**Predicting Company’s Financial Failure or Bankruptcy Using Financial Ratios and Ensemble Techniques**

*A Project Report Submitted by*

**Aarthiee U K (M20MA001)**

*In partial fulfillment of the requirements for the award of the degree of*

**M.Tech**

**Declaration**

I hereby declare that the work presented in this Project Report titled Predicting Company’s Financial Failure or Bankruptcy Using Financial Ratios and Ensemble Techniques Report –M.Tech submitted to the Indian Institute of Technology Jodhpur in partial fulfilment of the requirements for the award of the degree of M.Tech., is a bonafide record of the research work carried out under the supervision of Dr.Manish Agarwal. The contents of this Project Report in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

Signature

Aarthiee UK (M20MA007)

**INTRODUCTION:**

Multiple research has been done on financial stability or financial failure of an organization since 1930[1]. For the prediction of Bankruptcy, different research have been in place since late 1960’s to use financial ratios. Bankruptcy is filed when a company that has no operational source of money to operate the business and is in no position to repay the debts it owes to its creditors. Bankruptcy is a difficult situation to be as the company comes down to a stand-still state and leaves the employees and suppliers or customers in a high and financial instability state. When the competitors are increased and uncertainty in the global economy these days which is the cause to drive large organizations towards bankruptcy. The recent example of the British travel company, Thomas Cook which suddenly declared bankruptcy and left 21,000 people out of work and million dollars of investor money down the drains. Thus, it is impossible to overstate the damage caused in terms of the financial loss.[6]

The process of bankruptcy begins with a petition which is filed by the debtor or on behalf of creditors. The assets of the debtor are measured and their assets are used to repay a portion of their outstanding debt. To free itself from debt obligations, Bankruptcy filing is undertaken by the company. Debts that are not paid to creditors are forgiven for the owners.Bankruptcy filing varies in different countries. If you file for the bankruptcy in India, it will not go down well with our credit rating which means that it is tough to get a new loan if you plan to start afresh. However, it would save us from any financial trouble.

To assess and analyze the performance of firms, Financial ratios are used. Based on the various parameters, we compare and analyze a firm’s overall health. The Financial ratios are used by the Investors to get an insight to the firm’s profitability and its investment prospects. The Financial ratios are helpful in forecasting the firm’s future health by using the historical data from financial statements. Also they help to compare and contrast financial performance of firms among their competitors additionally. Inspite of many research in this area, several attempt towards exploring this relationship has made which led to limited success to the variability existence in every stock markets. The variance in volatility of these markets makes it difficult to obtain a uniform measure[5].

The Prediction of Bankruptcy is a very important task for financial institutions. The main aim is to predict the likelihood that a firm may go bankrupt or not. Financial institutions are in need of effective prediction models in order to make appropriate lending decisions [2], stock price prediction and more. Some of the Machine learning algorithms and techniques that are used for prediction of Bankruptcy are Support Vector Machines (SVM), Naïve Bayes Classifier (NB), Classification and Regression Trees(CART), k-nearest neighbor (KNN),and MLP(Multi-Layer Perceptron). Features for the machine learning models are stepwise discriminant analysis (SDA) (Fisher, 1936), stepwise logistic regression (SLR) (Fisher & Yates, 1963), and t-testing (Zimmerman, 1997)3 and the wrapper based methods of the genetic algorithm (GA) (Holland, 1975) and recursive feature elimination (RFE) (Guyon, Weston, Barnhill, &Vapnik, 2002). We also calculate the model score, Precision, Recall, F1 score and ROC score and compare which model gives highest F1 score. We are evaluating the models using Sum of Squared Errors (SSE), Root Mean Squared Errors (RMSE), Mean Absolute Error (MAE), Relative Absolute Error (RAE).

The Prediction of Bankruptcy is a very important task for financial institutions. Their main aim is to predict the likelihood that a firm may go bankrupt or not. Financial institutions are in need of effective prediction models in order to make appropriate lending decisions [2], stock price prediction and more. Some of the Machine learning algorithms and techniques that are used for prediction of Bankruptcy are Support Vector Machines (SVM), Naïve Bayes Classifier (NB), Classification and Regression Trees(CART), k-nearest neighbor (KNN), and MLP(Multi-Layer Perceptron). Features for the machine learning models are stepwise discriminant analysis (SDA), stepwise logistic regression (SLR) , and t-testing and the wrapper based methods of the genetic algorithm (GA) and recursive feature elimination (RFE) .

**Background**

The Features that are selected such as financial ratios and techniques such as statistical or machine learning techniques are the important factors which strongly affects the performance of the prediction of Bankruptcy. While many novel prediction techniques have proposed, very few of them have analyzed the discriminatory power of the features which are related to bankruptcy prediction. The financial ratios (FRs) and the corporate governance indicators (CGIs) have been found to be important type of input variable[2]. However, the Bankruptcy prediction performance is obtained by combining CGIs and FRs has not been fully examined. Only some selected CGIs and FRs have been used in related studies and the chosen features differed from study to study. In this, we will explore all the Financial Ratio’s and Corporate Governance Indicators *through SMOTE and Machine Learning methods with ensemble techniques*

The prediction performance of Bankruptcy is obtained by combining seven different categories of Financial Ratio’s and five different categories of corporate governance indicators. The experimental results which is based on the real-world dataset from Taiwan proves that FR categories of solvency and profitability and the CGI categories of board structure and ownership structure are the most important features in bankruptcy prediction.

**DATA:**

Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange. The Taiwan Stock Exchange is a financial institution located in Taipei, Taiwan. It has over 900 listed companies. It was established in 1961 and began operating as a stock exchange on 9 February 1962. The data were collected from the Taiwan Economic Journal for the years 1999 to 2009. The Taiwanese Bankruptcy Prediction data was obtained from UCI Machine Learning Repository.

The Taiwanese Bankruptcy Dataset contains:

* 95 features (X1-X95, business regulations of Taiwan Stock Exchange)
* 1 Vector of labels

Our aim is to use these features to understand their impact on the selected models and how they can help us recognizing the companies that are close to bankrupcty.All the variables that are part of this dataset are numeric and complete in nature.The data includes a majority of numerical attributes that help understand the possibility of bankruptcy.We will use various predictive models to see how accurately we predict which companies will face bankruptcy in the future.This is the same data set that was used by Liang D and Co[2] for research and donated to UCI community.

|  |  |
| --- | --- |
| X1 Cost of Interest-bearing Debt | X49 Fixed Assets Per Employee |
| X2 Cash Reinvestment Ratio | X50 total assets to GNP price |
| X3 Current Ratio | X51 Return On Total Assets(C) |
| X4 Acid Test | X52 Return On Total Assets(A) |
| X5 Interest Expenses/Total Revenue | X53 Return On Total Assets(B) |
| X6 Total Liability/Equity Ratio | X54 Gross Profit /Net Sales |
| X7 Liability/Total Assets | X55 Realized Gross Profit/Net Sales |
| X8 Interest-bearing Debt/Equity | X56 Operating Income /Net Sales |
| X9 Contingent Liability/Equity | X57 Pre-Tax Income/Net Sales |
| X10 Operating Income/Capital | X58 Net Income/Net Sales |
| X11 Pretax Income/Capital | X59 Net Non-operating Income Ratio |
| X12 Working Capital to Total Assets | X60 Net Income-Exclude Disposal Gain or Loss/Net Sales |
| X13 Quick Assets/Total assets | X61 EPS-Net Income |
| X14 Current Assets/Total Assets | X62 Pretax Income Per Share |
| X15 Cash/Total Assets | X63 Retained Earnings to Total Assets |
| X16 Quick Assets/Current Liability | X64 Total Income to Total Expenses |
| X17 Cash/Current Liability | X65 Total Expenses to Assets |
| X18 Current Liability to Assets | X66 Net Income to Total Assets |
| X19 Operating Funds to Liability | X67 Gross Profit to Sales |
| X20 Inventory/Working Capital | X68 Net Income to Stockholder's Equity |
| X21 Inventory/Current Liability | X69 One if Net Income is Negative for the Last Two Years; Zero Otherwise |
| X22 Current Liabilities/Liability | X70 (Inventory +Accounts Receivables) /Equity |
| X23 Working Capital/Equity | X71 Total Asset Turnover |
| X24 Current Liabilities/Equity | X72 Accounts Receivable Turnover |
| X25 Long-term Liability to Current Assets | X73 Days Receivable Outstanding |
| X26 Current Liability to Current Assets | X74 Inventory Turnover |
| X27 One if Total Liability exceeds Total Assets; | X75 Fixed Asset Turnover |
| X28 Equity to Liability | X76 Equity Turnover |
| X29 Equity/Total Assets | X77 Current Assets to Sales |
| X30(Long-term Liability+Equity)/Fixed Assets | X78 Quick Assets to Sales |
| X31 Fixed Assets to Assets | X79 Working Capital to Sales |
| X32 Current Liability to Liability | X80 Cash to Sales |
| X33 Current Liability to Equity | X81 Cash Flow to Sales |
| X34 Equity to Long-term Liability | X82 No-credit Interval |
| X35 Liability to Equity | X83 Cash Flow from Operating/Current Liabilities |
| X36 Degree of Financial Leverage | X84 Cash Flow to Total Assets |
| X37 Interest Coverage Ratio | X85 Cash Flow to Liability |
| X38 Operating Expenses/Net Sales | X86 CFO to Assets |
| X39 (Research and Development Expenses)/Net Sales | X87 Cash Flow to Equity |
| X40 Effective Tax Rate | X88 Realized Gross Profit Growth Rate |
| X41 Book Value Per Share(B) | X89 Operating Income Growth |
| X42 Book Value Per Share(A) | X90 Net Income Growth |
| X43 Book Value Per Share(C) | X91 Continuing Operating Income after Tax Growth |
| X44 Cash Flow Per Share | X92 Net Income-Excluding Disposal Gain or Loss Growth |
| X45 Sales Per Share | X93 Total Asset Growth |
| X46 Operating Income Per Share | X94 Total Equity Growth |
| X47 Sales Per Employee | X95 Return on Total Asset Growth |
| X48 Operation Income Per Employee |  |

**SMOTE:**

SMOTE stands for  **Synthetic Minority Oversampling Technique** is a machine learning technique that solves problems when using an **imbalanced dataset**. To address the imbalanced datasets, we do oversample the minority class. To counteract the class imbalance, Undersampling is used which means that **discarding the number of data points of the class that is too often present**. The disadvantage of undersampling is valuable data is lost. Another solution to imbalanced data is Oversampling which means making duplicates of the data that is the least present in our dataset. Then we add those duplicates to our dataset. Based on the original data points, SMOTE algorithm does data augmentation by creating **synthetic data points**. SMOTE is an advanced version of oversampling for data augmentation. The advantage of SMOTE is creation of synthetic data points that are **slightly different** from the original data points instead of duplicates generation. The **SMOTE algorithm**  draws a random sample from the minority class. Identifying the k nearest neighbors for the observations in this sample. Then one of those neighbors is chosen and identify the vector between the current data point and the selected neighbor. Next multiply the vector by a random number between 0 and 1. To obtain the synthetic data point, add this to the current data point. This is like **slightly moving the data point in the direction of its neighbor.** We ensure that our synthetic data point is **not an exact copy**of an existing data point and it is**also not too different** from the known observations in our minority class.

**LOGISTIC REGRESSION:**

Logistic regression is a machine learning algorithm which comes under the Supervised Learning technique and it is used for solving the classification problems. Predicting the categorical dependent variable using a given set of independent variables in logistic regression. So the outcome must be a categorical or discrete value. It gives the probabilistic values which lie between 0 and 1 instead of giving the exact value as 0 and 1. In logistic regression, we fit an S shaped logistic function which predicts two maximum values that is 0 or 1, instead of fitting a regression line. Using the different types of data, we classify the observations and easily determine the most effective variables for the classification in logistic regression. Logistic regression is used for probability calculation of a binary event and to deal with classification. Used to predict the likelihood of “yes” or “no” outcomes. Logistic regression helps [data analysts](https://careerfoundry.com/en/blog/data-analytics/difference-between-data-scientist-and-data-analyst/) to make informed decisions by predicting such outcomes. This helps to minimize the risk of loss and to optimize spending in order to maximize profits.

**RANDOM FOREST CLASSIFIER:**

Random Forest is a machine learning algorithm that belongs to the supervised learning technique and is used for both Classification and Regression problems. It is based on ensemble learning which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. **It uses bagging and feature randomness when building each individual tree to create an uncorrelated forest of trees**whose prediction is more accurate than that of any individual tree.Random Forest classifier has number of decision trees on various subsets of the given dataset and we take the average to improve the predictive accuracy. The random forest takes the prediction from each tree and based on the majority votes of predictions, it predicts the final output instead of depending on one decision tree,. The greater number of trees in the forest leads to higher accuracy and prevents overfitting. By Combining the multiple trees to predict the dataset class in random forest, some decision trees may predict the correct output, while others may not predict correct. Finally all the trees together predict the correct output.

**XGBOOST:**

XGBoost stands for Extreme Gradient Boosting. XGBoost is a tree based ensemble machine learning algorithm useful for tree boosting. It uses more accurate approximations to find the best tree model. XGBoost is an optimized distributed gradient boosting which is designed to be portable, flexible and highly efficient. It provides a parallel tree boosting to solve the problems in a fast and accurate manner. XGBoost which has higher predicting power and performance and it is achieved by improvisation on Gradient Boosting framework. XGBoost performs faster when compared to other algorithms because of its parallel and distributed boosting technique. XGBoost is developed in terms of systems optimization and principles in machine learning.

**CATBOOST:**

“CatBoost” comes from two words “Category” and “Boosting”. CatBoost is an open sourced machine learning algorithm developed from Yandex. It can be easily integrated with frameworks like Google’s TensorFlow and Apple’s Core ML. Their idea is to serve multi-functional purposes such as [Recommendation systems](https://dataaspirant.com/recommendation-engine-part-1/), Personal assistants, Self driving cars, Weather prediction, etc. "Boosting" in CatBoost refers to the [gradient boosting machine learning](https://dataaspirant.com/gradient-boosting-algorithm/) technique for  [regression and classification](https://dataaspirant.com/classification-and-prediction/)  problems. It has a good handling technique for categorical data. CatBoost supports numerical, categorical and text features. In the processing stage, The CatBoost algorithm has number of [parameters to tune the features](https://dataaspirant.com/hyperparameter-tuning-with-keras-tuner/). CatBoost can improve the **performance** of the model while [**reducing overfitting**](https://dataaspirant.com/handle-overfitting-deep-learning-models/) and the time spent on tuning.  The CatBoost algorithm has good performance and **greedy novel** gradient boosting implementation. CatBoost uses one hot encoding technique for categorical features with a small number of different values in most of the modes.

**Validation:**

Several metrics can be used to measure the performance of any classifier, computed by combining the results obtained in the confusion matrix. Four categories are composing this matrix:

1) True Positives (TP): number of samples correctly classified as bankrupt.

2) False Positives (FP): number of samples incorrectly classified as bankrupt.

3) True Negatives (TN): number of samples correctly classified as solvent.

4) False Negatives (FN): number of samples incorrectly classified as solvent.

Thus, since the binary classification accuracy results are not reliable while the data considered is extremely imbalanced (the classifiers always tend to predict the majority class and ignore the minority class), several metrics have been computed to make a better judgment about each classifier’s performance and reliability. These metrics are:

Accuracy: Performance of the classifier in terms of assigning the correct class to each instance.

Accuracy =(TP + TN)/(TP + TN + FP + FN)

• Type I error [58]: Also called as False Positive Rate (FPR). It represents the failure of the classifier to assign bankrupt companies to the ‘bankrupt’ class (wrong prediction), while its actual class is ‘bankrupt’ (real status).

Type 1 error = FP/(TN+FP)

Type II error [58]: Also called as False Negative Rate (FNR), represents the failure of the classifier in assigning solvent companies to ‘solvent’ class (wrong prediction), while its actual class is ‘solvent’ (real status).

Type II Error = FN/(TP+FN) = 1 - Recall

**Experiment and Results:**

Backward Elimination feature selection technique is used to select the features for the machine learning models and entire lot of features have been sent to the deep learning models. But SMOTE method is used to oversample the examples in the minority class and train and test dataset has been generated.

We are comparing the result generated by Liang D & Co [2] with our results as the underlying data set remains the same. FC is combination of Financial Ratio’s and Corporate Governance Indicators and he compared performance of the model with FC and Financial Ratio’s. It proved that FC are little better than Financial Ratio’s alone. For this experiment we are considering FC’s along with SMOTE sampling method and going to evaluate some of the Boosting, Bagging and ANN algorithms. Results of our experiment with comparison with results of Ling D[2] is below, by the way step wise linear regression filter(SLR) for the features performed well than others for the author.

Fig1: ROC Curve

Chart, line chart

Description automatically generated

Fig2: Precision Recall Curve

A picture containing line chart

Description automatically generated

Table 2: Results

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Type I Error | Type 2 Error |
| Support Vector Machine -SLR | 81.3 | 16.7 | 20.8 |
| K- Nearest Neighbor -SLR | 77.9 | 20.1 | 24.1 |
| CART – SLR | 78.2 | 22.2 | 21.4 |
| Multi Layer Perceptron – SLR | 80.5 | 18.8 | 20.3 |
| Naïve Bayes – SLR | 77.1 | 12.00 | 33.80 |
| **Logistic Regression-SMOTE** | **84.0** | **15.79** | **30.77** |
| **Random Forest Classifier-SMOTE** | **97.0** | **1.74** | **48.72** |
| **XG Boost-SMOTE** | **95.0** | **3.95** | **56.41** |
| **CAT Boost-SMOTE** | **97.0** | **1.57** | **25.64** |

Comparing various supervised algorithms with different types of feature selection methods and sample handling methods with ensemble techniques

**References:**

1. J. L. Bellovary, D. E. Giacomino, and M. D. Akers, ‘‘A review of bankruptcy prediction studies: 1930 to present,’’ J. Financial Educ., vol. 33, pp. 1–42, Dec. 2007.
2. [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.](https://isslab.csie.ncu.edu.tw/download/publications/1.pdf)
3. USE OF FINANCIAL RATIOS TO MEASURE THE QUALITY OF EARNINGS [Gullett, Nell S](https://www.proquest.com/indexinglinkhandler/sng/au/Gullett,+Nell+S/$N;jsessionid=33D95995239DD71173E40EA37D5426BC.i-03bc6ea303836550a); [Kilgore, Ronald W](https://www.proquest.com/indexinglinkhandler/sng/au/Kilgore,+Ronald+W/$N;jsessionid=33D95995239DD71173E40EA37D5426BC.i-03bc6ea303836550a); [Geddie, Mary F](https://www.proquest.com/indexinglinkhandler/sng/au/Geddie,+Mary+F/$N;jsessionid=33D95995239DD71173E40EA37D5426BC.i-03bc6ea303836550a" \o "Click to search for more items by this author). [Academy of Accounting and Financial Studies Journal](https://www.proquest.com/pubidlinkhandler/sng/pubtitle/Academy+of+Accounting+and+Financial+Studies+Journal/$N/29414/OpenView/2124417228/$B/75040B6CC80246EDPQ/1;jsessionid=33D95995239DD71173E40EA37D5426BC.i-03bc6ea303836550a); Arden[Vol. 22, Iss. 2,](https://www.proquest.com/indexingvolumeissuelinkhandler/29414/Academy+of+Accounting+and+Financial+Studies+Journal/02018Y04Y01$232018$3b++Vol.+22+$282$29/22/2;jsessionid=33D95995239DD71173E40EA37D5426BC.i-03bc6ea303836550a) (2018): 1-12
4. E. I. Altman, ‘‘Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,’’ J. Finance, vol. 23, no. 4, pp. 589–609, Sep. 1968
5. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer SMOTE: Synthetic Minority Over-sampling Technique Journal of Artificial Intelligence Research 16 (2002) 321–357
6. Ferris, Erin. (2018). Dividend Taxes and Stock Volatility. International Tax and Public Finance, 25(2), 377–403. <https://doi.org/10.1007/s10797-017-9455-2>
7. Corporate Bankruptcy Prediction using Machine Learning Techniques, Shantanu Deshpande, Dr. Vladimir Milosavljevic, [National College of Ireland](http://norma.ncirl.ie/4312/1/shantanudeshpande.pdf)
8. J. A. Ohlson, ‘‘Financial ratios and the probabilistic prediction of bankruptcy,’’ J. Accounting Res., vol. 18, no. 1, pp. 109–131, Apr. 1980. [32] J. Brozyna, G. Mentel, and T. Pisula, ‘‘Statistical methods of the bankruptcy prediction in the logistics sector in Poland and Slovakia,’’ Transformations Bus. Econ., vol. 15, no. 1, pp. 80–96, 2016.
9. P. P. M. Pompe and A. J. Feelders, ‘‘Using machine learning, neural networks, and statistics to predict corporate bankruptcy,’’ Comput.-Aided Civil Infrastruct. Eng., vol. 12, no. 4, pp. 267–276, Jul. 1997. [35]
10. J. Min and Y. Lee, ‘‘Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters,’’ Expert Syst. Appl., vol. 28, no. 4, pp. 603–614, May 2005.
11. N. Ghatasheh, H. Faris, R. Abukhurma, P. A. Castillo, N. Al-Madi, A. M. Mora, A. M. Al-Zoubi, and A. Hassanat, ‘‘Cost-sensitive ensemble methods for bankruptcy prediction in a highly imbalanced data distribution: A real case from the Spanish market,’’ Prog. Artif. Intell., vol. 9, no. 4, pp. 361–375, Dec. 2020.
12. Comparing the Performance of Deep Learning Methods to Predict Companies’ Financial Failure H. ALJAWAZNEH 1 , A. M. MORA 2, P. GARCÍA-SÁNCHEZ 1 AND P. A. CASTILLO-VALDIVIESO 1