Background

The data were collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange. You have been provided with 95 signals to predict company bankruptcy. Here is the list of variable names in the data:

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

Task

- Split the data into training sample and testing sample. i.e., use the first 70% of the data as training sample and the remainder as testing sample.
- Do a preliminary covariance analysis on all variables. Plot the heatmap to show the correlation structure of all variables. Clean the data if necessary. Briefly comment on your findings.
- Use the training sample and a simple logistic regression model, including all predictors, to train the model. Use the testing sample to predict company bankruptcy (if estimated probability of bankruptcy is greater or equal to 0.5, then we predict this company will be bankrupt) and show the confusion matrix. Report the accuracy rate for the out-of-sample (OOS) prediction.
- Use the training sample and a logistic regression model which only included the 5 most correlated predictors with the y variable, to train the model. Then, similarly, report the OOS confusion matrix and the accuracy rate.
- Use a boosted classification tree to train the model and then, similarly, report the OOS confusion matrix and the accuracy rate.
- Use a random forest to train the model and then, similarly, report the OOS confusion matrix and the accuracy rate.
- Compare the results from different models. Comment on your findings.

Instructions:

- The data is provided as CSV file.
- Try different values for the hyperparameters in ML models. Find the one that gives the best OOS performance.
- Submit the .m file and the PDF report. At the end of your PDF file, attach your codes as the appendix.
- Save figures as .png file, and use them in your PDF report.
- Save all required variables in one .mat file.
- Your PDF report shouldn’t exceed 4 pages, excluding figures and appendix. So, it is crucial to keep the sentences and paragraphs short and informative.
- Copy the coding script (i.e. the content in .m file) and attach it at the end of the reporting PDF file as the appendix.
- In the PDF report, set the font size of the main body text, as well as the appendix, as 12.
- Structure of the PDF report - it can consist of the following sections:
    - (a) Section One: Introduction
      give a very brief introduction on the problem you are investigating, and the models you will consider to tackle this problem. Give a general “big picture” of your findings and your conclusion. (should be around a page or less);
    - (b) Section Two: Methodologies
      give detailed accounts of methods you used to import data, clean data, as well as models you employed to analyse the dataset. briefly explain your models and the underlying theories. Point out pros and cons of each model, as well as any issues you have encountered with various models, and how you solved those problems.
    - (c) Section Three: Main findings
      in this section you should present your estimation results for each model and compare between them and draw your conclusion. Include all your figures and tables you may have obtained here.
    - (d) Section Four: Conclusion
      Very BRIEFLY summarise the problem you are investigating, draw your con- cluding remarks based on your findings. This section should be very short and mainly serves to give an emphasis to your main findings.