INDIAN INSTITUTE OF TECHNOLOGY KANPUR

MTH416A: REGRESSION ANALYSIS

A PROJECT REPORT ON

Bankruptcy Prediction Using Logistic Regression

Created by

HARSHIT PANDEY

Contents

1 ABSTRACT

In corporate sectors estimating the risk of bankruptcies is of large importance for the safe and sound investment of investors. Due to this bankruptcy prediction is an important area of finance and accounting research.It is important because the investor is always keen to know about the likeliness of a company or firm to be bankrupt. In recent years AI and ML methods have achieved astonishing results in corporate bankruptcy prediction settings. Therefore, in this study of project, we explore, build, and compare the different classification models to get an accurate model regarding bankruptcy. We have chosen the 'Taiwan Stock Exchange.' bankruptcy data set and begin by carrying out data preprocessing and exploratory analysis where we address outlier issue and we delete the data which are outliers from the dataset. For data imbalance issue, we apply Synthetic Minority Oversampling Technique (SMOTE) to oversample the minority class labels and also under-sampling method to get the best way out. Then, we build an appropriate model using Logistic Regression on our cleaned dataset and finally, analyse and evaluate the performance of the models on the validation datasets using several metrics such as accuracy, precision, recall, etc., and rank the models accordingly.

2 INTRODUCTION

2.1 Context

Bankruptcy prediction is a way of forecasting company financial distress of both public and firms. The purpose of predicting bankruptcy is very basic aspect in assessing the financial condition of a company and its future prospects. The financial soundness of a company is of great importance to the various investors and creditors of the company. The participants which get affected by the company's prospect are majorly the policymakers, investors, banks, internal management,and the general public referred to as consumers. Accurate prediction of the financial performance of companies is of great importance to various stakeholders in making important and significant decisions. Financial distress is a global phenomenon that affects companies across the global economy. Additionally, the credit lenders and investors need to evaluate the financial bankruptcy risk of a company before making an investment decision to avoid a significant loss. A company's suppliers or retailers always conduct credit transactions with the company so they also require to fully understand the company's financial status and make decisions on the credit transaction. Problems concerning bankruptcy have necessitated the need for studies to establish some methodologies for predicting bankruptcies to aid investors in making wise investment decisions. Corporate failures in significant economic companies have spurred research for better understanding to develop prediction capabilities that guide decision making in investments. Available accounting ratios are a vital indicator or signal to indicate danger. Typically, companies are quantified by many indicators that describe their business performance based on mathematical models constructed from past observations based on evidence from data. There are many different concerns that are associated with the bankruptcy prediction. Two main problems are the econometric indicators describing the firms condition are proposed by domain experts and the historical observations used to train a model are usually induced by imbalanced data phenomenon as there are lot of successful companies against bankrupt companies. As a consequent, the trained model tends to predict companies as successful (majority class) even when some of them are almost bankrupt firms. Both these issues mostly inhibit the predictive capability of the model. In the

recent times to overcome the above problems now the new survival methods are being applied. Methods like option valuation which involves stock price variability have been developed. Under structural models, a default event is almost certain to occur for a company when its assets reach a drastically low level compared to its liabilities. Neural network models and other sophisticated models have been tested on bankruptcy prediction. Modern methods applied by business information companies surpass the annual accounts content and consider current events like age, judgements, bad press, payment incidents and payment experiences from investors.

2.2 Aim Of The Study

In this project, we wish to predict Bankruptcy of taiwan stock exchange using Logistic Regression. The key steps involved are as follows:

1. To handle the outlying data using suitable techniques of removing outliers.
2. To tackle the problem of Data Imbalance using suitable Resampling Methods.
3. To build the model using Logistic Regression.
4. To compare the accuracy of the model.

3 METHODOLOGY

In the previous section of the report, we have just introduced the problem statement of bankruptcy prediction. In this section, we will try to explain our step-by-step solution of how we achieved results for bankruptcy prediction. We started with introducing the Taiwan stock exchange bankruptcy dataset and explaining the details of the dataset like features, instances,etc. Then, we dive deep into data pre-processing steps, where we state the problems present with the data like outlying data and data imbalance, and explain how we dealt with them. Next, we introduce the classification models we have considered and explain how we train our data using these models for getting an accurate model. Later, we analyse and evaluate the performance of these models using certain metrics like accuracy, precision and recall.

3.1 Dataset Description

The dataset is about bankruptcy prediction of taiwan companies.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.

The dataset is very apt for our research about bankruptcy prediction because it has highly useful econometric indicators as attributes (features) and comes with a huge number of samples of taiwan companies.

dataset characteristics	multivariate
no. of instances	6819
no. of attributes	96
associated tasks	classification
missing values	NA
date donated	28-06-2020
feature characteristics	real values

Table 1: dataset descriptions

The dataset contains of 96 features and 6820 instances. The features of the dataset are as follows:

Y - Bankrupt?: Class label

X1 - ROA(C) before interest and depreciation before interest: Return On Total Assets(C)

X2 - ROA(A) before interest and

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

As written above, there are 95 features labelled X1 through X95, and each feature is a synthetic feature. A synthetic feature is a combination of the econometric measures using arithmetic operations (addition, subtraction, multiplication, division). Each synthetic feature is as a single regression model that is developed in an evolutionary manner. The purpose of the synthetic features is to combine the econometric indicators proposed by the domain experts into complex features.

3.2 Dataset Quality Assessment

Before moving on to assessing the quality of our dataset, we first standardize our dataset as we can see that our dataset have large differences between their ranges which can hamper our process to build an accurate model. So, to prevent this problem, transforming features to comparable scales using

standardization is the solution. As we have mentioned earlier, the dataset suffers from outlying data and data imbalance.

3.2.1 Data Standardisation

3.2.2 Data Im-balancing

When we look at the numbers of Bankrupt class label, we can figure out that they are a minority when compare with the non bankrupt class label. In our dataset 96.77% companies are financially stable, that is they are not bankrupt. We can also see this through the bar graph provided below.

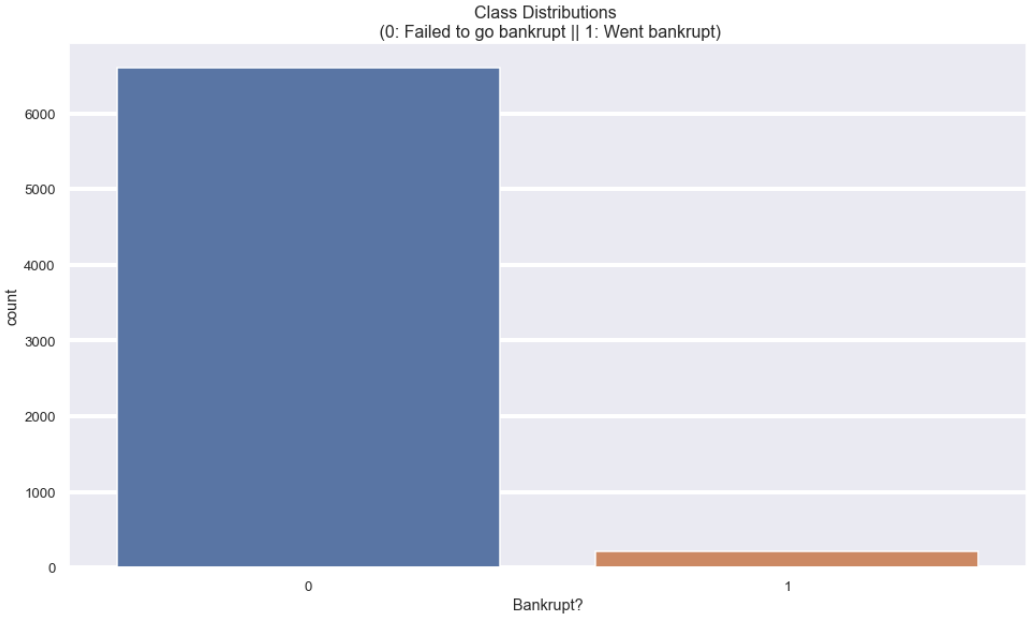

Figure 1: graph between no of bankrupt vs non bankrupt companies

We will use techniques like undersampling and SMOTE(systematic minority ovesampling technique) to deal with it.

3.2.3 Outlying Data

In our dataset we search for the outliers using the method of box plot. We make the box plot for all the data and then see what are the outlying data present. For example, we plot the graph of bankrupt or not with our feature variable X37 that is debt ratio and find the outliers. similarly we use this plot with feature variables such as X38 that is net worth assests and other interested variables.

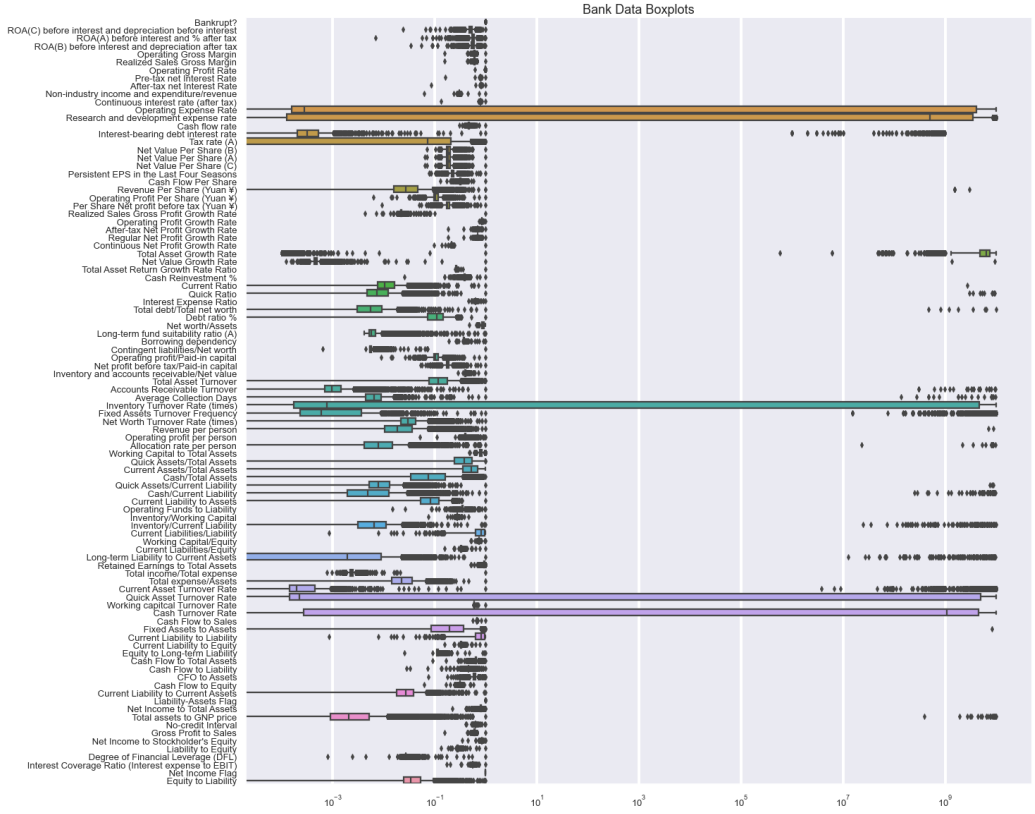

Figure 2: boxplot graph of the dataset

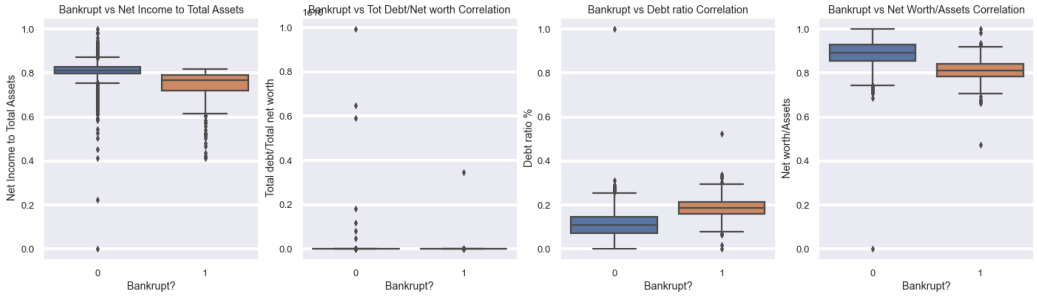

Figure 3: boxplot of responce variable with few interested variables

Figure 4: boxplot of responce variable with few interested variables

We also check the distribution of the interested variables to re check whether there are outliers in dataset or not. If the dataset has outliers then the distribution will not be normal.

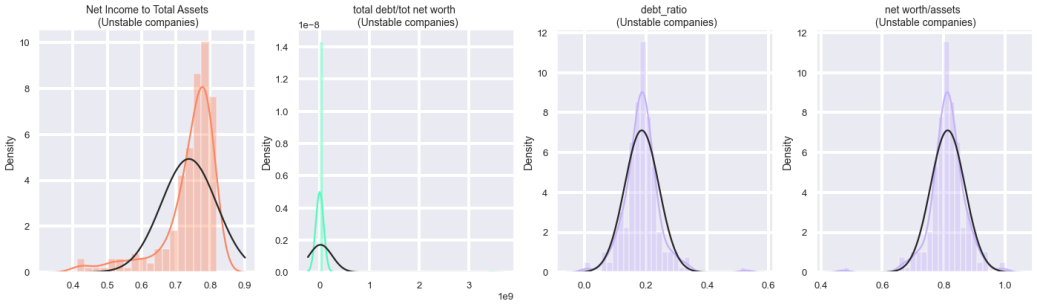

Figure 5: distribution graph of interested variables

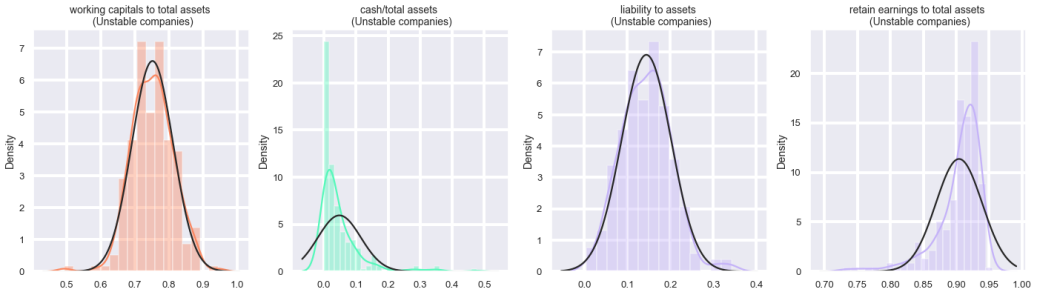

Figure 6: distribution graph of interested variables

Now we remove the outliers and again check the boxplot and again check the distribution functions of the interested variables.

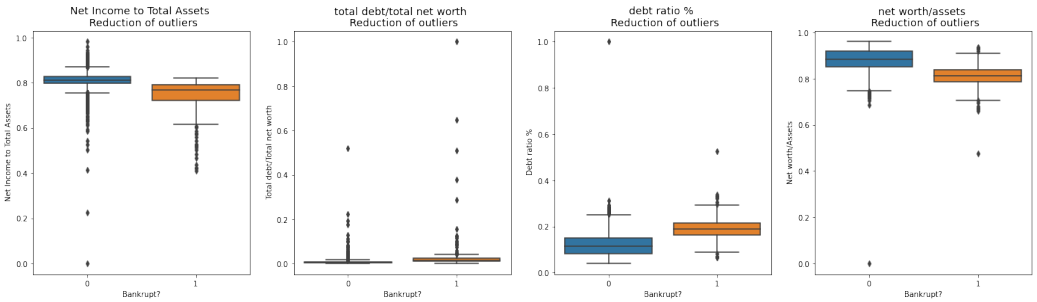

Figure 7: boxplot of variables after outlier removal

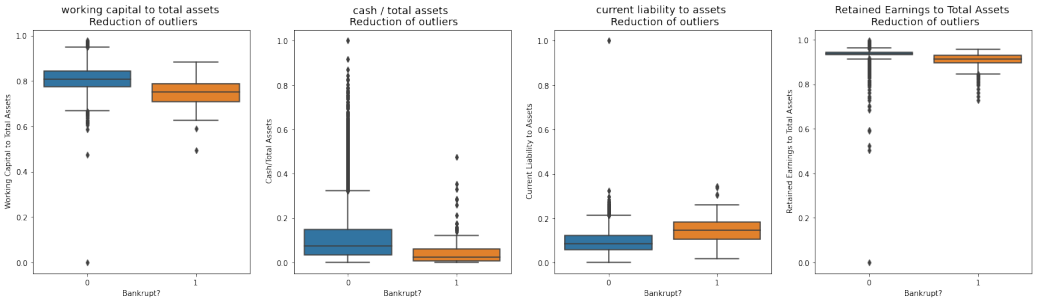

Figure 8: boxplot of variables after outlier removal

We check the distribution graphs of interested variables to see whether they show a normal distribution or not as the distribution should be normal if there are no outliers in the dataset.

Figure 9: distribution graph of variables after outlier removal

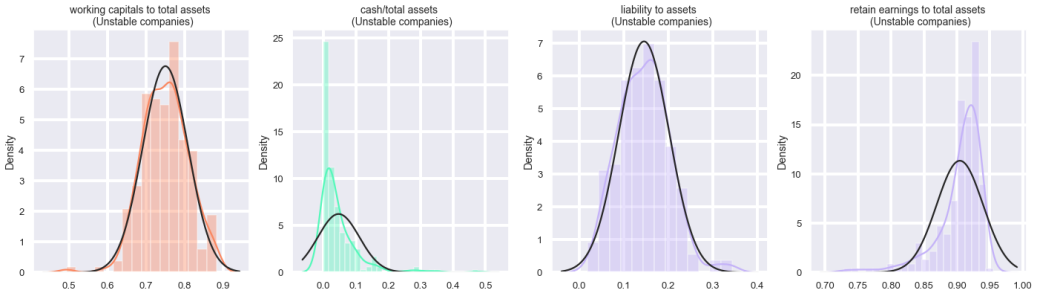

Figure 10: distribution graph of variables after outlier removal

We can observe that after the removal of outliers the graph has slightly shifted to the bell shape. At least for the few of the interested variables.

After doing the outlier removal we check the cardinality for all the variables present in our dataset.

3.2.4 MULTICOLLINEARITY

Next we move on to another important step of Model Adequacy Checking, that is, Checking Multicollinearity in our dataset. The basic assumption in multiple linear regression model is that the rank of the matrix of observations on explanatory variables is the same as the number of explanatory variables. In other words, such a matrix is of full column rank. This implies that all the explanatory variables are independent, i.e., there is no linear relationship among the explanatory variables.In this type of situation we say that the explanatory variables are orthogonal. But in real life situations due to many different kinds of reasons and conditions the variables depend on each other and are not independent and hence they have correlation between them. The situation where the explanatory variables are highly intercorrelated is referred to as **multicollinearity**.

In order to check the multicollinearity in our dataset we plot the spearman correlation heatmap. We can observe that darker the shade in the plot the higher is the correlation between those explanatory variables.

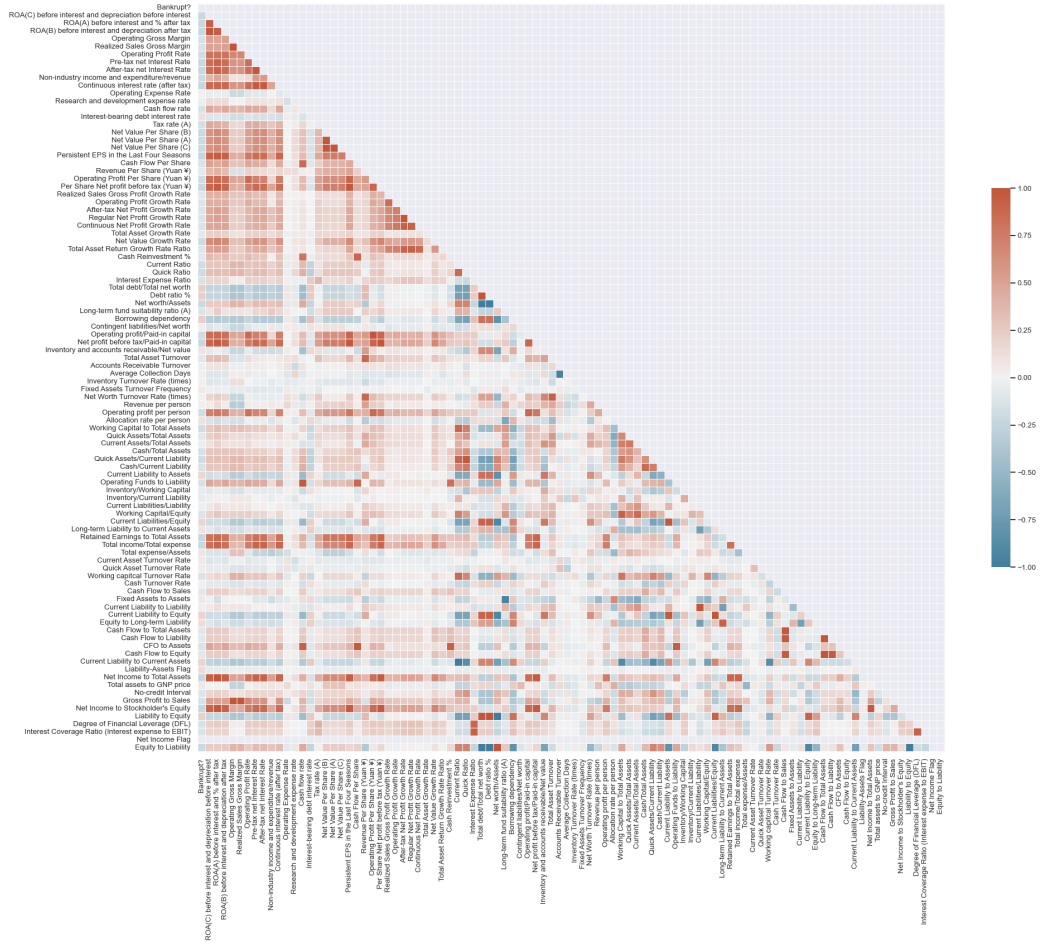

Figure 11: spearman correlation heatmap

4 DEALING WITH MUTLICOLLIENARITY AND DATA IMBALANCE

4.1 Dealing With Data Imbalance

As we have seen above that our dataset is highly imbalanced so we need to balance it using some techniques. Data Imbalance can be treated with Oversampling or Undersampling. In data analysis, Oversampling and Undersampling are roughly equivalent techniques of dealing with Data Imbalance, where they adjust the class distribution of a data set. Oversampling is

increasing the class distribution of the minority class label whereas Undersampling is decreasing the class distribution of the majority class label. In our project, we explored Synthetic Minority Oversampling Technique(SMOTE) and undersampling both to compare the results of both.

4.1.1 SMOTE or (Systematic Minority Oversampling Technique)

Synthetic Minority Oversampling Technique (SMOTE) is a widely used oversampling technique. For making this technique work we first take a training dataset which has s samples, and f features in the feature space of the data. As an example, let us consider a dataset of a health report for clarity. The feature space for the minority class for which we want to oversample could be cholestrol level, bone density, and eyesight. To oversample it we first take a sample from the dataset, and consider its k nearest neighbours in the feature space. To create a synthetic data point, take the vector between one of those k neighbours, and the current data point. Multiply this vector by a random number X which lies between 0, and 1. Adding this to the current data point will create the new synthetic data point. The imbalanced-learn library was used to implement SMOTE on our dataset. Synthetic Minority Oversampling Technique (SMOTE) is a widely used oversampling technique. To illustrate how this technique works consider some training data which has s samples, and f features in the feature space of the data. For simplicity, assume the features are continuous. As an example, let us consider a dataset of birds for clarity. The feature space for the minority class for which we want to oversample could be beak length, wingspan, and weight. To oversample, take a sample from the dataset, and consider its k nearest neighbors in the feature space. To create a synthetic data point, take the vector between one of those k neighbors, and the current data point. Multiply this vector by a random number x which lies between 0, and 1. Adding this to the current data point will create the new synthetic data point. SMOTE was implemented from the imbalanced-learn library. Before moving further we divided the dataset into two parts, namely, the training dataset and the testing dataset. We will build the model using the training dataset and will evaluate the performance of the model on the test dataset.

4.1.2 Undersampling

Undersampling is a technique to balance uneven datasets by keeping all of the data in the minority class and decreasing the size of the majority class. It is one of several techniques data scientists can use to extract more accurate information from originally imbalanced datasets. The difference between two or more classes is a class imbalance, and imbalanced classifications can be slight or severe. For example, the difference between two classes could be 3:1 or the difference could be 1000:1.

Undersampling techniques remove examples from the training dataset that belong to the majority class in order to better balance the class distribution, such as reducing the skew from a 1:100 to a 1:10, 1:2, or even a 1:1 class distribution. This is different from oversampling that involves adding examples to the minority class in an effort to reduce the skew in the class distribution The result of undersampling is a transformed data set with less examples in the majority class — this process may be repeated until the number of examples in each class is equal. Using this approach is effective in situations where the minority class has a sufficient amount of examples despite the severe imbalance. We have used undersampling on the whole dataset and then we have divided it into training and testing groups. If we divide the dataset before hand then the training data the class difference of the training data was getting lower than that of the testing data and ideally it should not happen.

4.2 Dealing With Multicollinearity

A lot of multicollinearity diagnostic approaches are present there, and each of them is based on a particular approach. In our case, we have used the approach "Variance Inflation Factor" **VIF** in order to detect the problem of multicollinearity. Based on this concept, the variance inflation factor for the j^{th} explanatory variable is defined as:

$$VIF_j = 1/(1 - R_j^2)$$

where R_j^2 denotes the coefficient of determination when X_j is regressed on the remaining $(k-1)$ variables excluding X_j .

In practice, usually, a VIF > 5 indicates that the associated regression coefficients are poorly estimated because of multicollinearity. Hence, in order to deal with it, we have used an iterative algorithm that drops variable

with highest VIF and then checks VIF again and then drop until VIF of all variables is less than 5. We have done VIF process on both our models which we obtained using the process of SMOTE and undersampling. The model which we have obtained using SMOTE is left with 58 variables after using iterative VIF method and the model which we obtained using Undersampling method is left with 27 variables after using the VIF method.

S.NO	SMOTE	Undersampling
1	Operating Gross Margin	Operating Expense Rate
2	Operating Expense Rate	Research and development expense rate
3	Research and development expense rate	Interest-bearing debt interest rate
4	Cash flow rate	Tax rate (A)
5	Interest-bearing debt interest rate	Revenue Per Share (Yuan ¥)
6	Tax rate (A)	Total Asset Growth Rate'
7	Net Value Per Share (C)	Net Value Growth Rate
8	Cash Flow Per Share	Current Ratio
9	Revenue Per Share (Yuan ¥)	Quick Ratio
10	Operating Profit Per Share (Yuan ¥)	Total debt/Total net worth
11	Realized Sales Gross Profit Growth Rate	Total debt/Total net worth
12	Operating Profit Growth Rate	Inventory Turnover Rate (times)
13	Regular Net Profit Growth Rate	Fixed Assets Turnover Frequency
14	Continuous Net Profit Growth Rate	Revenue per person
15	Total Asset Growth Rate	Cash/Total Assets
16	Net Value Growth Rate	Cash/Total Assets
17	Total Asset Return Growth Rate Ratio	Cash/Total Assets
18	Cash Reinvestment %	Long-term Liability to Current Assets
19	Current Ratio	Total expense/Assets
20	Quick Ratio	Current Asset Turnover Rate
21	Interest Expense Ratio	Current Asset Turnover Rate
22	Long-term fund suitability ratio (A)	Current Asset Turnover Rate
23	Contingent liabilities/Net worth	Current Asset Turnover Rate
24	Inventory and accounts receivable/Net value	Current Asset Turnover Rate
25	Accounts Receivable Turnover	Liability-Assets Flag
26	Average Collection Days	Total assets to GNP price'
27	Inventory Turnover Rate (times)	Degree of Financial Leverage (DFL)

Table 2: remaining variables after applying iterative VIF

S.NO	SMOTE	Undersampling
28	Fixed Assets Turnover Frequency	NA
29	Net Worth Turnover Rate (times)	NA
30	Revenue per person	NA
31	Operating profit per person	NA
32	Allocation rate per person	NA
33	Quick Assets/Total Assets	NA
34	Cash/Total Assets	NA
35	Quick Assets/Current Liability	NA
36	Cash/Current Liability	NA
37	Inventory/Working Capital	NA
38	Inventory/Current Liability	NA
39	Long-term Liability to Current Assets	NA
40	Total income/Total expense	NA
41	Total expense/Assets	NA
42	Current Asset Turnover Rate	NA
43	Quick Asset Turnover Rate	NA
44	Working capitcal Turnover Rate	NA
45	Cash Turnover Rate	NA
46	Fixed Assets to Assets	NA
47	Current Liability to Liability	NA
48	Cash Flow to Liability	NA
49	Cash Flow to Equity	NA
50	Current Liability to Current Assets	NA
51	Liability-Assets Flag	NA
52	Total assets to GNP price	NA
53	No-credit Interval	NA
54	Net Income to Stockholder's Equity	NA
55	Degree of Financial Leverage (DFL)	NA
56	Interest Coverage Ratio(Interest expense to EBIT)	NA
57	Equity to Liability'	NA
58	Total Asset Growth Rate	NA

Table 3: remaining variables after applying iterative VIF

5 VARIABLE SELECTION

Regression analysis depends on the explanatory variables present in the model. It is understood in the regression analysis that only correct and important explanatory variables appear in the model. In real life problems the analyst usually obtains a set of explanatory variables which may effect the experiment. Generally, all such candidate variables are not used in the regression modelling, but a subset of "best" explanatory variables is chosen from this pool.In our project, we have used two variable selection method, that is, Stepwise Regression and Lasso Regression.

5.1 Step-wise Regression

A combination of forward selection and backward elimination procedure is the stepwise regression. It is a modified version of forward selection procedure and has the following steps.

- We start with the intercept model and compute the AIC(Akikie Information Criteria) for the model.
- We then compute AIC for all the possibilities of adding one more variable in our intercept only model. We select the variable with the smallest AIC if it has a lower AIC than intercept only model.
- We then again compute AIC for adding one more variable in the model along with the AIC for removing the already added variable. We sort the values by ascending order and variables are either added or the existing variable is subtracted depending on the value of AIC.
- We continue performing these steps until any further action - addition or subtraction, results in increase of AIC of the model.

Therefore, we have selected variables on the basis of Akaikie Information Criterion(AIC).

For **UNDERSAMPLING dataset**, we have the following variables in our Model:

Research and development expense rate; Total expense Assets;Revenue Per Share Yuan A;Total debt Total net worth;Total Asset Growth Rate ;QuickAsset Turnover Rate;Cash Total Assets;Contingent liabilities Net worth ;Cash Turnover Rate ;Cash Turnover Rate ; Interest bearing debt interest rate; Fixed Assets Turnover Frequency ;Net Value Growth Rate ;Total assets to GNP price;Quick Ratio;Long term Liability to Current Assets;Revenue per person ;Inventory Current Liability ;Liability Assets Flag ;Tax rate A ;Average Collection Days;Cash Current Liability; Operating Expense Rate ;Current Asset Turnover Rate ;Current Liability to Current Assets;Current Ratio

For **SMOTE dataset**, we have the following variables in our Model:

Degree of Financial Leverage DFL.;Operating profit per person; Total Asset Growth Rate;Inventory and accounts receivable Net value; Long term Liability to Current Assets ;Working capitcal Turnover Rate ;Inventory Current Liability ;Total assets to GNP price ;Quick Ratio;Current Liability to Current Assets ;Operating Profit Per Share Yuan Â ;No credit Interval ;Interest Expense Ratio; Contingent liabilities Net worth ;Operating Profit Growth Rate;Revenue Per Share Yuan Â ;Revenue per person; Current Liability to Liability; Quick Asset Turnover Rate; Interest Coverage Ratio Interest expense to EBIT; Inventory Working Capital; Research and development expense rate; Continuous Net Profit Growth Rate;Quick Assets Current Liability;Net Value Growth Rate.

6 LOGISTIC REGRESSION

Since we have cleaned our dataset, now we are well set to define our model. As our response variable 'Y ' is a categorical variable with only two nominal categories, that is, 0 which indicates that there is no bankruptcy and 1 which indicates there is a bankruptcy. Logistic regression makes use of the canonical link function, $\ln p/1-p$. .The logistic regression model is given as:

$$Y_i \sim Binomial(N_i, p_i)$$

$$ln(p_i/(1-p_i)) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} = X\beta \text{ for i=1,2,3....}$$

where x_{ij} is the element in the i^{th} row and j^{th} column of the model matrix X. To evaluate the accuracy of the classification or the logistic model, we use a confusion matrix.

6.1 Confusion Matrix

Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model.The confusion matrix shows the ways in which your Classification model is confused when it makes predictions. A confusion matrix is a technique for summarizing the performance of a classification algorithm.Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset. There are four important components of confusion matrix:

- **True positives :** The cases in which we predicted YES and the actual output was also YES.
- **True Negatives :** The cases in which we predicted NO and the actual output was NO.
- **False Positives :** The cases in which we predicted YES and the actual output was NO.
- **False Negatives :** The cases in which we predicted NO and the actual output was YES.

6.2 Metrices of Confusion Matrix

6.2.1 Accuracy score

Accuracy score is a parameter which can tell the analyst whether a model is being correctly trained and how it may perform generally. It does not give a detailed information regarding its application to the problem, it merely answers our question that how often our model is correct. The major problem

with using accuracy as your main performance metric is that it does not do well when you have a severe class imbalance.

$$Accuracy = \frac{no.\ of\ correct\ predictions}{total\ no\ of\ predictions\ made}$$

6.2.2 Precision Score

It is the number of correct positive results divided by the number of positive results predicted by the classifier. Precision talks about how precise/accurate your model is out of those predicted positive, how many of them are actual positive.Precision is a good measure to determine, when the costs of False Positive is high.

$$precision = \frac{TRUE\ POSITIVE}{TRUE\ POSITIVE + FALSE\ POSITIVE}$$

6.2.3 Recall score

It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).It is also known as the sensitivity. Recall actually calculates how many of the Actual Positives our model capture through labelling it as Positive (True Positive).

$$recall = \frac{TRUE\ POSITIVE}{TRUE\ POSITIVE + FALSE\ NEGATIVE}$$

6.2.4 Specificity

Specificity, also known as the true negative rate(TNR), measures the proportion of actual negatives that are correctly identified as such. It is the opposite of the recall.

$$recall = \frac{TRUE\ NEGATIVE}{TRUE\ NEGATIVE + FALSE\ POSITIVE}$$

6.2.5 F1 score

The F1 Score is a measure of a test's accuracy, that is, it is the harmonic mean of precision and recall. It can have a maximum score of 1 and a minimum of 0. Overall, it is a measure of the preciseness and robustness of the model.

$$F1\ Score = \frac{2(PRECISION\ \times\ RECALL)}{PRECISION\ +\ RECALL}$$

$$= \frac{2{\cdot}TRUE\ POSITIVE}{2{\cdot}TRUE\ POSITIVE\ +\ FALSE\ POSITIVE\ +\ FALSE\ NEGATIVE}$$

6.2.6 Precision Recall Curve

The precision-recall curve is constructed by calculating and plotting the precision against the recall for a single classifier at a variety of thresholds. For example, if we use logistic regression, the threshold would be the predicted probability of an observation belonging to the positive class.

6.2.7 ROC curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate. The ROC curve is produced by calculating and plotting the true positive rate against the false positive rate for a single classifier at a variety of thresholds. For example, in logistic regression, the threshold would be the predicted probability of an observation belonging to the positive class.

6.3 Model Building And Comparision