Introduction

Data from the Taiwan Economic Journal for the years 1999–2009 representing company bankruptcy based on the business regulations of the Taiwan Stock Exchange.

The data description is shown in here. The outcome of this machine learning problem is binary it means there are two outcomes. Figure.1 (High resolution) represents the histogram of all features just to show their distribution. As seen in the histogram the data has skewed distribution and many data outliers that's why this data is interesting to me. I am going to examine different techniques of improving a model to see the result.

For better understanding of the outlier data Figure.2 can be helpful. Figure.2 shows box plot of all the data and the outliers are shown outside the IQR range.

The models the data is going to be trained on are SVM, Decision Tree, Neural network and K-Nearest Neighbor also we are going to boost the Decision tree with AdaBoosting. The models will be trained on both processed and original data to see the difference.

I have made a function to evaluate the models using scikit-learn module. And the results of each model will be compared using that function. The function gives train accuracy, test accuracy and prediction latency.

First we eliminate the outliers and also apply a method called SMOTE -Synthetic Minority Oversampling Technique – for normalizing the data. Figure.3 is a box plot showing the result of our augmentations.

We keep the original data to see the difference in model results.

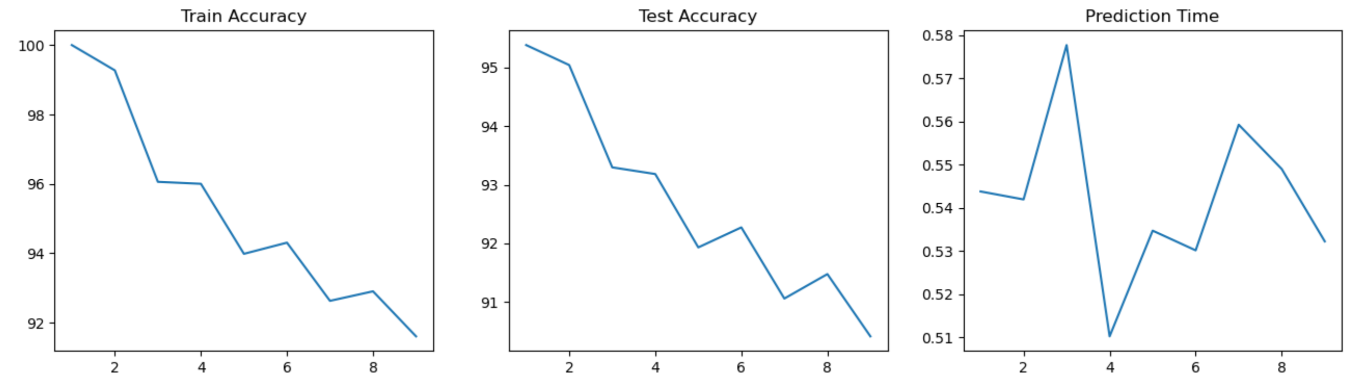
KNN

The first model is KNN and since we have to provide a k for the training we are going to give a range of k from 1 to 10 and test the results. As we can see the accuracy drops with higher k and the best test accuracy is in k=1.

Best result:

The Training Accuracy of the KNeighborsClassifier is 0.9997158552756203

The Testing Accuracy of the KNeighborsClassifier is 0.9753787878787878

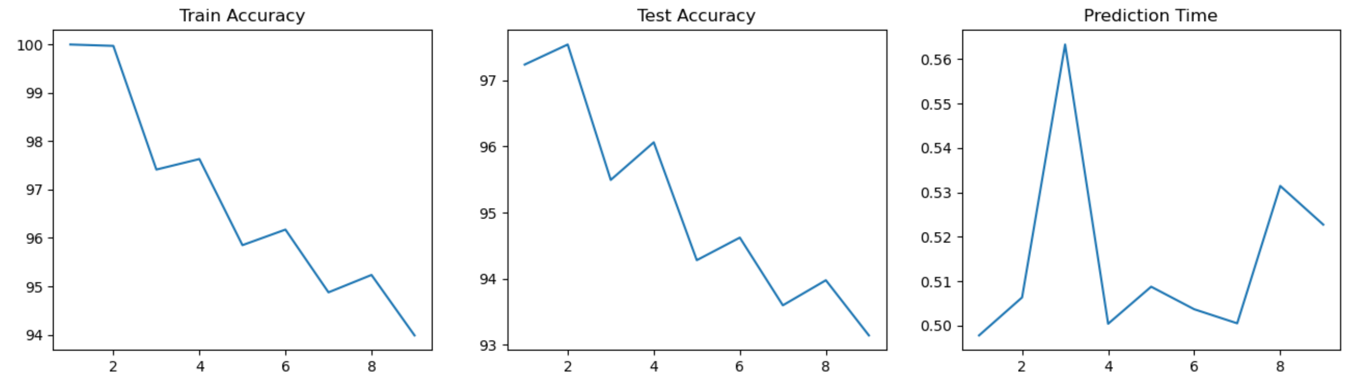


And these plots show the same model with processed data is the same training model.

Best result:

The Training Accuracy of the KNeighborsClassifier is 0.9927069520742565

The Testing Accuracy of the KNeighborsClassifier is 0.9503787878787879

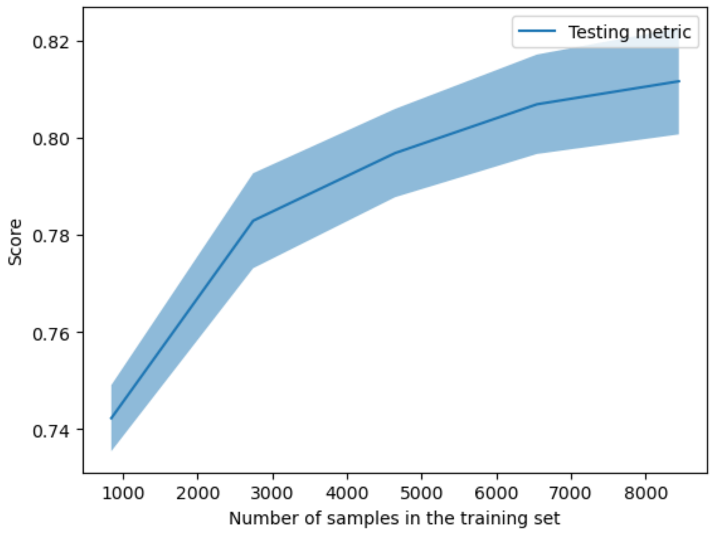


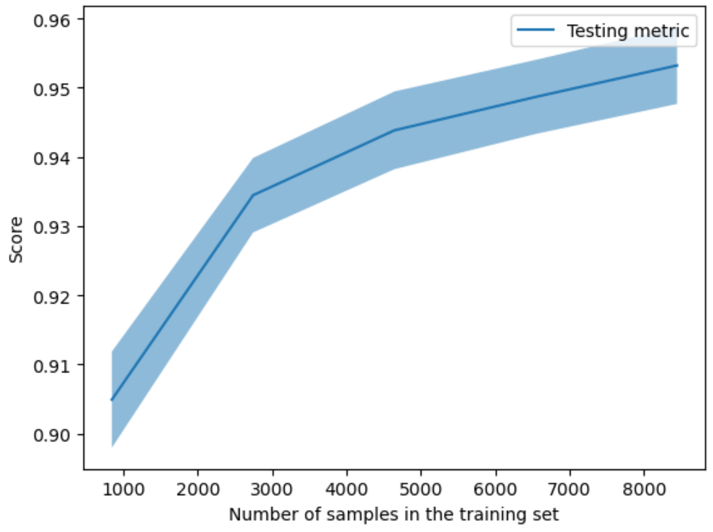
With k =1 the training accuracy becomes 100 which is understandable, but test accuracy is lower because it cannot generalize well.

The highest testing accuracy over all is achieved by k = 2 in the processed data which shows that normalizing data helps a little.

SVM

Next, we train our Support vector model with both data and plot the learning curve. Let's see the results first:





Original Data

```
The Training Accuracy of the SVC is   0.788122750520932
The Testing Accuracy of the SVC is    0.8125
function timer completed in           1.58321022987  s
```

Processed Data

```
The Training Accuracy of the SVC is   0.961356317484372
The Testing Accuracy of the SVC is    0.9549242424242425
function timer completed in           2.349263668 s
```

As we can see the normalization of the data improved the model significantly, also shown in the learning curve the more we have samples the better the performance. I guess this is due to the high number of features.

Decision Tree

The results of decision tree training for the original data is :

The Training Accuracy of the DecisionTreeClassifier is 1.0

The Testing Accuracy of the DecisionTreeClassifier is 0.9462121212121212

function timer completed in 0.003997564315 seconds

And for the processed data**:**

The Training Accuracy of the DecisionTreeClassifier is 1.0

The Testing Accuracy of the DecisionTreeClassifier is 0.9587121212121212

function timer completed in 0.00599765 seconds

As we learned decision trees are not affected by outliers and the result didn't change even after we removed the outliers.

BOOSTING

After feeding out model to Ada boosting with different number of weak learners and different learning rates the accuracy of the models only improved by 1 percent while learning rate was 0.2 and the number of weak learners was 100.

Since the model didn't improve by Ada boost I tried Gradient Boosting with different depth and I could improve the model to 98% !! The first try with depth of 1 didn't improve the model but as the depth grow so did the model accuracy.

The training time is high(almost 1 minute) but the accuracy makes up for it. Results for max_depth = 4 are :

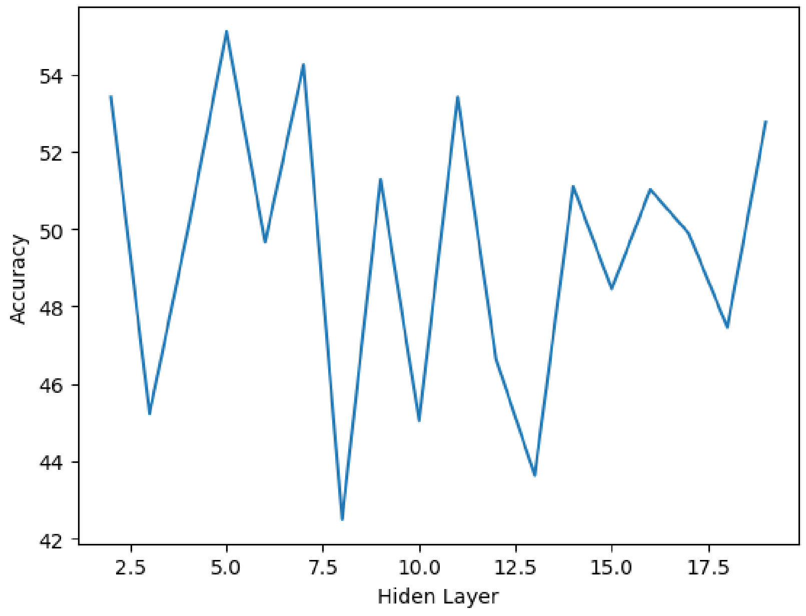
The Training Accuracy of the GradientBoostingClassifier is 1.0

The Testing Accuracy of the GradientBoostingClassifier is 0.9803030303030303

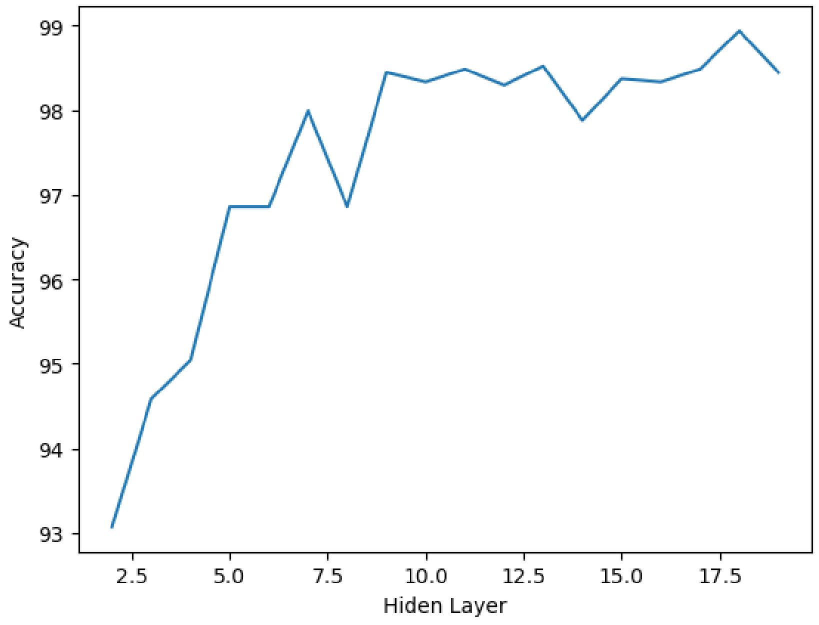
function timer completed in 0.015996694564819336 seconds

Neural Network

Last model is neural network using lbfgs and relu with different layers to test which one can be better for our results:



The model is not good at all with very low accuracy. Then I trained the model with the processed data:



The data that has been normalized has a lot of impact on the result with accuracy as high as 99% we can see the importance of feature engineering for neural networks.

The Training Accuracy of the MLPClassifier is	1.0
The Testing Accuracy of the MLPClassifier is	0.98939393939394
function timer completed in	0.0050039294 seconds

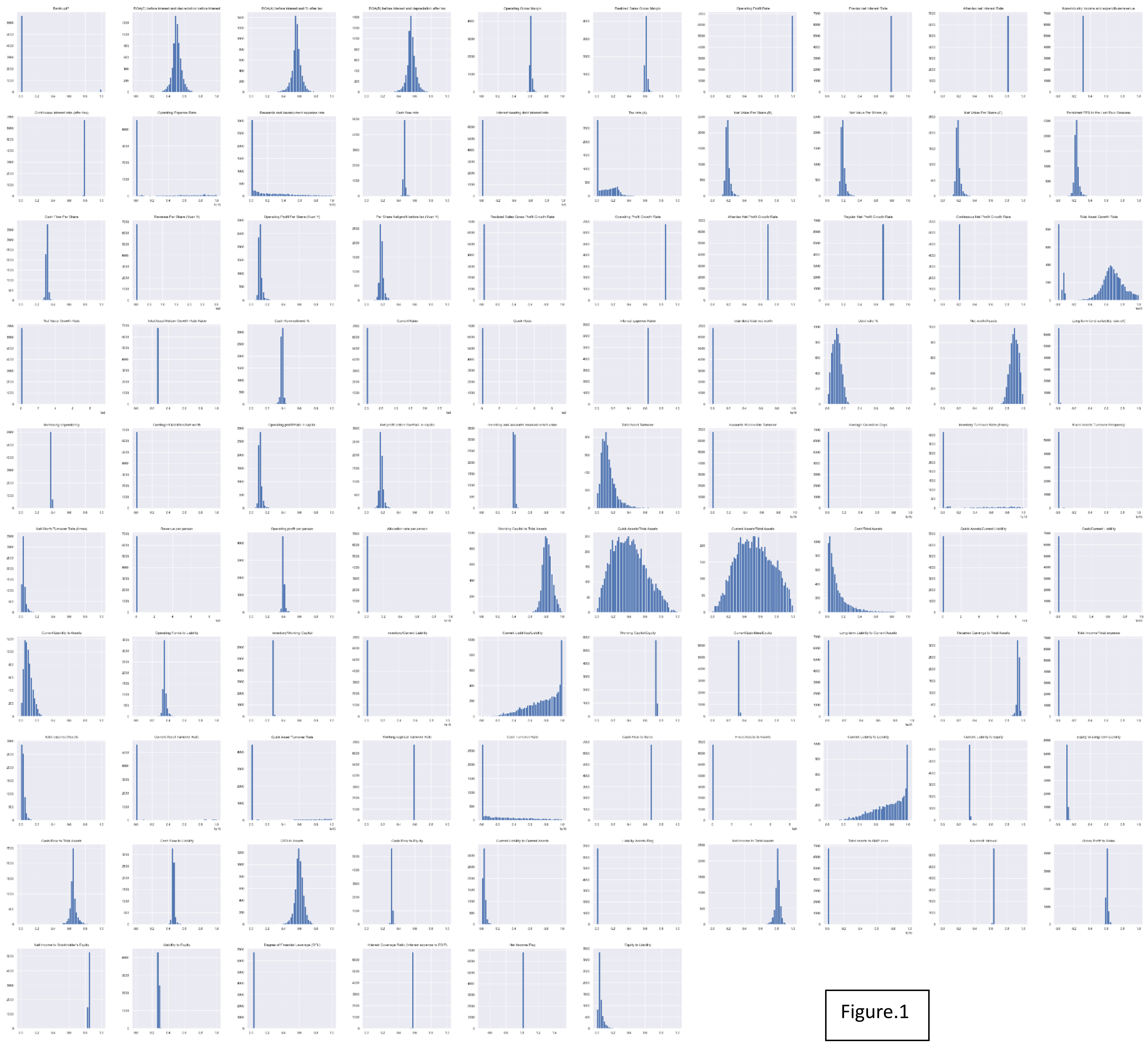

Bankrupt?
ROA(C) before interest and depreciation before interest
ROA(A) before interest and % after tax
ROA(B) before interest and depreciation after tax
Operating Gross Margin
Realized Sales Gross Margin
Operating Profit Rate
Pre-tax net Interest Rate
After-tax net Interest Rate
Non-industry income and expenditure/revenue
Continuous interest rate (after tax)
Operating Expense Rate
Research and development expense rate
Cash flow rate
Interest-bearing debt interest rate
Tax rate (A)
Net Value Per Share (B)
Net Value Per Share (A)
Net Value Per Share (C)
Persistent EPS in the Last Four Seasons
Cash Flow Per Share
Revenue Per Share (Yuan ¥)
Operating Profit Per Share (Yuan ¥)
Per Share Net profit before tax (Yuan ¥)
Realized Sales Gross Profit Growth Rate
Operating Profit Growth Rate
After-tax Net Profit Growth Rate
Regular Net Profit Growth Rate
Continuous Net Profit Growth Rate
Total Asset Growth Rate
Net Value Growth Rate
Total Asset Return Growth Rate Ratio
Cash Reinvestment %
Current Ratio
Quick Ratio
Interest Expense Ratio
Total debt/Total net worth
Debt ratio %
Net worth/Assets
Long-term fund suitability ratio (A)
Borrowing dependency
Contingent liabilities/Net worth
Operating profit/Paid-in capital
Net profit before tax/Paid-in capital
Inventory and accounts receivable/Net value
Total Asset Turnover
Accounts Receivable Turnover
Average Collection Days
Inventory Turnover Rate (times)
Fixed Assets Turnover Frequency
Net Worth Turnover Rate (times)
Revenue per person
Operating profit per person
Allocation rate per person
Working Capital to Total Assets
Quick Assets/Total Assets
Current Assets/Total Assets
Cash/Total Assets
Quick Assets/Current Liability
Cash/Current Liability
Current Liability to Assets
Operating Funds to Liability
Inventory/Working Capital
Inventory/Current Liability
Current Liabilities/Liability
Working Capital/Equity
Current Liabilities/Equity
Long-term Liability to Current Assets
Retained Earnings to Total Assets
Total income/Total expense
Total expense/Assets
Current Asset Turnover Rate
Quick Asset Turnover Rate
Working capitcal Turnover Rate
Cash Turnover Rate
Cash Flow to Sales
Fixed Assets to Assets
Current Liability to Liability
Current Liability to Equity
Equity to Long-term Liability
Cash Flow to Total Assets
Cash Flow to Liability
CFO to Assets
Cash Flow to Equity
Current Liability to Current Assets
Liability-Assets Flag
Net Income to Total Assets
Total assets to GNP price
No-credit Interval
Gross Profit to Sales
Net Income to Stockholder's Equity
Liability to Equity
Degree of Financial Leverage (DFL)
Interest Coverage Ratio (Interest expense to EBIT)
Net Income Flag
Equity to Liability

Figure.1

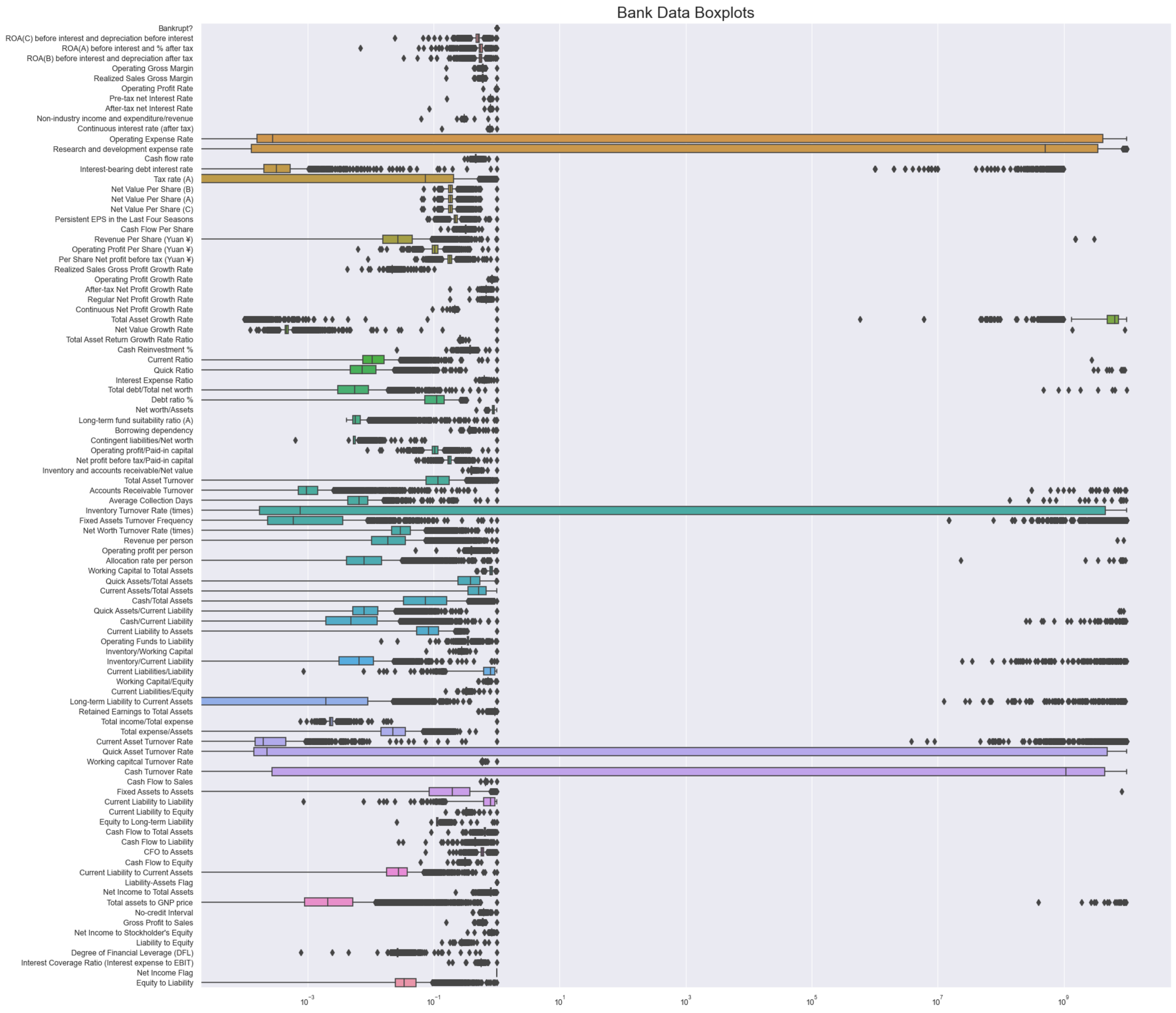

Bank Data Boxplots
Bankrupt?
ROA(C) before interest and depreciation before interest
ROA(A) before interest and % after tax
ROA(B) before interest and depreciation after tax
Operating Gross Margin
Realized Sales Gross Margin
Operating Profit Rate
Pre-tax net Interest Rate
After-tax net Interest Rate
Non-industry income and expenditure/revenue
Continuous interest rate (after tax)
Operating Expense Rate
Research and development expense rate
Cash flow rate
Interest-bearing debt interest rate
Tax rate (A)
Net Value Per Share (B)
Net Value Per Share (A)
Net Value Per Share (C)
Persistent EPS in the Last Four Seasons
Cash Flow Per Share
Revenue Per Share (Yuan ¥)
Operating Profit Per Share (Yuan ¥)
Per Share Net profit before tax (Yuan ¥)
Realized Sales Gross Profit Growth Rate
Operating Profit Growth Rate
After-tax Net Profit Growth Rate
Regular Net Profit Growth Rate
Continuous Net Profit Growth Rate
Total Asset Growth Rate
Net Value Growth Rate
Total Asset Return Growth Rate Ratio
Cash Reinvestment %
Current Ratio
Quick Ratio
Interest Expense Ratio
Total debt/Total net worth
Debt ratio %
Net worth/Assets
Long-term fund suitability ratio (A)
Borrowing dependency
Contingent liabilities/Net worth
Operating profit/Paid-in capital
Net profit before tax/Paid-in capital
Inventory and accounts receivable/Net value
Total Asset Turnover
Accounts Receivable Turnover
Average Collection Days
Inventory Turnover Rate (times)
Fixed Assets Turnover Frequency
Net Worth Turnover Rate (times)
Revenue per person
Operating profit per person
Allocation rate per person
Working Capital to Total Assets
Quick Assets/Total Assets
Current Assets/Total Assets
Cash/Total Assets
Quick Assets/Current Liability
Cash/Current Liability
Current Liability to Assets
Operating Funds to Liability
Inventory/Working Capital
Inventory/Current Liability
Current Liabilities/Liability
Working Capital/Equity
Current Liabilities/Equity
Long-term Liability to Current Assets
Retained Earnings to Total Assets
Total income/Total expense
Total expense/Assets
Current Asset Turnover Rate
Quick Asset Turnover Rate
Working capitcal Turnover Rate
Cash Turnover Rate
Cash Flow to Sales
Fixed Assets to Assets
Current Liability to Liability
Current Liability to Equity
Equity to Long-term Liability
Cash Flow to Total Assets
Cash Flow to Liability
CFO to Assets
Cash Flow to Equity
Current Liability to Current Assets
Liability-Assets Flag
Net Income to Total Assets
Total assets to GNP price
No-credit Interval
Gross Profit to Sales
Net Income to Stockholder's Equity
Liability to Equity
Degree of Financial Leverage (DFL)
Interest Coverage Ratio (Interest expense to EBIT)
Net Income Flag
Equity to Liability
10^{-3}
10^{-1}
10^{1}
10^{3}
10^{5}
10^{7}
10^{9}

Figure.2

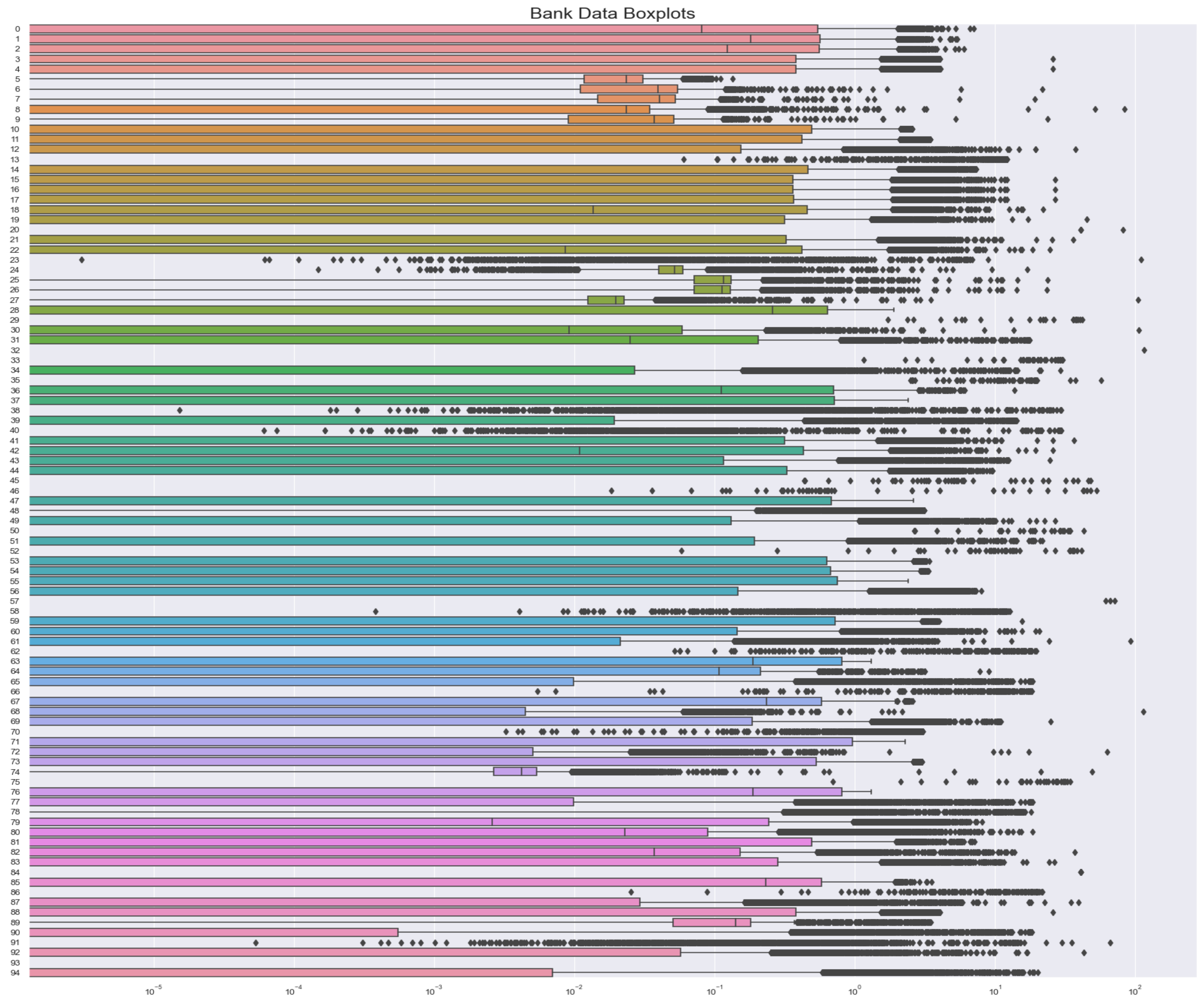

Bank Data Boxplots
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

Figure.3

References :

https://www.kaggle.com/code/gcmadhan/bankruptcy-perdiction-96-accuracy