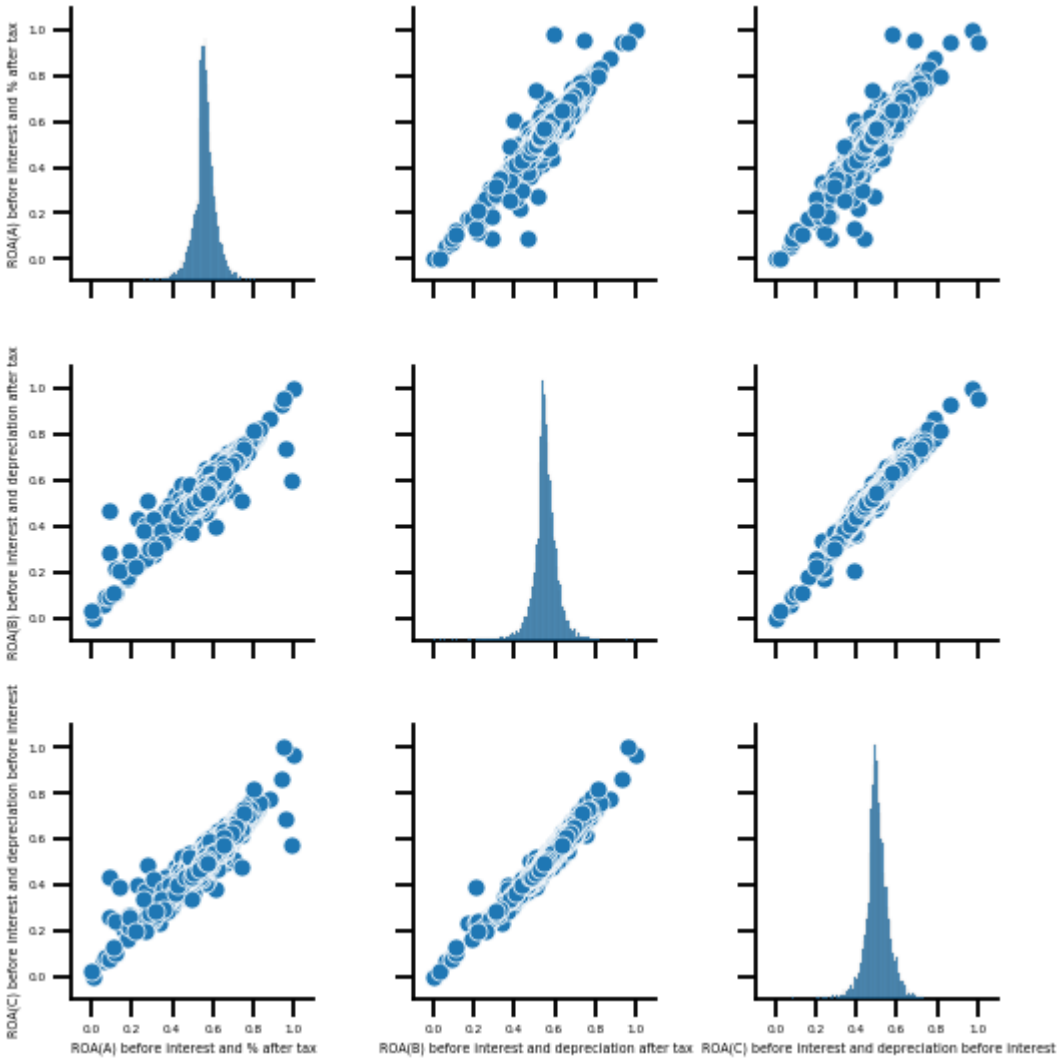

Figure 2. ROA comparison

Furthermore, near perfect linear dependance was discovered between *'Regular Net Profit Growth Rate'* and *'After-tax Net Profit Growth Rate'* (Figure 3). The first one was therefore dropped from the data.

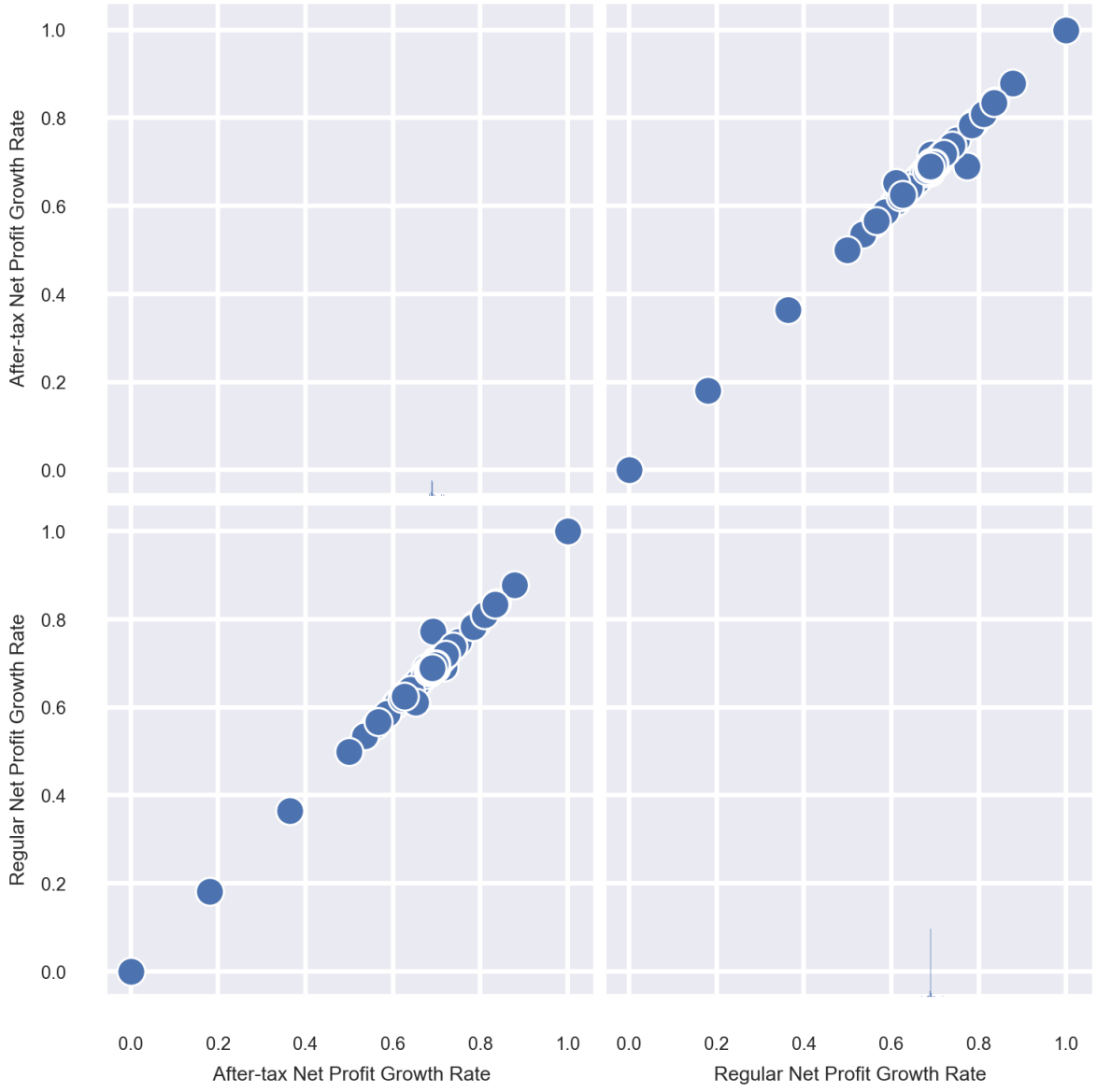

Figure 3. Net Profit Growth Rate Comparison

Finally, during different visual checks it was found that two data rows are mostly containing zeros or outlier values and, as a consequence, were removed from the data.
Additionally, skewness of the data was checked. Columns exceeding skew limit of 0,75 were log transformed.

Data hypotheses

Basing on the gathered data 3 hypotheses were formulated:

1. *'Cash Flow to Sales'* feature **does not** affect Bankruptcy of the company
2. *'Operating Profit Rate'* feature **does not** affect Bankruptcy of the company
3. *'Working Capital Turnover Rate'* feature **does not** affect Bankruptcy of the company

Significance test

Detailed assessment of *'Operating Profit Rate* feature **does not** affect Bankruptcy of the company' null hypothesis was performed as a result of significance test. Alternative hypothesis could be, therefore, formulated as *'Operating Profit Rate* feature **does** affect Bankruptcy of the company'.

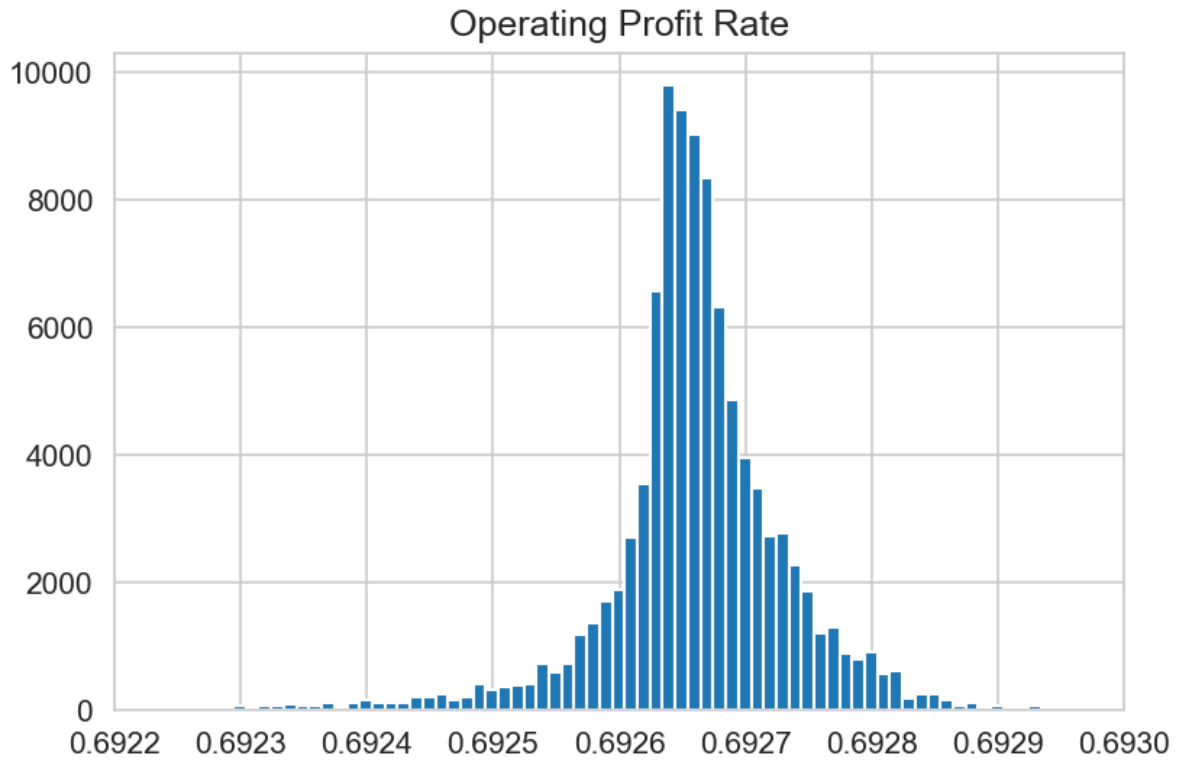

Figure 4. Operating Profit Rate distribution

Cutoff probability was assumed as 5%, therefore values lower than **0.6925075** (5% of empirical cumulative density function) or higher than **0.69277855** (95% of empirical cumulative density function) would result in rejecting the null hypothesis. Mean value for '1' class of 'Bankruptcy' feature accounts for **0.69251629.** Probability of such an event is around 5.4%, therefore the null hypothesis cannot be rejected. On the other hand, the result was extremely close to the cutoff threshold and additional data might be useful for more detailed evaluation.

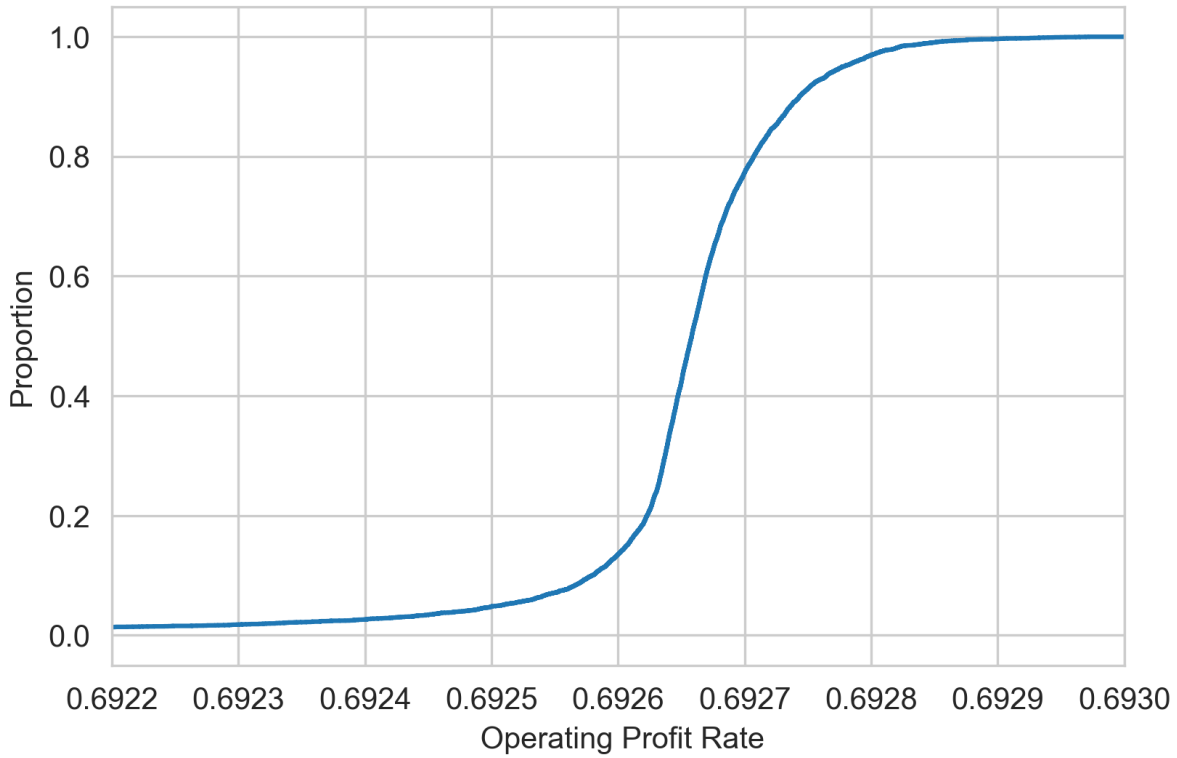

Figure 5. Operating Profit Rate empirical cumulative density function

Summary and suggestions

Dataset provides a wide range of different financial parameters in order to predict a company's bankruptcy. CSV set was introduced to Pandas dataframe, allowing for data analysis using Python.

It was discovered that there were no missing values or duplicated rows in the dataset. On the other hand, duplicated columns had to be removed. One-hot encoding was not necessary, as no variables with string values were present in the dataset. Two rows containing mostly zeros and outliers were deleted from the dataframe.

Furthermore, several features turned out to be linearly dependent from each other or even with almost exactly the same values, therefore were excluded from the data as duplicated information.

Based on the calculated mean and median for each variable, a significance test of *'Operating Profit Rate* feature does not affect Bankruptcy of the company' hypothesis was conducted. Null hypothesis could not be rejected, however the result was close to the cutoff threshold and additional data might be useful for more detailed assessment.

Overall, the dataset can be viewed as a valuable source of information. Unfortunately, there is an imbalance between negative and positive cases of Bankruptcy class, which has to be taken into account in further steps (for example using the F-score method).

ANNEX A

List of the data features:

Y - Bankrupt?: Class label

X1 - ROA(C) before interest and depreciation before interest: Return On Total Assets(C)
X2 - ROA(A) before interest and % after tax: Return On Total Assets(A)
X3 - ROA(B) before interest and depreciation after tax: Return On Total Assets(B)
X4 - Operating Gross Margin: Gross Profit/Net Sales
X5 - Realized Sales Gross Margin: Realized Gross Profit/Net Sales
X6 - Operating Profit Rate: Operating Income/Net Sales
X7 - Pre-tax net Interest Rate: Pre-Tax Income/Net Sales
X8 - After-tax net Interest Rate: Net Income/Net Sales
X9 - Non-industry income and expenditure/revenue: Net Non-operating Income Ratio
X10 - Continuous interest rate (after tax): Net Income-Exclude Disposal Gain or Loss/Net Sales
X11 - Operating Expense Rate: Operating Expenses/Net Sales
X12 - Research and development expense rate: (Research and Development Expenses)/Net Sales
X13 - Cash flow rate: Cash Flow from Operating/Current Liabilities
X14 - Interest-bearing debt interest rate: Interest-bearing Debt/Equity
X15 - Tax rate (A): Effective Tax Rate
X16 - Net Value Per Share (B): Book Value Per Share(B)
X17 - Net Value Per Share (A): Book Value Per Share(A)
X18 - Net Value Per Share (C): Book Value Per Share(C)
X19 - Persistent EPS in the Last Four Seasons: EPS-Net Income
X20 - Cash Flow Per Share
X21 - Revenue Per Share (Yuan ¥): Sales Per Share
X22 - Operating Profit Per Share (Yuan ¥): Operating Income Per Share
X23 - Per Share Net profit before tax (Yuan ¥): Pretax Income Per Share
X24 - Realized Sales Gross Profit Growth Rate
X25 - Operating Profit Growth Rate: Operating Income Growth
X26 - After-tax Net Profit Growth Rate: Net Income Growth
X27 - Regular Net Profit Growth Rate: Continuing Operating Income after Tax Growth
X28 - Continuous Net Profit Growth Rate: Net Income-Excluding Disposal Gain or Loss Growth
X29 - Total Asset Growth Rate: Total Asset Growth
X30 - Net Value Growth Rate: Total Equity Growth
X31 - Total Asset Return Growth Rate Ratio: Return on Total Asset Growth
X32 - Cash Reinvestment %: Cash Reinvestment Ratio
X33 - Current Ratio
X34 - Quick Ratio: Acid Test
X35 - Interest Expense Ratio: Interest Expenses/Total Revenue
X36 - Total debt/Total net worth: Total Liability/Equity Ratio
X37 - Debt ratio %: Liability/Total Assets
X38 - Net worth/Assets: Equity/Total Assets
X39 - Long-term fund suitability ratio (A): (Long-term Liability+Equity)/Fixed Assets
X40 - Borrowing dependency: Cost of Interest-bearing Debt
X41 - Contingent liabilities/Net worth: Contingent Liability/Equity
X42 - Operating profit/Paid-in capital: Operating Income/Capital
X43 - Net profit before tax/Paid-in capital: Pretax Income/Capital
X44 - Inventory and accounts receivable/Net value: (Inventory+Accounts Receivables)/Equity
X45 - Total Asset Turnover
X46 - Accounts Receivable Turnover
X47 - Average Collection Days: Days Receivable Outstanding
X48 - Inventory Turnover Rate (times)
X49 - Fixed Assets Turnover Frequency

X50 - Net Worth Turnover Rate (times): Equity Turnover
X51 - Revenue per person: Sales Per Employee
X52 - Operating profit per person: Operation Income Per Employee
X53 - Allocation rate per person: Fixed Assets Per Employee
X54 - Working Capital to Total Assets
X55 - Quick Assets/Total Assets
X56 - Current Assets/Total Assets
X57 - Cash/Total Assets
X58 - Quick Assets/Current Liability
X59 - Cash/Current Liability
X60 - Current Liability to Assets
X61 - Operating Funds to Liability
X62 - Inventory/Working Capital
X63 - Inventory/Current Liability
X64 - Current Liabilities/Liability
X65 - Working Capital/Equity
X66 - Current Liabilities/Equity
X67 - Long-term Liability to Current Assets
X68 - Retained Earnings to Total Assets
X69 - Total income/Total expense
X70 - Total expense/Assets
X71 - Current Asset Turnover Rate: Current Assets to Sales
X72 - Quick Asset Turnover Rate: Quick Assets to Sales
X73 - Working capitcal Turnover Rate: Working Capital to Sales
X74 - Cash Turnover Rate: Cash to Sales
X75 - Cash Flow to Sales
X76 - Fixed Assets to Assets
X77 - Current Liability to Liability
X78 - Current Liability to Equity
X79 - Equity to Long-term Liability
X80 - Cash Flow to Total Assets
X81 - Cash Flow to Liability
X82 - CFO to Assets
X83 - Cash Flow to Equity
X84 - Current Liability to Current Assets
X85 - Liability-Assets Flag: 1 if Total Liability exceeds Total Assets, 0 otherwise
X86 - Net Income to Total Assets
X87 - Total assets to GNP price
X88 - No-credit Interval
X89 - Gross Profit to Sales
X90 - Net Income to Stockholder's Equity
X91 - Liability to Equity
X92 - Degree of Financial Leverage (DFL)
X93 - Interest Coverage Ratio (Interest expense to EBIT)
X94 - Net Income Flag: 1 if Net Income is Negative for the last two years, 0 otherwise
X95 - Equity to Liability

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