Name of the Students	: Nikita Rathi Shivani Mahendra Vedija Pillay
Title of Report	: Credit Risk Modelling for SME loans.
Area of the project	: Banking and Financial Services

Abstract

Credit models are useful to evaluate the risk of consumer loans. The application of the technique with greater precision of a prediction model will provide financial returns to the institution. In this study a sample set of SME loan applicants from a large financial institution was focused on in order to develop Seven models each one based on one of the alternative techniques: Logistic Regression, Lasso Regression, Ridge Regression, Random Forest and Support Vector Machine.

The quality and performance of these models are evaluated and compared to identify the best one. Results obtained by the Logistic Regression, Lasso Regression, Ridge Regression, Random Forest and Support Vector Machine models are good and very similar.

This study illustrates the procedures to be adopted by a financial institution in order to identify the best credit model to evaluate the risk of SME loans and thereby get increasing profits.

Background

Indian Small and Medium Enterprises (SME) sector has emerged as a highly vibrant and dynamic sector of the Indian economy over the last five decades. SMEs not only play a crucial role in providing large employment opportunities at comparatively lower capital cost than large industries but also help in industrialization of rural areas. SMEs are complementary to large industries as ancillary units and contribute enormously to the socio-economic development of the country. India currently has more than 48 million SMEs. These SMEs contribute more than 45% of India's industrial output, 40% of the country's total exports and create 1.3 million jobs every year. Various estimates, including that of DBS Bank, puts India's GDP growth rate in FY15/16 at 7.4%. Of the many engines that are powering this growth, the SME sector is a powerful driving force

The SME sector in general and more specifically the small and micro enterprises rely heavily on the banks for finance and as such banks have to recognize the vast potential that exists in responsible lending to the SME segment. Sensitivity on the issue needs to be developed at various hierarchical levels.

Nonetheless, despite the increase in financing to the sector there is still a considerable credit gap which needs to be bridged. The extent of financial exclusion and the vicious cycle of credit issues as graphically shown in the Report of the Inter-Ministerial Committee for Accelerating Manufacturing in the SME Sector (Chairman: Shri Madhav Lal, Secretary Ministry of SME) shows that the cycle starts from lack of access to formal sources of finance which leads to tapping alternate sources of funds that are costly- higher cost of credit results into poor net cash inflow- which increases the risk profile of the small unit-and reduces their credit worthiness- which in turn further aggravates lack of access to formal sources of finance.

Thus access to timely and adequate credit from banks is critical for the sector. The ability of SMEs (especially those involving innovations and new technologies) to access alternate sources of capital like equity finance, angel funds/risk capital is extremely limited. At present, there is almost negligible flow of equity capital into this sector, which poses serious challenge to development of knowledge-based industries, particularly those that are promoted by first-generation entrepreneurs with the requisite expertise and knowledge. In the absence of alternate sources of finance, the SMEs' reliance on debt finance is very high. The high reliance on debt, combined with high cost of credit adversely impacts the financial viability of start-ups, particularly in the initial years, thereby threatening their long-term survival and sustainability.

Credit is a crucial input for promoting growth of SME sector, particularly the SME sector, in view of its limited access to alternative sources of finance. Various estimates on the credit availability to the SME sector, however, indicate a serious credit gap. Though the heterogeneous and unorganized nature of the sector poses inherent challenges for a credible estimate, the fact remains that there is considerable credit gap, which is a matter of serious concern and needs to be bridged if the sector has to foray into the next level of growth trajectory.

Information opacity arising from SME's lack of accounting records, inadequate financial statements

or business plans also makes it difficult for potential creditors to assess the creditworthiness of SME applicants. Besides, high administrative/ transaction cost of lending small amounts also queers the pitch for banks insofar as SME financing is concerned. Further, the existing system of banks' credit appraisal and related processes also need to be geared to appraise the financial requirements of SME sector.

Credit risk model offers an alternative method of evaluating loans for small businesses. The credit risk model approach, which is based on use of computer technology and mass production methods, was originally designed to handle consumer loans, but are now being used effectively for lending to small businesses by predicting their potential loan delinquency.

Credit risk model, is a model applied by banks in their assessment and approval or decline of the loan requests by SMEs. As there is a strong link between the payment behavior of the business owner and that of the business, SME credit scores usually include financial characteristics from both the business and the business owner. Credit risk model is based upon information like how the repayment of the previous loans has gone, what is the current income level of the enterprise, what are the outstanding debts, if any? It focuses on the credit history of the enterprise. As part of the process, the lenders see whether the enterprise/business owner has the reliability and honesty to repay the loan. It also examines how the enterprise has used credit before, its record for repayment of bills, including utility bills, how long the enterprise has been in existence, assets possessed by the enterprise and sustainability and viability of the activities that the unit is engaged in. Credit risk model draws inputs from historical information on the performance of loans with similar characteristics.

When used appropriately, credit risk model can benefit multiple stakeholders, including lenders, borrowers, and the overall economy. For the lender, risk model leads to process automation which facilitates process improvements leading to many by-products such as improved management information, control and consistency. It also increases the profitability of SME lending by reducing the time and cost required to approve loans and increasing revenues by expanding lending opportunities as lenders can safely approve marginal applicants that an individual underwriter might reject. International evidence has shown that credit risk model can assist in overcoming the inherent benefit/cost trade-off that banks face when deciding whether or not to invest in obtaining information regarding a potential borrower. A study that was meant to test the credit risk model situation in US estimated that the cost of evaluating micro loan applications in the US using credit risk model was reduced to around $100 compared to a range of $500-$1800 prior to the introduction of credit risk model. The time saving involved meant that banks could focus more time on marginal applications, existing loans that are showing signs of distress and processing more loan applications. Use of credit risk model has also meant that the marginal benefits of taking and maintaining collateral are not justified for small loans.

Given the extent of exclusion in the SME sector and the criticality of the sector for the economy, banks urgently need to step up lending to the sector. For evaluating loan proposals and for facilitating SME financing, banks would need to employ low cost and quick decision making alternatives. The use of credit risk model can go a long way in facilitating lending decisions by reducing costs and increasing service levels, which can deliver great benefits for both the lenders and SME borrowers.

Description of the project work

The aim of this project was to acquire hands-on experience with the tools for working with actual (real-life) data. Specifically learning to organize different available tools in a successful application are central. The project consists of four main steps, each of which are documented in the report.

1. Visualize the data, point out characterizing properties and state the solution. Be creative in the economic use of time plot, histogram or frequency plots.

2. Do some simple simulations: e.g. what is the best constant prediction. What is the best we can do using standard techniques implemented in the ident tool?

3. Based on experience gathered during the previous phase, what is a proper method for identification of the system? Perform the simulations trying to get the best result possible. Most importantly, verify the result: why is this result satisfactory? How does it compare to the estimates obtained in (2)?

4. Summarize the contribution in an ’abstract’ and “conclusions” of the report. Which contributions to standard approaches can be claimed, and how do the models support such clams?

Introduction

Ability to classify companies into different predefined groups is an important business research issue, which can be utilized a strong risk management tool. Default prediction has been an important area of business interest for many researchers, from the theoretical and practical aspect, as it is an integral part of the credit risk, which is considered to be one of the most important banking risks. Expert default prediction interests a wide range of stockholders such as banks, microcredit organizations, insurance companies, other creditors, auditors and more. The increase of default occurrence can be linked to the latest global financial crisis and appropriate credit risk management.

Many economists consider the latest global financial crisis to be the worst crisis since the Great Depression. Many European countries like Greece, Portugal, Italy and Ireland are facing severe financial and liquidity crisis, which are likely to lead to further issues s like mass demonstrations, European currency crisis, further jobs reduction. Financial crisis is defined in the relevant literature as a "disturbance to financial markets that disrupts the markets capacity to allocate capital – financial intermediation and hence investment come to a halt". It is believed that one of the main causes of the crisis lays in the collapse of large financial institutions, generally banks around the world.

International Convergence of capital measurement and capital standards segments three main parts of the minimal capital requirements for banks. The main three parts used for calculation of minimal capital requirements are credit, operational and market risks.

Assessing the probability of occurrence of credit default is the main interest of this research. According to the definition given by Basel Committee on Banking Supervision credit default occurs when one or more of the following takes place:

- "It is determined that the obligor is unlikely to pay its debt obligations (principal, interest, or fees) in full;
- A credit loss event associated with any obligation of the obligor, such as charge-off, specific provision, or distressed restructuring involving the forgiveness or postponement of principal, interest, or fees;
- The obligor is past due more than 90 days on any credit obligation or The
- obligor has filed for bankruptcy or similar protection from creditors."

In predicting credit default existing literature uses several classification techniques such as multiple discriminant analysis, linear probability, logit analysis, probit analysis, multinomial logit, decision trees, and artificial neural networks.

The main purpose of this study is to assess the probability of default occurrence on SME Loans in the banking sector. In other words, the main purpose of the study is to predict credit default, or to create a prediction model that distinguishes defaulted and non-defaulted companies, based on the financial data obtained from their financial statements.

Objective

The main objective of this Project Report is to build an effective Credit Risk model for SME Loans in order to minimize the loss due to default.

In addition, models may offer:

(a) the incentive to improve systems and data collection efforts;

(b) a more informed setting of limits and reserves;

(c) more accurate risk- and performance-based pricing, which may contribute to a more transparent decision-making process; and

(d) a more consistent basis for economic capital allocation.

Data

The SME Loans Dataset consists information about 8470 SME's and 28 Variables.

Data Dictionary

Srno	**Unique Id Assigned To The Company**
Coname	Name Of The Small And Medium Enterprise
Company	Undefined Variable
Rating	Rating As Per NSCI
City	City Location
Ownership	Ownership Type
Lob	Line Of Business
Industry	Industry It Belongs To
Ownoffice	The Ownership Status Of The Office Location
Years	Number Of Years In Business
Employees	Employee Strength Of The Company
Supplier	Supplier Base
Utilization	Capacity Utilization
Buyers	Buyer Base
Marketing_Method	Marketing Method Used By Company
Qualification	Qualification Level Of Owner
Experience	Business Experience (In Years) In Similar Line Of Business
Generations	Generations In Business
Sales	Annual Sales
Trendsales	Trend Of Sales
Netprofit	Net Profit
Trendprofit	Trend Of Profit
Icr	Interest Coverage Ratio
Invturnover	Inventory Turnover Ratio
Days	Collection In Days
Dte	Debt To Equity Ratio
Roa	Return On Assets Ratio
Netmargin	Net Margin
Default	Defaulter "Yes" Or Non-Defaulter "No"

Data Manipulation and Transformation

1. Deleted columns which would not benefit our model based on intuition such as Customer ID and Company name.
2. Gave all the columns score between 1 to 10 as provided by the client to bin the data into a standardized format.
3. Deleted columns which still had more than 30% NA values.
4. Deleted rows with NA values that were remaining.
5. Feature Selection was done using Boruta
6. Created a new data set of just scores.[Refer Appendix 2 for Scores]
7. Converted the complete data set to dummy variables to get the final dataset on which models would be applied.

Methodology

Tools Used

R and Tableau

Techniques Used for Model Building

Credit Risk model was built using Predictive Analytics Techniques. Since the Dependent Variable "Default" had a Binary output i.e. Yes, and No. where we were classifying Obligor as "Defaulter" and "Non Defaulter" the following Supervised Learning techniques were used for classification:

- Regression
  - Logistic Regression
  - Lasso
  - Ridge
- Decision Tree
- Random Forest
- Support Vector Machine

Techniques Used for Evaluating the Models

- Cross Validation
- Confusion Matrix
- ROC Curve

Techniques Used for Model Building

Logistic Regression

We applied a statistical model to historical data on SME characteristics. The particular form of statistical model is a discrete-time event history model. This model is designed to predict the risk of an event occurring, as a function of specified variables measured before the event occurs. The linear regression (a discrete time model) can be used to predict the risk of an event within a certain time period. This is equal vent and estimated by applying a logistic regression to issuer-year of data. The logistic regression takes the following form

$$\log\left(\frac{p}{1-p}\right) = \sum_{k=1}^{K} \beta_k x_k$$

where p is the probability of the event occurring, and K independent variables, x, are each weighted by a coefficient, β.

LASSO Regression

The above logistic regression model can be further extended into Lasso logistic regression model by imposing an L1constraint on β parameters and the problem is to minimize the negative log-likelihood function with the penalty

$$\sum_{i=1}^{n}\left\{\log(1+e^{\beta_0+x_i^T\beta}) - y_i(\beta_0+x_i^T\beta)\right\} + \lambda\sum_{j=1}^{p}|\beta_j|$$

Due to the above constraint, making λ sufficiently large will cause some coefficients in β be-come zero. Depending on the value of λ, a Lasso model can include any number of variables. In this way, both shrinkage and feature (variable) selection are done simultaneously and it is also this property that makes Lasso generally much easy to interpret and a very popular algorithm.

RIDGE Regression

Ridge regression is a variant to least squares regression that is sometimes used when several explanatory variables are highly correlated. The "usual" ordinary least squares (OLS) regression produces unbiased estimates for the regression coefficients (in fact, the Best Linear Unbiased Estimates). However, when the explanatory variables are correlated, the OLS parameter estimates have large variance. It might be desirable to use a different regression technique, such as ridge regression, in order to obtain parameter estimates that have smaller variance. The trade-off is that the estimates for the ridge regression method are biased.

More specifically, we use a logistic regression model with a quadratic penalty function, i.e. a ridge logistic regression.

The likelihood is expressed as the following logistic function:

$$l(\beta) = \sum_i [Y_i \log p(X_i) + (1 - Y_i) \log(1 - p(X_i))] \quad , \quad p(X_i) \stackrel{\text{def}}{=} \frac{\exp(X_i\beta)}{1 - \exp(X_i\beta)}$$

The objective function is ( ) 2 l β − λ β where λ is the ridge parameter. The objective function is minimized with a quasi-Newton method.

Decision Tree

The Decision Tree procedure assumes that:

The appropriate measurement level has been assigned to all analysis variables.

For categorical (nominal, ordinal) dependent variables, value labels have been defined for all categories that should be included in the analysis.

Using a tree model, you can analyze the characteristics of the two groups of customers and build models to predict the likelihood that loan applicants will default on their loans. The Decision Tree Procedure offers several different methods for creating tree models. But typically, in a decision tree, at each node a variable is selected for the split/partition and the best split of the variable. This process is repeated at each node unless split criteria are met at each of the node.

Random Forest

Random Forest Trees are based on a number of prediction trees that are less tolerant to noise compared to "Adaboost" and utilize random selection of features in splitting the trees. "Random Forests" is a voting procedure for the most popular class among a large number of trees. Thus, a random forest classifier is composed of a set of tree-structured classifiers [20- 22]. Equation [1] represents the classifiers where Θi represents a number of independent random vectors distributed identically, such that every tree has a vote for most popular class of input X.

Using Random Forests for prediction has many advantages such as their immunity to overfitting, an appropriate selection of randomness type leads to accurate classification or regression, the correlation and strength of predictors makes a good estimate of the ability for prediction, faster than boosting and bagging, better estimation of internal errors, not complicated, and can perform well in parallel processing. Based on the empirical results in Random Forests could compete with similar approaches in terms of accuracy.

Support Vector Machine

Support Vector Machines is an efficient and effective solution for pattern recognition problem whereas a following quadratic optimization problem has to be solved:

$$\text{minimize } W(\alpha) = -\sum_{i=1}^{l} \alpha_i + \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

where the number of training examples is denoted by l, training vectors $X_i \in R, i = 1,..,l$ and a vector $y \in R^l$ such as $y_i \in [-1;1]$. α is a vector of l values where each component i corresponds to a training example (xi, yi).

If training vectors xi are not linearly separable, they are mapped into a higher (maybe infinite) dimensional space by the kernel function

$$K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j).$$

Techniques Used for Evaluating the Models

Cross Validation

Steps:

a) At model development start, the whole data sample is split randomly (70/30, 75/25, 80/20...) The bigger sample is used for model development, while the smaller sample is used for cross-validation.
b) Model's predictive power (Gini index) is measured on the independent, validation sample
c) Done to avoid overfitting
d) The predictive power shouldn't be much lower on the validation sample than it is on the development – that's when the validation is considered successful

Confusion Matrix

The confusion matrix function evaluates classification accuracy by computing the confusion matrix.

By definition, entry i, j in a confusion matrix is the number of observations actually in group i, but predicted to be in group j.

		Classified as	
		Good Credit	Bad Credit
Actual	Good Credit	TP (A)	FN (C)
	Bad Credit	FP (B)	TN (D)

Figure 2. Confusion Matrix

This confusion matrix illustrates the classification's "confusion" or what is called classification error, in which the rows represent the actual classes and the columns represent the predicted classes. Total number of correctly classified instances will be represented diagonally in True Positive (TP) and

True Negative (TN) cells. TP represents “Actual Good” classified as “Good” while TN represents “Actual Bad” classified as “Bad”. The higher TP and TN the better is the performance of the classification algorithm. Incorrectly classified instances go to False Positive (FP) and False Negative (FN) that are “Actual Good” classified as “Bad” and “Actual Bad” classified as “Good” respectively. Total sum of the correctly and incorrectly classified classes should match the total number of the input instances.

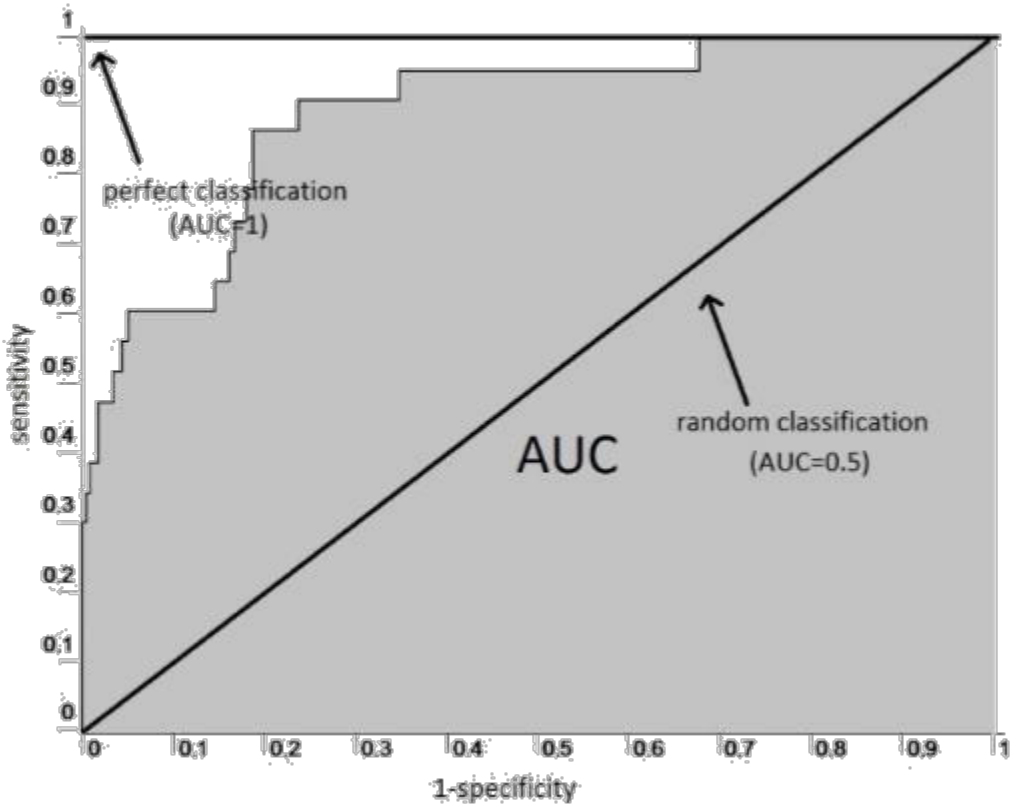
Several mathematical measures based on the confusion matrix make it easier to assess deeply the performance of the classification algorithm, also make it easier to compare the performance of different algorithms. Due to the variety of measures most of the researchers use and for better comparison, this study will report a number of measures that are:

Total Accuracy (Correctly Classified Instances) = TP + TN / (TP + TN + FP + FN) Sensitivity (Recall, Hit Rate, TP Rate, or Type II Error) = TP / (TP + FN) Precision (Confidence or Type I Error) = TP / (TP + FP) F-Measure = (2 * Precision * Sensitivity) / (Precision * Sensitivity)

Area Under Receiver Operating Characteristics Curve (AUC), for some of the algorithms. It is calculated automatically in R.

Receiver Operating Characteristic (ROC)

In statistics, **ROC curve**, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm and can be calculated as (1 −specificity). The ROC curve is thus the sensitivity as a function of fall-out. In general, if the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting the cumulative distribution function (area under the probability distribution from – (infinity) to the discrimination threshold) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability in x-axis.


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
sensitivity
perfect classification
(AUC=1)
AUC
random classification
(AUC=0.5)
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
1-specificity

Analysis

The Exploratory Data Analysis was done using Tableau.

Tableau Dashboard

default by ofcownership

Ownoffice		
Leased	443	23
NA	730	22
Occupied	620	24
Owned	4,469	253
Rented	1,796	90

default by employees

Employees		
21-50	2,940	169
51-70	1,409	68
71-100	563	41
100-150	1	
101-150	2,132	96
151 and above	589	25
Adequate	4	1
Inadequate	1	
NA	420	11

default by indusrty

Industry		
Auto Ancillary	332	25
Cement	24	1
Ceramic & Stone	128	12
Chemical	384	25
Dealers	70	1
Electrical & Engineering Go..	891	37
Food & Agro	309	21
Garments	183	8
Gems & Jewellery	47	2
Infrastructure	134	3
It & Ites	76	4
Leather & Leather Products	161	2
Manufacturing - Sundry	377	17
Manufacturing Others	125	3
Mechanical	606	21
Metal & Metal Products	1,736	82
Others	130	8

default by generation

Generations		
Null	268	2
First Generations	3,819	198
Second Generations	1,999	104
Third Generations	1,972	108

default by experience

Experience		
0 - 1	22	1
0-1	10	1
15-20	3,284	163
21-25	1,435	73
21-50	1	
26-Above	2,753	154
NA	554	19

default by rating

Rating		
SE 3A	2	
SE 3B	3	
SE 4C	1	
NA	22	1
SE 1A	54	
SE 1B	19	
SE 2A	242	
SE 2B	829	8
SE 2C	17	
SE 3A	267	4
SE 3B	3,348	46
SE 3C	1,673	54
SE 4A	1	
SE 4B	651	48
SE 4C	929	137
SE 5B	2	
SE 5C	112	

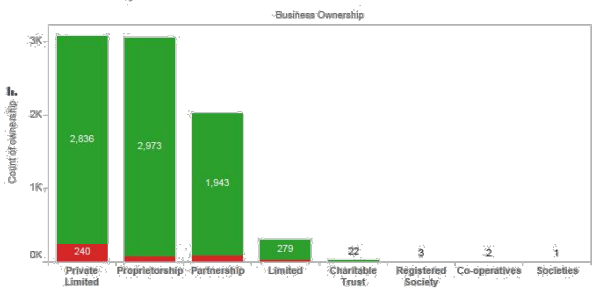
Default
N
Y

default by number of yrs in business

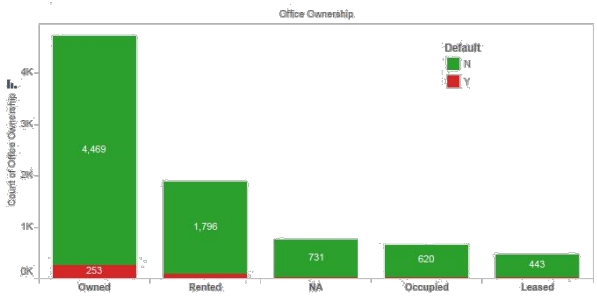
Years		
0 - 1	173	4
2-3	29	1
2-Mar	3,617	193
13 - 20	1,999	104
21 - Above	1,972	108
NA	268	2

default by city

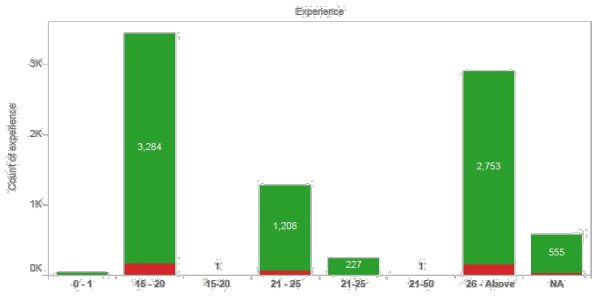
Business Ownership



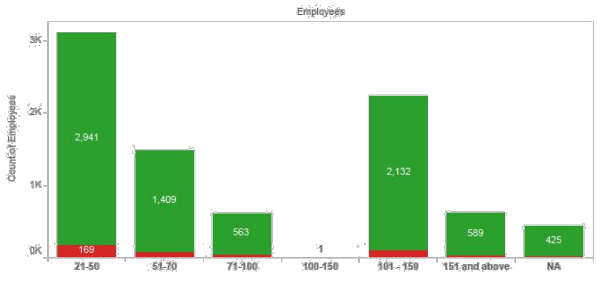
Office Ownership



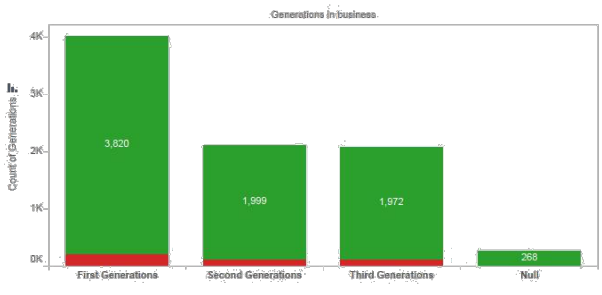
Experience



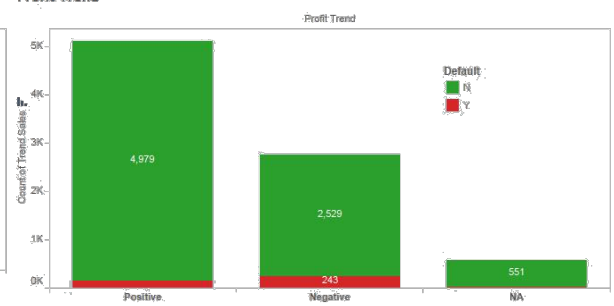
Employees



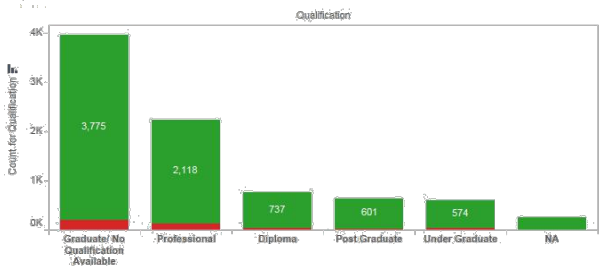
Generations



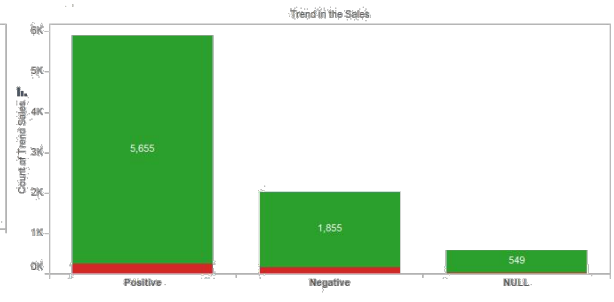
Profit Trend



Qualification



Sales Trend



Feature engineering:
We split NSIC rating in form of two features as Performance Capability and Financial Strength and give rating to each feature according to the given scores. Each feature was converted in to factor with certain scores respectively.On executing Boruta function we found the following features as most important:

[1] ”yearsRating" "empRating"
[3] "qualRating" "genRating"
[5] "saletrn_Rating" "profittrn_Rating"
[7] "icrRating" "roaRating"
[9] "nmRating" "dteRating"
[11] "SalesRating" "profitRating"
[13] "Perf_cap" "fs_rating”

We build a model using above features.

Model Building

Step 1: Split Dataset into Training (0.75) and Testing (0.25).

Step 2: Build Model on Training dataset and evaluate on Testing dataset.

REGRESSION:

Model1: Logistic regression using all independent variables in training dataset. model1 = glm (training$default ~icrRating+empRating,data = training, family = "binomial")
summary(model1)

AIC	643.45

Predicting using Test Dataset

pred0 = predict.glm(model1,newdata = test,type = "response")
pred_cat0=rep("N",1866)
pred_cat0[pred0 > 0.75]="Y" t0 <-
table(pred_cat0,test2) t0

	N	Y
N	1736	12
Y	41	26

Model2: Lasso Regression Model
cv.model.lasso <- cv.glmnet(sub_data3,y=k, family = "binomial",alpha=1)

Note: Only below mentioned features were considered by lasso regression as important rest features coefficient was found zero.

Lambda min:0.000436021
(Intercept) 2.915538658
profittrn_Rating -0.008965085
icrRating -0.212372578
roaRating -0.028364896
nmRating -0.226969676
dteRating 0.019720430
profitRating -1.746732851
Perf_cap -0.479890396
fs_rating -0.018389961

Predicting using Test Dataset

```
predicted = predict(model, s='lambda.min', newx=test1, type="class")
t2 <- table(actual,predicted)
t2
```

	N	Y
N	1736	17
Y	36	31

Model3: Ridge Regression Model

```
model.ridge = cv.glmnet(sub_data3,y=k,family = "binomial",
type.measure = "class",alpha =0) coef(model.ridge)

plot(cv.model.ridge)
```

Predicting using Test Dataset

```
predicted = predict(cv.model.ridge, newx = test1, s = "lambda.min",type = "class")
t2 <- table(actual,predicted)
t2
```

	N	Y
N	1743	10
Y	45	22

Model 4a: Decision Tree

```
tree.2 <- rpart(form,training,parms = list(loss
=matrix(c(0,10,1,0),ncol=2)),control=rpart.control(cp=0.001))
fancyRpartPlot(tree.2)
```

Predicting using Test Dataset

```
pred.tree = predict(tree.2,test)
pred.tree
predicted =rep("N",1815)
predicted[pred.tree[,2] > 0.80]="Y"
t3=table(predicted,actual)
t3
```

	N	Y
N	1741	26
Y	7	41

Model 4b: Decision Tree (all features without scores)

```
training1 = training_tree[,c(1:21)]
form1 <- as.formula(training1$Default ~ .)
tree.4 <- rpart(form1,training1)
prp(tree.4)
```

Predicting using Test Database

```
pred.tree.2 = predict(tree.4,test_tree)
predicted =rep("N",1712)
predicted[pred.tree.2[,2] > 0.917]="Y"
t4 <- table(actual_tree,predicted)
t4
```

	N	Y
N	1762	
Y	90	

Note: It was overfitted with True Positive and False Negative.

Model5: Random Forest

```
fit <- randomForest(Default ~ .,training,ntree=400)
```

Predicting using Test Dataset

```
predicted= predict(fit,test)
t6 <- table(predicted,actual)
t6
```

	N	Y
N	1740	23
Y	13	44

Model6: Support Vector Machine

```
svm.model = svm(Default~., data = training, cost =100,gamma =.01)
```

Predicting using Test Dataset

```
pred.svm=predict(svm.model,test)
t5 = table(actual,pred.svm)
t5
```

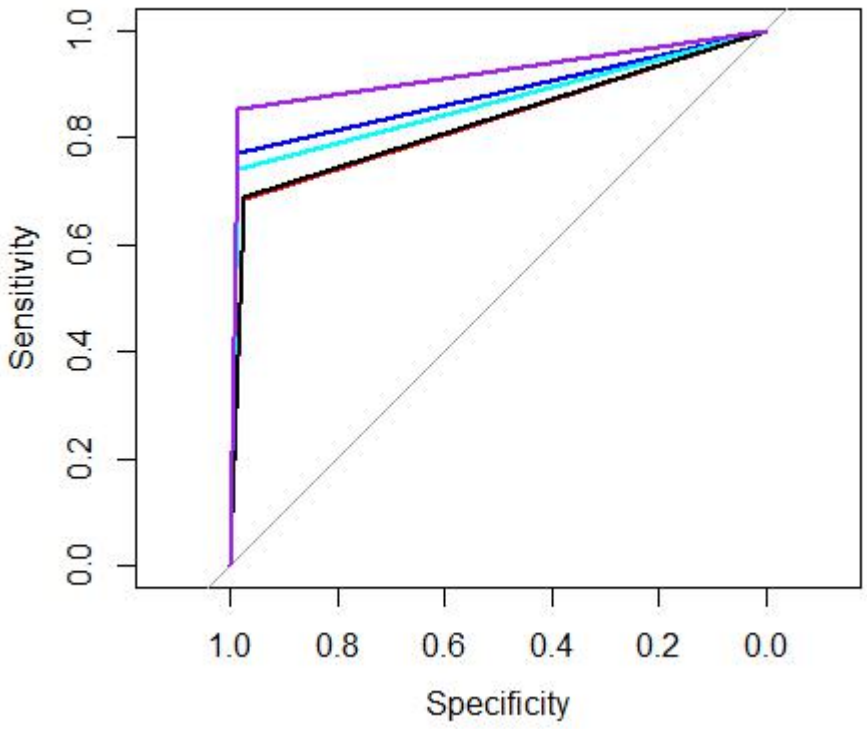
	N	Y
N	1738	15
Y	24	43

Step3: Model Evaluation

Evaluation Technique1: Confusion Matrix

Total Accuracy (Correctly Classified Instances) = TP + TN / (TP + TN + FP + FN)
Misclassification Error Rate (Incorrectly Classified Instances) = [(FP+FN)/TP+TN+FP+FN]
Sensitivity (Recall, Hit Rate, TP Rate, or Type II Error) = TP / (TP + FN)
Specificity = (TN/TN+FP)

Evaluation Technique2: ROC Curve



Combined ROC Curve
Logistic Regression = "red"

LASSO = "purple"

RIDGE = "black"

Random Forest = "blue"

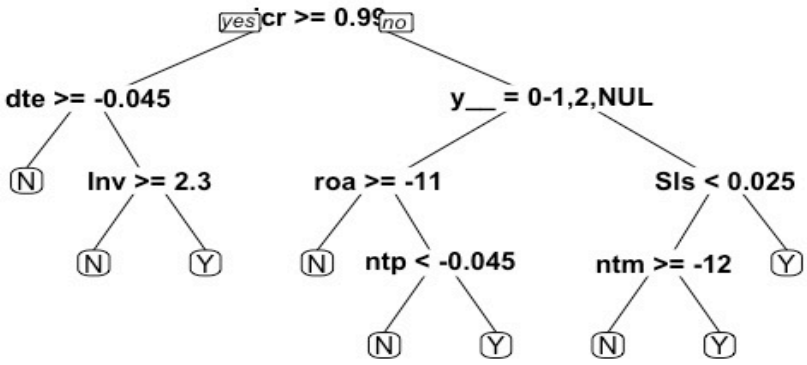
SVM = "cyan"
Decision Tree = "green"

Evaluation Matrix						
Parameters	**Method**					
	Logistic Regression (Model1)	**Lasso Regression Model**	**Ridge Regression Model**	**Decision Tree Model**	**Random Forest**	**SVM Model**
AIC	643.45					
AUC	0.8306	0.9197	0.9197	0.8541667	0.8794	0.8639
Error Rate	0.0292011	0.0291208	0.0302197	0.01818182	0.01978022	0.02142857
Accuracy	0.9707989	0.9708791	0.9697802	0.9818182	0.9802198	0.9785714
Sensitivity	0.6842105	0.4626866	0.3283582	0.9852858	0.7719298	0.9914432
Specificity	0.9769274	0.9903023	0.9942955	0.8541667	0.9869541	0.641791

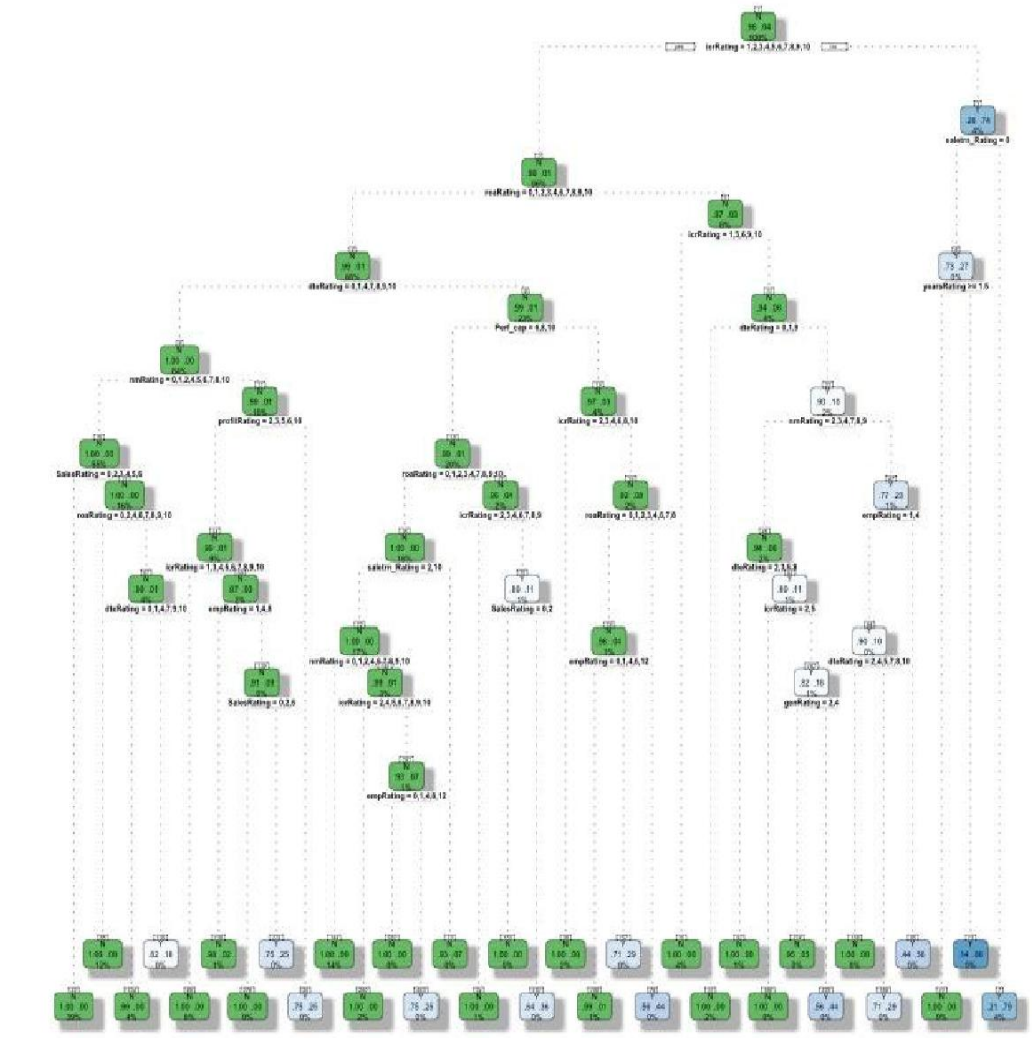
Conclusion

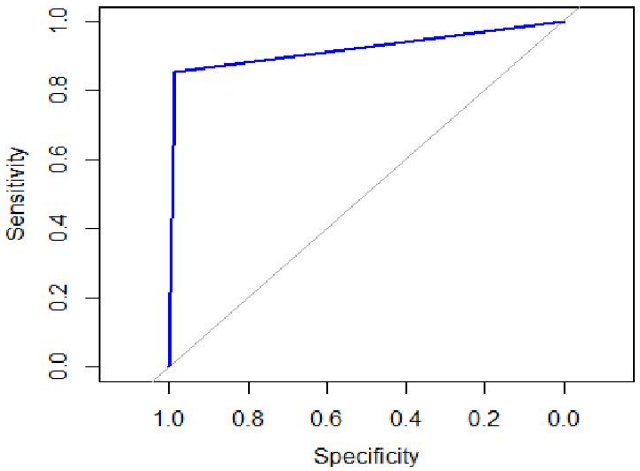
As per our analysis of the data we can deduce that LASSO & Decision Tree are the best model to predict if a customer would default or not. It has least number of misclassifications and maximum AUC.

Tree Plot without rating

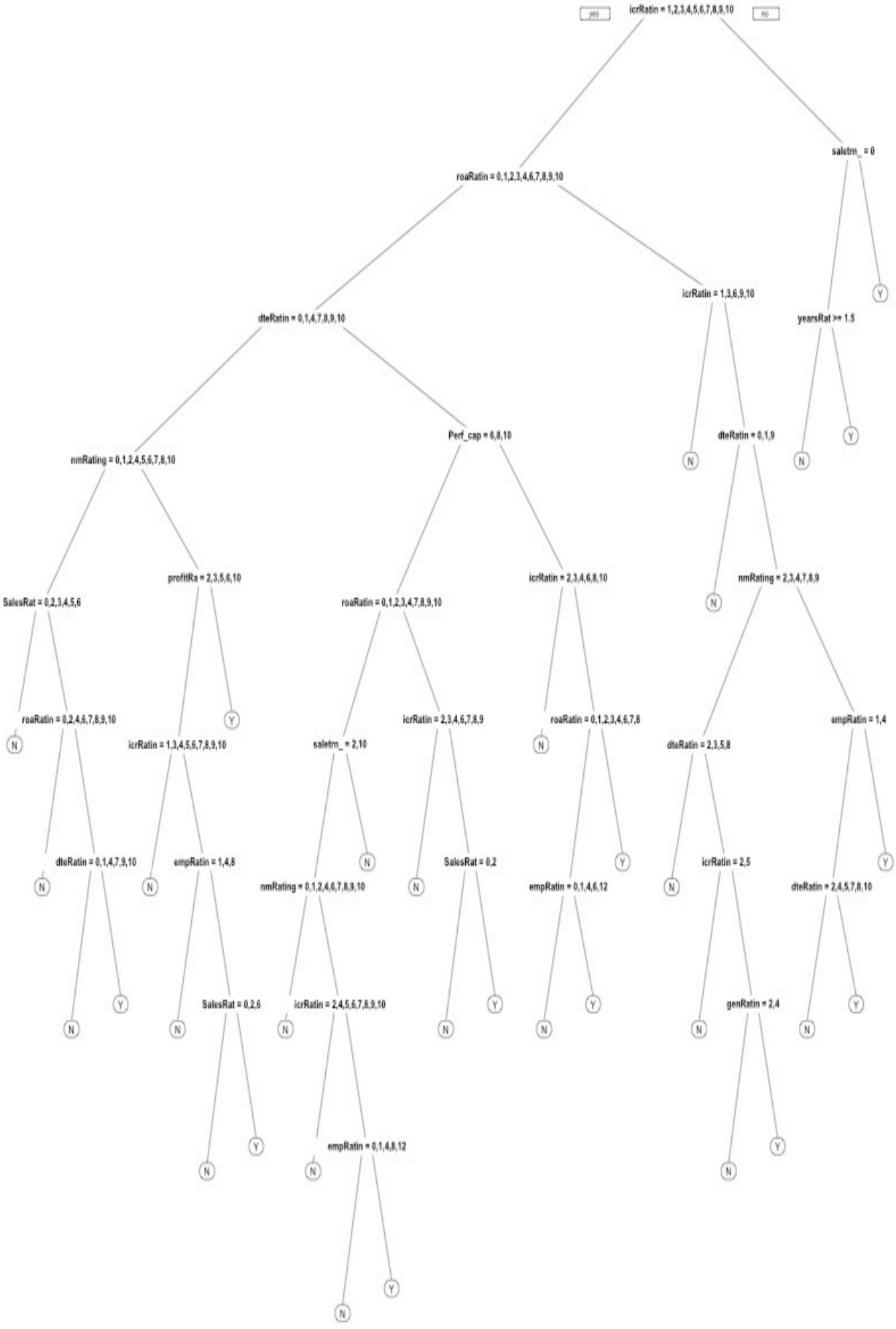


Decision Tree Plot with ratings





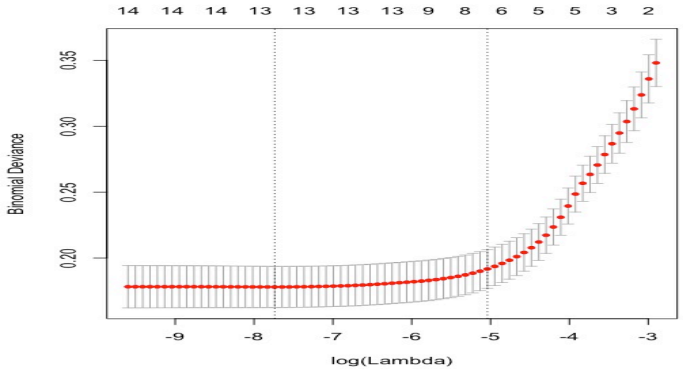
Tree plot 2:


yes
icrRatin = 1,2,3,4,5,6,7,8,9,10
no
saletrn_ = 0
roaRatin = 0,1,2,3,4,6,7,8,9,10
Y
icrRatin = 1,3,6,9,10
yearsRat >= 1.5
dteRatin = 0,1,4,7,8,9,10
Perf_cap = 6,8,10
N
dteRatin = 0,1,9
Y
nmRating = 0,1,2,4,5,6,7,8,10
N
Y
profitRa = 2,3,5,6,10
SalesRat = 0,2,3,4,5,6
roaRatin = 0,1,2,3,4,7,8,9,10
icrRatin = 2,3,4,6,8,10
nmRating = 2,3,4,7,8,9
N
roaRatin = 0,2,4,6,7,8,9,10
Y
icrRatin = 1,3,4,5,6,7,8,9,10
saletrn_ = 2,10
icrRatin = 2,3,4,6,7,8,9
roaRatin = 0,1,2,3,4,6,7,8
dteRatin = 2,3,5,8
empRatin = 1,4
N
dteRatin = 0,1,4,7,9,10
empRatin = 1,4,8
N
nmRating = 0,1,2,4,6,7,8,9,10
SalesRat = 0,2
empRatin = 0,1,4,6,12
icrRatin = 2,5
dteRatin = 2,4,5,7,8,10
Y
SalesRat = 0,2,6
icrRatin = 2,4,5,6,7,8,9,10
genRatin = 2,4
empRatin = 0,1,4,8,12

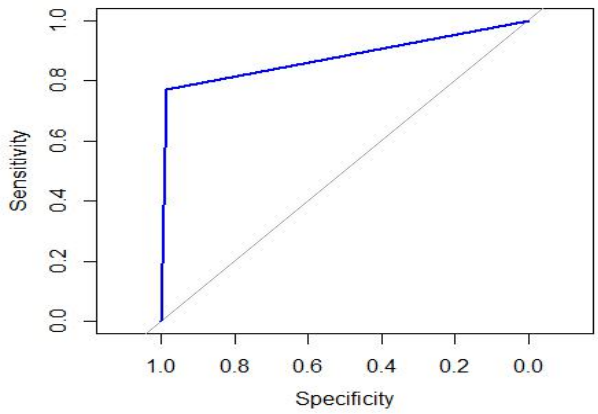
Note: From without and with ratings we come up with a conclusion that icr rating plays a key role in deciding the default rate for respective customer. Also we came up with set of rules for identifying the default rate. Rules can be interpreted from the tree plot graph.
For e.g: If the icr rating is high >0.98 then the next deciding factor will be Return on asset ratio , if that too is high then we will check debt to equity ratio , net margin and sales trend .
In case the icr rating is low tree leaves on the right will be deciding factors. We will check the sales trend and business experience. If Business experience found >1.5 then with 98% accuracy we can take a chance of approving loan. In order to minimize the risk we can suggest loan of minimum amount with certain duration to that SME client.

Note: Right branches means no and left branches means yes

LASSO Lambda Curve



LASSO ROC Curve



References

1. http://www.mba-berlin.de/fileadmin/user_upload/MAIN-dateien/1_IMB/Working_Papers/vor_2007/WP_23.pdf
2. http://ir.knust.edu.gh/bitstream/123456789/4255/1/Appiah%20Naana.pdf
3. http://www.intellegrow.com/images/download/publication/Publication%20-%20IFC%20MSME%20Report.pdf
4. https://www.ifc.org/wps/wcm/connect/b4f9be0049585ff9a192b519583b6d16/SMEE.pdf?MOD=AJPERES
5. https://www.researchgate.net/publication/251342271_Screening_Creditworthiness_of_SME's_The_Case_of_Small_Business_Assistance_in_Turkey

Appendix1

final_script.R