# Classifying Company Bankruptcy:

# A Machine Learning Approach

Alexa Aguirre, Bernardo Arambula, Yumi Jin

## 1. Executive Summary

The main goal of this project is to be able to correctly classify companies as bankrupt or not based upon key financial attributes. This is valuable to allow investors to make informed decisions about money investments and companies to understand what financial metrics may lead to bankruptcy to address concerns before they may become detrimental. In order to do this, data has been fitted to multiple models to optimize classification accuracy and minimize errors that prove the most risk in incorrect classification. After conducting exploratory analysis of our data, we used k-means clustering to understand how observations could be grouped based upon financial metric similarities, disregarding their classification of bankruptcy or not. Next, we began to fit our data to multiple models: logistic regression full model, logistic regression reduced model using BH procedure, logistic regression using lasso cross validation, decision tree, random forest, and random forest with an adjusted threshold. Out of sample metrics were calculated for each model to identify the model with the best performance and minimal amounts of costly risks.

## 2. Background, Context, Domain Knowledge

Bankruptcy prediction is an important aspect of financial analysis that aims to assess the financial health of a company by forecasting whether a company is likely to face financial turmoil or insolvency in the future. Bankruptcy prediction models can help financial institutions, investors and government agencies make investment decisions and monitor the stability of the market. By accurately predicting bankruptcy, stakeholders can proactively mitigate risk by adjusting lending rates or amounts.

Our project aims to apply various machine learning techniques to analyze a dataset sourced from Kaggle. The dataset contains 95 dependent variables including financial metrics and ratios, with the binary ‘Bankruptcy?’ variable indicating whether a company went bankrupt (1) or not (0). To tackle this binary classification we decided to use logistic regression techniques paired with variable selection methods as well as apply other machine learning concepts including decision trees and random forest

models. Fortunately, the dataset required very little preprocessing, there were no missing values (NA) and contained only numerical variables.

Understanding financial accounting and corporate governance is essential to interpret the financial indicators and ratios chosen during model development. Additionally, industry specific knowledge provides valuable insight for interpreting metrics and identifying potential risk factors associated with bankruptcy. (see appendix for data dictionary)

## 3. Discussion of Traditional Problem-Solving Methods

Predicting bankruptcy is vital for all stakeholders in a company to gain insights about how to make better decisions and reduce costly losses. Businesses can make necessary adjustments before consequences are detrimental, and investors can accurately assess risk. Traditionally, a combination of analyzing financial metrics, looking at historical data trends, and employing individuals with financial expertise to understand the company's financial stability have been used to understand the probability of bankruptcy. Analytical techniques and machine learning models have also increased in popularity to make early on predictions of bankruptcy.

It is important for companies to be aware of their financial metrics to evaluate performance and try to address areas that may be hindering their success. Financial documents such as income statements, balance sheets, and cash flow statements are tools that financial analysts may use to gather key metrics. Evaluating historical values by using trend analysis is a common strategy to identify firms that are at risk of bankruptcy. Looking at revenue and profitability trends that show constant declines may indicate poor financial performance. Additionally, looking at debt-to-equity ratio, profit margins, cash flows, and payment trends historically are also key indicators. Certain values may lead to concerns in financial health and can be warnings against financial crisis. Detecting these signs early may be imperative to save a business and reduce systemic risks.

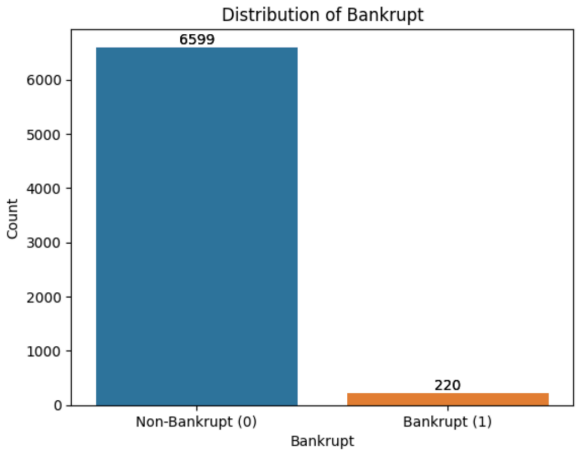
Companies also employ individuals with expertise in the domain of finance. They are better equipped to understand the overall financial climate of the industry to assess financial performance

compared to competitors. Additionally, they may be able to identify qualitative factors associated with the particular company such as quality of management or other structural problems that could be leading the company to underperform that are not directly reflected in numeric financial indicators. However, solely using a person with expertise can lead to subjectivity and bias in predictions.

Specific strategies that are used to predict bankruptcy include both analytical and machine learning methods. The Altman Z-Score is one of the most widely and historically used methods that leverages five important financial ratios and predicts bankruptcy off of weighted coefficients. Machine learning techniques offer a flexible approach to predict bankruptcy because they do not rely on historical data and can account for large volumes of data. They can also take into account complex relationships between variables that may have a significant influence on prediction of bankruptcy. Starting in the 1960's multivariate discrimination analysis became the most popular model and then was overtaken by logit and probit analysis in the 1980's. Neural networks and random forest models have more recently been employed and have shown success in their predictive power.

## 4. Analysis

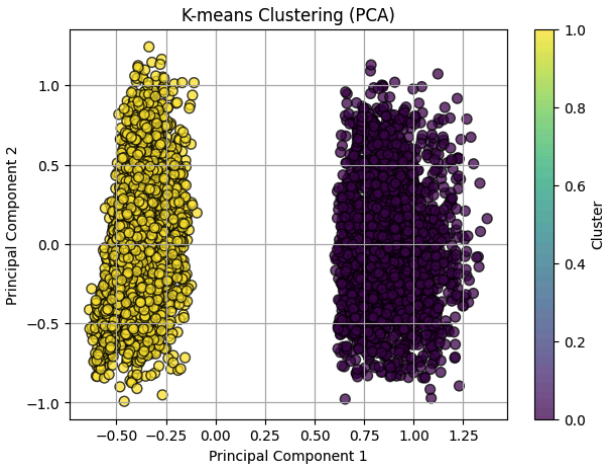
### 4.1 Exploratory Data Analysis




Since we are focused on predicting bankruptcy, we want to first understand the distribution of observation by class, 1 meaning a company went bankrupt and 0 indicating it didn't. There is a large imbalance in the data of bankrupt companies, so this will need to be taken into account when building our predictive models. If we do not take this imbalance into account, the models will be biased in that they will predict the non-bankrupt instances well, but fail to correctly identify bankrupt cases since they are the minority class.

### 4. 2 K-Means Clustering

K-means clustering is not a direct technique to use for predictive modeling, but can provide insight into financial similarities between distinct groups present in the data. Using the elbow method we classified observations into two different clusters. Dimensions were then reduced to 2 using principal component analysis to plot the clusters on a 2 dimension axis. Cluster zero has higher values in principle component 1 and a large range of values in principle component 2, while the opposite is true for cluster 1. We then determined that cluster zero contained 130 bankrupt observations and cluster one contained 90. This means that clusters had financial similarities, but weren't able to detect differences in observations to determine if a company was going to go bankrupt. This implies that there may not be large differences in variable values between companies that went bankrupt compared to those that didn't. Here we can see the comparison of mean values of each variable between cluster 0 and cluster 1. On average it seems that cluster 1 has slightly higher averages in more areas compared to that of cluster 0.




### 4. 3 Model Performance Metrics

For every model that developed, we created a classification report that contains the following metrics calculated on our test data:

- Precision: ratio of true positive predictions to the total number of positive predictions
- Recall (sensitivity)- ratio of true positive predictions total number of actual positive occurrences
- F1: the harmonic mean of precision and recall, higher indicates better performance
- Support: number of occurrences of each class in the dataset
- Accuracy: how correct the model is overall
- Macro Average: unweighted average of precision, recall, and F-1 score

- Weighted Average: average of precision, recall, and F-1 score weighted by the influence from each class

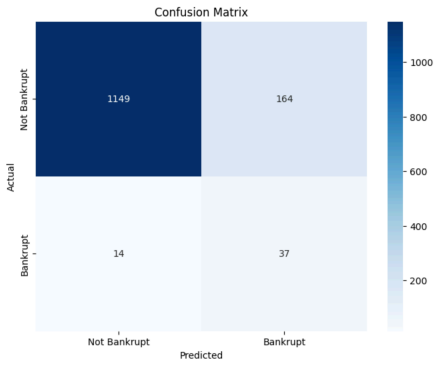
Since we are interested in classifying companies as bankrupt, the overall accuracy of our model is important to maximize, but it will also be important to maximize sensitivity to minimize the number of classifying non bankrupt when they are bankrupt. We want to minimize the number of false negatives because the error is more detrimental. For example, if a company were to be predicted as non-bankrupt, an investor may wrongly allocate their money.

Additionally, confusion matrices for each model illustrate the following:

- True Positives (classified as bankrupt and are bankrupt)
- True Negatives (classified as not bankrupt and aren't bankrupt)
- False Positives (Classified as bankrupt and aren't bankrupt)
- False Negatives (Classified as not bankrupt and are bankrupt)

### 4.4 Logistic Regression Model

The most basic model that we will begin with to fit our data for prediction is a logistic regression model. We chose a logistic regression model since our dependent variable, bankruptcy, is a binary variable. We used the hyperparameter class_weight= “balanced” to assign class weight proportional to class size to account for the class imbalance and scaled all of the dependent variables to ensure that they all have equal influence on bankruptcy. We built out the full logistic regression model using all of the variables which could lead to the potential of overfitting since the model may fit perfectly to the training set of data. After fitting the model, we predicted bankruptcy using the tests data and obtained fairly good results for out of sample deviance. The model’s out of sample accuracy shows that the model accurately classifies companies as

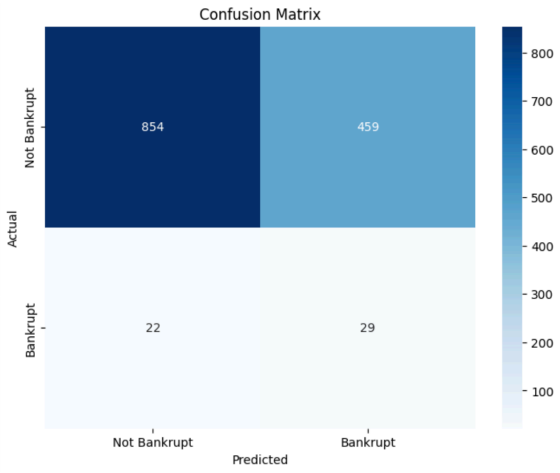



| Feature | Coefficient |
|---|---|
| Borrowing dependency | 1.628461 |
| Operating profit/Paid-in capital | 1.284512 |
| Debt ratio % | 1.064729 |
| Operating Profit Per Share (Yuan ¥) | 1.061313 |
| Net Value Per Share (C) | 0.781526 |

bankrupt or not 87% of the time. The confusion matrix above shows that exact number of out of sample instances' predicted and actual classes. Additionally, the top five variables that increase the odds of bankruptcy can be seen in the table. Each variable will increase the odds of bankruptcy by the corresponding coefficient.

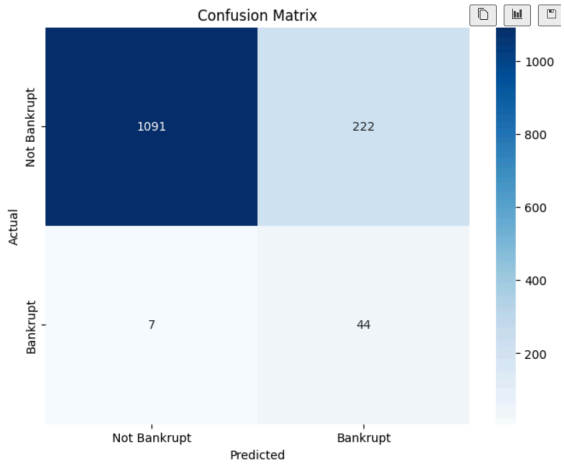
### 4. 5 False Discovery Rate- BH Procedure

After analyzing the confusion matrix from the logistic regression model, we see a false discovery rate of about 13%. We decided to use a false discovery rate(FDR) control using the Benjamin-Hochberg procedure, setting the threshold (q) at 10%. This allowed us to control for the observed 13% false discovery rate. Using this procedure we identified 23 variables as true discoveries. We used these findings to rerun our logistic regression model and update it to only include the 23 significant variables and see an accuracy score of 65% which is a reduction from the 87% achieved by our full logistic regression model. However, when using cross validation techniques we observe a cross validation accuracy score of 97%. This could indicate that our model didn't perform well on our subset of original testing data but when generalized across five different subsets of training and testing subsets became a very powerful model. This shows the reduced model robustness and predictive power when applied to new and unseen data.




### 4. 6 Regularization-Lasso CV

Our dataset contains 95 variables which led us to explore various variable selection techniques. Lasso regularization was an approach we applied due to its ability to apply penalty terms to less important variables based on the strength of our alpha parameter. We ran our model to select the optimal alpha and select significant variables. We found that our model's optimal



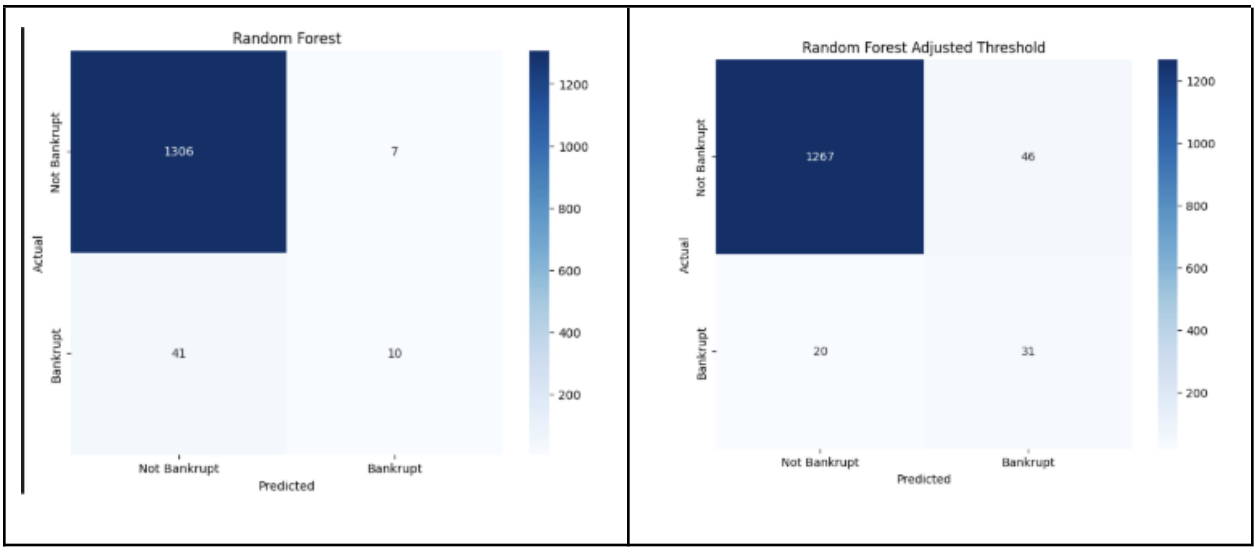

alpha was 100 and the model shrunk all coefficients to 0 except for five: ROA(B) Before interest and depreciation after tax, Debt ratio %, Borrowing dependency and Net income to total assets. Our new model displayed a strong performance, with an in sample cross validation score of 97% and an out of sample cross validation score of 96% and a mean squared error (MSE) of .03. These results show the effectiveness of lasso regularization and its ability to identify a parsimonious model while maintaining predictive accuracy.

### 4. 7 Classification Decision Tree

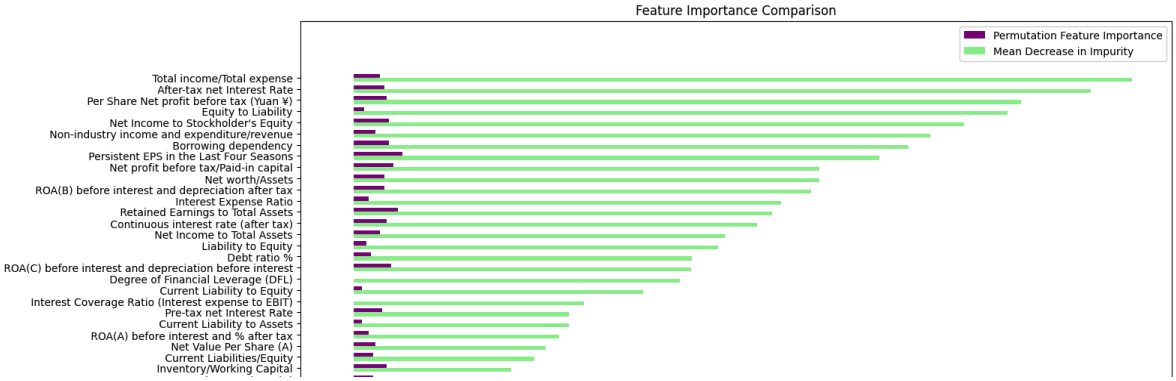

We also ran a decision tree regression on our data. The tree model would be capable of capturing any nonlinear relationships or interactions between variables. It is also an easily interpretable model, does not require assumptions about data distributions and is robust to irrelevant variables. When analyzing our model performance output we see that the decision tree regression is another powerful predictor of bankruptcy prediction. The accuracy of the model on the testing data is 95%, the average accuracy on in sample 5-fold cross validation score is 95% and the average out of sample cross validation score is 95%. The decision tree model's consistent scores provide evidence of the models robustness and generalizability.

### 4.8 Random Forest

Next, we fit a random forest classification model. This model will solve any issues of overfitting that could have resulted from the decision tree by aggregating many trees obtained using a bootstrap method. The random forest model also captures any interactions between variables which may lead to better predictions than our previous models that did not account for such interaction effects between variables. We accounted for the class imbalance to obtain better predictions and got an out of sample accuracy of 96%. However, the model's ability to correctly classify bankrupt cases is quite low with a precision score of 0.59, recall of 0.20 and f1-score of 0.29. Since misclassifying a bankrupt company as not bankrupt has a higher cost, we tuned the hyperparameters by finding a better classification threshold. The decision boundary default is 0.5, but after testing different thresholds we were able to improve the out of sample precision to 0.4, recall to 0.61, and f-1 score to 0.48 with an accuracy of 95%.




Using permutation feature importance and mean decrease in impurity we can see which variables are most important in decreasing the deviance between each node in the tree.



## 5. Recommendations and Business Values

After fitting the data to multiple models, we can see that the random forest model has the highest out of sample accuracy score of 96%, however the recall score is very low meaning that the model does not do a sufficient job at classifying bankrupt cases correctly. After adjusting the threshold, we determined the tuned hyperparameter model is the best. Although we reduced the overall model accuracy by .01%, we significantly increased the recall score which is the most costly risk that we want to maximize. Additional hyperparameter tuning may need to be conducted in the future to increase the recall even more. This model can be useful to detect early signs of company bankruptcy to inform decisions and mitigate investing risks. The top variables identified that lead to the most deviance reduction using mean decrease in impurity, thus indicating their importance in classifying companies as bankrupt include total income/total expense, after-tax net interest rate, per share net profit before tax, and equity to liability. Total income/total expense illustrates a company's ability to generate income relative to its expenses. A declining ratio indicates poor performance and is of high concern to address to prevent bankruptcy. A low after-tax net interest may suggest difficulty in meeting debt obligations. Per share net profit before tax is the company's profitability per share before taxes, indicating poor financial performance if it declines. A declining equity to liability ratio may indicate a

| Model | Precision | Recall (sensitivity) | F1-Score | Support |
|---|---|---|---|---|
| Random Forest | | | | |
| 0 | 0.97 | 0.99 | 0.98 | 1313 |
| 1 | 0.59 | 0.2 | 0.29 | 51 |
| accuracy | | | 0.96 | |
| | | | | |
| Random Forest (Adjusted Threshold) | | | | |
| 0 | 0.98 | 0.96 | 0.97 | 1313 |
| 1 | 0.4 | 0.61 | 0.48 | 51 |
| accuracy | | | 0.95 | |
| | | | | |
| Decsion Tree | | | | |
| 1 | 0.97 | 0.98 | 0.98 | 1313 |
| 0 | 0.36 | 0.31 | 0.33 | 51 |
| accuracy | | | 0.95 | |
| | | | | |
| Logistic Regression (full model) | | | | |
| 0 | 0.99 | 0.88 | 0.93 | 1313 |
| 1 | 0.18 | 0.73 | 0.29 | 51 |
| accuracy | | | 0.87 | |
| | | | | |
| Logistic Regression (Lasso with CV) | 0.99 | 0.83 | 0.91 | 1313 |
| | 0.17 | 0.86 | 0.28 | 51 |
| | | | 0.83 | |
| | | | | |
| Logistic Regression (FDR Variables) | | | | |
| 0 | 0.97 | 0.65 | 0.78 | 1313 |
| 1 | 0.06 | 0.57 | 0.11 | 51 |
| accuracy | | | 0.65 | |

company’s inability to cover obligations. Overall, assessing and monitoring metrics in these areas may help identify concerns within a company's financial performance that need to be addressed to prevent bankruptcy or investing in companies that are underperforming.

## 6. Summary and Conclusions

The objective of this project was to accurately classify companies as bankrupt or not based on key financial attributes. These classifications are invaluable for stakeholders and regulatory bodies to make informed decisions based on the financial wellbeing of a company. Despite the random forest model's high accuracy, its low recall score suggests it may not adequately classify bankrupt cases. However, after adjusting the threshold, we observed a better balance between precision and recall, making it our preferred model. Key variables identified, such as total income/total expense, after-tax net interest rate, and equity to liability ratio, offer valuable insights into a company's financial health. Monitoring these metrics can help identify potential risks and inform proactive decision-making to prevent bankruptcy. Overall, the project highlights the capabilities of machine learning techniques in predicting bankruptcy and emphasizes the importance of continuous monitoring and adjustment to mitigate financial risks. Further research and fine-tuning of the model may enhance predictive accuracy and provide even greater value to stakeholders.

## 7. References

Bellovary, Jodi L., et al. “A Review of Bankruptcy Prediction Studies: 1930 to Present.” *Journal of Financial Education*, vol. 33, 2007, pp. 1–42. *JSTOR*, http://www.jstor.org/stable/41948574. Accessed 20 Mar. 2024.

*Taiwanese bankruptcy prediction*. UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

## 8. Appendix: 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 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**