PREDICTING BANKRUPTCY

Data Science report on how to predict when a company will go bankrupt, based on past data

MIGUEL FREITAS

What is Bankruptcy?

When an individual or business is unable to pay their outstanding **debts** or **obligations**, bankruptcy is a legal procedure that can be initiated to provide them with a **solution** to re-think their business approach.

The process typically begins with the debtor filing a petition, although in some cases creditors may initiate the process instead. The debtor's **assets** are then assessed and valued, and in some cases, they may be used to partially repay the outstanding debt.

The data

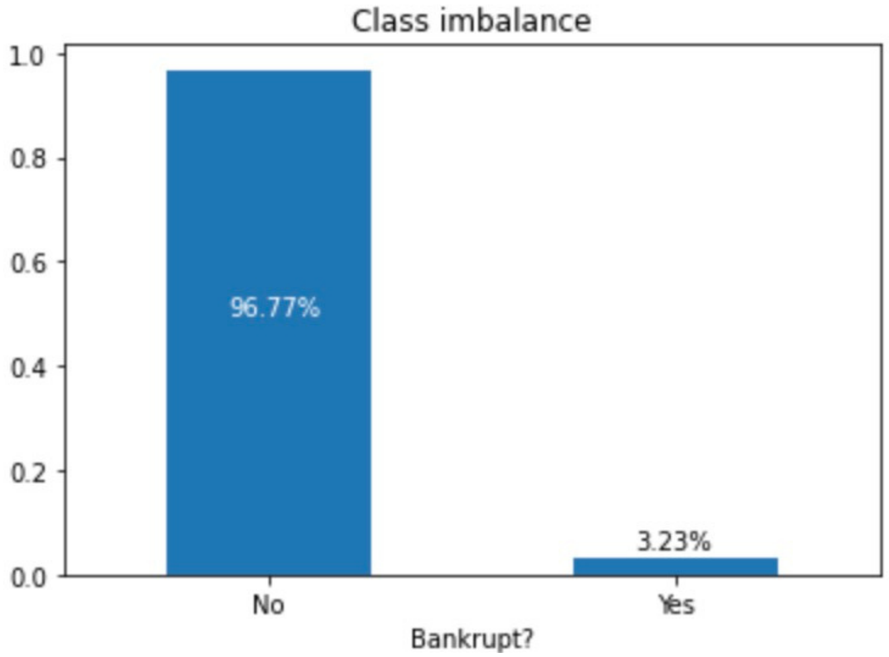
The data used in this report was uploaded by the user "FEDESORIANO", on Kaggle, upon the following link: https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction

Below you can find a brief overview:

- **6819** observations/ rows
- **96** features/ columns
- **0** missing values
- **All** features are numeric
- File in **CSV** format

The data

We can observe, as per the graph below, that the data used on this problem is severely **imbalanced**:



Methodology

The data has the following characteristics:

- Highly **correlated** values
- High value **ranges**
- High number of variables/ **dimensions**

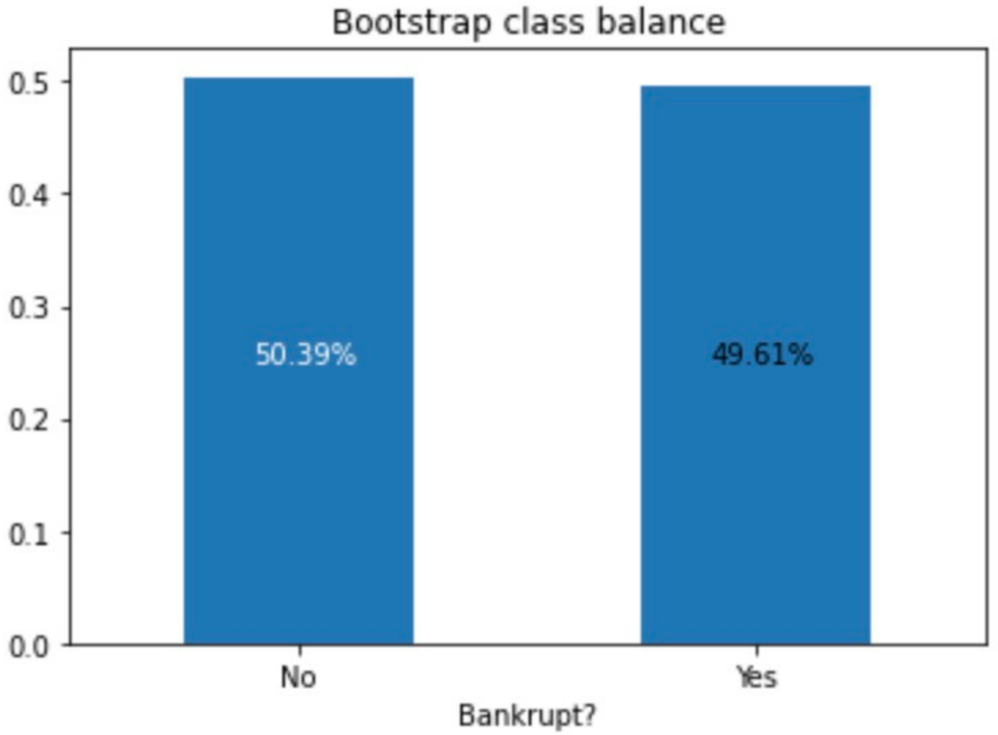
Before proceeding with the model, due to the class imbalance, a bootstrap sample method was applied to bypass this problem, where sample with replacement occurs with higher frequency in the lesser class.

A relevant number of models are **sensitive** to these characteristics, however, Classification and Regression Trees (CART) based models are robust and perform well.

Since Decision Trees can suffer from the curse of dimensionality (high number of variables), I chose a **Random Forest model**, which is comprised of several individual trees. Since it uses only a handful of variables each time a tree is generated in the ensemble, it is prone to perform better. While the better performance of the Random Forest is not solely due to this factor.

Results

As said, before deploying a Forest Tree model, the data set was bootstrapped to increase the class of interest ("Yes") and in turn, allowing for better predictions:



Results

However, this comes with a **tradeoff between prediction capabability and risk of overfitting**. Since the class of interest ("Yes"/ 1) is sampled with replacement, it will exist far too many repeated observations, both on the training and testing set. This will logically result in very high accuracy levels of predictions, since the exact same observations on the training set, exist as well on the test set:

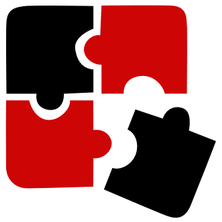
	precision	recall	f1-score	support
0	1.00	0.98	0.99	687
1	0.98	1.00	0.99	677
accuracy			0.99	1364
macro avg	0.99	0.99	0.99	1364
weighted avg	0.99	0.99	0.99	1364

Results

After training the model on the bootstrapped data set, the original data set was fed to the model (on the left), while a Random Forest base model applied directly on the original data set was deployed as a **performance comparison** (on the right):

	precision	recall	f1-score	support
0	1.00	0.98	0.99	6599
1	0.69	1.00	0.81	220
accuracy			0.99	6819
macro avg	0.84	0.99	0.90	6819
weighted avg	0.99	0.99	0.99	6819

	precision	recall	f1-score	support
0	0.97	1.00	0.98	1320
1	0.56	0.11	0.19	44
accuracy			0.97	1364
macro avg	0.76	0.56	0.59	1364
weighted avg	0.96	0.97	0.96	1364

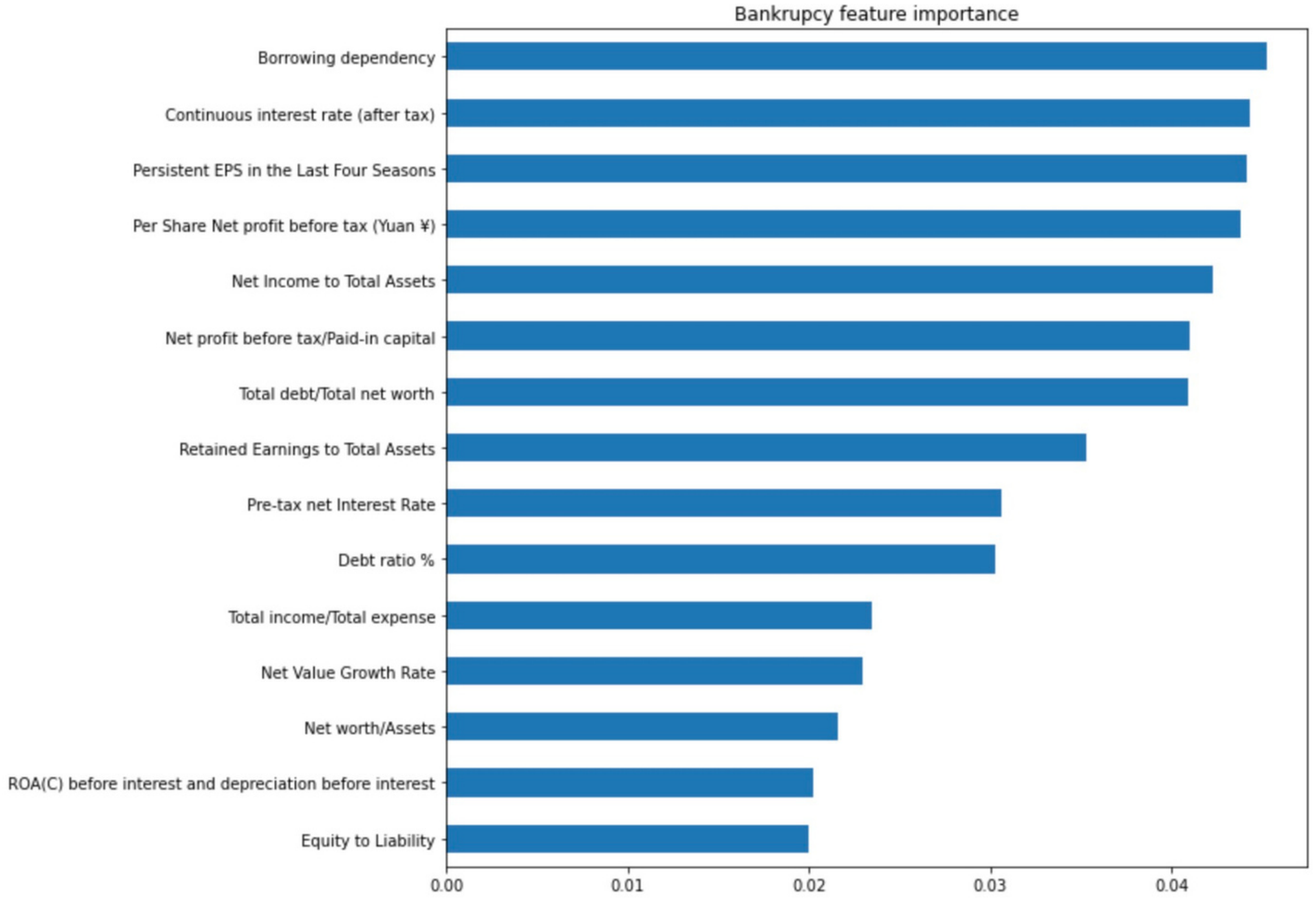


The model trained on the bootstrapped data performed far better with an F1-score of **81%** while the comparison model has an F1-score of **19%**. Looking to the class of interest ("Yes"/ 1) metrics, the best model predicted **no false negatives** (actual bankruptcy, predicted no bankruptcy), while **some false positives** were detected (actual no bankruptcy, predicted bankruptcy), with 69% of the class of interest being correctly labeled.

Understanding Bankruptcy

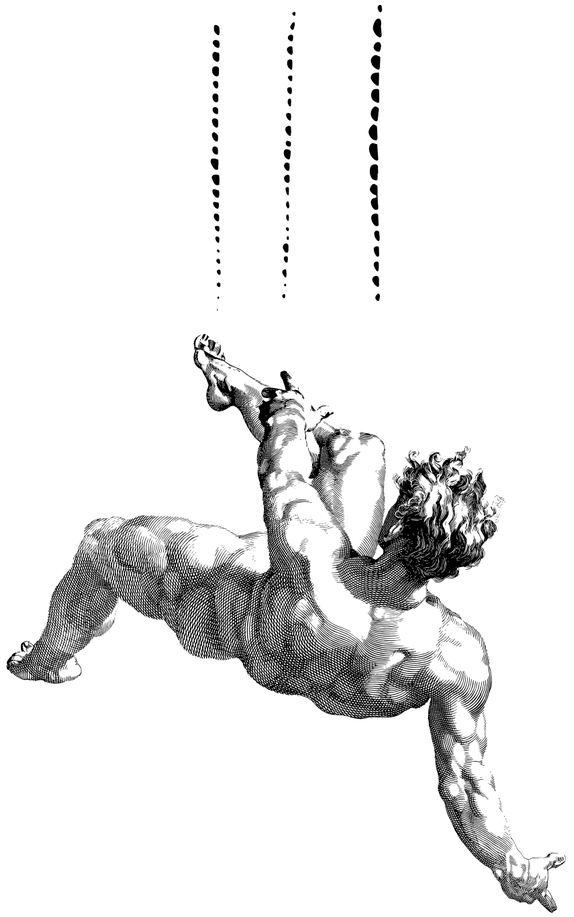
To be able to predict when a company will file for bankruptcy, is not simple math neither something that is linearly defined. In fact, there are several variables (or dimensions) that **need to be analyzed together** to understand a given company's financial situation. This is why the business data analysis team, in pair with advanced software, perform such a vital role to ensure an organization's vitality.

On your right you can see **15** out of the **95** variables that influence a firm's financial health. Yet, this podium alone comprises of approximately **50% of the prediction power of the model**. In other words, with a 50% threshold, these features are the most important that the model uses to make it's predictions.



“A bankruptcy judge can fix your balance sheet, but he cannot fix your company.”

— Gordon Bethune



Attachements

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

Attachements

Data features

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

Attachements

References

1. https://www.investopedia.com/terms/b/bankruptcy.asp