**Predicting Corporate Bankruptcy Case**

**Answers and Executive Summary**

**Answers to Questions:**

1. Clustering is the data mining technique that would be appropriate in assessing whether there are groups of variables that convey the same information since it is a technique that works by partitioning data into meaningful sub groups or clusters in such a way that highly similar data are clustered together (intra cluster) and are well separated from other data. Since the given data set provides a lot of ratios in which some appear very similar to another this helps us separate them and use a reduced set of parameters for analysis thereby eliminating multicollinearity.
2. **Profiling Characteristics of Bankrupt Firms:**

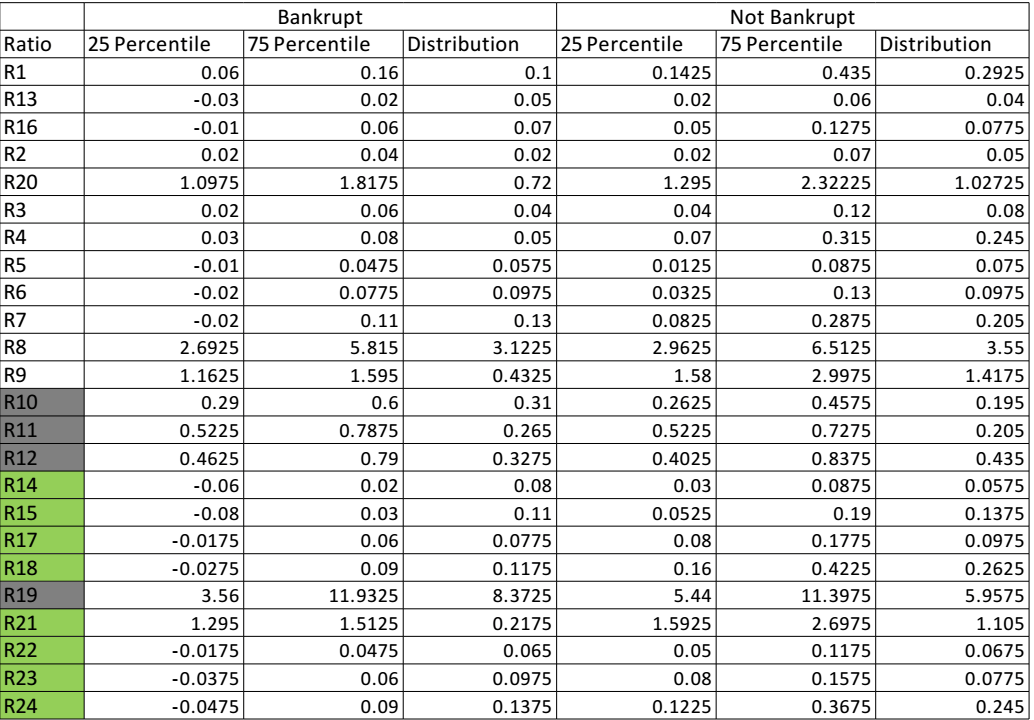
* Helps us understand how different data are dependent on/independent from each other.
* Helps us get a preliminary view of which parameters significantly affect a firm’s chances of bankruptcy and its closely associated variables and helps us find out the variables least affecting the outcome.
* Helps us find collinearity between data which will help us operate with a reduced set of parameters thereby reducing computational complexity and also reduce the amount of processing needed.

**Simply Predicting (Black Box Style):**

* Training the model when using a black box style of approach can be difficult. Due to the black box style of work, the visibility of the work is greatly limited
* Different parameters have to be tweaked to the right measure for this approach to achieve efficient results. Model also becomes largely over fitted.
* Increased Computational complexity and needs longer time to process as every variable is given importance

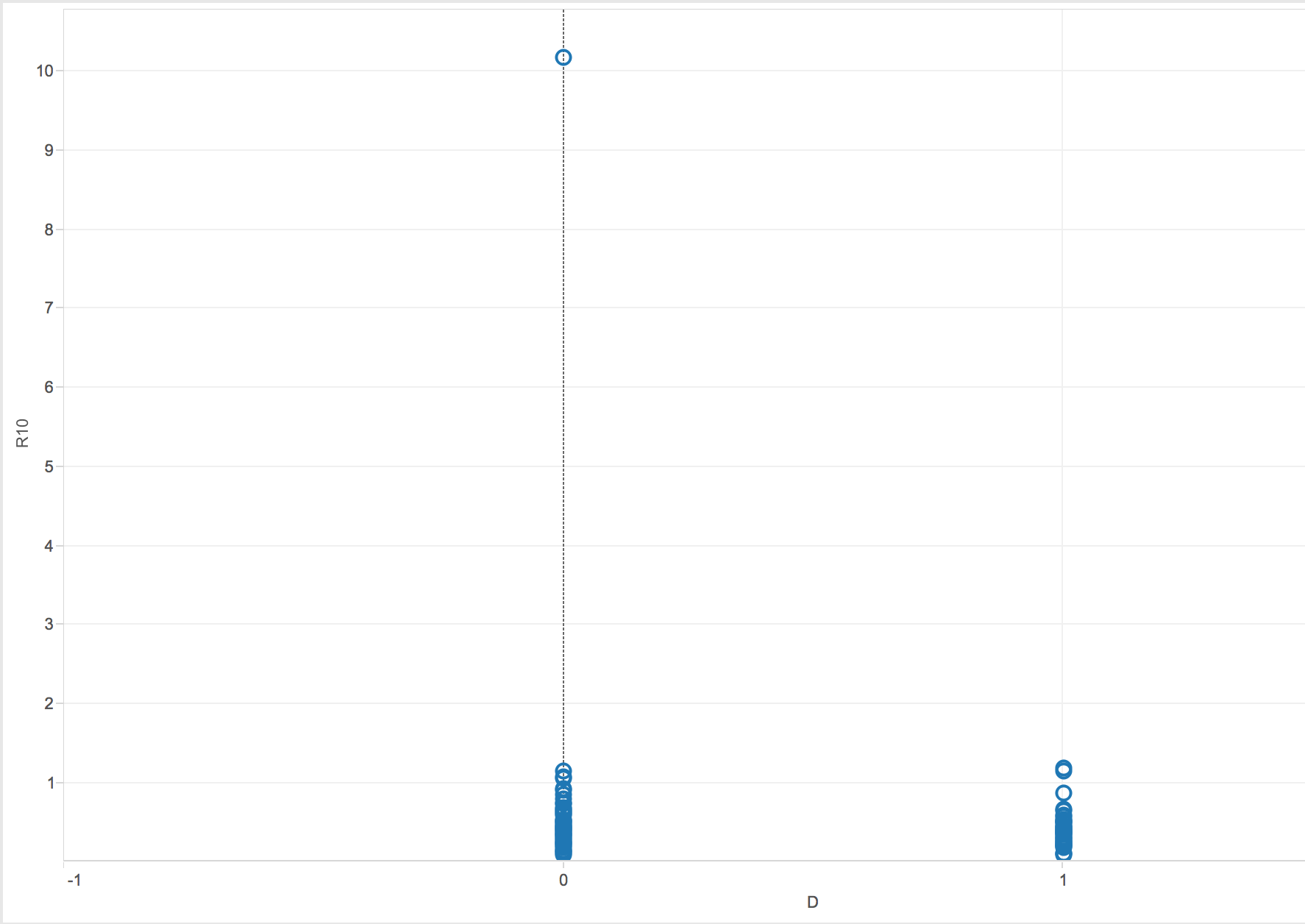
Looking at both the methods mentioned above it can be said that the black box style of approach is always the second best method to approach the analysis given that the understanding of the problem (Profiling Characteristics) is first completed. So a combination of both goals would work better. CART and Random Forrest are some classification methods for the first approach and neural networks is a good method for the black box style approach.

1. The following table was compiled after studying each of the ratio’s performance against the Bankrupt/Healthy Variable (D) in a boxplot. The 25th and 75th percentile values of each ratio is listed below. From these the ratios that significantly affect the result of a firm becoming bankrupt or staying healthy are highlighted in Green and those which least affect are marked in grey.

****

Also we analyzed if a given ratio will affect the outcome significantly or not by studying its scatterplot analysis. When the points are very closely located to each other we found the parameter to least affect the outcome of the output variable.

Example of one such variable **(R10)** is shown below.



1. We chose to use the following three models to predict whether a firm goes bankrupt or stays healthy. They are:

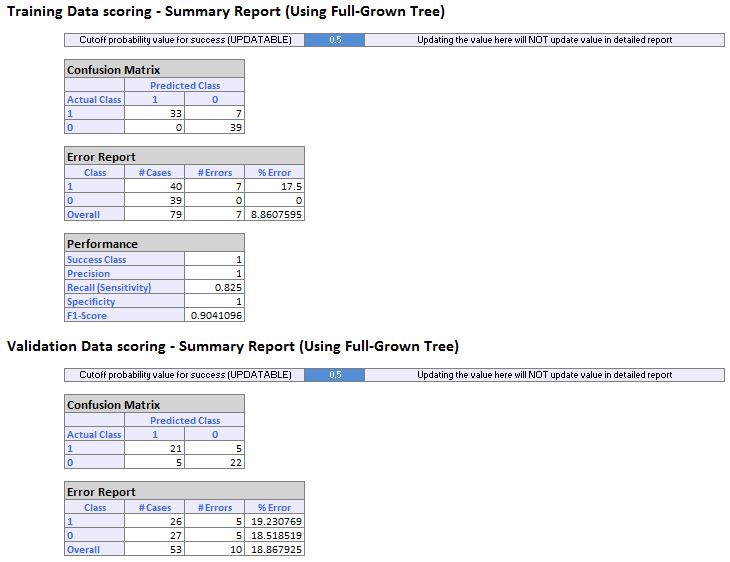
**Methods that use profiling characteristics**  
- CART  
- Random Forrest  
  
**Black box approach method**  
Neural Networks

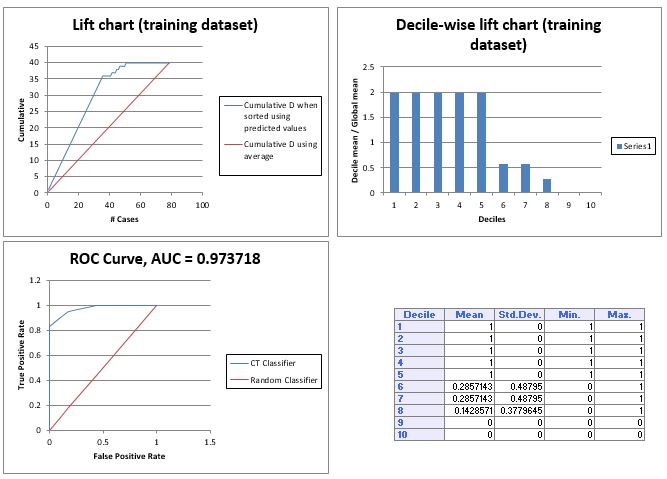
As we had already found out the parameters that significantly affect the outcome of the firm being bankrupt or remaining healthy using the boxplot analysis in XLMiner. We chose to use only those parameters in the model for the first two methods which allows us to use a reduced set of parameters for better results and faster processing.

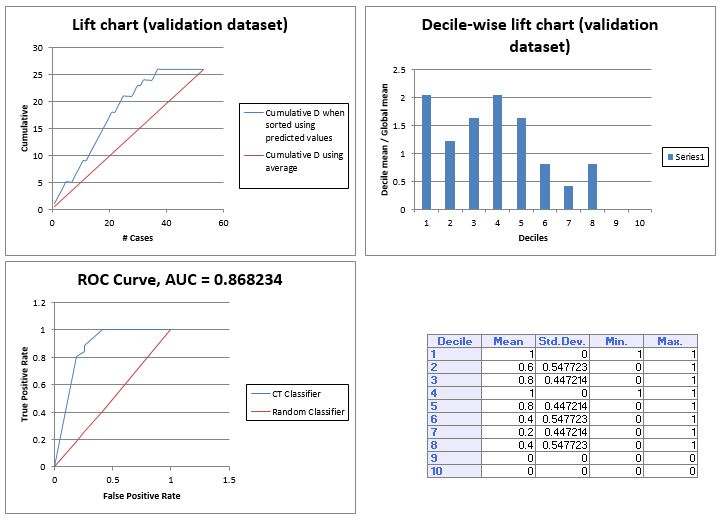
The results from each method (Error Report, Training and Validation Lift Charts, Decile plot analysis) is given below.

**Results from CART:**

CART Algorithm was run using the classification principle with only the outcome significant parameters selected for modelling. From the results shown below, we can say that the CART algorithm which builds a single decision tree selecting the highest significant parameter as the starting node and then building the tree using the other parameters on descending order of significance has performed well.  
  
With this studied, we go into the next algorithm, Random forest which aims to improve the results provided by CART Algorithm.

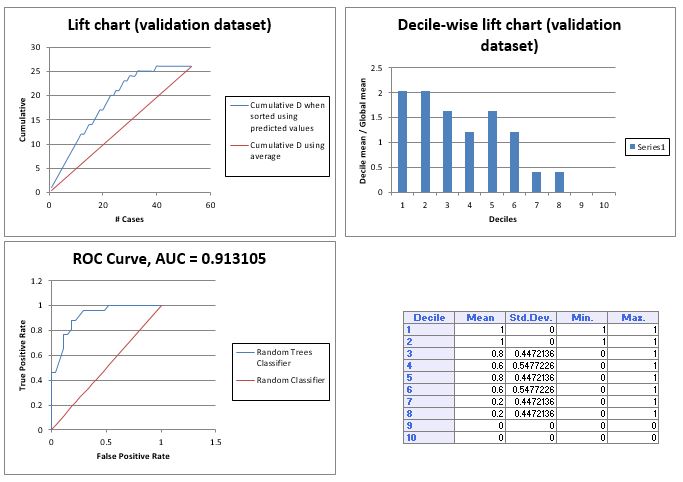
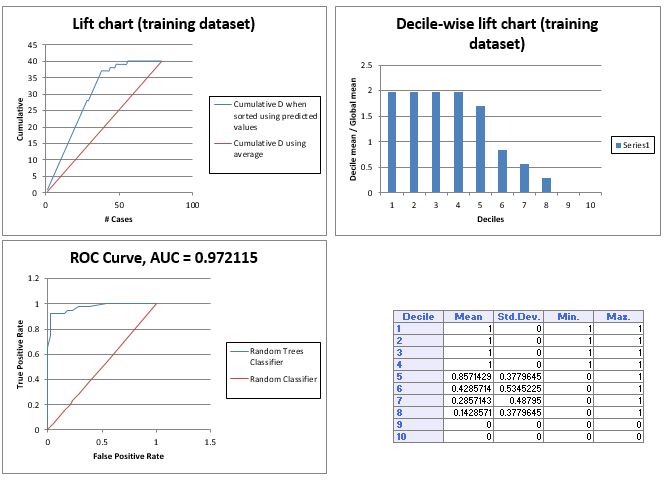
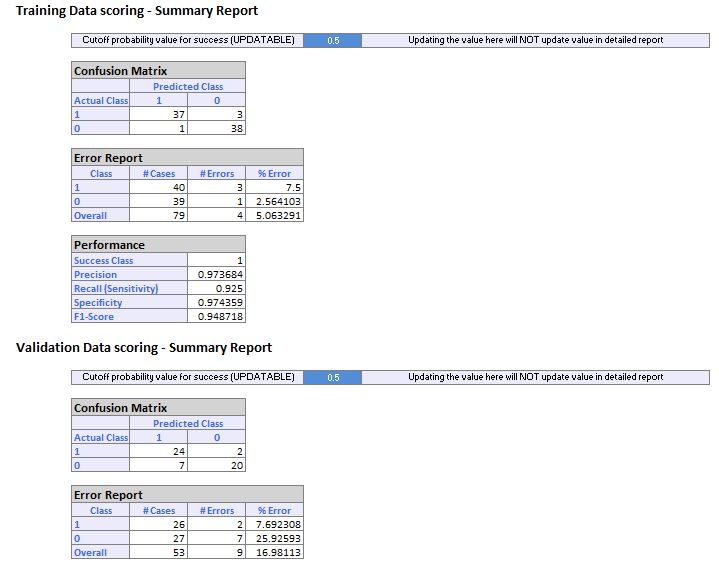
****

****

****

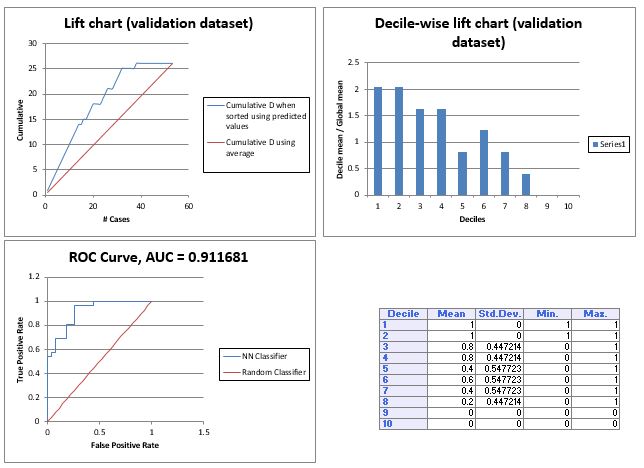
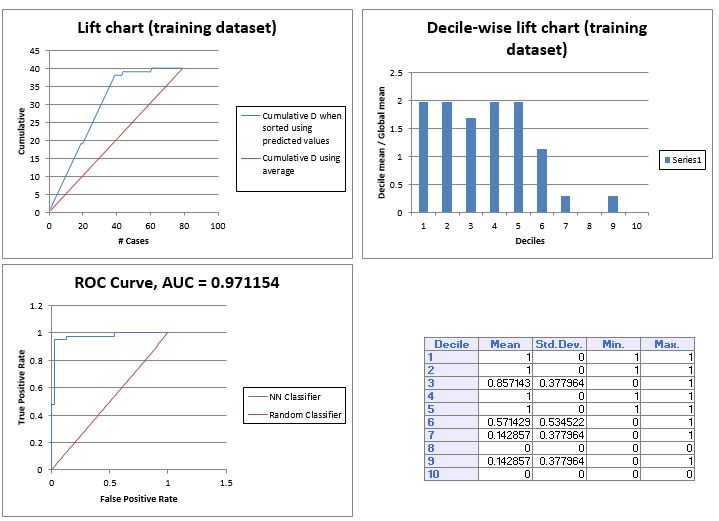
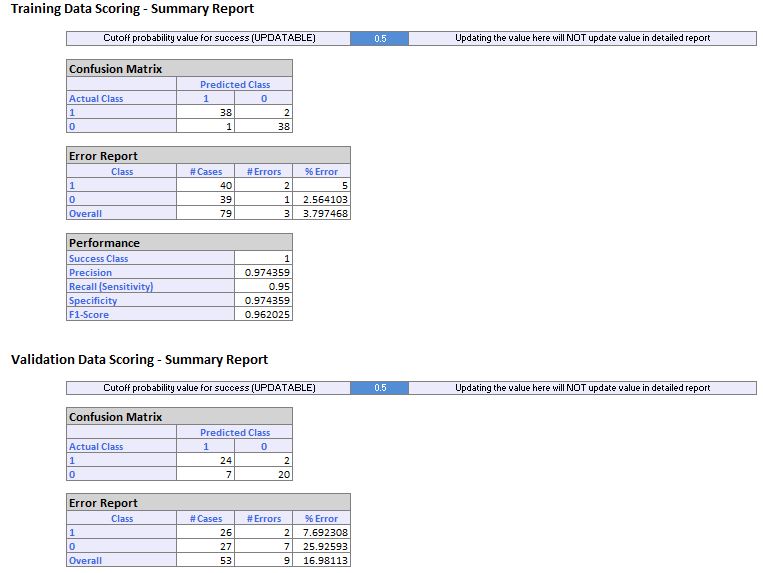
**Results from Random Forrest Algorithm:**

Random Forrest algorithm was run using the classification principle with only the outcome significant parameters selected for modelling. From the results shown below, we can say that the RF algorithm which builds multiple decision trees using the bootstrap principle has provided a better result when compared to CART. This is not surprising as the working of Random forest which involves creating multiple decision trees into a group (Forest) and then running each data point through them giving the outcome based on a polling basis between the trees thereby reduces bias and variance and improves accuracy of prediction.

The results are shown below.  
  
****

**Black box approach method**  
Neural Networks  
  
As discussed earlier, the black box method goes into analysis with very limited to no understanding of the dataset and the relationship between the different parameters.   
So no matter how good the performance of the model, they suffer significantly from issues such as over fitting, taking longer time to run, and taking an even bigger effort to train.

When the Neural networks method is run on the current dataset setting the gradient from 0.1 to 0.3 to have a larger gradient step that covers more data and also increasing the Epochs from 30 to 50, the following results were observed.  
  
*From these results it is easily evident that the black box method run here using the neural networks method suffers from over fitting when we look at the performance and more specifically the error percentage difference them.*

Also looking at the values in the training data set for the percentage error in NN and comparing it with similar values of CART and RF. NN returns the lowest error percentage as the data is largely over fitted to the training dataset. The point where this approach fails is the difference between the error percentages in training and validation  
  


**5.** Looking at the results shown above, we do an analysis of which parameters have affected the outcome more than the others in detail. The ratio R14 (Income/Assets) is a significant parameter as for a firm to remain healthy the amount of income earned per total assets should be high and its values can easily tell whether a firm will remain healthy or go into bankruptcy. The ratio R15 which looks at the income earned per debts should be high i.e. the income should be more and the debts as low for this ration to have a higher value as higher the value of this ratio the healthier the firm. The ratios R17, R18 which look at the Income plus depreciation against the Assets and debts a firm has have to be higher which means the firm must have a high income even after depreciation is factored into it to remain steady. R21 which studies the assets per debts a firm has it has to be high as lower the debts a firm has the more chance it has of not falling into bankruptcy. The ratios R22, R23, R24 which look at Working capital from operations against sales, debts and assets are also significant ratios as a firm that has a high working capital from its operations will remain healthy and avoid bankruptcy.

**EXECUTIVE SUMMARY**

**Problem Statement:**The given dataset provides us information about various ratios that help to determine the health of a firm financially with which we can classifying a given firm as financially bankrupt   
or healthy. The dataset lists 132 firms from which a statistical data model has to be built.

**Approach:**   
The data set was looked at first for irregularities in the provided data and an exploratory analysis was conducted to understand the dependence and independence of any given variable with another as a lot of financial ratios going into the model would increase its complexity and time taken to run. Also the ratios which significantly affect the outcome are studied by conducting a box plot analysis with the ratio going against the prediction parameter “D” and only those which significantly affect are selected for modelling.  
  
From this analysis of identifying which parameters have affected the outcome more than the others in detail, We find that the ratio R14 (Income/Assets) is a significant parameter as for a firm to remain healthy the amount of income earned per total assets should be high and its values can easily tell whether a firm will remain healthy or go into bankruptcy. The ratio R15 which looks at the income earned per debts should be high i.e. the income should be more and the debts as low for this ration to have a higher value as higher the value of this ratio the healthier the firm. The ratios R17, R18 which look at the Income plus depreciation against the Assets and debts a firm has have to be higher which means the firm must have a high income even after depreciation is factored into it to remain steady. R21 which studies the assets per debts a firm has it has to be high as lower the debts a firm has the more chance it has of not falling into bankruptcy. The ratios R22, R23, R24 which look at Working capital from operations against sales, debts and assets are also significant ratios as a firm that has a high working capital from its operations will remain healthy and avoid bankruptcy.

**Methods:**The analysis of the dataset was conducted using two different methods of approach:  
  
**Methods that use profiling characteristics Black box approach method**  
- CART - Neural Networks  
- Random Forrest

The Advantages and disadvantages of each method are listed below:   
  
**Advantages of Profiling Characteristics of Bankrupt Firms:**

* Helps us understand how different data are dependent on/independent from each other.
* Helps us get a preliminary view of which parameters significantly affect a firm’s chances of bankruptcy and its closely associated variables and helps us find out the variables least affecting the outcome.
* Helps us find collinearity between data which will help us operate with a reduced set of parameters thereby reducing computational complexity and also reduce the amount and speed of processing needed.
* There is high visibility of model working which helps us conduct sanity checks on performance and make fixes wherever necessary.

**Disadvantages of Profiling Characteristics of Bankrupt Firms:**

* Requires domain knowledge and time to understand data behavior.

**Advantages of Simply Predicting (Black Box Style):**

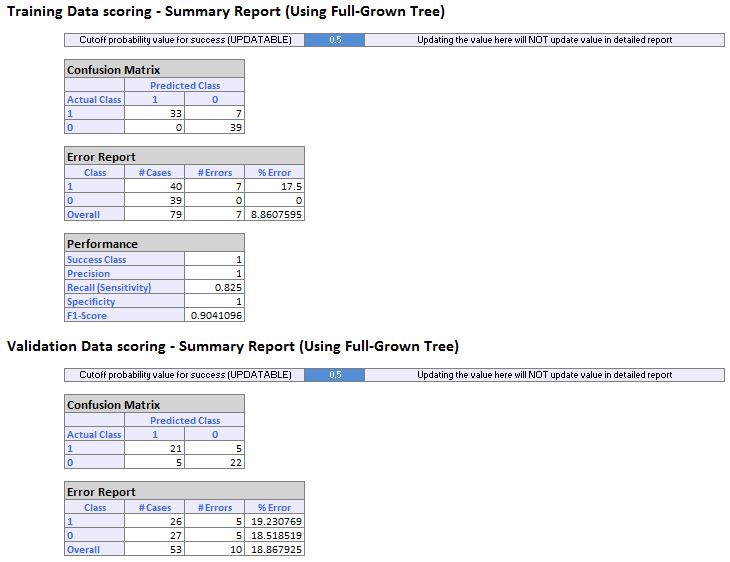
* Quick to start. Has an improve-as-you-go method of working
* Model can be tweaked
* Runs multiple iterations to provide a better model

**Disadvantages of Simply Predicting (Black Box Style):**

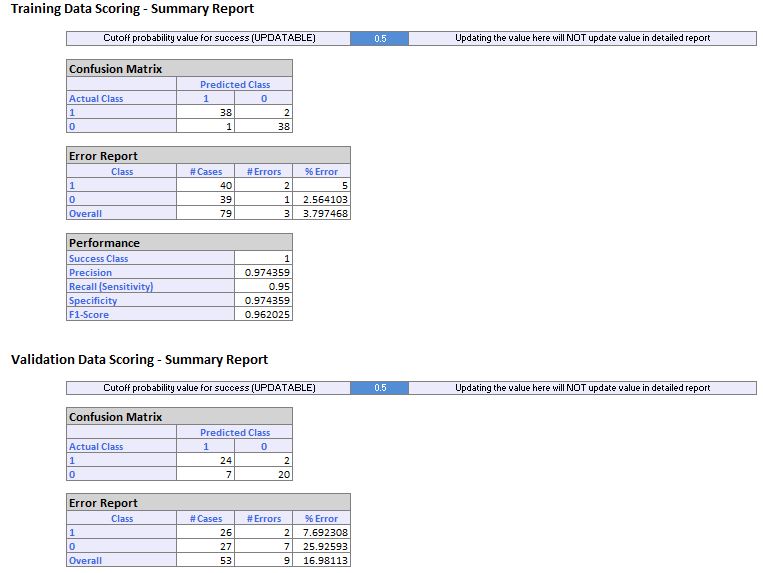
* Training the model when using a black box style of approach can be difficult. Due to the black box style of work, the visibility of the work is greatly limited
* Different parameters have to be tweaked to the right measure for this approach to achieve efficient results. Model also becomes largely over fitted.
* Increased Computational complexity and needs longer time to process as every variable is given importance

**Result:***When the data analysis of the corporate bankruptcy was conducted using the three different algorithms* ***CART vs Random Forrest vs Neural Networks,*** *It was found that random forest gave the best results in terms of assessment conducted on the model’s performance on a validation partition. The results of which are shown below.*

**CART Results**



**Random Forrest- Results**



**Neural Networks Results:**

