Predicting Bankruptcy with Financial Statements

By : Tamanna Baig and Tim Mango

Project Motivation

Bankruptcy prediction is the task of predicting whether a firm will go bankrupt given information from the firm's financial statements. This analysis is important to creditors and investors as they evaluate bankruptcy risk.

Description of Data Set

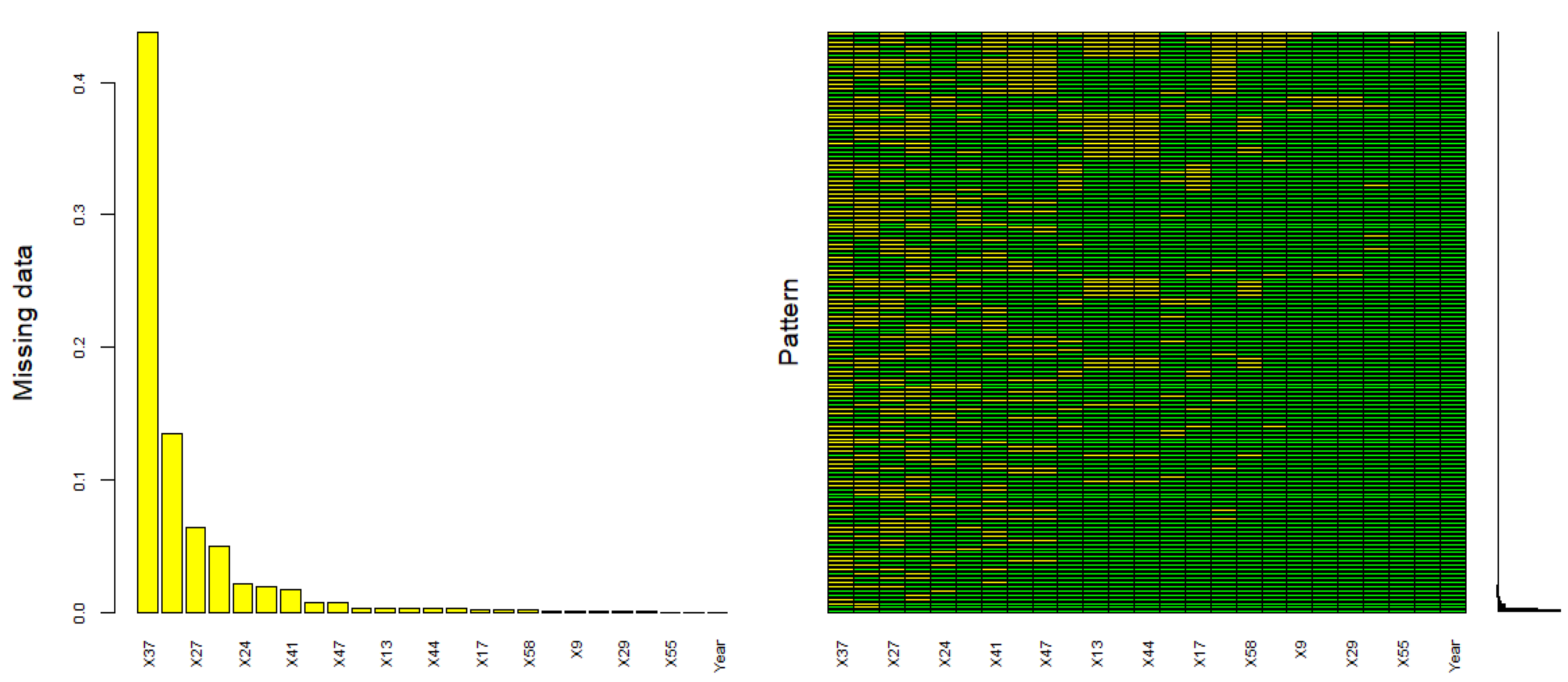
The dataset was created by Sebastian Tomczak and contains bankruptcy information for Polish companies. The data was collected from Emerging Markets Information Service (EMIS), which is a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013.

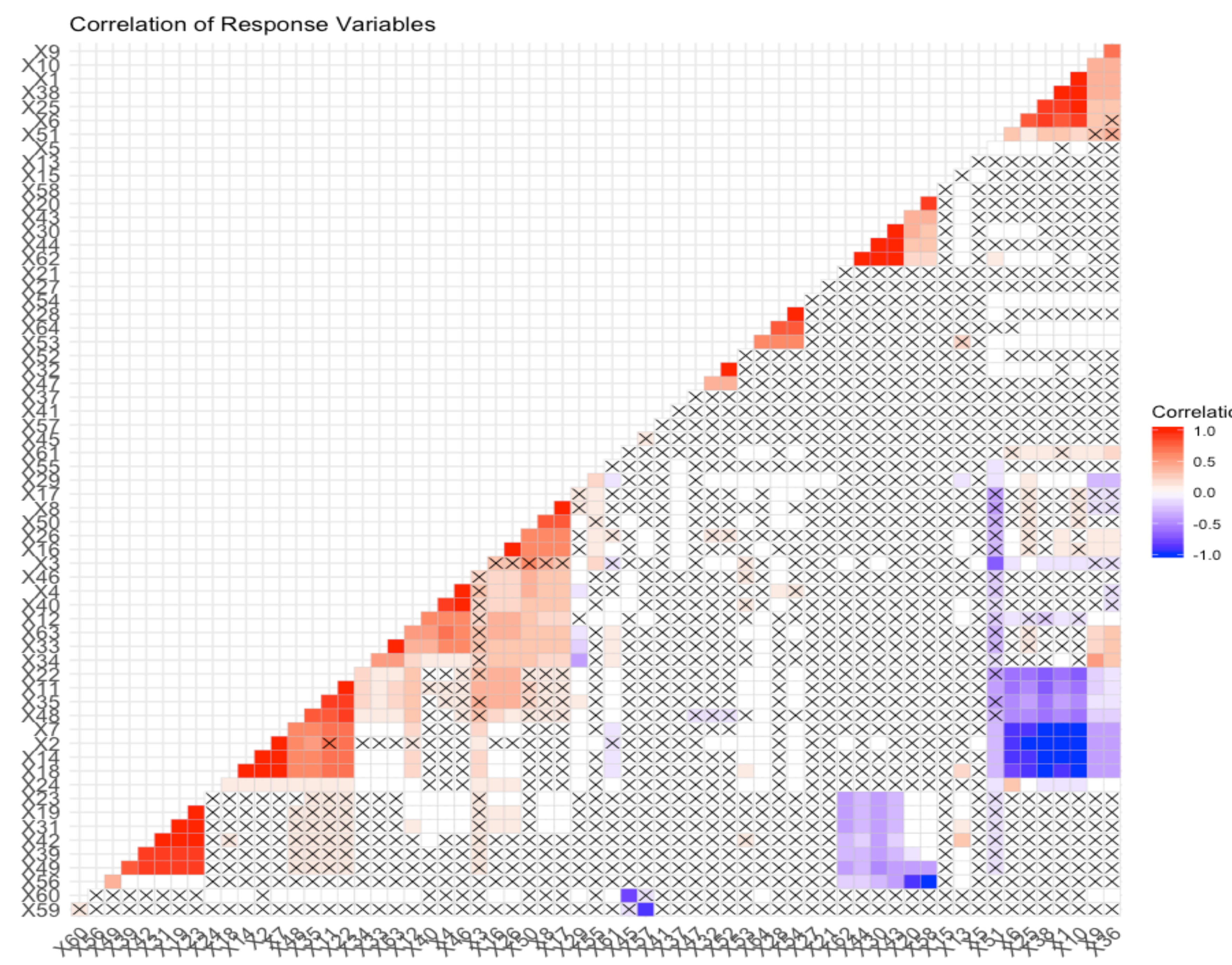
The set of features considered in classification process.

ID	Description	ID	Description
X1	net profit / total assets	X33	operating expenses / short-term liabilities
X2	total liabilities / total assets	X34	operating expenses / total liabilities
X3	working capital / total assets	X35	profit on sales / total assets
X4	current assets / short-term liabilities	X36	total sales / total assets
X5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365,	X37	(current assets - inventories) / long-term liabilities
X6	retained earnings / total assets	X38	constant capital / total assets
X7	EBIT / total assets	X39	profit on sales / sales
X8	book value of equity / total liabilities	X40	(current assets - inventory - receivables) / short-term liabilities
X9	sales / total assets	X41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
X10	equity / total assets	X42	profit on operating activities / sales
X11	(gross profit + extraordinary items + financial expenses) / total assets	X43	rotation receivables + inventory turnover in days
X12	gross profit / short-term liabilities	X44	(receivables * 365) / sales
X13	(gross profit + depreciation) / sales	X45	net profit / inventory
X14	(gross profit + interest) / total assets	X46	(current assets - inventory) / short-term liabilities
X15	(total liabilities * 365) / (gross profit + depreciation)	X47	(inventory * 365) / cost of products sold
X16	(gross profit + depreciation) / total liabilities	X48	EBITDA (profit on operating activities - depreciation) / total assets
X17	total assets / total liabilities	X49	EBITDA (profit on operating activities - depreciation) / sales
X18	gross profit / total assets	X50	current assets / total liabilities
X19	gross profit / sales	X51	short-term liabilities / total assets
X20	(inventory * 365) / sales	X52	(short-term liabilities * 365) / cost of products sold)
X21	sales (n) / sales (n-1)	X53	equity / fixed assets
X22	profit on operating activities / total assets	X54	constant capital / fixed assets
X23	net profit / sales	X55	working capital
X24	gross profit (in 3 years) / total assets	X56	(sales - cost of products sold) / sales
X25	(equity - share capital) / total assets	X57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
X26	(net profit + depreciation) / total liabilities	X58	total costs /total sales
X27	profit on operating activities / financial expenses	X59	long-term liabilities / equity
X28	working capital / fixed assets	X60	sales / inventory
X29	logarithm of total assets	X61	sales / receivables
X30	(total liabilities - cash) / sales	X62	(short-term liabilities *365) / sales
X31	(gross profit + interest) / sales	X63	sales / short-term liabilities
X32	(current liabilities * 365) / cost of products sold	X64	sales / fixed assets

Exploratory Data Analysis

Correlation of Variables is represented using 'ggcorrplot' library, and the most correlated variables are clustered together in the visualization. Variables were ultimately dropped that had correlation values of greater than 0.5. The updated data set has few missing data cells, therefore, Data Imputation is done by using **'Multivariate Imputation by Chained Equations (MICE)'** method.





Naïve Bayes Summary

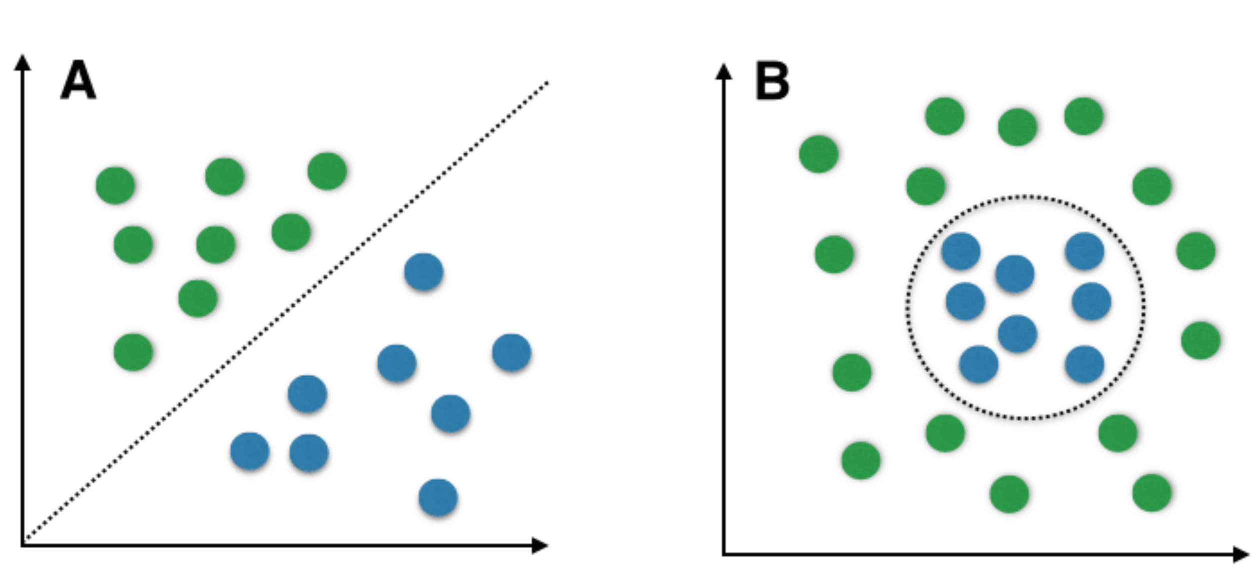
The Naïve Bayes classifier is a probabilistic classifier that relies on the Bayes theorem, which assumes that explanatory variables are completely independent of each other. The classifier is "Naïve" as real explanatory variables almost always share some amount of correlation. This assumption of independence means that all the predictors are expected to have an equal effect on the outcome. The Naïve Bayes formula is shown below:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Example: Fires are rare (1%) but smokey air is fairly common (10%), 90% of fires make smokey air:

P(Fire|Smoke) =P(Fire) P(Smoke|Fire) =1% x 90% = 9%P(Smoke)10%

9% of the time expect smoke to mean a dangerous fire.

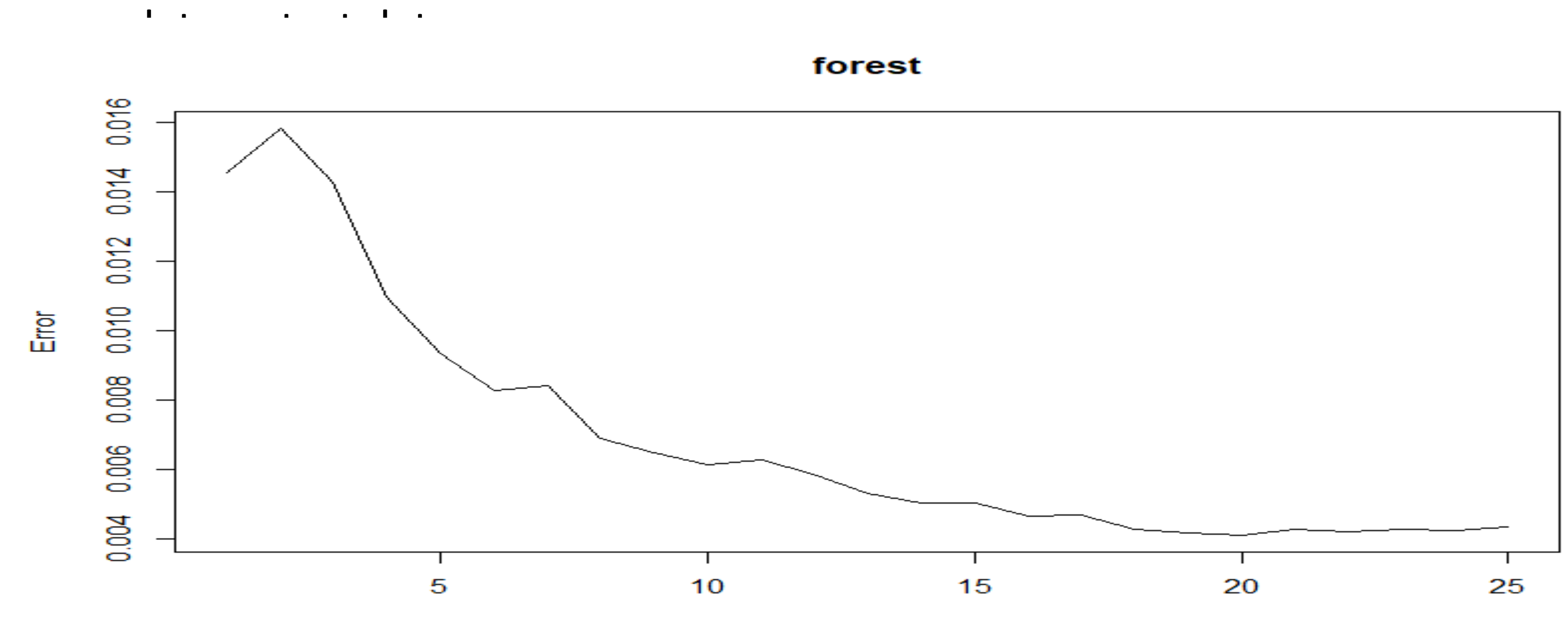


Model	Accuracy	Precision	Recall
Naïve Bayes	0.9799	0.9720	0.7126
Random Forest	0.9827	0.8483	0.7671
Naïve Bayes Year 5	0.9810	0.9578	0.8798
Random Forest Year 5	0.9780	0.9920	0.8065

Random Forest Summary

Random Forest is a type of Supervised learning technique that extends from decision trees. Random Forestl is designed to combat correlated trees and overfitting.

Several trees are generated on different bootstrapped samples from training data and then they are averaged. This helps to reduce the variance and also improves the performance of decision trees on new



Best Model:

Each of the models performed similarly when observing model accuracy. The Naïve Bayes Model is simple and computationally inexpensive while the Random Forest Model is a supervised learning method and is computationally expensive. The Naïve Bayes Model performed worse than Random Forest Model on recall, but better on precision with randomly generated training and testing sets. The Naïve Bayes 5 Year Model performed worse on precision but better on recall than the Random Forest 5 Year model. The last two models test training data from years 1-4 on the 5th year testing set. These models performed better than the original models as they included time series related information. The best model depends on whether recall or precision is most important.