Project Summary

Batch details	DSE November 2020 - Online
Team members	Apoorva Garg , Pranavi Krishnamsetty, Vishal Pandey, Yash Vahi
Domain of Project	BFSI
Proposed project title	Bankruptcy Prediction
Group Number	3
Team Leader	Yash Vahi
Mentor Name	Mr. Animesh Tiwari

Date: 10th June 2021

Signature of the Mentor

Signature of the Team Leader

Table of Contents

Interim Report

Industry Review

The rationale for developing and predicting the financial distress of a company is to develop a predictive model used to forecast the financial condition of a company by combining several econometric variables of interest to the researcher. For the longest time, the Altman Z-score was the king of prediction when it came to bankruptcy, it was the best model we had that would tell us which exact behaviours of a company could lead them to bankruptcy. Ever since its last modification in 1991, the use of Altman Z-score has declined over the years as much more complex and robust predictive models have been made since. Studies have examined the causes of business failure indicated by values of bankruptcy scores established during the decline stage of the business. In a survey of the 70 Estonian manufacturing firms, the researcher obtained the causes of bankruptcy from court judgments. The firms classified the reasons and the types of failure, that is, internal factors, that are different from management deficiencies and external factors to the firm. Ohlson's model and a local (Grünberg's) bankruptcy prediction model were used to calculate bankruptcy scores for the first and second pre-bankruptcy years. Applying median tests form independent samples to examine whether the different failure types are associated with different failure risk. The findings revealed that multiple causes have a significantly higher bankruptcy risk than single reasons for the year before the declaration of bankruptcy. The results indicate that numerous reasons lead to a considerably higher insolvency risk as compared with a single cause for the year before bankruptcy disclosure.

Machine learning techniques represent one type that has been used for decades in issuing loans. Banks use prudential choices of protecting the performance of companies by accessing corporate loan applicants. One of the methods they use is data envelopment analysis (DEA) to evaluate several decisions making units (DMU) ranked based on the best practice in their sector. Linear programming is imperative as it is

used in calculating corporate efficiency, used as a measure of differentiating between financially sound companies and those that are economically distressed. The results based on a study that sampled 742 listed Chinese companies over ten years suggest that Malmquist DEA offers discernments into the competitive position of a company in addition to accurate financial distress predictions based on the DEA efficiency measures. Ratio analysis financial indicators are the most popular variables used in bankruptcy prediction models. They often exhibit heavily skewed results owing to the presence of outliers. It is not very clear how different approaches affect the predictive power of models that predict bankruptcy. One of the challenges faced in models is the lack of a clear cut way of how to handle outliers and extremes that affect the power of models—two ways of reducing outlier bias by omission and winsorization. The categorisation of financial ratios is an effective way of handling outliers concerning the predictive performance of bankruptcy prediction models.

Predicting financial distress in empirical finance has received a lot of attention from researchers throughout the globe. Sampling small and medium enterprises in France using the Logit model, artificial neural networks, support vector machine techniques, partial least squares, and a hybrid model integrating support vector machine with partial least squares, it has been established that, within a year of financial distress, support vector machine should be preferred because it is the best and most accurate method for predicting for bankruptcy. In the case of two years, then the hybrid model outperforms the support vector machine, Logit model, partial least squares, and artificial neural networks with 94.28% overall accuracy of prediction. Financially distressed firms are found to be smaller, more leveraged, and with lower repayment capacity. In addition, they have lower profitability, liquidity, and solvency ratios. Creditors should, therefore, correctly evaluate the financial position of firms and be keen on any signs that may lead to negative growth to avoid capital loss and costs-related risks.

Dataset and Domain

The dataset that we're working on consists of financial information about companies in Taiwan and our aim is to build a model that predicts whether a company is or will go bankrupt in the future. Our dataset has 6819 rows and 95 columns.

Data Dictionary:-

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 capital 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

Why bankruptcy:-

Bankruptcy is a hot topic in the financial and corporate world. No company would want to go bankrupt, and 90 times out of a 100, no one really benefits from it. Banks loose money, people loose their jobs and worst of all, it’s the metaphorical and literal end of an idea, irrespective of the quality of it. Analysing what causes bankruptcy and laying out the symptoms can help stop the disease of bankruptcy in its tracks and even reverse some of its effects. For eg. we know through empirical research and countless predictive models that high cholesterol causes heart diseases but if you are diagnosed with high cholesterol at an early stage by a doctor, he can help you stop it in its tracks by recommending you to eat healthy foods and take your medication on time, and if you listen to his recommendations/prescriptions, your chances of not getting heart disease

greatly increase. The same applies to this problem, if we know the symptoms and telltale signs of bankruptcy, we can detect them at an early stage and deploy countermeasures to ensure that they don't progress further. This could lead to more number of healthy companies in the market space resulting in a healthier economy which in turn could result in better living standards or quality of life on average.

Exploratory data analysis and data preparation

Exploratory data analysis is the first step after problem selection and understanding. It comprises of 2 parts that are usually performed one after the other. The first part includes finding out information(data analysis) about the data, its shape, data types, null value detection, outlier detection and various univariate and multivariate analysis to find relationships between features. The second part entails using the information gathered in the first part and rectifying it to prepare the data for modeling(data preparation), for eg. missing value, outlier treatment, transformation, scaling etc. The quality of our data dictates how our models will fare. Below are the steps we took for EDA and the insights we gained.

Data analysis:

1. We checked the shape of the dataset and found that our dataset had 6819 rows and 96 columns.
2. We looked at the datatypes and found that all features in our dataset were numerical. Our target variable was the only categorical feature and the values 0 & 1.
3. We looked at the summary statistics and found that columns like "Net income flag" and "Liability assets flag" were insignificant for further analysis as they were not rich in information(same values throughout), so we removed them from the dataset. We also found that most of our features were in the range 0-1.
4. We then moved onto analysing our target feature "Bankrupt?"and found that it was severely imbalanced with heathy firms occupying 96.77% of the dataset pointing to fact that there might be a need for upsampling(*Figure 1.1*).

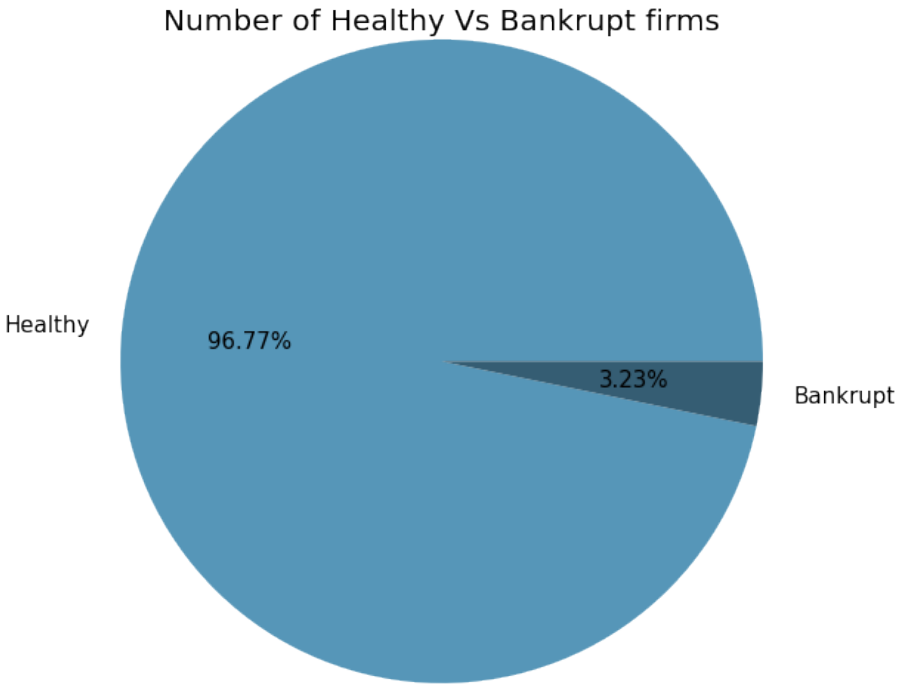


Figure 1.1

5. We started our univariate analysis by analysing the distribution of all the features using distribution plots and metrics like skewness and kurtosis and found that majority of the datasets were highly skewed and leptokurtic.
6. After our distribution analysis, we visualised outliers in each feature using box plots and found out that there were severe outliers presents in almost all the features.
7. Our null and duplicate value analysis told us that there were none present in the dataset.
8. We started our bivariate and multivariate analysis by analysing correlation between variables using the correlation matrix and a heat map. We found 77 distinct pairs of features that had correlation greater than 0.70(*Figure 1.2*).

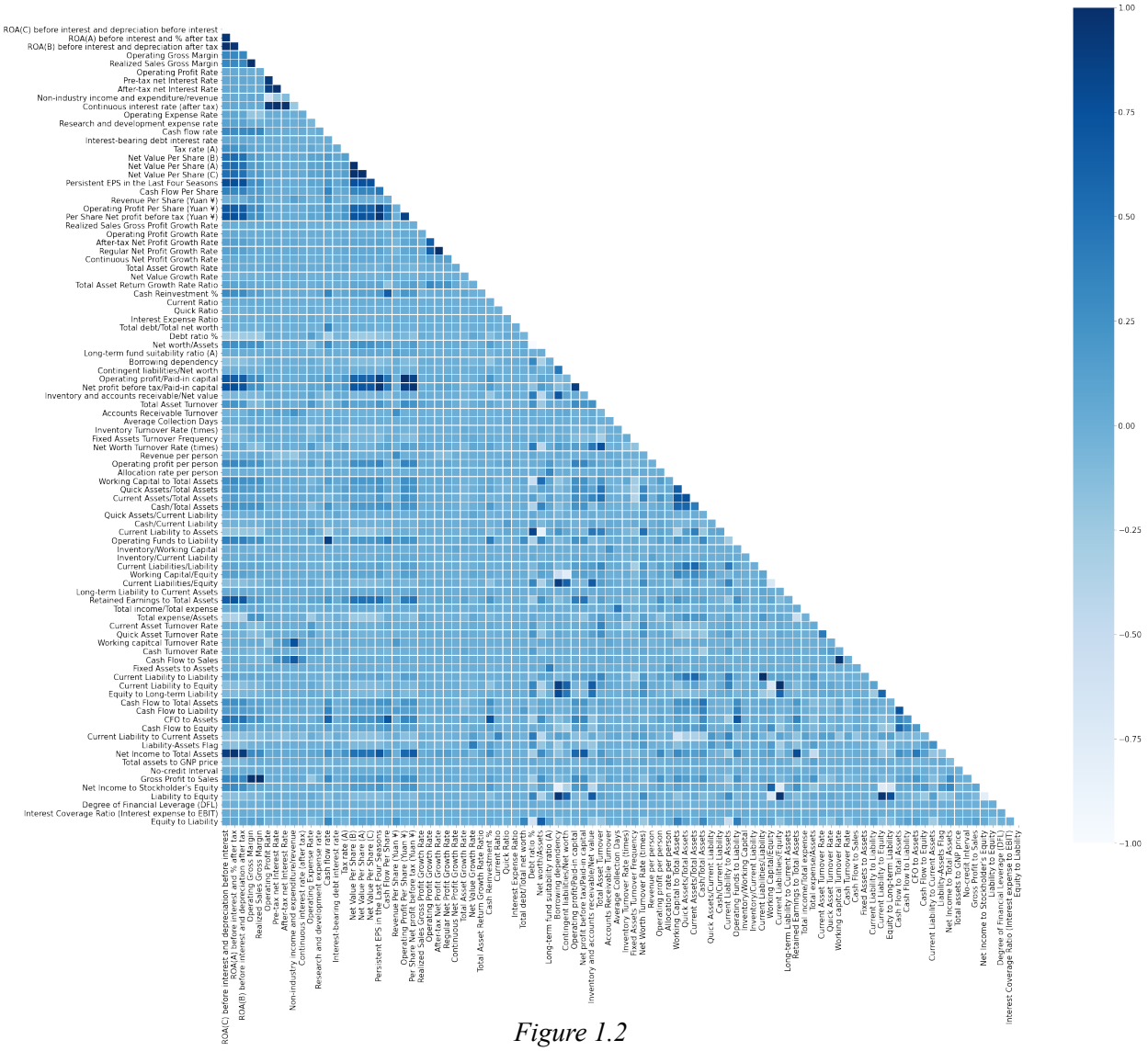


Figure 1.2

9. We plotted bar charts with ”Bankrupt?” on x-axis and median of each independent feature on y-axis separately to see whether there were any clear relationships we could identify, we found that

- Companies with low Research and development expense rate,Tax rate (A),Quick Ratio,Cash/Total Assets,Quick Assets/Current Liability and Cash/Current Liability tend to go bankrupt.
- Companies with high Total debt/Total net worth, Inventory Turnover Rate (times),Fixed Assets Turnover Frequency,Allocation rate per person,Current Liability to Assets,Long-term Liability to Current Assets,Quick Asset Turnover

Rate,Current Liability to Current Assets and Total assets to GNP price1 tend to go bankrupt(*Figure 1.3*).

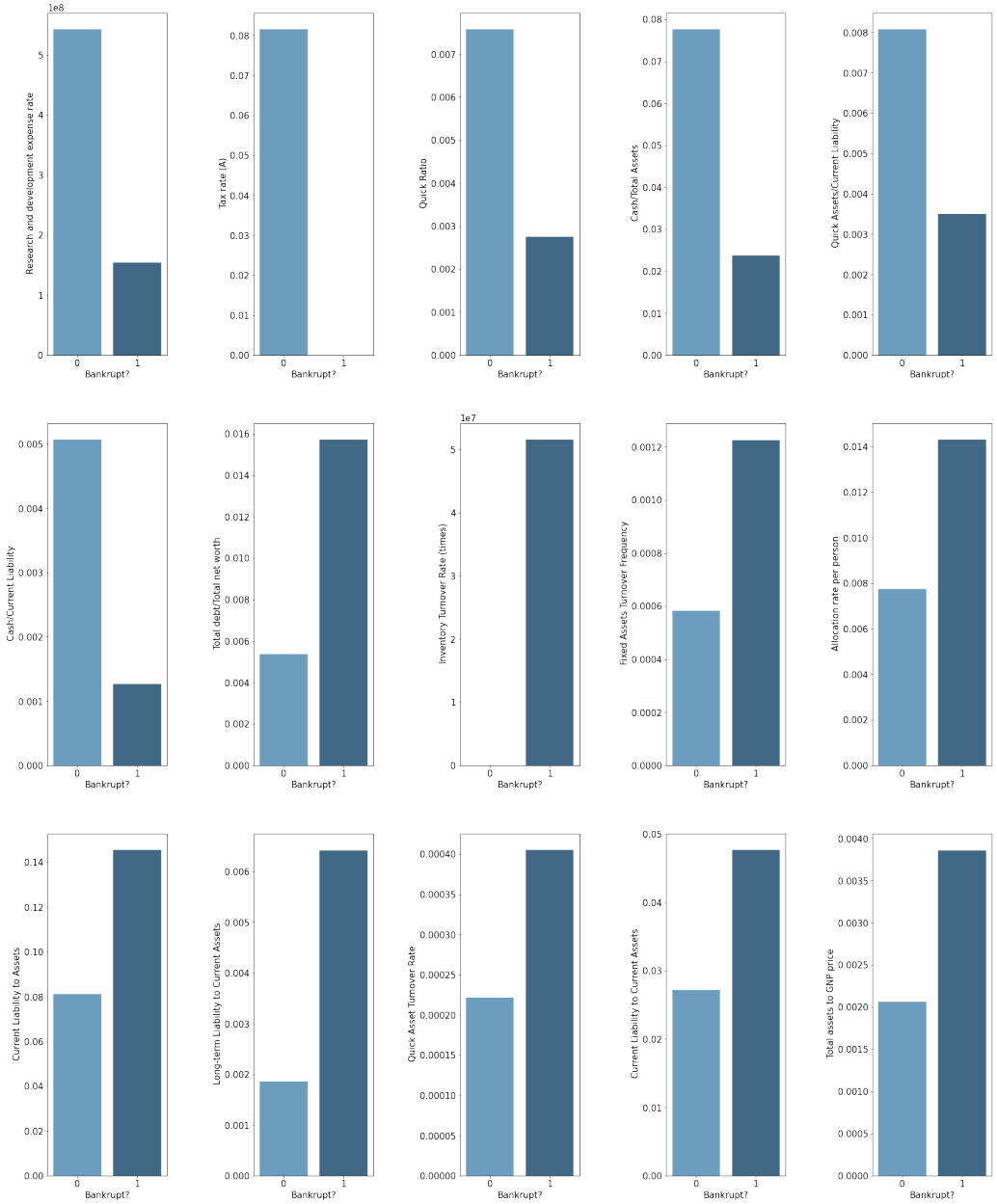


Figure 1.3

10. We plotted scatterplots for extremely correlated features(corr > .90) with hue as bankruptcy to see whether there are any evident clusters, we found that

- Companies with low Nett value per share(A,B&C) tend to go bankrupt
- Companies with low Persistent EPS in the Last Four Seasons,Per Share Net profit before tax (Yuan ¥),Net profit before tax/Paid-in capital tend to go bankrupt.(*Figure 1.4)*

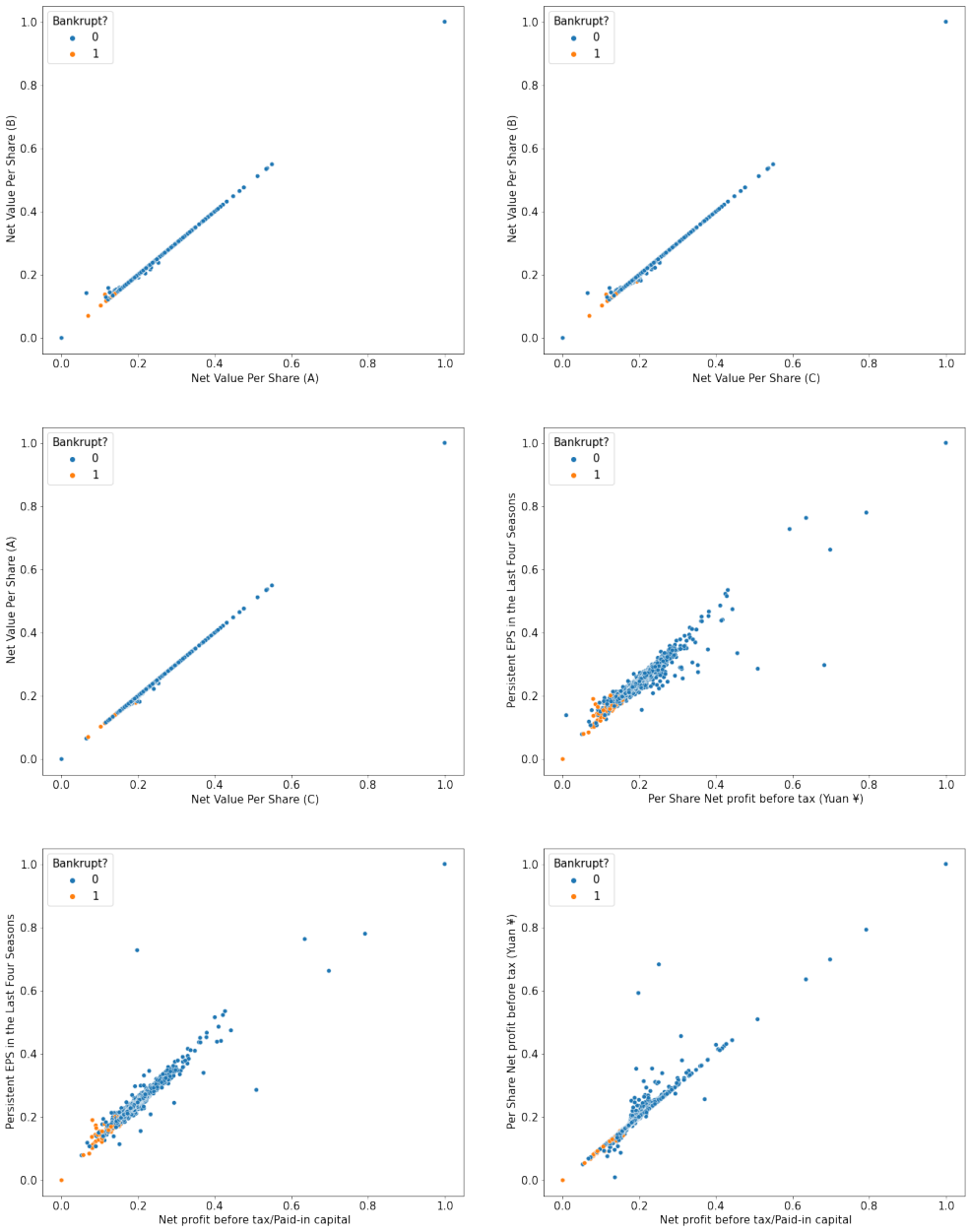


Figure 1.4

Data Preparation/cleaning:

1. We first removed all the redundant columns based on our domain knowledge i.e. columns that had information already contained in other columns. We were left with 81 columns after dropping all those columns.
2. We removed all the outliers using the IQR method.
3. We applied square root transformation on columns with more than absolute 1 skewness.

4. We decided to impute all the null values resulting from outlier removal using the KNN imputer. So, we first normalised all the features and imputed the null values using the KNN imputer.
5. Using the VIF filtering process, we reduced multicollinearity by removing all columns with VIF > 5.
6. Our last step was performing the train-test split, we performed a 70:30 split.

Modeling

Modeling is the stage where we can start applying methods to construct predictive models, we get to this stage after our data is ready i.e. meeting all the assumptions. We will build the following models and compare performances across the board:-

1. Logistic Regression(Maximum Likelihood Estimation)
2. Decision Tree
3. Random Forest Classifier
4. K-Nearest Neighbours
5. Naïve Bayes

Logistic Regression

Logistic regression is a classification algorithm built on top of linear regression. We will be going for the Maximum likelihood estimation approach which will help us to interpret the coefficients easily and allow us to establish relationships between the independent and target variable. Our aim is to maximise recall as letting unhealthy companies slip by is more costly than classifying some of the healthy companies as unhealthy.

We will follow the following steps for Logistic Regression:-

1. Build a full Logistic Regression model using Maximum likelihood estimation.
2. Plot Receiver Operating Characteristic(ROC) and get Area under Curve(AUC)
3. Build scorecard to see how different metrics like Precision, Recall, F1-score and kappa are performing under different thresholds.
4. Perform recursive feature elimination(RFE).

5. Build a model with features that survived RFE.
6. Repeat steps 2 and 3 for RFE model.
7. Compare both models.

Full model Performance:-

1. Our full Logistic Regression(MLE) model gave us an ROC of 0.9555.(*Figure 2.1*)
2. Maximum precision was 0.833 at thresholds 0.9.
3. Maximum Recall was 0.948718 at threshold 0.020431 through Youdens' index.
4. Maximum F1 was 0.483516 at threshold 0.2.
5. Maximum kappa was 0.459989 at threshold 0.2.

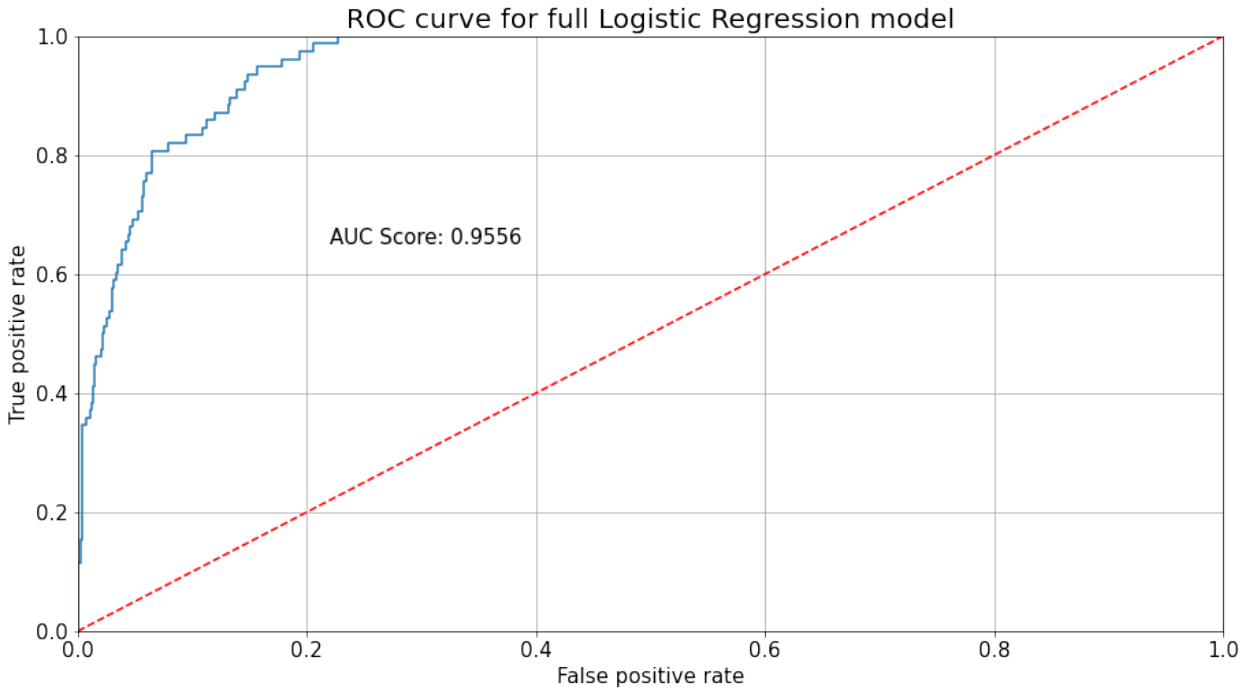


Figure 2.1

RFE model performance:-

1. Our Logistic Regression model after RFE gave us an ROC of 0.9593, which is barely above our full model.(*Figure 2.2*)
2. Maximum precision was 0.8333 at threshold 0.9.
3. Maximum Recall was 0.974359 at threshold 0.020815 obtained through Youdens' index.
4. Maximum F1 was 0.5 at threshold 0.3.

5. Maximum kappa was 0.480183 at threshold 0.3.

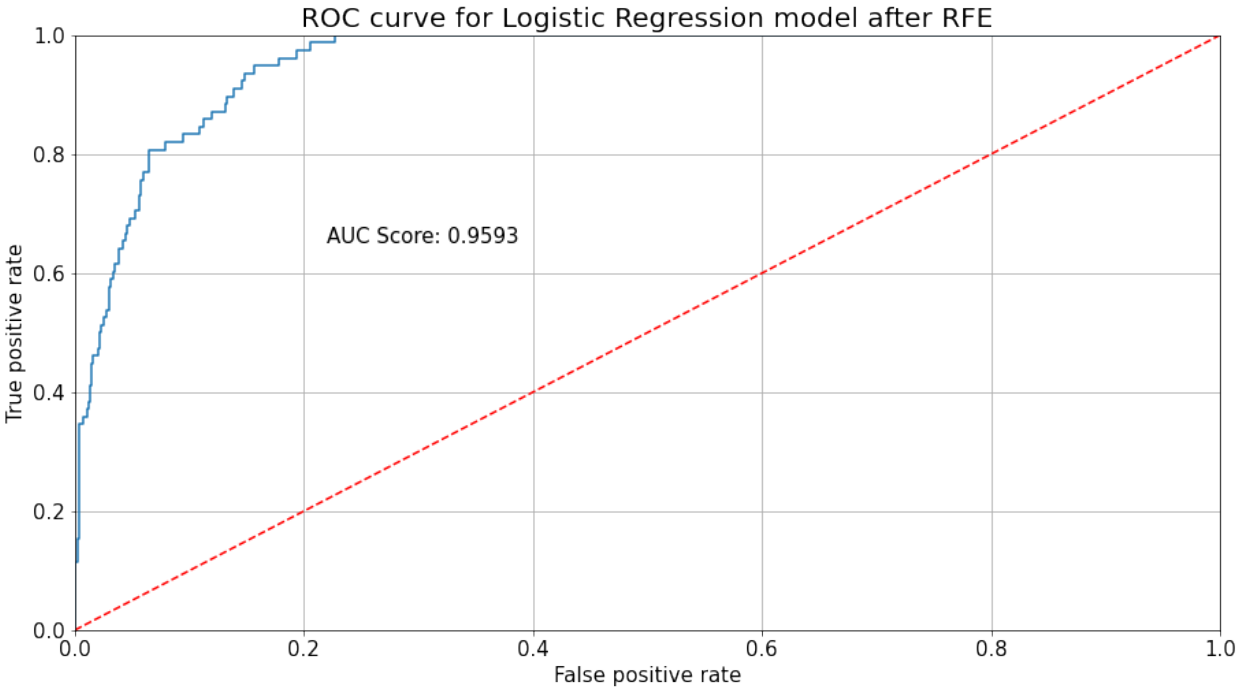
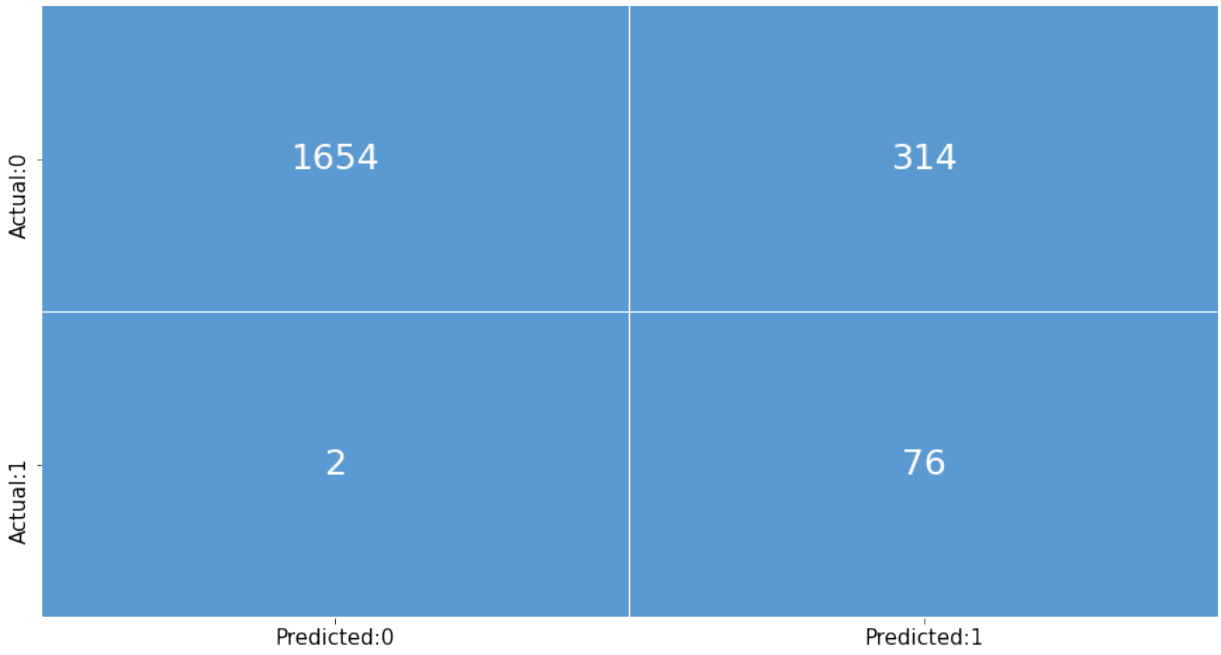


Figure 2.2

Model Comparison:-

As the purpose of our project is to maximise recall and detect as many bankrupt firms as possible, we will select the RFE model as it maximises our recall at 0.974359 at threshold 0.020815. Here's the confusion matrix for it,



Next steps:-

Build main model scorecard, where best models from different algorithms can be compared. Add the Logit RFE model to the scorecard.

Build rest of the models, compare best from each algorithm and pick the best.

References:

https://www.mdpi.com/1911-8074/13/3/47/pdf

Notes For Project Team

Original owner of data	Taiwan Economic Journal
Data set information	The data was 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.
Any past relevant articles using the dataset	-
Reference	Kaggle
Link to web page	https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction

****