**Company Bankruptcy Prediction**

**CUNY School of Professional Studies**

**Data 698 Final Project**

**Professor Nasrin Khansari**

1. **Abstract**

Effective bankruptcy prediction is critical for financial institutions to make appropriate lending decisions. In general, the input variables (or features), such as financial ratios, and prediction techniques, such as statistical and machine learning techniques, are the two most important factors affecting the prediction performance. Bankruptcy prediction is important for financial institutions to make proper business decisions.

The input variables such as financial ratios and prediction techniques such as statistical and machine learning techniques are the two most important factors affecting the prediction performance. In this paper, we will be analyzing which different categories of financial ratios are of the most importance in predicting bankruptcy.

1. **Data Source**

In this study, data were collected from the Taiwan Economic Journal for the years 1999–2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange. In addition, there were two criteria used in collecting the data samples. First, the sample companies had to have at least three years of complete public information before the occurrence of the financial crisis. Second, there should be a sufficient number of comparable companies of similar size in the same industry for comparison of the bankrupt and non-bankrupt cases. The resultant sample includes companies from the manufacturing industry composed of industrial and electronics companies (346 companies), the service industry composed of shipping, tourism, and retail companies (39 companies), and others (93 companies), but not financial companies.

It should be noted that if there is a significant difference between the number of bankrupt and non-bankrupt cases, this results in a class imbalance problem, which is likely to lead to a degradation in the final prediction performance. Therefore, we use the method of stratified sampling (Altman, 1968) to collect the same number of bankrupt and non-bankrupt cases.

1. **Introduction**

Bankruptcy or business failure can have a negative impact in both the industry and the global economy. Many business leaders, researchers, investors and governments have studied many ways to identify the risk of bankruptcy failures in business. Bankruptcy prediction is important to the companies itself because the business leaders can utilize the data that is driving their companies down and try to avoid it. Financial ratios can be classified into seven categories: solvency, profitability, cash flows ratios, capital structure ratios, turnover ratios, growth, and etc. 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.

1. **Literature Review**

Accounting-based vs. market-price indicators.

There have been several studies comparing the relative importance of accounting-based ratios and market-price variables. Some experimental studies indicate that market-price variables have stronger predictive power in bankruptcy analysis. However, the evidence is mixed. Shumway (2001) suggests a model based on both accounting-based and market-price variables that produces quite accurate forecasts. On the other hand, Hillegeist et al. (2004) evaluate whether Altman’s (1968) Z-Score and Ohlson’s (1980) O-Score effectively summarize publicly available information about the probability of bankruptcy. They conclude that market-based probability estimates of corporate bankruptcy (based on the Black-Scholes-Merton option-pricing model) are superior, even recommending that future research should focus exclusively on market-price indicators.

Alternative bankruptcy predictors.

A smaller number of studies have investigated other potentially important bankruptcy predictors, including corporate governance indicators (such as stockholder concentration/structure), analyst estimates/forecasts, credit ratings changes, macroeconomic factors, and other industry and firm-specific factors (Jones et al. 2015, 2017).

The use of accounting-based measures (such as financial ratios) to predict bankruptcy

has a long history (Beaver 1966; Altman 1968; Zmijewski 1984). Beaver et al. (2005) observe that accounting-based predictors have proven remarkably robust over a 40-year period (1962–2002). Across many empirical studies, a range of accounting based measures have been shown to have predictive power in corporate bankruptcy. However, measures associated with working capital, cash flow, earnings, and leverage have surfaced as key predictors of firm financial distress in many studies (Altman 2002; Beaver et al. 2005; Jones and Hensher 2008).

As noted by Beaver et al. (2005, 95–96): The precise combination of ratios used

seems to be of minor importance with respect to overall predictive power, because the explanatory variables are correlated. However, conventional bankruptcy models such as logit/probit and LDA are low dimensional models; that is, they are severely limited by the multicollinearity condition (and other statistical problems), which restricts the number of input variables that can be tested in the model. These models have limited capacity to extract signals from other potentially important variables and related interaction effects. The gradient boosting model enables the testing of a much wider range of financial ratios, irrespective of their correlation with other variables and without any cost to model stability or performance. Other three models, Support Vector classifiers, K nearest neighbors and logistic regression models would be used to observe as well.

Corporate governance indicators.

There are some studies where Corporate governance indicators are included alongside financial ratios to increase the predictive power in bankruptcy analysis. Many corporate governance indicators (CGIs) have been identified in the literature which have been used for solving bankruptcy or financial crisis problems. These can be classified into five categories including board structure, ownership structure, cash flow rights, key persons retained, and others. However not all CGIs used for bankruptcy analysis in related works are the same. For example, Lee and Yeh (2004) used 6 FRs relating to solvency, profitability, and other categories and 10 CGIs in the board structure and ownership categories. What they found was that the model performance could be enhanced by using a combination of CGIs and FRs. However, in this research project, only financial ratios will be analyzed to determine the bankruptcy predictors.

1. **Methods**

Similar to feature selection, there are many common and well-known techniques which can be employed to develop prediction models. In this project, four techniques are compared, Logistic Regression, Support Vector Classifier, Gradient Boosted Classifiers, and K- means Nearest Neighbors. The kernel selected for Support Vector Classifier will be linear.

**5.1) Predictive Models**

1. **Logistic Regression**

Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature. Logistic regression is most commonly used when the data in question has binary output, so when it belongs to one class or another, or is either a 0 or 1.

1. **Support Vector Classifiers**

Support vector classifiers (SVCs) are a set of supervised learning methods used for classification, regression and outlier detection. The advantages of support vector classifiers are effective in high dimensional spaces, still effective in cases where number of dimensions is greater than the number of samples, uses a subset of training points in the decision function (called support vectors), so it is also memory efficient, and versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector classifiers include if the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial and SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).

1. **Gradient Boosted Classifiers**

Gradient Boosted Classifiers build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. When combined, these many weak successive trees produce a powerful “committee” that are often hard to beat with other algorithms. The advantages of GBM are that it provides predictive accuracy. Boosting works by adding new models to the ensemble sequentially. At each particular iteration, a new weak, base-learner model is trained with respect to the error of the whole ensemble learnt so far. GBM starts with base-learning models, then training weak models and followed up with sequential training with respect to errors. The following is explained more in detail on the process.

Much bankruptcy research has relied on parametric models, such as multiple discriminant analysis and logit, which can only handle a finite number of predictors (Altman in The Journal of Finance 23 (4), 589–609, 1968; Ohlson in Journal of Accounting Research 18 (1), 109–131, 1980). The gradient boosting model is a statistical learning method that overcomes this limitation. The model accommodates very large numbers of predictors which can be rank ordered, from best to worst, based on their overall predictive power (Friedman in The Annals of Statistics 29 (5), 1189–1232, 2001; Hastie et al. 2009).

While unscaled measures can introduce significant heteroscedasticity into the dataset, the gradient boosting model is largely insensitive to the shape and structure of data. Outliers do not affect the analysis. (Friedman 2001).

1. **K means nearest neighbors**

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

The KNN Algorithm first loads the data and initializes K to your chosen number of neighbors. For each example in the data, it calculates the distance between the query example and the current example from the data and adds the distance and the index of the example to an ordered collection. Then it sorts the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances. Then it picks the first K entries from the sorted collection and returns the labels of the selected K entries. If regression, it returns the mean of the K labels. If classification, it returns the mode of the K labels.

**5.2) Results and Discussion**

A brief summary of the dataset, all features are numeric, one feature has 0 variance that is constant throughout, major features are in the range of 0-1, and there are outlier infected features. However, there is a severe class imbalance of bankruptcy rate of the samples. There are 220 counts of bankruptcy to 6599 counts of non bankruptcy. To solve this, SMOTE will be used which oversamples the minority class, the bankrupt counts in this case. It duplicates examples in the minority class, yet they won't add any new information to the models. The new examples can be synthesized from existing examples.

Multicollinearity is also used to check if there is perfect correlation between any of the columns. If the correlation is more than 0.95 or less than -0.95 then remove one of the columns to avoid multicollinearity in the dataset. A total of 17 columns dropped as a result of this.

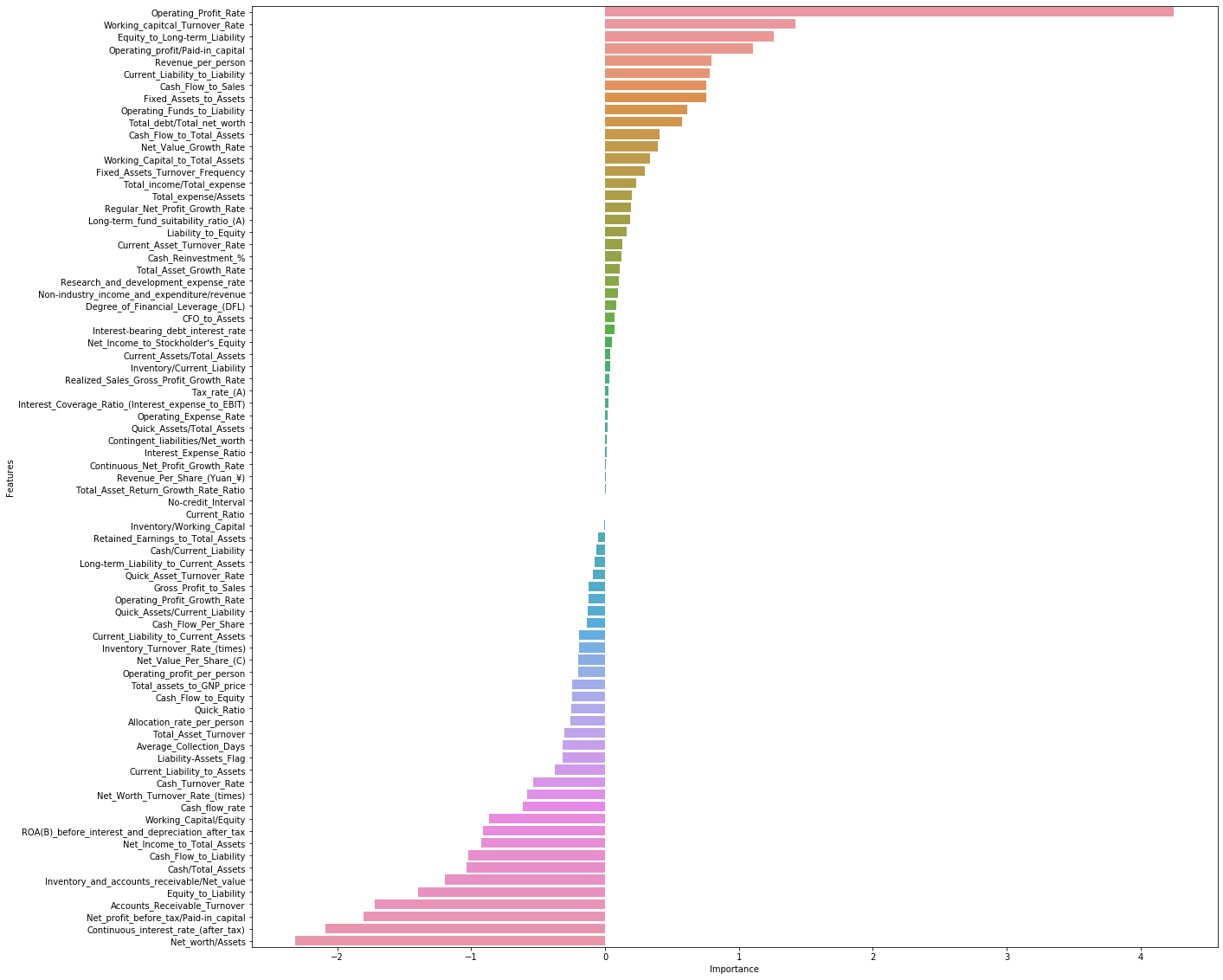
Next, split the dataset into a training set of 0.70 and testing set of 0.30. Then normalize the dataset with the scalar function. Using GridSearchCV, identify the optimal parameters to use for the four models, Logistic regression, Support Vector Classifiers, Gradient Boosted Classifiers, and K nearest neighbors (see following Table).

|  |  |  |  |
| --- | --- | --- | --- |
|  | model | best\_score | best\_params |
| 0 | svm | 0.996515 | {'C': 20, 'kernel': 'rbf'} |
| 1 | GBM | 0.995302 | {'learning\_rate': 0.1, 'max\_depth': 8, 'n\_estimators': 100, 'random\_state': 100, 'subsample': 0.85} |
| 2 | logistic\_regression | 0.920443 | {'C': 5, 'solver': 'liblinear'} |
| 3 | KNN | 0.999848 | {'algorithm': 'auto', 'n\_neighbors': 5} |

1. **Logistic Regression**

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 1759 | 221 |
| 1 | 148 | 1832 |

**Figure 1.**

Logistic Regression shows that Operating Profit Rate is a top bankruptcy predictor. This is true because if a company is not making profit it is bound to become bankrupt in the future. There are also negative importance values shown for this model that shows those features have negative importance in predicting bankruptcy.

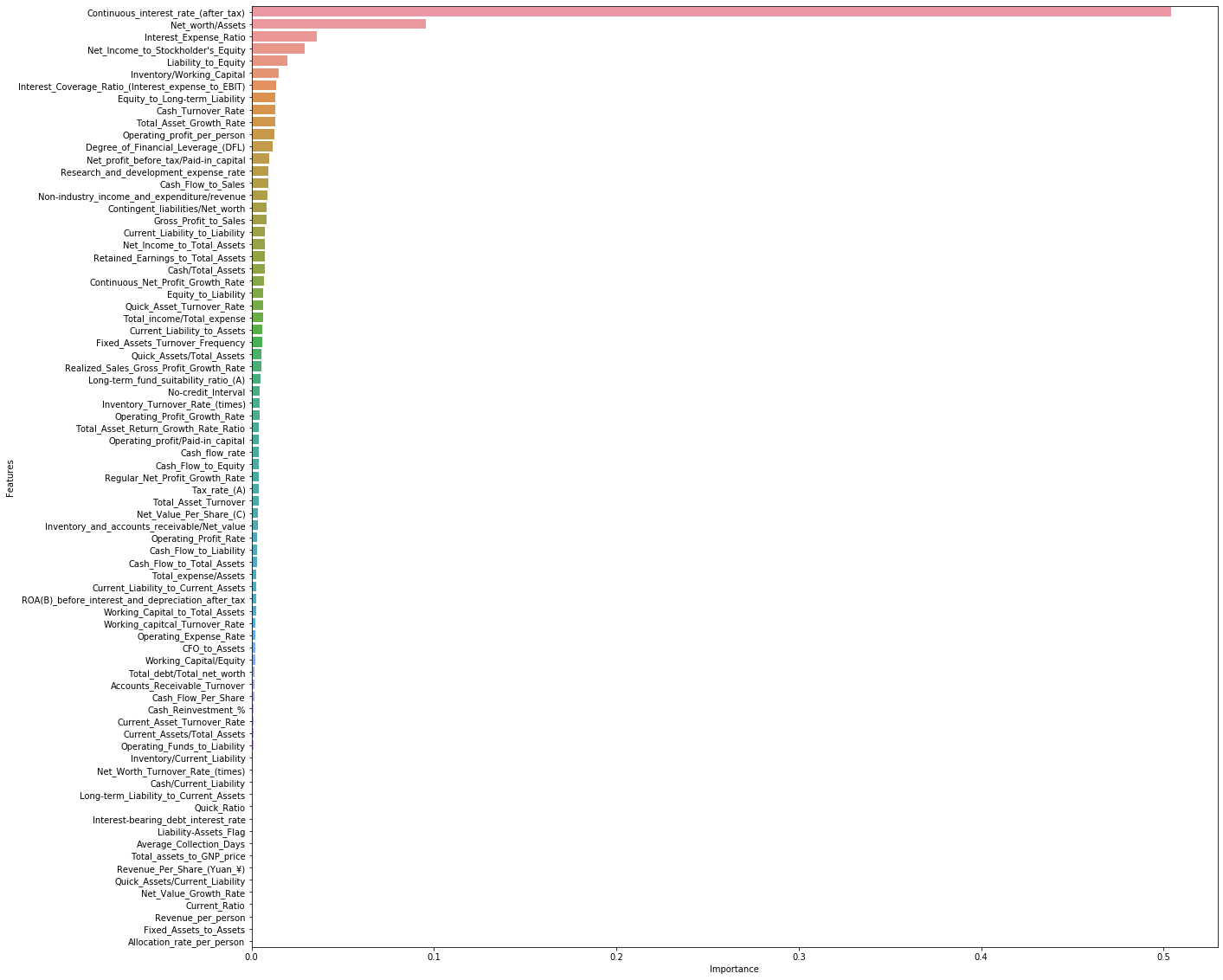
1. **Support Vector Classifiers**

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 1899 | 81 |
| 1 | 8 | 1972 |

1. **Gradient Boosted Classifiers**

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 1917 | 63 |
| 1 | 14 | 1966 |

**Figure 2.**



1. **K-nearest neighbors**

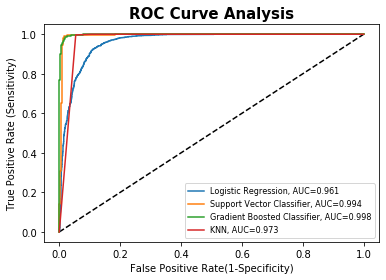
|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 1726 | 254 |
| 1 | 0 | 1980 |

1. **Conclusions**

Out of the four models used were Gradient Boosted Classifiers which performed best with AUC of 99.8%, 98% accuracy, 97% recall for 0: Fin. Stable and 99% for 1: Fin. Unstable. The GBC model has a 99.8% chance to distinguish between positive class and negative class. Continuous interest rate (after tax) is a big predictor for company bankruptcy because if a company has high debt and interest rates it becomes harder to stay afloat. Networth/asset is of second importance in predicting company bankruptcy.

KNN seems like the worst model in distinguishing between positive class and negative class of the bankruptcy predictors.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | AUC | Accuracy | Recall of 0: Fin. Stable | Recall of 1: Fin. Unstable |
| Logistic Regression | 0.961 | 0.91 | 0.89 | 0.93 |
| Support Vector Classifier | 0.994 | 0.98 | 0.96 | 1.00 |
| Gradient Boosted Classifiers | 0.998 | 0.98 | 0.97 | 0.99 |
| KNN | 0.973 | 0.94 | 0.87 | 1.00 |



This graph is a comparison of roc curve analysis of the three models, Logistic Regression, Support Vector Classifier k-nearest neighbors and Gradient Boosted Classifier.

1. **Code**

<https://github.com/Sizzlo/Data-698/blob/main/Data698BankruptcyPredictor.ipynb>

1. **References**

<https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction>

Liang, D., Lu, C.-C., Tsai, C.-F., and Shih, G.-A. (2016) Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study. European Journal of Operational Research, vol. 252, no. 2, pp. 561-572.

Liang, Deron, et al. “Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study.” *European Journal of Operational Research*, North-Holland, 13 Jan. 2016, www.sciencedirect.com/science/article/pii/S0377221716000412.

“1.4. Support Vector Machines¶.” Scikit, scikit-learn.org/stable/modules/svm.html.

Jones, Stewart, et al. “Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Frameworks.” *Wiley Online Library*, John Wiley & Sons, Ltd, 27 Oct. 2016, onlinelibrary.wiley.com/doi/abs/10.1111/jbfa.12218.

1. **Appendix**

Column names and description to make the data easier to understand (Y = Output feature, X = Input 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

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

Add Literature Review

<https://link.springer.com/article/10.1007/s11142-017-9407-1#Sec5> - another literature review.

Add confusion matrix and accuracy chart of all the 4 methods. Compare the numbers.