FRA\_NitinYadav

Nitin Yadav

14/02/2020

## Business Problem

We have been given the data for various companies in terms of financial information which includes Assets, Capital, Income, Sales and key financial ratios. Based on the Networth “Next Year” we have to build a predictive model to identify if the company will default next year or not. We will use the Logistic regression model to build the model and check the model performance on the Test set for the accuracy.

## Observation from data

1. Total of 3541 entries with 53 variables
2. Dependant variable is Default wwhich is derived from “Net worth Nextyear”
3. All independent variables are numeric except the dependent which is categorical
4. Missing values are present in most of the variables
5. Outliers are present in the key predictor variabes like “Income”, “Assets”, “Net worth” and other variables
6. Default rate is around 6.9% for the overall datapoints

## Outlier and Missing value Treatment

Check histogram and box plot for income and Sales as they have highest maximum amount and consider only values that are less than 50000. Still after doing this we notice that there are lot of outliers present in multiple variables in dataset, but still we will go ahead with the data as we will lose important data aspect.

Also we notice that there are lot of NA values present in the dataset,almost in many of the columns , so we will treat the NA values.

We have option to do the imputation using mean, median or using the MICE package. We will not use Mice package as couple of variable names in the dataset have space in between them, hence the other preferred method is KNN imputation which will use the nearby variables to calculate the imputation value.

We will also take the variables of interest in our dataset and the final dataset after treating for outliers and missing value. The final dataset consist of 3302 entries and 22 columns. The test data has 670 entries and 22 columns

## New Variable creations

We will create the below new variables to be included in the dataset. Liquidity ratio are already calculated as part of dataset, so we will not crete variables for liquidity.

Profitability

1 . Return on Asset 2. Return on Equity

Leverage

1. Debt to Equity ratio - This ratio already given as part of dataset
2. Debt to Capital ratio

Liquidity Ratio

1. Current Ratio - given as part of dataset
2. Quick Ratio - given as part of dataset
3. Cash ratio - given as part of dataset

Company Size Ratio

1. Income to Sales Ratio
2. Profit to Sales Ratio
3. Profit to Income Ratio

## Multicolinearity Analysis

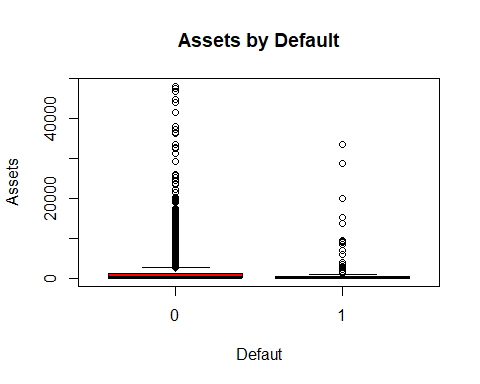
We notice that the financial ratios have been derived from the individual variables , hence if we check the correlation plot , we will surely notice the collinearity among variables,so we will build the logistic regression model and after that we will check multicolinearity using VIF and will also create a correlation plot.

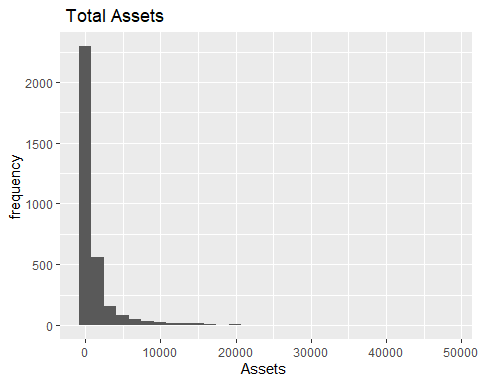
The same has been covered as part of LR model building

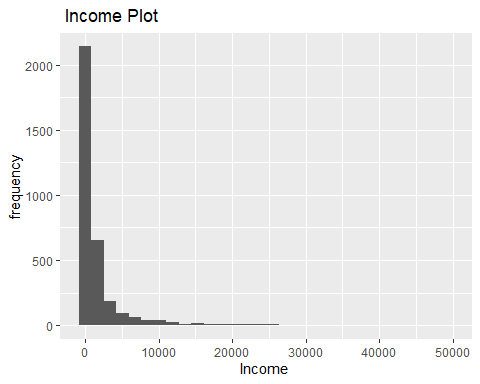
## Univariate and Bivariate analysis

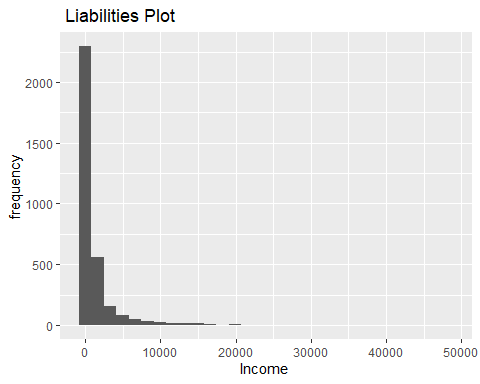
For Train and Test Data

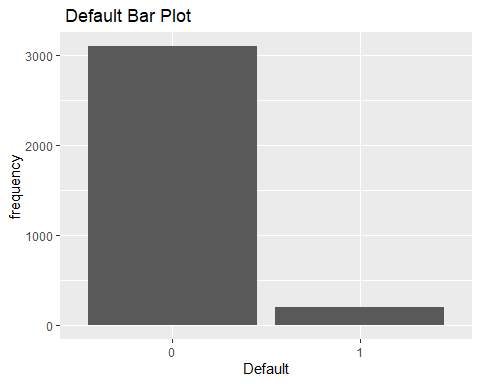
1. Defaulting companies loses profit by 6 unit per 100 units and non defaulting companies earn profit by 3 units per 100 units of income.
2. Non Defaulting companies have more assets compared to defaulting companies
3. PAT , Cash profit and Return on Equity is more for Non defaulting companies compared to Defaulting ones
4. Debt to Equity ratio for defaulting companies is higher means they have more debt to finance their operations and hence have higher chance of no repayment and default.
5. Total liabilities are more and Total sales are less for Defaulting companies compared to Non Defaulting companies.

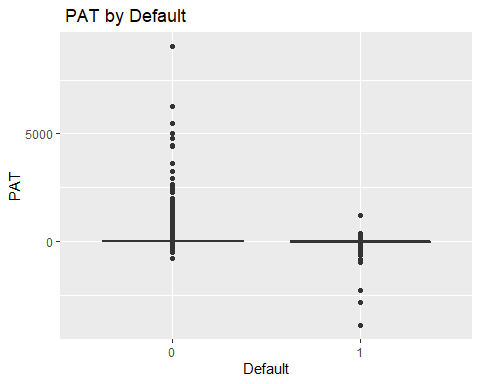


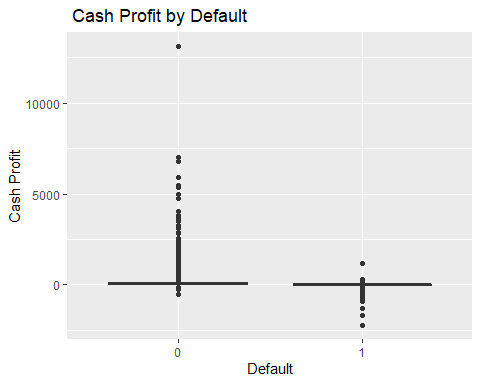


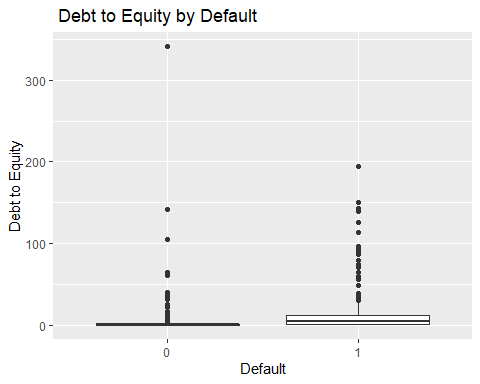


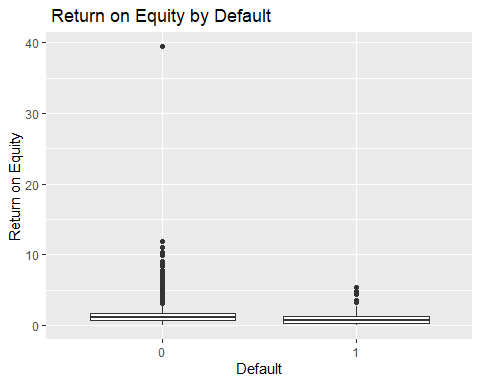


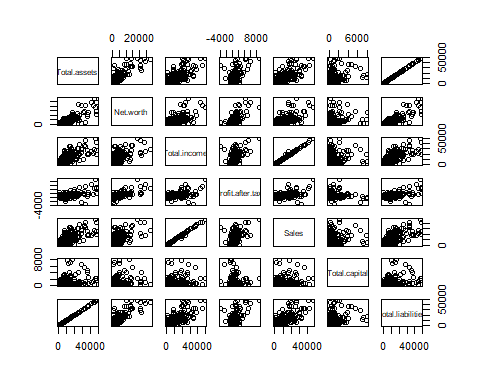


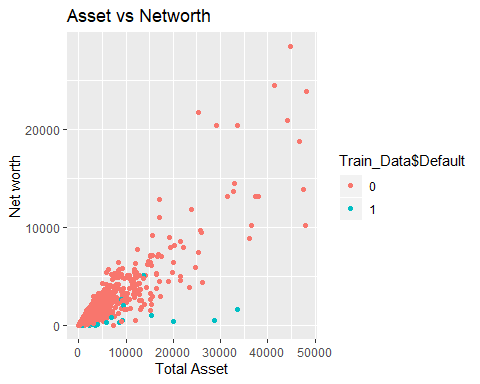


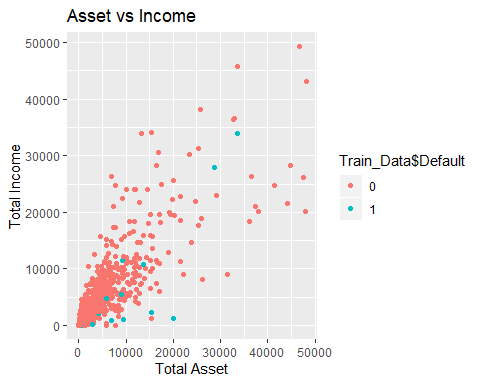


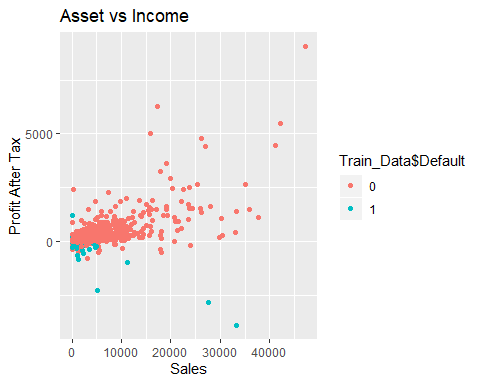


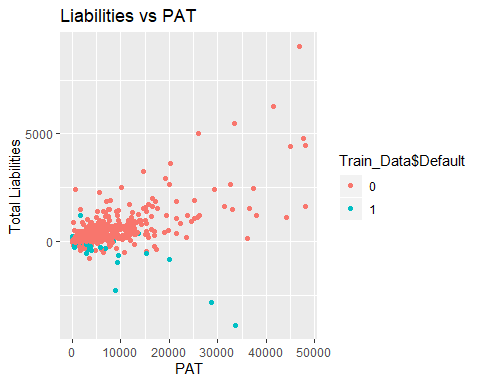


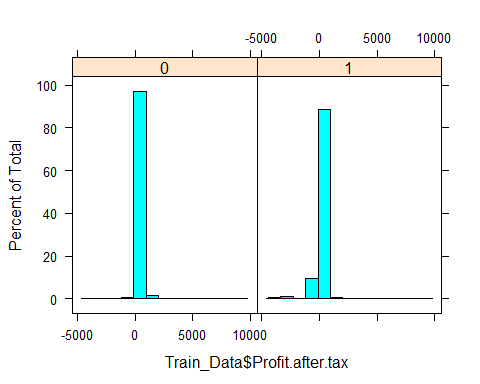












## Logistic Regression model on important variables

We build the model on the below variables

1. Profit to Income
2. Income to Sales
3. Current Ratio
4. Debt to Equity
5. Debt to Capital
6. Return on Equity
7. Return on Asset

We notice that Probability of default is a logit function with coefficient and intercept. The positive sign in the coefficient indicates that as the value increases , the probability of default increases and negative sign means they are inversly proportional ie as the value increases chance of default decrease..for eg as PAT increase , risk of default decrease

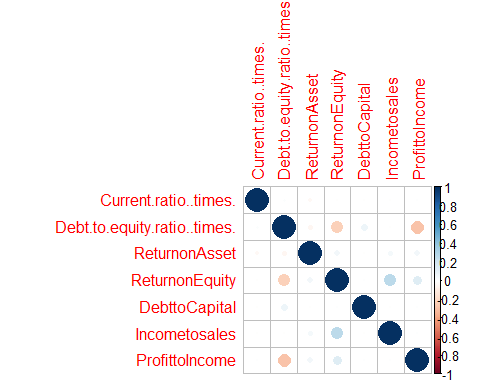
These coffiencients are meanigful and they are signficantly different from zero and hence are important predictors in identifying default

when we check the VIF for multicolinearity we find that all the variables have scores less than 3 so they are are independent. The same is visible from correlation plot.

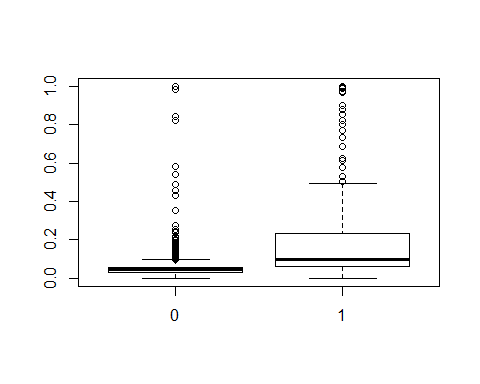
## glm(formula = Train\_Data$Default ~ ProfittoIncome + Incometosales +   
## Current.ratio..times. + Debt.to.equity.ratio..times. + DebttoCapital +   
## ReturnonEquity + ReturnonAsset, family = binomial, data = Train\_Data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -7.2591 -0.3444 -0.2962 -0.2384 4.5381   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.265359 0.151902 -14.913 < 2e-16 \*\*\*  
## ProfittoIncome -0.065495 0.023558 -2.780 0.00543 \*\*   
## Incometosales 0.012269 0.005013 2.447 0.01439 \*   
## Current.ratio..times. -0.039414 0.015975 -2.467 0.01361 \*   
## Debt.to.equity.ratio..times. 0.087283 0.011978 7.287 3.17e-13 \*\*\*  
## DebttoCapital -0.015525 0.006725 -2.308 0.02098 \*   
## ReturnonEquity -0.071093 0.033580 -2.117 0.03425 \*   
## ReturnonAsset -0.644777 0.126185 -5.110 3.23e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1498.2 on 3301 degrees of freedom  
## Residual deviance: 1260.7 on 3294 degrees of freedom  
## AIC: 1276.7  
##   
## Number of Fisher Scoring iterations: 9

# These coffiencients are meanigful and they are signficantly different from zero and hence are important predictors in identifying default  
# Positive values of coefficient tells that higher the value , higher is chance of default i.e more debt means more chances of default. For negative values , higher the value lesser is chance of default i.e more profit means lesser chance of default  
# when we check the VIF for multicolinearity we find that all the variables have scores less than 3 so they are are independent.  
  
vif(Tr\_L7)

## ProfittoIncome Incometosales   
## 1.059584 1.020419   
## Current.ratio..times. Debt.to.equity.ratio..times.   
## 1.188565 1.502964   
## DebttoCapital ReturnonEquity   
## 1.252656 1.108094   
## ReturnonAsset   
## 1.068494



## Prediction for Train Set



## Model performance for Train Set

We take threshold of 8% to find that below this threshold there is minimum prob. of default and above this threshold the risk of default increases.

OVerall Accuracy - 90% ie in identifying the 1s and 0s correctly which means 10% is missclassification

Sensitivity - 63% ie ability of model to predicting 1s correctly. We can increase this value by reducing the threshold but that will also increase the Loss rate and missclasification.

Specificity - 92% , the ability to identify 0s correctly

Loss - 2.5% current loss stands at 2.5% i.e it is the rate for which the model actually said , company will not default but actually company defaulted.

ROC - 82%

KS Calculation -(Difference between %Cumulative Right and % Cumulative Wrong) KS values intrepret that the model is moderatley good in terms of separating the “1”s and “0”s and we can interpret that the model holds a good prediction ability.

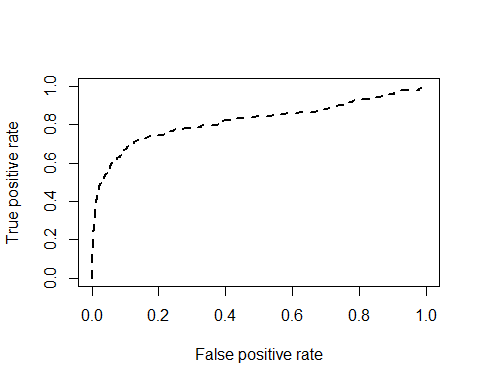
KS Score - 58%

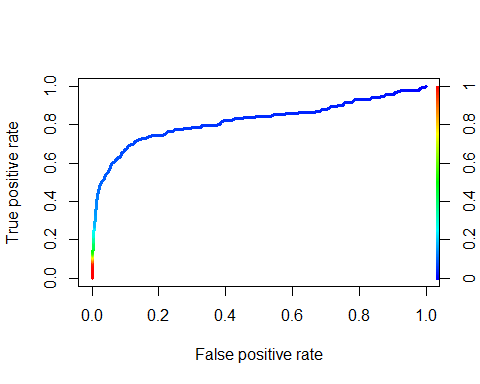
AUC calculation -Larger the AUC and larger Gini better the model is ..Gini = 2 AUC-1),AUC is the % of Box that is under the ROC curve

AUC - 82%

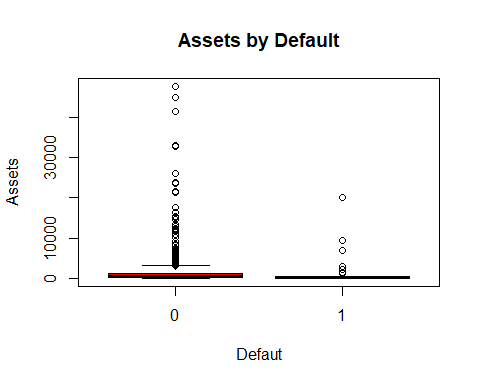
Gini - 42%

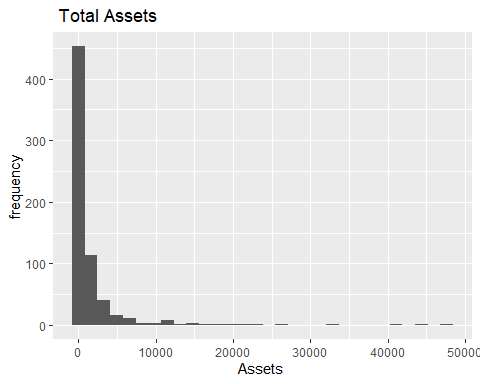
Concordance - 82%

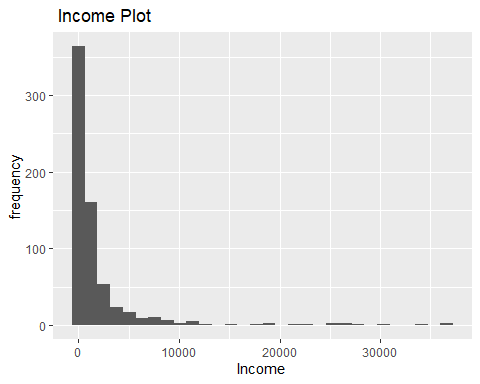


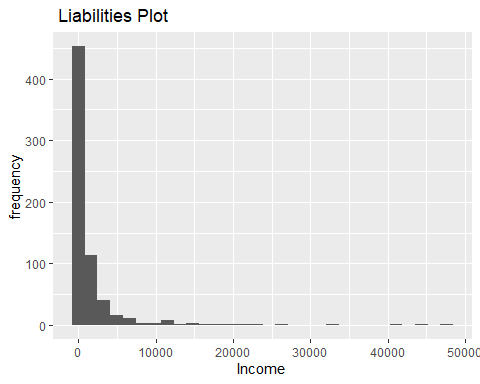


## EDA for Test Data Set

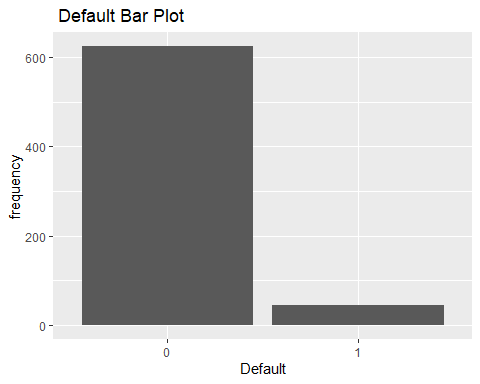


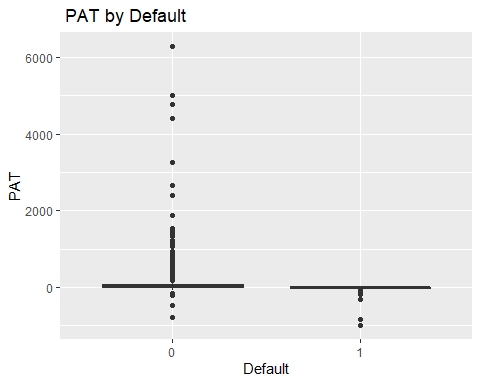


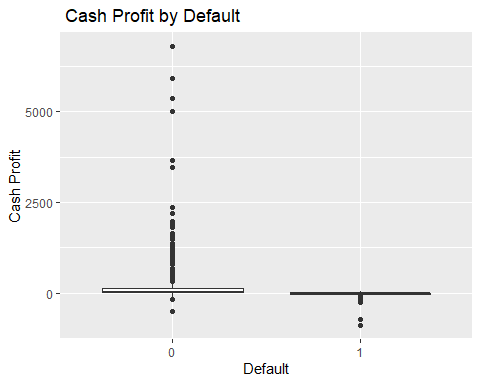


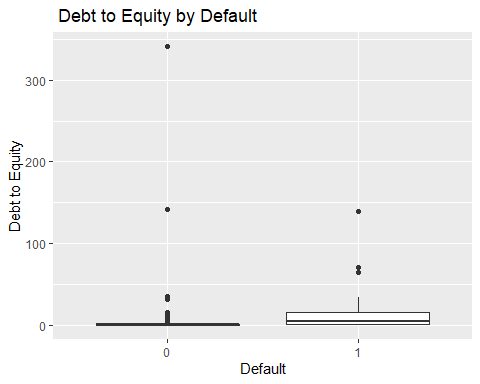


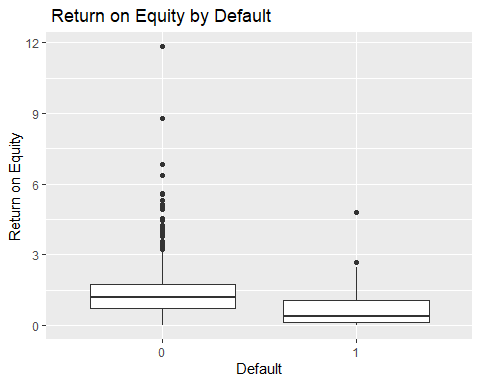
# Categorical

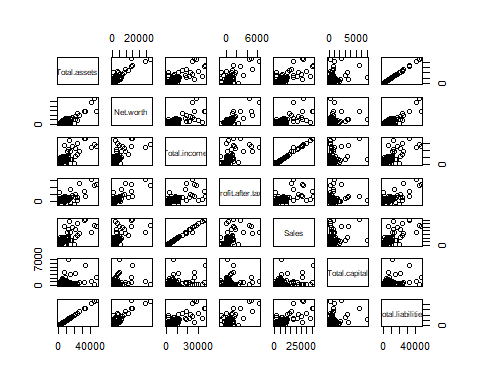


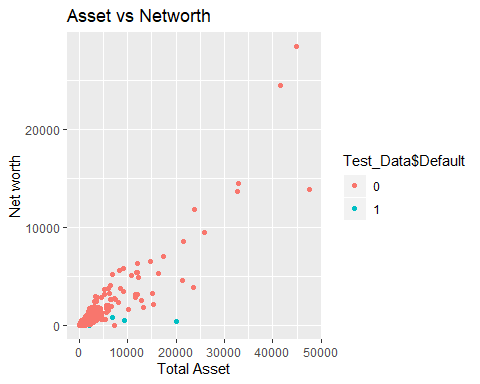


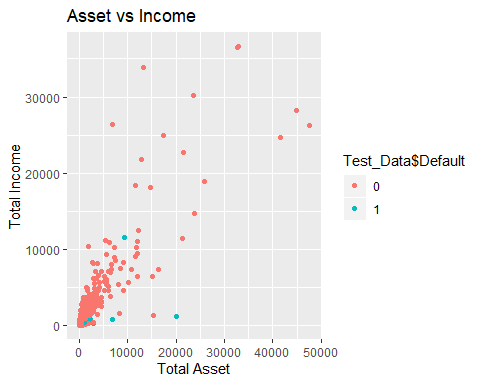


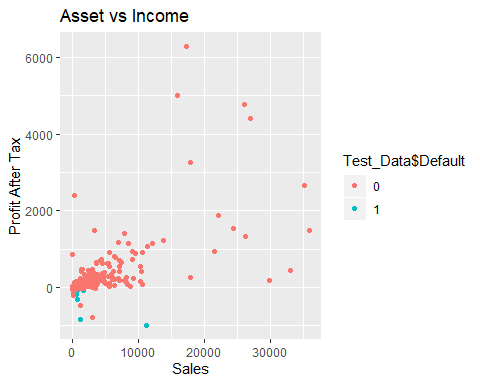


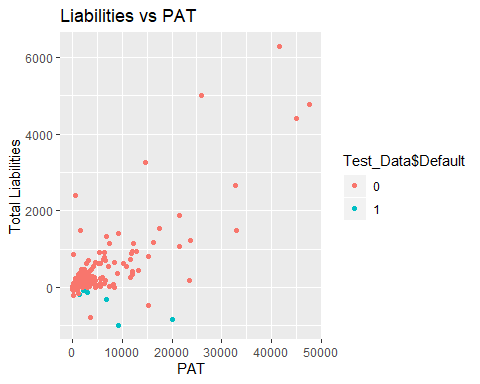


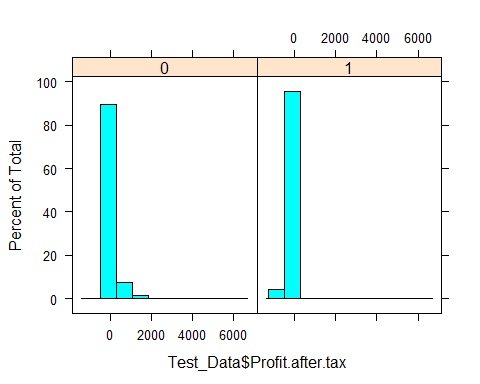




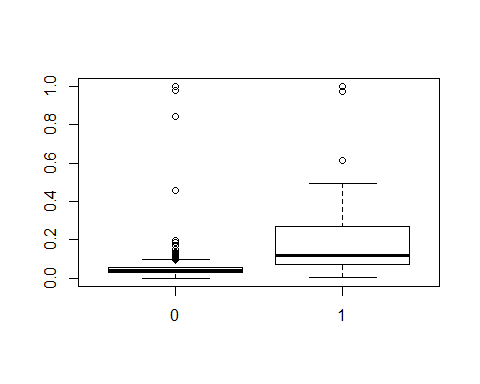








## Doing the prediction using the Test Data



## Model performance for Test Set

We take threshold of 8% to find that below this threshold there is minimum prob. of default and above this threshold the risk of default increases.

OVerall Accuracy - 91% ie in identifying the 1s and 0s correctly which means 10% is missclassification

Sensitivity - 71% ie ability of model to predicting 1s correctly. We can increase this value by reducing the threshold but that will also increase the Loss rate and missclasification.

Specificity - 92% , the ability to identify 0s correctly

Loss - 2.1% current loss stands at 2.5% i.e it is the rate for which the model actually said , company will not default but actually company defaulted.

ROC - 86%

KS Calculation -(Difference between %Cumulative Right and % Cumulative Wrong) KS values intrepret that the model is moderatley good in terms of separating the “1”s and “0”s and we can interpret that the model holds a good prediction ability.

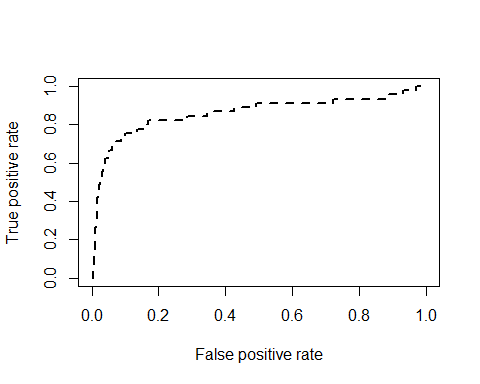
KS Score - 65%

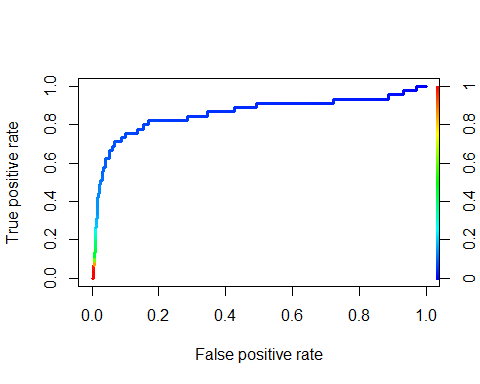
AUC calculation -Larger the AUC and larger Gini better the model is ..Gini = 2 AUC-1),AUC is the % of Box that is under the ROC curve

AUC - 86%

Gini - 46%

Concordance - 86%





## Sorting the data in Descending order for 10 Deciles both for train and Test

When we arrange the train and test data based on 10 deciles we found that 20% of top data for Train set gives a cumulative response rate of 73% , however for test set 20% of top data gives a response rate of 80%.

The data on validation set has performed better in terms of the overall response rate.

## Conclusion

We build the model for both train and test set and check the accuracy and model performance measures , we found that test/validation data model performance measures are better compared to train set. Also when we arranged the data in descending order based on the cumulative response rate and Prob. of default, we found that validation test performed better.

We took threshold in our case as 8%, to increasr the sesnsitivity (ability to increase TP), we can lower the threshold, however that will also increase the FP and FN and hence chances of getting loss and opportunity loss will be more, so we have to carefully balance out the threshold to get the right sensitivity and minimum loss.

Train Set

## Decile cnt cnt\_tar1 cnt\_tar0 Rrate cum\_resp cum\_nonresp  
## 1: (0.0831,1] 331 121 210 36.56 121 210  
## 2: (0.0663,0.0831] 330 24 306 7.27 145 516  
## 3: (0.0573,0.0663] 330 9 321 2.73 154 837  
## 4: (0.0511,0.0573] 330 4 326 1.21 158 1163  
## 5: (0.0456,0.0511] 330 8 322 2.42 166 1485  
## 6: (0.0402,0.0456] 330 4 326 1.21 170 1811  
## 7: (0.0348,0.0402] 330 4 326 1.21 174 2137  
## 8: (0.0284,0.0348] 330 10 320 3.03 184 2457  
## 9: (0.0194,0.0284] 330 6 324 1.82 190 2781  
## 10: [7.37e-13,0.0194] 331 8 323 2.42 198 3104  
## cum\_relresp cum\_relnonresp KS  
## 1: 61.11 6.77 54.34  
## 2: 73.23 16.62 56.61  
## 3: 77.78 26.97 50.81  
## 4: 79.80 37.47 42.33  
## 5: 83.84 47.84 36.00  
## 6: 85.86 58.34 27.52  
## 7: 87.88 68.85 19.03  
## 8: 92.93 79.16 13.77  
## 9: 95.96 89.59 6.37  
## 10: 100.00 100.00 0.00

Test Set

## Decile cnt cnt\_tar1 cnt\_tar0 Rrate cum\_resp cum\_nonresp  
## 1: (0.0856,1] 67 30 37 44.78 30 37  
## 2: (0.0671,0.0856] 67 6 61 8.96 36 98  
## 3: (0.0572,0.0671] 67 1 66 1.49 37 164  
## 4: (0.0498,0.0572] 67 2 65 2.99 39 229  
## 5: (0.0441,0.0498] 67 1 66 1.49 40 295  
## 6: (0.0387,0.0441] 67 1 66 1.49 41 361  
## 7: (0.0338,0.0387] 67 0 67 0.00 41 428  
## 8: (0.0279,0.0338] 67 1 66 1.49 42 494  
## 9: (0.0176,0.0279] 67 1 66 1.49 43 560  
## 10: [2.38e-10,0.0176] 67 2 65 2.99 45 625  
## cum\_relresp cum\_relnonresp KS  
## 1: 66.67 5.92 60.75  
## 2: 80.00 15.68 64.32  
## 3: 82.22 26.24 55.98  
## 4: 86.67 36.64 50.03  
## 5: 88.89 47.20 41.69  
## 6: 91.11 57.76 33.35  
## 7: 91.11 68.48 22.63  
## 8: 93.33 79.04 14.29  
## 9: 95.56 89.60 5.96  
## 10: 100.00 100.00 0.00