PREDICTING BANKRUPTCY USING MACHINE LEARNING ALGORITHMS

Abhishek Karan[1] and Preetham Kumar[2]

[1]Department of Information & Communications Technology, Manipal Institute of Technology, Manipal University, Manipal, Karnataka, India
[2] Professor & Head Department of Information & Communications Technology, Manipal Institute of Technology, Manipal University, Manipal, Karnataka, India

ABSTRACT

This paper is written for predicting Bankruptcy using different Machine Learning Algorithms. Whether the company will go bankrupt or not is one of the most challenging and toughest question to answer in the 21st Century. Bankruptcy is defined as the final stage of failure for a firm. A company declares that it has gone bankrupt when at that present moment it does not have enough funds to pay the creditors. It is a global problem. This paper provides a unique methodology to classify companies as bankrupt or healthy by applying predictive analytics. The prediction model stated in this paper yields better accuracy with standard parameters used for bankruptcy prediction than previously applied prediction methodologies.

KEYWORDS

Machine Learning, Classification, Regression, Correlation, Error Matrix, ROC

1. INTRODUCTION

Bankruptcy is a legal status of a firm that cannot repay the debts it owes to creditors. The latest research within the field of Bankruptcy and Predictive Analytics compares various different approaches, modelling techniques, and individual models to ascertain whether any one technique is superior to its counterparts and if so then which all parameters will help better predict the outcome of bankruptcy.

Bankruptcies affect all stakeholders: from employees to regulators, investors or managers. Therefore, it is very interesting to understand the phenomenon that leads to a company going bankrupt in order to take advantage over their competitors. Companies are never protected against going bankrupt. Either in an economic expansion or in a recession, firms are likely to go bankrupt.

Financial ratios are a tool to determine the operational & financial efficiency of business undertakings.

2. LITERATURE REVIEW

2.1 Machine Learning

Machine learning comes under artificial intelligence that provides computers with the ability to learn without being explicitly programmed. It makes the machine learn on its own in overdue

course of time and also with increase in volume of datasets. It focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data in a trained environment.

It is a mature and well-recognized research area of information technology, mainly concerned with the discovery of models, patterns, and other regularities in data. Machine learning is closely related to and often overlaps with computational statistics: a discipline that also specializes in prediction-making. It has strong ties to mathematical optimization, which deliver methods, theory and application domains to the field.

Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible. Example applications include spam filtering, search engines and computer vision.

2.2. Predictive Model Approach

There are 5 approaches to follow and create the best model possible with given dataset.

2.2.1. Model Creation

Models are basically algorithms and approaches to deal with certain predictive environment. Many different software models are available which allows to create models to run one or more algorithms on the data set and predict the answers with certain percentage of probability.

2.2.2. Model Testing

Testing models on new and unknown dataset helps us to estimate the accuracy of the constructed model. In most cases, the testing is done on past data to see how best the model predicts.

2.2.3. Model Validation

Validation is also done on previously unknown datasets to tweak and optimize the existing model and hence create an airtight one. Results of these tests are run on visualization tools and business data as well.

2.2.4. Model Evaluation

Evaluate the best fit model from different models used and choose the model right fitted for the dataset. This approach is taken when more than one model have to be compared which is most often than not a general situation. It uses tools like Receiver Operating Curves, Confusion Matrices, Accuracy and Precision.

2.2.5. Deploying of Model

The model created to be deployed in forms of graphs, trees or equations. In real world a simple GUI (Graphical User Interface) is created of the model and presented to the clients.

2.3. Tools Used

1.2.1 R & RStudio
1.2.2 Rattle
1.2.3 Microsoft Excel 2013

R Libraries: Library(rattle), Library(rpart), Library(MASS), Library(e1071), Library(caret) & Library(corrgram)

3. METHODOLOGY

3.1. Dataset

3.1.1. Collection

The Financial Ratios, Bankruptcy Prediction dataset was received from Alto University's Application of Machine learning dataset [5].

3.1.2. Description of Dataset

The dataset consists of 500 samples or instances based on 41 attributes. The Output is a class type variable describing the instance as 'Bankrupt' (250 cases) & 'Healthy' (250 cases). The attributes are financial indicators from several anonymized companies of 2002 for bankruptcy prediction task.

3.1.3. Data Cleaning

3.1.4. The NULL factor

Removing NULLs

If less number of NULL values are present in dataset as compared to actual number of records, then those records can be removed.

Domain Expertise

Domain experts can contribute valuable knowledge with respect of data quality, i.e. selection of attributes /data quality.

Mark as N.A

Certain NULL values can be replaced with 'N.A' values for easier calculations in statistics [13]. This dataset doesn't contain any missing values.

3.1.5. The Outlier Factor

Removing outlier through Box Plot

Box Plot of all attributes are drawn. The values which are irrelevant, outliers and are not making sense to the dataset are removed.

Sorting

Each attribute is sorted in any order (Ascending/Descending). Values which seems to be absurd or out of bounds are removed.

After applying the Outlier methodology, 59 records were removed. Now, the dataset contains 441 samples based on 41 attributes and 1 Target variable (Bankrupt).

3.1.6.The Correlation Factor

Correlation is statistical tool which is used to measure the degree to which the movements of independent variables or attributes are related. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. The correlation coefficient, often denoted 'ρ' or 'r', measures the degree of correlation. The most common of these correlations is the Pearson correlation coefficient (used in this paper as well), which is sensitive only to a linear relationship between two variables. The correlation is defined as:

$\rho_{x,y} = E\,[(X-\mu_x)\,(Y-\mu_y)] \,/\, (\sigma_x\,\sigma_y)$ Where μ_X and μ_Y are expected values & σ_X and σ_Y are standard deviations. E is the expected value operator. The correlation coefficient 'r' ranges from -1.0 to +1.0. The closer it is to +1 or -1, the more closely the two variables are related.

1. If 'r' is close to 0, it means there is no relationship between the variables.
2. If 'r' is positive, it means that as one variable gets larger the other gets larger.
3. If 'r' is negative it means that as one gets larger, the other gets smaller (an "inverse" correlation).

3.1.7. Correlation Matrix

Correlation Matrix is a matrix giving the correlations between all pairs of data sets. The Correlation between the target variable and other predictors is viewed in this matrix. These 2 predictors have the least correlation with Bankrupt (1 = Healthy, 2 = Bankrupt)

Total Sales/Total Assets: -0.007816907
Labor Expenses/Total Sales: -0.002093413

Correlation Plot

The Correlation Plot helps us to visualize the data in correlation matrices. The correlation values of predictors Total Sales/Total Assets & Labor Expenses/Total Sales is the least with target variable (Bankrupt) from the correlation matrix as well as the correlation plot. After applying the Correlation Factor, 2 columns (Total Sales/Total Assets & Labor Expenses/Total Sales) were removed from dataset. Now, the dataset contains 441 samples based on 39 attributes and 1 Target variable (Bankrupt).

3.1.8. The Mean & Median Factor

The summary of entire dataset is found which contains the Mean, Median, Minimum, Maximum values along with 1st and 3rd quartiles in R. Predictors having very different Mean and Median are removed as that data seems to be faulty. But as a judgement call certain predictors are not removed because they share high affinity with those of real world bankruptcy factor.
Duplication of Column names

There are 2 columns with same names (LI7 Quick Assets/Total Assets & LI9 Quick Assets/Total Assets). We remove the L19 column. Now, the dataset contains 441 samples based on 38 attributes and 1 Target variable (Bankrupt).

3.2. Distribution of Dataset

In building a model 70% subset of all of the available data is used. We call this 70% sample as training dataset. The remainder is split equally into a validation dataset (15%) and a testing dataset (15%).

1. The validation dataset is used to test different parameter settings or different choices of variables. It is important to note that this dataset should not be used to provide any error estimations of the final results from machine learning algorithms since it has been used as part of the process of building the model.
2. The testing dataset is only to be used to predict the unbiased error of the final results. It is also a new and previously unknown dataset.

In our Model the dataset is divided into 70/30. There is no validation dataset. Only Training (70%) & testing (30%) dataset is present. The division of samples into testing and training datasets are done randomly, keeping in mind the seed value for the random function used for distribution.

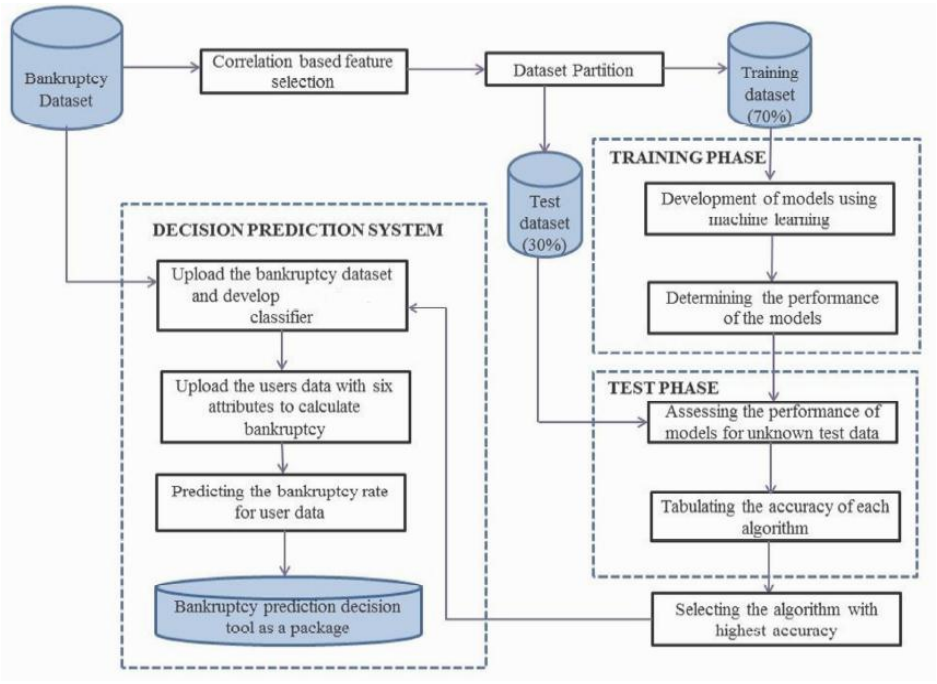

Figure 1: Decision Support System for predictive Bankruptcy [1]

3.3. Building a Model

3.3.1. Classification Algorithm

Following algorithms are used in the paper

1. Decision Tree
2. SVM (Support vector machines)
3. Logistic Regression

Decision Tree

The building of a decision tree starts with a description of problem which should specify the variables, actions and logical sequence for a decision-making [10]. In a decision tree, a process leads to one or more conditions that can be brought to an action or other conditions, until all conditions determine a particular action, once built you can have a graphical view of decision-making[4]. The tree model is created using the "rpart" library of R.

Which variables are best Classifiers?

The answer to this lies in Entropy and Information Gain.

Entropy: Entropy characterizes the purity or impurity of an arbitrary collection of examples. It is defined as:

$$Entropy(S) = (-p_{+}log_2 p_{+}) + (- p_{-} log_2 p_{-}) \quad [16]$$

Where S is a collection. Entropy is very common in Information Theory.

Information Gain: Information Gain measures how well a given attribute separates the training examples according to their target Classification. Entropy is a fundamental attribute in finding this gain. It is defined as:

$$Gain\ (S, A) = Entropy\ (S) - \sum |S_v| / |S| * Entropy\ (S_v)$$ [16] Where S is a collection and A is a chosen attribute.

The best Classifiers for this dataset are:

1. Accounts Receivable/ TotalSales
2. Cash Total Debt
3. Net Income Shareholder's Funds
4. Profit before Tax/Shareholder's Funds
5. Shareholder's Funds/Total Assets

After getting these parameters from R we create a decision tree. This tree may be subjected to overfitting so we chose to prune this tree on basis of certain parameters as discussed below

Table 1: 'rpart' Result of Decision Tree

CP	*NSPLIT*	*REL_ERROR*	*XERROR*	*XSTD*
0.753695	0	1.000000	1.00000	0.051561
0.093596	1	0.246305	0.28571	0.034962
0.019704	2	0.152709	0.21182	0.030687
0.014778	4	0.113300	0.23645	0.032218
0.010000	5	0.098522	0.23153	0.031921

Table 2: Confusion Matrix and Statistics before pruning

Predicted	**Actual**		
		1	2
	1	57	5
	2	3	67

Pruning is a technique in decision tree model of machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting [8].

The ‘rpart’ function in R gives 3 important columns:

- rel_error,
- xerror
- xstd

Cross-validation error will actually grow as the tree gets more levels (at least, after the 'optimal' level). The optimum value is usually the lowest level where the (rel error + xstd) < xerror [15]. As per table 1, lowest level is level 3 thus advisable to prune at level 3.

Also, after plotting the graph of xerror (cross validation error) as shown in figure 2, it was very clear that the least error was found at 3rd level where CP value is: 0.019704. Hence, the optimal level of tree is till 3rd.

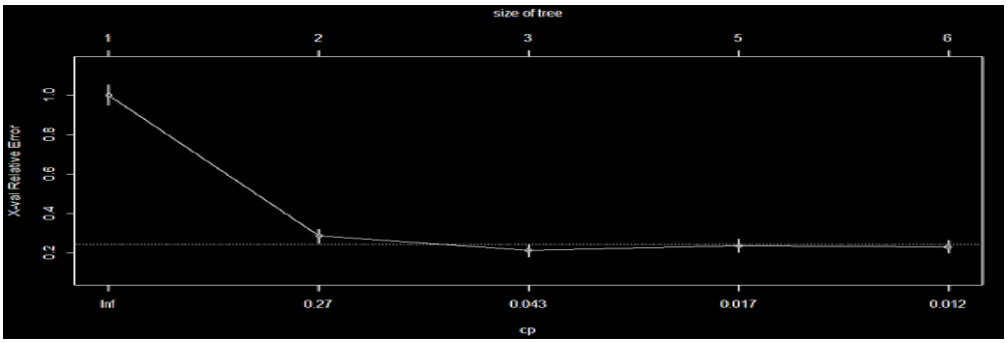

Figure 2: ‘xerror’ Plot of Decision Tree

Based on the above 2 factors, optimal level of pruning is at level 3.

Table 3: Decision Tree confusion matrix after Pruning

Predicted	**Actual**	
	1	2
1	58	6
2	2	66

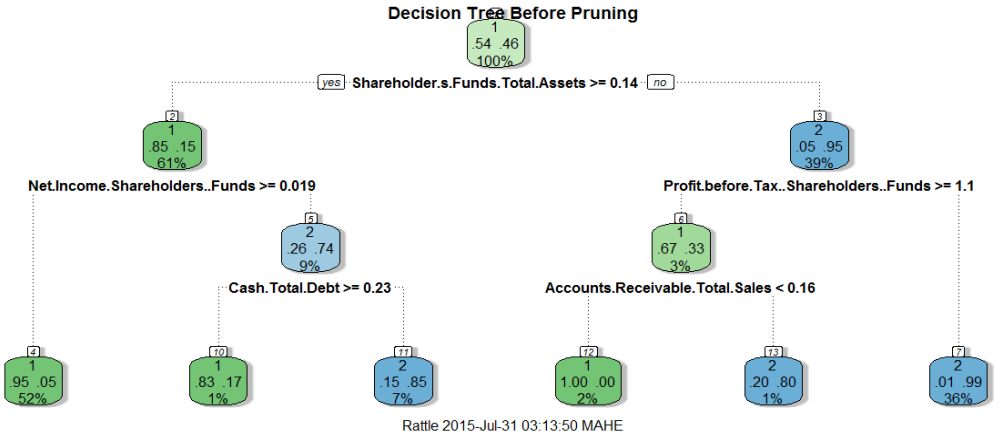

Figure 3: Decision Tree before Pruning

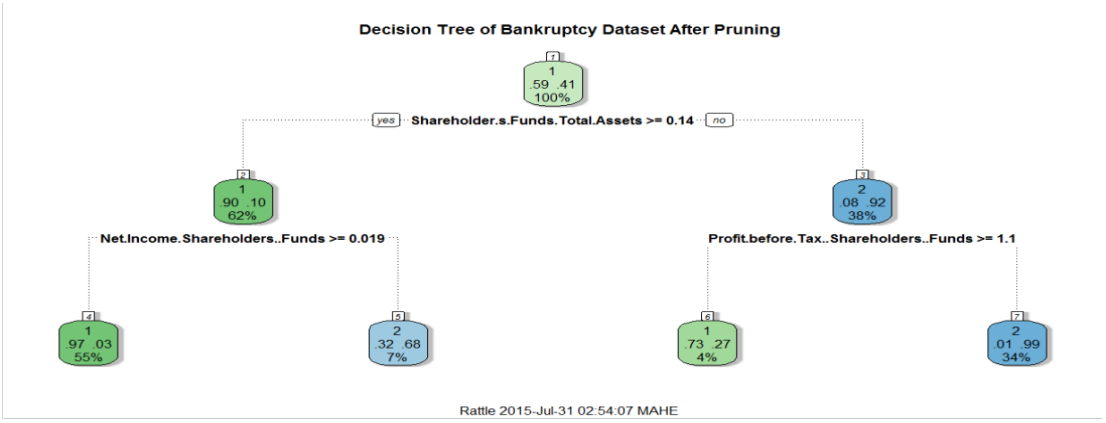

Figure 4: Decision Tree after Pruning

Decision Tree Traversal Rules from Rattle:

Rule #7:
Bankrupt=2 cover=132 (34%) prob = 0.99
Shareholder's Funds / Total Assets < 0.14
Profit before Tax/Shareholder's Funds < 1.1

Rule #5:
Bankrupt=2 cover=16 (7%) prob = 0.68
Shareholder's Funds / Total Assets >= 0.14
Net Income Shareholder's Funds < 0.01906

Rule #6:
Bankrupt=1 cover=12 (4%) prob = 0.27
Shareholder's Funds / Total Assets < 0.14
Profit before Tax/Shareholder'sFunds >=1.1

Rule #4:
Bankrupt=1 cover=148 (55%) prob = 0.03
Shareholder's Funds / Total Assets >= 0.14
Net Income Shareholder's Funds >= 0.01906

Support Vector Machine (SVM)

Basic idea of support vector machines:

An optimal hyperplane for linearly separable patterns is used. Extend the patterns that are not linearly separable by transformations of original data to map into new space – Kernel function.

- SVMs Support vectors Maximize margin.
- SVMs maximize the margin around the separating hyperplane.
- The decision function is fully specified by a subset of training samples, the support vectors.

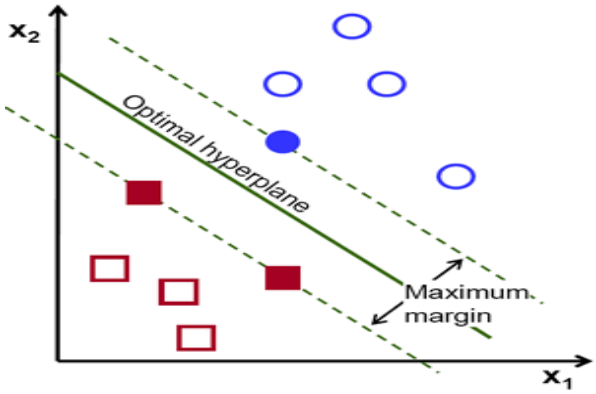

Figure 5: SVM Hyperplane [7]

Factors that influence SVM Model [9]

1. Cost (c): Measure of misclassification.

'c' is the parameter for the soft margin cost function, which controls the influence of each individual support vector. This process involves trading error penalty for stability. A large 'c' gives low bias and high variance. Low bias because it penalizes the cost of misclassifications a lot.

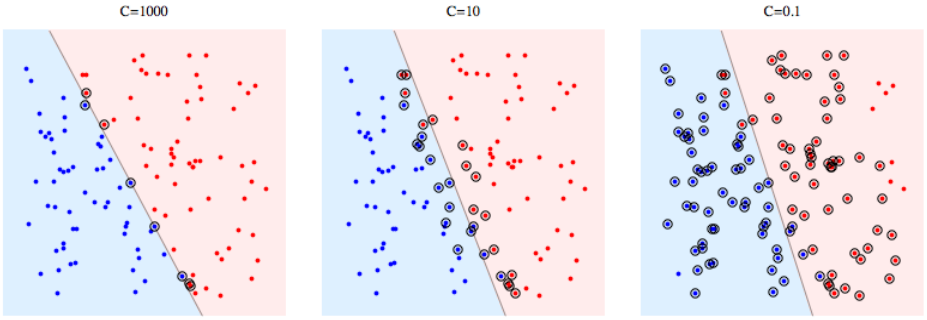

Figure 6: Cost Representation

2. Gamma:

Raise a hyperplane between blue circles and red squares from figure 8. Peak measures Gamma. It is the parameter of a Gaussian Kernel (to handle non-linear classification). They are not linearly separable in 2D so it should be transformed to a higher dimension where it'll will be linearly separable. Imagine "raising" the blue points, then it can be separated from the red points with a plane (hyperplane). To "raise" the points use the RBF kernel function. Gamma controls the shape of the "peaks" where the points are raised. A small gamma gives a pointed bump in the higher dimensions, a large gamma gives a softer, broader bump. So a small gamma will give low bias and high variance while a large gamma will give higher bias and low variance.

- The optimal gamma and cost factors are chosen.
- The SVM Model is created using the "e1071" library in R.
- Error Matrix is created to check the accuracy of the model.

Table 4: Snapshot of SVM Performance Results

S.NO.	*GAMMA*	*COST*	*ERROR*	*DISPERSION*
1	0.001000	1	0.11241808	0.02928308
2	0.010000	1	0.07524877	0.02914474
3	0.100000	1	0.11797534	0.0274428

Summary Result of SVM Function Model from R:

1. Best parameters: Gamma=0.01, Cost=1
2. Best performance: 0.07524877

3.3.2. Regression Model

Logistic Regression
Logistic Regression seeks to:

1. Model the probability of an event occurring depending on the values of the independent variables, which can be categorical or numerical.
2. Estimate the probability that an event occurs for a randomly selected observation versus the probability that the event does not occur.
3. Predict the effect of a series of variables on a binary response variable(1 & 0)
4. Classify observations by estimating the probability that an observation is in a particular category (such as approved or not approved).

To understand Logistic Regression first understand Odds.

- Odds = P(Occurring of an event) / P(Not Occurring of the event)

Odds Ratio:

- Odds Ratio = odds1 / odds2

The dependent variable in Linear Regression follows Bernoulli distribution having an unknown probability 'p'. In Logistic Regression we are estimating an unknown 'p' for any given linear combination of independent variables. Thus we need to link together our independent variable to Bernoulli distribution, this link is called LOGIT. Bernoulli is a special case of Binomial distribution where n=1.

Now, when the points on graph follows an approximately straight line. $\boldsymbol{Logit\ (p) = \beta_0 + \beta_{1x}}$ Although this model looks similar to a simple linear regression model, the underlying distribution is binomial and the parameters β_0 and β_1 cannot be estimated in exactly the same way as for simple linear regression. Instead, the parameters are usually estimated using the method of Maximum Likelihood Estimation (MLE).

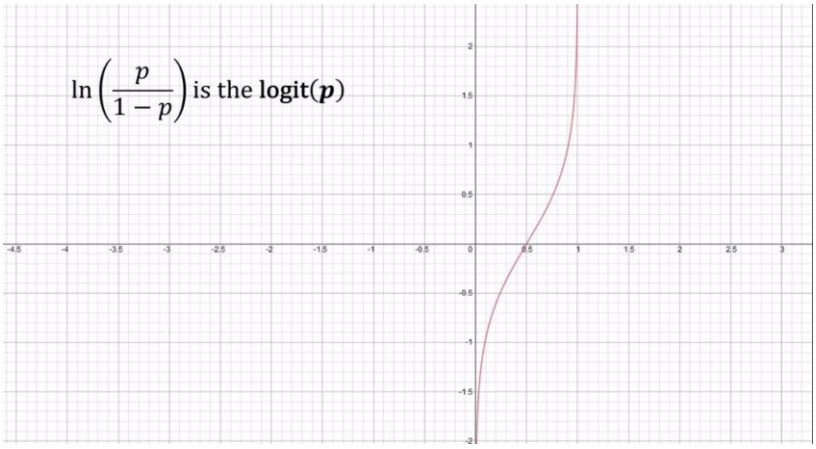

Figure 7: Log Curve

To get Probabilities on Y-Axis we take its inverse.

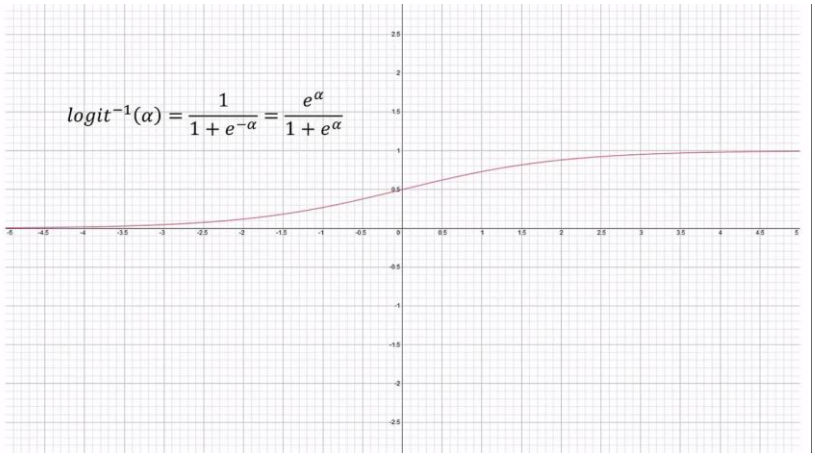

Figure 8: Inverse Logit Curve

Where $\boldsymbol{a = \beta_0 + \beta_{1x}}$ Thus, the Logistic function can now be written as: $\boldsymbol{F(x) = 1/(1 + e^{-(\beta 0 + \beta 1x)})}$
Where β_0 and β_1 are coefficients for input variable x. F(x) is the probability of the event to occur.

The value of F(x) ranges from zero to one.

Approach Taken for Logit Regression

Using all 38 variables:

Firstly, the Logit Model was made using all available 38 variables:

1. This gave very low AUC (Area Under Curve) value of 0.9097 compared to other models.
2. Also, this gave high p-value test results for a lot of variables.
3. Using 16 Variables

Finally only 15 variables were chosen which has acceptable p-values of less than or equal to 0.05 Same, Logit Model test was done in this new dataset of 15 variables.

Following are results.

- High AUC of 0.9697
- High Accuracy of 94.73%

3.4. Evaluating Model

After all the machine learning models are created they are then evaluated on the basis of test dataset.

3.4.1. Error Matrix

3.4.2.

An Error matrix is a table that is used to describe the performance of a classification model (or "classifier") or regression model on a set of test data for which the true values are known. It is also known as a contingency table or a confusion matrix.

Table 5: Decision Tree Error Matrix

Predicted	**Actual**		
	1	2
1	58	6
2	2	66

Table 6: SVM Error Matrix

Predicted	**Actual**		
	1	2
1	57	10
2	3	62

Table 7: Logistic Regression Error Matrix

Predicted	**Actual**		
	1	2
1	78	2
2	5	48

3.4.3. Receiver Operating Characteristic (ROC) [3]

3.4.4.

ROC curve, is a graphical plot that assesses predictive behaviour independent of error costs or class distributions. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

- True positive (TP): The actual negative class outcome is predicted as negative class from the model.
- False positive (FP): The actual negative class outcome is predicted as a positive class outcome.
- False negative (FN): The actual positive class outcome is predicted as negative class from the model.
- True negative (TN): The actual class outcome excluded is also predicted to be excluded from the model.

Positive is taken as the company going bankrupt (Bankrupt=2) and Negative is taken as company being healthy (Bankrupt=1).Based on these four parameters the performance of algorithms can be adjudged by calculating the following ratios.

- Accuracy (%) = (TP + TN) / (TP +FP +TN + FN)
- TPR (%) = TP / (TP + FN)
- FPR (%) = FP / (FP + TN)
- Precision (%) = TP / (TP +FP)

ROC curve represents the accuracy of a classifier or Model. Accuracy, Area under Curve & Precision were calculate for all models and represented in Table 10.

Table 8: Result of Bankruptcy prediction using machine learning algorithm

S.NO.	***RESULT***	***ACCURACY***	***AUC***	***PRECISION***
1	Decision Tree	93.94 %	0.9423	97.05 %
2	SVM	90.15 %	0.9495	95.38 %
3	Logistic Regression	94.73 %	0.9697	93.97 %

Summary from Rattle:

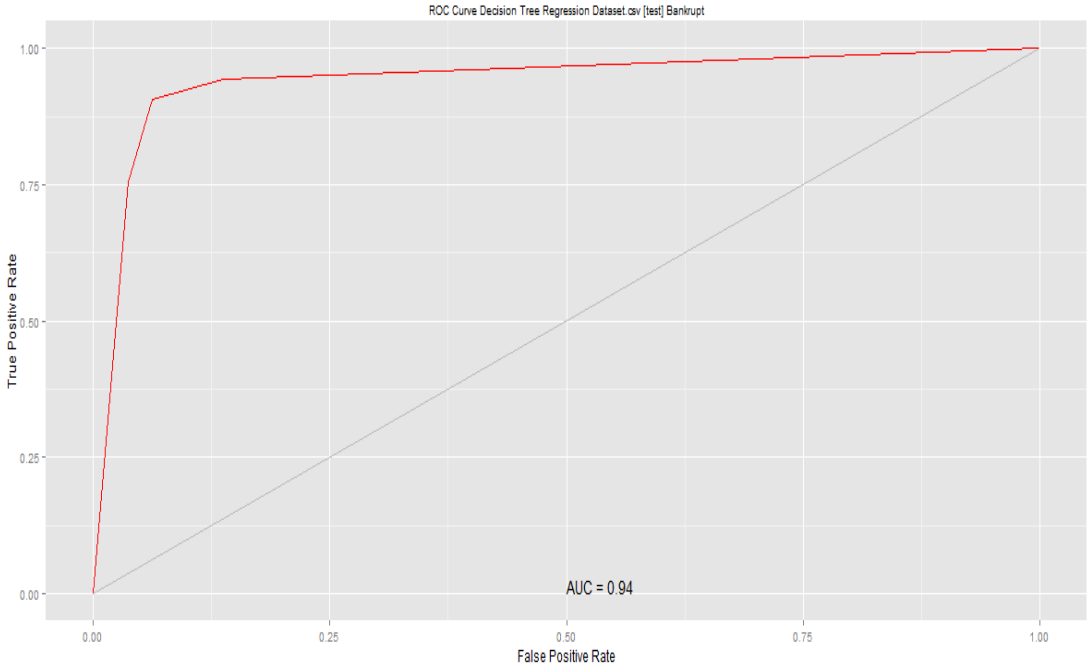

Figure 9: ROC of Decision Tree

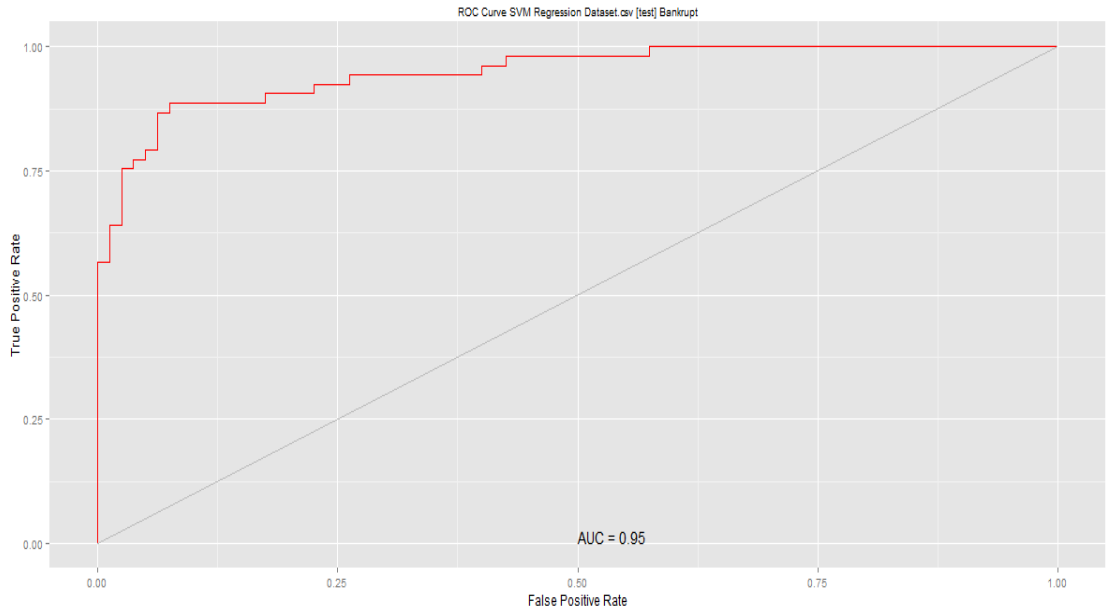

Figure 10: ROC of SVM

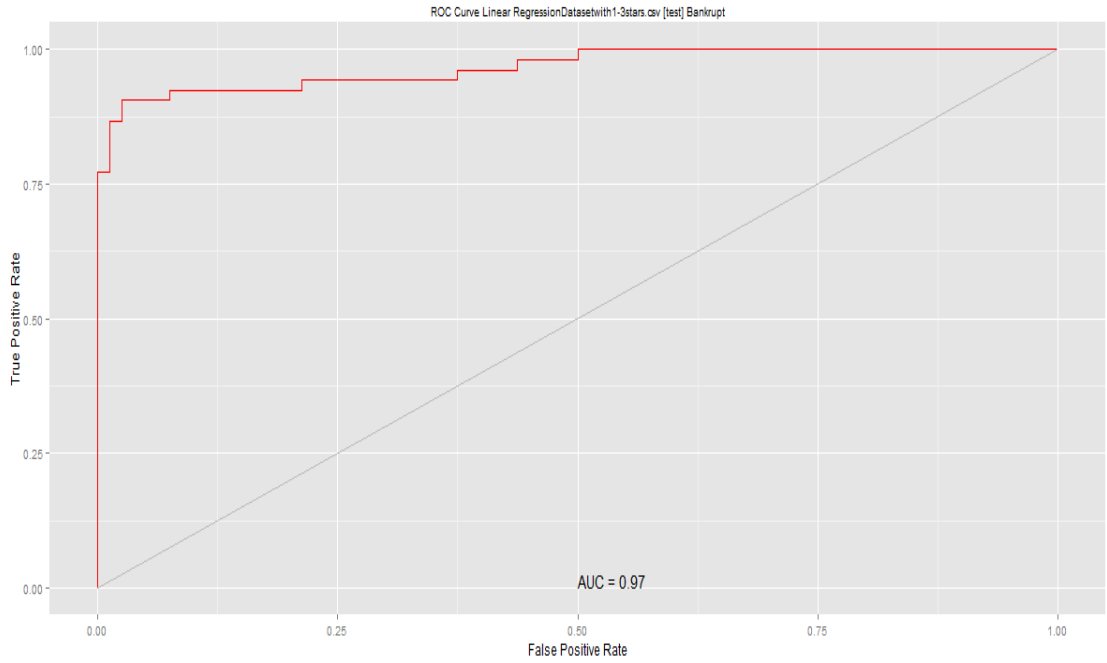

Figure 11: ROC of Logistic Regression

4. RESULT

Accuracy, Area under Curve (AUC) & Precision were calculated for all models and represented in the table 8. Logistic Regression has highest Accuracy and Precision and very high value of Area under Curve from ROC. Hence, Logistic Regression is chosen as it has outperformed other machine learning algorithms.

5. CONCLUSION

In this paper we presented an efficient classification model using Logistic regression for predicting Bankruptcy for an organization. The experimental result shows that after iteration of model with different number of variables improved the effectiveness and efficiency of the model. The experimental result suggests that Logistic regression seems to be more accurate machine learning algorithm for prediction of bankruptcy.

5.1. Limitation of study

- Data is not adjusted for Inflation and seasonal changes. Thus, Results may differ.
- This dataset includes many companies but for different business domains like manufacturing, finance, ecommerce and medical which usually have their own peculiarity and similarities, this model should be used with caution.

5.2. Suggestions

- Researchers should work on Industry based modeling.
- Three algorithms have been used in this paper, however more algorithms like Naïve Bayes, Genetic, and KNN exists.
- Each country has their own laws, culture, socio demographic patterns and economical status. To get reliability in predictions, country wise parameters must be included.
- The study should be done for longer periods.

REFERENCES

[1] Kalyan Nagaraj,Amulyashree Sridhar,(January 2015) "A Predictive System for Detection of Bankrupty using Machine Learning Tecniques", IJDKP-Vol.5, No.1.
[2] Graham Williams,(January 1, 2011) "Data Mining With Rattle and R_ The Art of Excavating Data for Knowledge Discovery".
[3] Sadiq Hussain, G.C. Hazarika, (2014) "Educational Data Mining Model Using Rattle" Vol. 5, No. 6
[4] Prajwala T R,(January 2015) "A Comparative Study on Decision Tree and Random Forest Using R Tool", IJARCCE,-Vol. 4, Issue 1
[5] Financial Ratios, Bankruptcy Prediction –"http://research.ics.aalto.fi/eiml/datasets.shtml"
[6] Eibe Frank ,(January 2000) "Pruning Decision Trees and Lists"
[7] R. Berwick, "An Idiot's guide to Support vector machines (SVMs)"
[8] Tree Based Model, 'http://www.statmethods.net/advstats/cart.html'
[9] SVM hard or soft Margins, 'http://stackoverflow.com/questions/4629505/svm-hard-or-softmargins'
[10] M.A. Sprengers (21st August 2005) "Bankruptcy Prediction Using Classification and Regression Trees".
[11] Charles X. Ling , Jin Huang, 'AUC: a Better Measure than Accuracy in Comparing Learning Algorithms'.
[12] Qi Yu, Amaury Lendasse and Eric S´everin,'Ensemble KNNs for Bankruptcy Prediction.
[13] Q. Yu, Y. Miche, A. Lendasse and E. Séverin,'Bankruptcy Prediction with Missing Data'.
[14] Correlation and dependence 'https://en.wikipedia.org/wiki/Correlation_and_dependence'
[15] 'http://stackoverflow.com/questions/29197213/what-is-the-difference-between-rel-error-and-x-error-in-a-rpart-decision-tree'-Stack OverFlow
[16] Tom Mitchell (1997), Machine Learning, McGraw Hill.