Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018)
IEEE Xplore Compliant Part Number: CFP18K74-ART; ISBN:978-1-5386-2842-3

A Bankruptcy Prediction Model Using Random Forest

Shreya Joshi[1], Rachana Ramesh [2] , Shagufta Tahsildar[3]
Vidyalankar Institute of Technology
Mumbai University
Mumbai , India
Email: [1] shreya16j@gmail.com, [2] rachana22.r@gmail.com, [3]shaguftatahsildar@gmail.com

***Abstract*—Bankruptcy Prediction is very important for an organization as well as decision makers such as financiers and investors. Selection of dataset for training the prediction model, the Machine Learning tool used for prediction and various other factors are essential in building an efficient prediction model. The dataset includes financial ratios as attributes that are derived from the financial statements of various companies. The most influencing ratios that are required for predicting bankruptcy are selected on the basis of the Genetic Algorithm which filters out the most important ones from different existing bankruptcy models. These ratios of different companies are fed as an input to train the model being implemented in R. The prediction algorithm used is Random Forest, which will enable us to differentiate between bankrupt and non-bankrupt companies.**

Keywords—bankruptcy prediction, bankruptcy models, financial ratios, genetic algorithm, random forest, decision trees, RStudio

I. INTRODUCTION

Bankruptcy is a legal status of a person or a company that cannot repay the debts it owes to creditors. Increase in business failures in various companies has led to the development of bankruptcy prediction models (BPM) which can be used by financers or banks to identify firms that are healthy enough to receive loans [1].The use of BPMs to identify potential business failure can help prevent failures as well as ensure credit or contracts are given only to healthy companies. Hence, bankruptcy prediction is regarded as a critical issue and is studied extensively in the accounting and finance literature [2]-[3].The bankruptcy prediction method has been developed by Altman [4] and Beaver [5] since about 1960. The performance of a bankruptcy prediction model is dependent on the tool employed to build the model and the size and type of data available among other factors.

Machine Learning (ML) can be used to develop a robust and updatable prediction system. The most popular ML tools include Multi Discriminant Analysis (MDA) and Logit Analysis (LA) while popular and promising AI tools include Support Vector Machines (SVM) [10], Case based reasoning (CBR) [11], Iterative Dichotomiser (ID3) etc. Also, the selection of data set for training the model is an essential factor for creating an effective prediction model. Hence, in this paper we also see the financial ratios which are influential in predicting bankruptcy of a company. The dataset will contain these ratios as attributes.

Typically, this can be solved as a classification problem to separate distressed companies from healthy ones based on the analysis of financial characteristics or ratios as we call them [12]. The prediction model will work on these financial factors of the company and will make a prediction whether the company will go bankrupt or not. The prediction results are, however, based on how well the model is trained. In this paper, we propose a system where we select financial ratios of about 50 bankrupt and 50 non-bankrupt companies as training data set for training the model. The ML tool used will be Random Forest which is an ensemble learning method for classification, regression and other tasks. Random Forest will then make a prediction for new test dataset.

The rest of the paper is organized as follows: Section 2 reviews prior research. Section 3 describes the proposed system using Random forest as the ML tool and explains why it is superior to traditional methods such as Decision Trees. Section 4 discusses about the experimentation and implementation details. Section 5 deals with the results of the prediction model and its accuracy. Section 6 concludes the paper with remarks and outcomes.

II. PRIOR RESEARCH

A. Data Set and Bankruptcy Models

Many bankruptcy prediction models have been developed and each of these models has a focal point which addresses the issues related to bankruptcy prediction. The focal points addressed in various models are, finding the accuracy, role of variables (financial variables & non-financial variables), identifying the types of failure, events that may affect the financial situation of a firm (non-payment of a debt, reduction

of dividend payments), and models on final bankruptcy dissolution(liquidation, reorganization, takeover). Bankruptcy prediction models have selected financial ratios from the previous literature or either using univariate statistical tests or data mining techniques. The ratios used in the prediction, form the key constituents of these models and they reflect the characteristics of stability, profitability, growth, activity and cash flow of a business. However, only limited bankruptcy prediction models have attempted to assess the contribution of financial ratios which affects the performance of a model.

Selection of the most influential financial ratios from these models is done by using GBRAT i.e. Genetic Bankruptcy Analysis Tool [6].This Genetic tool analyses the nonlinear relationship among financial ratios of bankruptcy models and classifies the ratios into influencing and non-influencing ratios. An influencing ratio is a ratio which plays a key role to obtain bankruptcy prediction accuracy among other financial ratios. Predominating bankruptcy models have been applied by various firms to predict bankruptcy and the same have been chosen in the present study to identify the influencing ratios using GBRAT. The bankruptcy models with accuracy levels of more than 80% are selected and ranked and from which, the top five bankruptcy models are chosen for ratio analysis. The selected models are Altman (1968) [4], Deakin (1972) [7], Edmister (1972) [8] , Springate (1978) [9] , and Fulmer (1984) [1]. In Table 2.1, the details of the chosen bankruptcy models are described.

Bankruptcy Model	Bankrupt Score	Ratio Range*
Altman Model	2.675	1.0 > WC < 2.0, TA < 1, RE +ve > 1, EBIT > 1, 1> MVOE < 2, TL > 1, 1> EQ < 2, SR +ve range
Deakin Model	1.5	NI > 1, TA < 1, CA > 2, CR > 1, CL > 1
Springate Model	0.862	1.0 > WC < 2.0, TA < 1, NPBIT > 1, NPBT > 1, CL > 1, 1> EQ < 2
Edmister Model	0.530	AF>1, CL>1, 1>EQ <2, NWC > 1, QR>1, AR>1
Fulmer Model	0	RE +ve > 1, TA < 1, EBIT > 1, TD > 1, CL > 1, 1.0 > WC < 2.0, 1> EQ < 2, EBIT > 1.

**CA-Current Asset, CR-Cash Ratio, CL-Current Liabilities, TL-Total liabilities, NI – Net Income, TA – Total Asset, WC-Working Capital, RE-Retained Earnings, EQ-Equity, MVOE- Market Value of Equity, AF- Annual Funds, NWC-Net working Capital, AR- Average Ratio, QR-Quick Ratio, SR-Sales Ratio, NPBT-Net Profit before Taxes, NPBIT- Net Profit before Interest and Taxes, TD-Total Debt, AR – accuracy rate, EBIT- Earnings before Interest and Taxes*

Table 2.1: Bankruptcy Models

All ratios fall under their specific ranges as specified in Table 2.1; Table 2.1 describes the selected bankruptcy models and their details, along with the specific range of each of the ratios applied in the respective bankruptcy models. For example, the parameter details of Deakin model, with its specified range, are described as follows, Net Income (NI), Cash Ratio (CR) and Current Liabilities (CL) should be greater than 1, Total Assets (TA) should be less than 1, Current Assets (CA) should be greater than 2 and Sales Ratio (SR) should be in positive range. Five different bankruptcy models have been considered for the analysis and identification of the most influencing ratios using GBRAT. Mutation and crossover operators have been applied on each selected model and the results are discussed in the following sections.

1. Analysis of Altman Bankruptcy Model Ratios

The bankruptcy equation of Altman bankruptcy model is given below,

$Z = 0.012\ X_1 + 0.014\ X_2 + 0.033\ X_3 + 0.006\ X_4 + 0.999\ X_5$….(1)

Where, X_1 = Working capital / Total assets, X_2 = Retained earnings / Total assets, X_3 = Earnings before interest and taxes/Total assets, X_4 = Market value of equity / Book value of total liabilities, X_5 = Sales / Total assets, Z = Altman Bankrupt Value. The performance of any business is assessed based on the value of Z. That is, when the value of Z is less than 2.675, Altman prediction would lead to bankruptcy, otherwise it assures the better performance of business.

2. Analysis of Deakin Bankruptcy Model Ratios

The bankruptcy equation of Deakin model is given by the following equation,

$I= -1.369 + 13.855X_1 + 0.060X_2 – 0.601X_3 + 0.396X_4 +0.194X_5$ ... (2)

Where, X_1 = Net income/ Total assets, X_2 = Current assets/ Total assets, X_3 = Cash/ Total assets, X_4 = Current assets/ Current liabilities, X_5 = Sales/ Current assets, I = Overall index (Deakin Bankrupt Value). For any given business, the value of the bankruptcy prediction variable I of the Deakin model should be greater than or equal to 1.5. Otherwise, it indicates that the business is likely to bankrupt.

3.Analysis of Springate Bankruptcy Model Ratios

The bankruptcy equation of the Springate model is as follows,

$Z= 1.03A + 3.07B + 0.66C + 0.4D$ ... (3)

Where, A -Working Capital/Total Assets, B-Net Profit before Interest and Taxes/Total Assets, C-Net Profit before Taxes/Current Liabilities, D-Sales/Total Assets, Z- Springate bankrupt value. For any given business, the value of the bankruptcy prediction variable Z of the Springate model should be greater than or equal to 0.865. A value lower than 0.865, indicates that the business is likely to bankrupt.

4. Analysis of Edmister Bankruptcy Model Ratios

Edmister developed a seven-variable, zero-one linear regression equation, which gives the bankruptcy value of Edmister model,

$Z= 0.951–0.523X_1 –0.293X_2 –0.482\ X_3 + 0.277\ X_4 – 0.452\ X_5 – 0.352\ X_6 - 0.924\ X_7$ ... (4)

Where, X_1 = Annual funds to Current liabilities, X_2 = Equity to Sales, X_3 = Net working capital to Sales, divided by RMA average ratio (average ratios for firms), X_4 = Current liabilities to Equity, divided by RMA average ratio, X_5 = Inventory to Sales, divided by RMA average ratio, X_6 = Quick ratio divided

by the trend in RMA quick ratio, X_7 = Quick ratio divided by RMA quick ratio, Z = Edmister variable. In this model, the minimum threshold value of an offspring should be 0.530. A threshold value lower than 0.530, indicates that the business is likely to bankrupt.

5. Analysis of Fulmer Bankruptcy Model Ratios

The Fulmer bankruptcy model takes the following form,

$H = 5.528 (V_1) + 0.212 (V_2) + 0.073 (V_3) + 1.270 (V_4) - 0.120(V_5) + 2.335 (V_6) + 0.575 (V_7) + 1.083 (V_8) + 0.894(V_9) - 6.075$ ... (5)

Where, V_1 = Retained Earning/Total Assets, V_2 = Sales/Total Assets, V_3 = EBT/Equity, V_4 = Cash Flow/Total Debt, V_5 = Debt/Total Assets, V_6 = Current Liabilities/Total Assets, V_7 = Log Tangible Total Assets, V_8 = Working Capital/Total Debt V_9 = EBIT/Interest, H = Fulmer bankrupt variable

The influencing ratios obtained using GBRAT for all the five models in this study are given in Table 2.2.

Influencing Ratio	Bankruptcy Model
X_1 = Working capital / Total assets	Altman
X_2 = Retained earnings / Total assets	
X_3 = Earnings before interest and taxes / Total assets X_5 = Sales / Total assets	
X_3 = Cash/ Total assets	Deakin
A= Working Capital/Total Assets	Springate
X_1 = Annual funds / Current liabilities	Edmister
X_2= Equity / Sales	
X_4= Current liabilities / Equity, divided by RMA avg. ratio	
V_5= Debt / Total Assets	Fulmer

Table 2.2: Most Influencing Ratios

From Altman model given in *equation (1)*, one can find the performance of a business using only the most influencing ratios X_1, X_2 X_3 and X_5. Deakin model given in *equation (2)* is suggested for those who want to assess the financial status of a business with cash and total assets (X_3) which is the only most influencing ratio in this model. Springate model represented in *equation (3)* is the best model to measure the business financial crisis using working capital and total assets. The most influencing ratios of Edmister model given in *equation (4)* are X_1, X_2 and X_4 which can be used to find the financial stability of a business. It is also observed that V_5 is the only most influencing ratio in Fulmer model given in *equation (5)* to predict the bankruptcy of a business. It is observed that the total assets is a key financial variable in designing a bankruptcy model as it is an influencing ratio in all the models. Amongst the five models considered for the study, Altman, Deakin, Springate and Fulmer models are designed based on total assets whereas the Edmister model is influenced by current liabilities.

The influential ratios thus selected from the models as seen in table 2.2 will be considered for the dataset in order to build an accurate prediction model. The ML algorithm will work on these ratios and do the classification.

III. SYSTEM PROPOSAL

The task of classification, which is one of supervised data mining technique, is to predict accurately the class to which the data samples belong to. In supervised learning, classes are predetermined. A certain segment of data will be labeled with these classification. The task is to search for patterns and construct mathematical model.

Bankruptcy Prediction is a classification problem which requires firms to be classified as failing or non-failing. The two predetermined classes here are the bankrupt class and the non-bankrupt class and the training data will already be labeled with these classes. We use Random Forest which is a learning method used for classification which will predict the class to which the given test data sample belongs to. The dataset used in our system will be the most influential ratios as seen in Section II of this paper.

A. Random Forest

Random forest is one of the most popular and most powerful machine learning algorithms. It is a type of ensemble machine learning algorithm called Bootstrap Aggregation or bagging which is a very simple and powerful method. An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model.

Bootstrap Aggregation is a general procedure that can be used to reduce the variance for those algorithm that have high variance. An algorithm that has high variance are decision trees, like classification and regression trees (CART). Decision trees are sensitive to the specific data on which they are trained. If the training data is changed (e.g. a tree is trained on a subset of the training data) the resulting decision tree can be quite different and in turn the predictions can be quite different. Bagging is the application of the Bootstrap procedure to a high-variance machine learning algorithm, typically decision trees.

Random Forests are an improvement over bagged decision trees. The methodology of Random Forest includes construction of decision trees of the given training data and matching the test data with these. Combining predictions from multiple models in ensembles works better if the predictions from the sub-models are uncorrelated or at best weakly correlated. Random forest changes the algorithm for the way that the sub-trees are learned so that the resulting predictions from all of the subtrees have less correlation. In CART, when selecting a split point, the learning algorithm is allowed to look through all variables and all variable values in order to select the most optimal split-point. The random forest algorithm changes this procedure so that the learning algorithm is limited
to a random sample of features of which to search.

B. Random Forest v/s Decision Trees

A Random forest is ensemble of decision trees so it helps in predicting data accurately. A random forest model is typically made up of tens or hundreds of decision trees. Redistribution error rate of random forest is less than decision

tree. Random Forest overcomes the problem with over fitting and also handles the problems of sparse data or missing data well. One important scenario where Random Forest stands out is when data is limited. The accuracy of decision tree becomes less when training data is less whereas Random Forest performs well with limited training dataset as well.

We have made a few observations on the performance of Decision Trees against the performance of Random Forest with limited training dataset using RStudio (an open source platform for executing R programs). Our dataset was limited to only 14 companies.

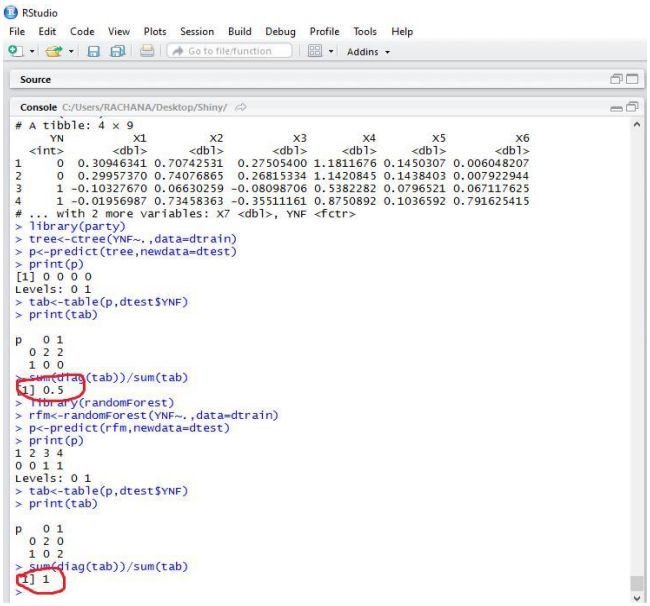

Figure 3.1: Confusion Matrix and Accuracy of Decision Trees and Random Forest

Figure 3.1 shows the confusion matrix of Decision Trees and Random Forest. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. With the help of the confusion matrix, one can see how the model does the prediction and thus the accuracy of the model can be calculated.

With limited dataset, that is, with only 14 companies, we split 80 % data for training and 20% for testing and observe the results of random forest and decision trees algorithms. In the figure 3.1, we observe that in the confusion matrix of decision tree out of 4 companies which are under testing, 2 companies are predicted properly whereas 2 companies are predicted incorrectly and hence the accuracy is 0.5 (50%) whereas Random Forest predicts all the 4 companies in their correct categories bringing its accuracy to 1(100%). Thus, Random Forest performs well even if the training data set is limited whereas Decision Tree fails with limited training data.

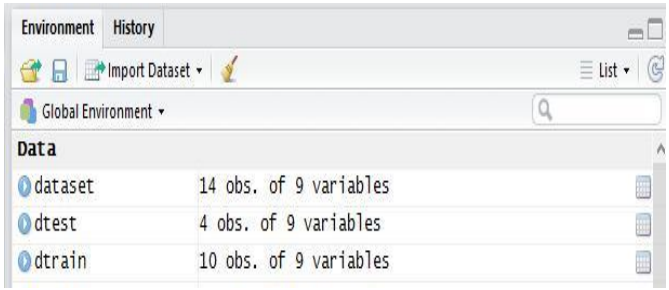

Figure 3.2: Splitting of the dataset into training data and testing data

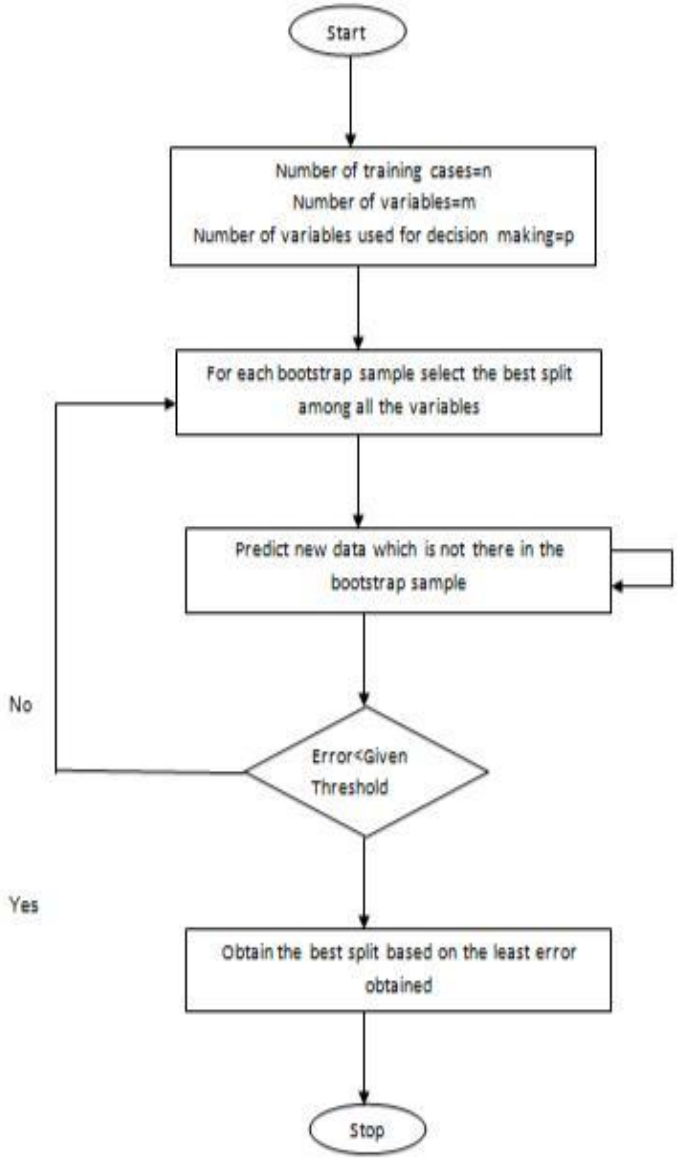

Figure 3.3: Flow chart for Random Forest

IV. EXPERIMENTATION DETAILS

The algorithm will be coded in R language which can be implemented with the help of RStudio. R is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and others. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made. For machine learning, R's advantages are linked mostly to R's strong ties to academia. Any new research in the field probably has an accompanying R package to go with it from the get-go. So in this respect, R stays at the cutting edge. R Programming applications compass the universe from hypothetical, computational statistics and the hard sciences. R has almost 5,000 packages (libraries of functions) large portions of which are committed to particular applications. R programming will be used for the purpose of running the prediction algorithm as it is the most feasible programming language which can be used for prediction.

As discussed previously, the algorithm which will be used in R is Random Forest algorithm. The final dataset of 50 bankrupt and 50 non-bankrupt companies containing the most influential ratios will be given as the training input to the algorithm coded in R. The dataset will be in the form of a .csv file. Actual data of the companies are obtained from their respective Annual Reports. The Random Forest algorithm will be trained on this data. We can now give the file of the company whose status is to be predicted to the already trained model. The prediction model will display the chances of the company becoming bankrupt in the near future. Thus, by using this machine learning algorithm in R, we can predict the bankruptcy of a company and investors or banks can decide whether the company is healthy enough to receive loan or bonds. This system can also be used by startups or growing companies to check their own performance.

This whole implementation will be displayed in a user friendly interface. This GUI will be designed using the Shiny functionality available in R. It is a web application framework in R. Shiny makes it easy for R users to turn analyses into interactive web applications that anyone can use. It lets the users choose input parameters using user friendly controls like sliders, drop-down menus, and text fields. We can easily incorporate any number of outputs like plots, tables, and summaries. Shiny is an open source R package that provides an elegant and powerful web framework for building web applications using R. Shiny helps you turn your analyses into interactive web applications without requiring HTML, CSS, or JavaScript knowledge. The Shiny platform used in conjunction with R has become increasingly popular in allowing for the design of interactive web applications that allow a non-R specialist to conduct statistical analysis with a more sophisticated UI. This is becoming increasingly important as the R language is increasingly gearing itself towards non-specialists who wish to interact with the program in a more simplistic and interactive way. The web interface will have a tab which will allow the users to browse through the files which they want to give in for prediction. It will display the chances of bankruptcy of the company and also how accurate can this result be. Based on this result, the banks can decide whether it will be acceptable to grant a loan to the company and will the company be able to repay the amount on time.

V. RESULTS

The results are displayed in the interface in a separate tab. The probability percentages will be displayed as to how probable the company is, in going into bankruptcy. The model has a good accuracy since only the most influential ratios of the companies are considered.

Figure 5.1: Prediction Results as seen in R Shiny

Figure 5.1 represents the prediction results of 3 companies. The first and the third ones are the actual data of Kingfisher Airlines, two and three years before it went bankrupt and the model correctly predicts them to be bankrupt giving them bankruptcy probabilities as 100% and 66% respectively, whereas the second company is a non-bankrupt company and its bankruptcy probability is predicted as 3%.(Note that binary 0 here indicates non-bankruptcy and 1 indicates bankruptcy).

VI. CONCLUSION

Bankruptcy of a company can be determined by considering various factors which include the financial ratios derived from the financial statements of the company. However, priority is given to those ratios which carry the most weightage while determining the bankruptcy. Only those influencing ratios are selected. Random Forest is the ML tool used for building the prediction model. Random Forest is trained on a dataset consisting the ratios and finally the model can predict accurate results on various test cases. This algorithm will be

implemented in RStudio. The Graphical User Interface required for the interaction with the user will be developed with the help of the Shiny Web App which is an integrated tool available in the RStudio. Accordingly one can use this method to design a Bankruptcy Prediction Model.

VII. REFERENCES

[1] Fulmer, Jr. G. John Moon, E. James Gavin, A. Thomas Erwin and J. Michael, A Bankruptcy Classification Model for Small Firms, Journal of Commercial Bank Lending ISSN 0021-986X (1984) 25-37.

[2] C. James Van Horne, Financial Management Policy (Twelfth Edition,

[3] Guoqiang Zhang, Michael Y. Hu b, B. Eddy Patuwo , Daniel C. Indro, Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis, European Journal of Operational Research 116(1) (1999) 16-32

[4] Financial ratios, discriminant analysis and the prediction of corporate Bankruptcy- Altman, Edward .I., 1968, Journal of Finance 23, 589-609.

[5] Financial ratios as predictors of failure- Beaver, William H., 1966, Journal Of Accounting research, 71-111.

[6] Martin Aruldoss, Miranda Lakshmi Travis, Prasanna Venkatesan Venkatsamy – A Genetic Bankrupt Ratio Analysis Tool using Genetic Algorithm to Identify Influencing Financial Ratios, IEEE Transactions On Evolutionary Computation.

[7] E. B. Deakin, A Discriminant Analysis of Predictors of Business Failure, Journal of Accounting Research 10(1) (1972) 167-179.

[8] R.O. Edmister, An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction, Journal of Financial and Quantitative Analysis 7 (1972) 1477-1493.

[9] Springate and L.V. Gordon, Predicting the Possibility of Failure in a Canadian Firm. Unpublished M.B.A. Research Project, Simon Fraser University. (1978).

[10] Sung-Hwan Mina, Jumin Leeb and Ingoo Hanb , Hybrid Genetic Algorithms And Support Vector Machines for Bankruptcy Prediction, Expert Systems with Applications 31(3) (2006) 652-660.

[11] Hui Li and Jie Sun, Case based reasoning ensemble for business failure Prediction: A computations from multiple case representations, Expert Systems with applications 39(3) (2012) 3298-3310.

[12] Chen and Du, Using neural networks and data mining techniques for the financial distress prediction models, Expert systems with Applications 36(2) (2009) 4075-4086.

[13] Credit risk evaluation in power market with Random Forest H Mori,Y Umezawa(2007)