BFSI Capstone Project

FINAL - SUBMISSION

Submitted By:-
Aviral Raj
Bharat Vashistha
Subhajit Das
Anargha Biswas

Business Understanding

- CredX a leading credit card provider gets thousands of credit card applicants every year but the company is experience a credit loss in the recent years.

- The company wants to acquire the right set of customers to mitigate the credit risk and to decrease the credit loss to the company.

- So, we have to identify the right set of customers for the company using Predictive Models thereby determining the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of the project.

Planned Approach

Solution approach

It's a binary supervised classification problem.

So we aim to build a predictive model to identify the customers who are at a risk of defaulting if offered a credit card using Logistic Regression & Random Forest.

Next we plan to evaluate the models using x-fold validation technique to find the best approach.

As the initial step we have followed Cross Industry Standard Process for Data Mining (CRISP–DM) framework. It involves a series of steps:

1. Business understanding
2. Data understanding
3. Data Preparation
4. Data Modeling
5. Model Evaluation
6. Model Deployment

Data Understanding

To begin with we have been provided 2 datasets for analysis:-

- **Demographic Data**:- The dataset provides details regarding the demographic information for the customers who have applied for the credit card. e.g. Age, Gender, profession, Education, Type of Residence etc.

- **Credit Bureau Data:-** The dataset provides details regarding the credit card utilization for the customers. i.e. Numbers of times customers have defaulted for credit card payment in last 6 months/1 year etc. This information can be obtained from **CIBIL.**

Data Dictionary

Demographic Data	
Variables	Description
Application ID	Unique ID of the customers
Age	Age of customer
Gender	Gender of customer
Marital Status	Marital status of customer (at the time of application)
No of dependents	No. of childrens of customers
Income	Income of customers
Education	Education of customers
Profession	Profession of customers
Type of residence	Type of residence of customers
No of months in current residence	No of months in current residence of customers
No of months in current company	No of months in current company of customers
Performance Tag	Status of customer performance (" 1 represents "Default")

Credit Bureau Data	
Variable	Description
Application ID	Customer application ID
No of times 90 DPD or worse in last 6 months	Number of times customer has not payed dues since 90days in last 6 months
No of times 60 DPD or worse in last 6 months	Number of times customer has not payed dues since 60 days last 6 months
No of times 30 DPD or worse in last 6 months	Number of times customer has not payed dues since 30 days days last 6 months
No of times 90 DPD or worse in last 12 months	Number of times customer has not payed dues since 90 days days last 12 months
No of times 60 DPD or worse in last 12 months	Number of times customer has not payed dues since 60 days days last 12 months
No of times 30 DPD or worse in last 12 months	Number of times customer has not payed dues since 30 days days last 12 months
Avgas CC Utilization in last 12 months	Average utilization of credit card by customer
No of trades opened in last 6 months	Number of times the customer has done the trades in last 6 months
No of trades opened in last 12 months	Number of times the customer has done the trades in last 12 months
No of PL trades opened in last 6 months	No of PL trades in last 6 month of customer
No of PL trades opened in last 12 months	No of PL trades in last 12 month of customer
auto loans)	Number of times the customers has inquired in last 6 months
auto loans)	Number of times the customers has inquired in last 12 months
Presence of open home loan	Is the customer has home loan (1 represents "Yes")
Outstanding Balance	Outstanding balance of customer
Total No of Trades	Number of times the customer has done total trades
Presence of open auto loan	Is the customer has auto loan (1 represents "Yes")
Performance Tag	Status of customer performance (" 1 represents "Default")

Data Quality Issues in Demographic Dataset

Column Name	Missing Data	Erroneous Data
Age	-	20 wrong data ranging from -3 to 0
Gender	2 rows doesn't have any value	-
Marital Status	6 rows doesn't have any value	-
Number of Dependents	3 rows doesn't have any value	-
Income	-	81 rows have income less than 0.
Education	119 rows doesn't have any value	-
Profession	14 rows doesn't have any value	-
Type of Residence	8 rows doesn't have any value	-
No of months in current residence	-	-
No of months in current company	-	-
Performance Tag	1425 rows doesn't have any value	

Data Quality Issues in Credit Bureau Dataset

Column Name	Missing Data	Erroneous Data
No of times 90 DPD or worse in last 6 months	-	-
No of times 60 DPD or worse in last 6 months	-	-
No of times 30 DPD or worse in last 6 months	-	-
No of times 90 DPD or worse in last 12 months	-	-
No of times 60 DPD or worse in last 12 months	-	-
No of times 30 DPD or worse in last 12 months	-	-
Avgas CC Utilization in last 12 months	1058 rows doesn't have any value	
No of trades opened in last 6 months	1 row doesn't have any value	-
No of trades opened in last 12 months	-	-
No of PL trades opened in last 6 months	-	-
No of PL trades opened in last 12 months	-	-
No of Inquiries in last 6 months (excluding home & auto loans)	-	-
No of Inquiries in last 12 months (excluding		

Handling of Data Quality Issues

- We have the application ID as the unique identifier for the both the data sheets.
- Performance Tag is common in both the sheets , so we have removed one of them and restored the other.
- After merging the two data sheets, we find that there are 3 duplicate application ID which present different information. For the sake of consistency in data we have removed the first instance of duplicate rows .
- The missing data is handled in different ways based on the features. Also WOE is calculated for each of the attributes and the missing values are treated with WOE based on the requirement.

- We have removed the rows where Performance tag is not entered as those records are considered as cases where the credit card application of the customer has been rejected; hence those are not considered in analysis as well.
- For some variables, we have treated the missing data with mean and median of the corresponding columns, whereas for some variables we have replaced the missing data with the corresponding WOE values as suggested in the problem description.

Exploratory Data Analysis (EDA) - Plots

- An comprehensive Exploratory Data Analysis for all the variables which have been analyzed have been attached for further reference.
- The plots show Univariate and Bivariate analysis of different variables along with the response rate and WOE plots.

EDA Plots

Information Value of the Attributes

- We calculated the IV values of the attributes and from the IV values we can conclude that parameters in the demographic data don't play any significant role in prediction.

- The significant variables are arranged from the top, in descending order.

	Variable	IV
10	No.of.months.in.current.residence	7.912793e-02
6	Income	4.227555e-02
11	No.of.months.in.current.company	2.162803e-02
2	Age	3.392713e-03
5	No.of.dependents	2.658040e-03
8	Profession	2.278417e-03
1	Application.ID	1.505471e-03
9	Type.of.residence	9.169435e-04
7	Education	7.650668e-04
3	Gender	3.195489e-04
4	Marital.Status..at.the.time.of.application.	9.221277e-05

	Variable	IV
8	Avgas.CC.Utilization.in.last.12.months	0.310115692
10	No.of.trades.opened.in.last.12.months	0.297978053
12	No.of.PL.trades.opened.in.last.12.months	0.296106047
14	No.of.Inquiries.in.last.12.months..excluding.home...auto...	0.295453840
16	Outstanding.Balance	0.246066844
4	No.of.times.30.DPD.or.worse.in.last.6.months	0.241771755
17	Total.No.of.Trades	0.236657390
11	No.of.PL.trades.opened.in.last.6.months	0.219828055
5	No.of.times.90.DPD.or.worse.in.last.12.months	0.214261624
3	No.of.times.60.DPD.or.worse.in.last.6.months	0.206141944
13	No.of.Inquiries.in.last.6.months..excluding.home...auto.l...	0.205243077
7	No.of.times.30.DPD.or.worse.in.last.12.months	0.198361443
9	No.of.trades.opened.in.last.6.months	0.186141572
6	No.of.times.60.DPD.or.worse.in.last.12.months	0.185843147
2	No.of.times.90.DPD.or.worse.in.last.6.months	0.160458118
15	Presence.of.open.home.loan	0.017660857
18	Presence.of.open.auto.loan	0.001665156

Model Building and Evaluation - Demographic Data Frame

The first model has been built using only Demographic Data . Following were the final variables predicted by the model :-

- Modified Income
- Binning Current Residence woe
- Binning Current Company woe

Results Obtained from Initial Model:

- Accuracy:- 57.3%
- Sensitivity:- 56.44%
- Specificity:- 57.35%

Note:- The significant variables predicted by the model is in accordance with the IV values; i.e. the top 3 variables according to Information values are predicted by model as significant as well.

Model Building : Combined Data Frame

- Combined Data Frame has been created using combination of two data frames:-
  - ✓ **Demographic Data Final:-** It contains all the variables of Demographic data frame including the imputed as well as the WOE variables along with dummy variables of demographic data frame.
  - ✓ **Credit Bureau Data WOE:** It contains the WOE variables of Credit Bureau data frame.

Note:- The reason of keeping dummy variables in the Combined Data frame is to build a logistic regression model on it. Since, logistic regression models require dummy variables for analysis, therefore, we have put it into the data frame for analysis to let the model predict how significant they are.

- Important Variables Predicted by the Model are:-
  - ✓ No_of_times_30_DPD_or_worse_in_last_6_months_woe
  - ✓ Binned_CC_Utilization_woe
  - ✓ No_of_PL_trades_opened_in_last_12_months_woe
  - ✓ Binned_No_of_Inquiries_in_last_12_months_woe
  - ✓ `BinningCurentResidence(0,20]`
  - ✓ Modified_No_of_dependents2

It is evident from the analysis that the variables predicted by the model as significant do contain dummy variables as well which justifies the reason of keeping dummy variables in combined data frame(**Unbalanced**).

- Results Obtained:-
  - ✓ Accuracy:- 64%
  - ✓ Sensitivity:- 63%
  - ✓ Specificity:- 64%

Note:- The aforementioned figures obtained after the model building have been considered as benchmark figures.

Sampling of Dataset

- Initial analysis of the dataset suggest that the dataset is very unbalanced which is evident from the figures below:
- Count of Performance Tag: 0:- 66919
- Count of Performance Tag 1:- 2948
- Response Rate:- 4.29%
- To overcome this, we have tried to sample the data using **Oversampling/Under sampling/Both Way sampling/ Synthetic Sampled Data** as applicable**.**

- **With Sampling,** the approach is to select, manipulate and analyze a representative subset of data points in order to identify patterns and trends in the larger **data set** being examined.

Random Forest

Since the data is skewed and unbalanced, we need to decide if we should go for a tree based algorithm or a non-tree based algorithm.

Performance. Tag : No Yes

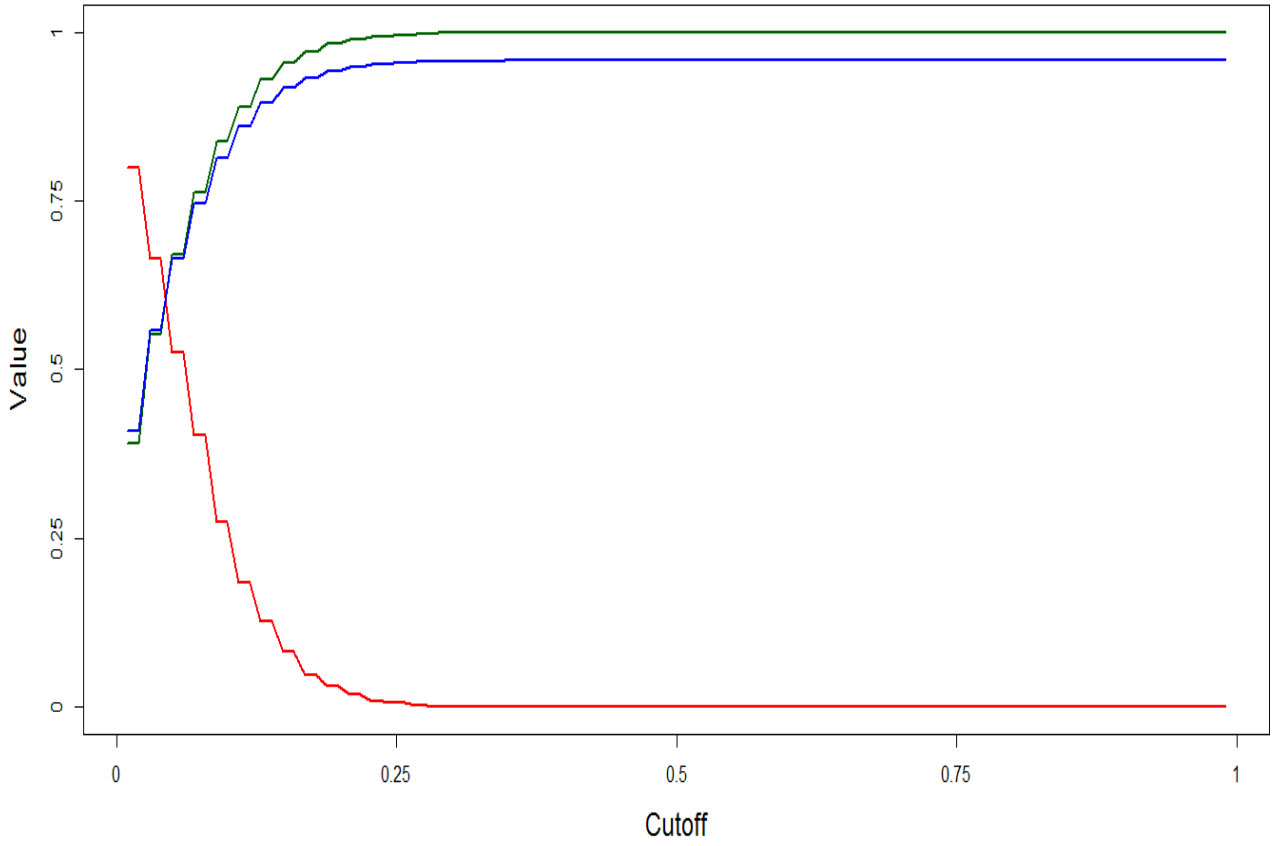
66919 2948

Lets have a look at the results of random forest modeling on unbalanced data set.

Sensitivity 0.524

Specificity 0.670

Accuracy 0.664



Random Forest contd..

- But before that we had to balance the data to see if the results were any thing better. We tried out the following techniques to balance the data.
  - ✓Over Sampling
  - ✓Under Sampling
  - ✓Both Sampling
  - ✓Synthetic data generation.

- We split the data into test and train data frames, using stratified Sampling, using the caret package.
- While Under Sampling reduces the original data to a huge extent, the accuracy is also very low when tested on test data frame.

- The synthetic data generation technique proves to be better than the rest of the sampling techniques and hence we go ahead with modeling of the synthetic data generation sampling technique. Using Random Forest, with cutoff probability of 0.2, we see the following accuracy scores. The cutoff value was chosen using Hyper Parameter Optimization Technique.

Random Forest Results

We present a comparison of results of random forest after different ways of sampling of data.

Over Sampling Technique.

Sensitivity 0.440

Specificity 0.616

Accuracy 0.609

Both Sampling Technique.

Sensitivity 0.504

Specificity 0.679

Accuracy 0.672

Under Sampling Technique with cutoff probability of 0.2

Sensitivity 0.933

Specificity 0.168

Accuracy 0.200

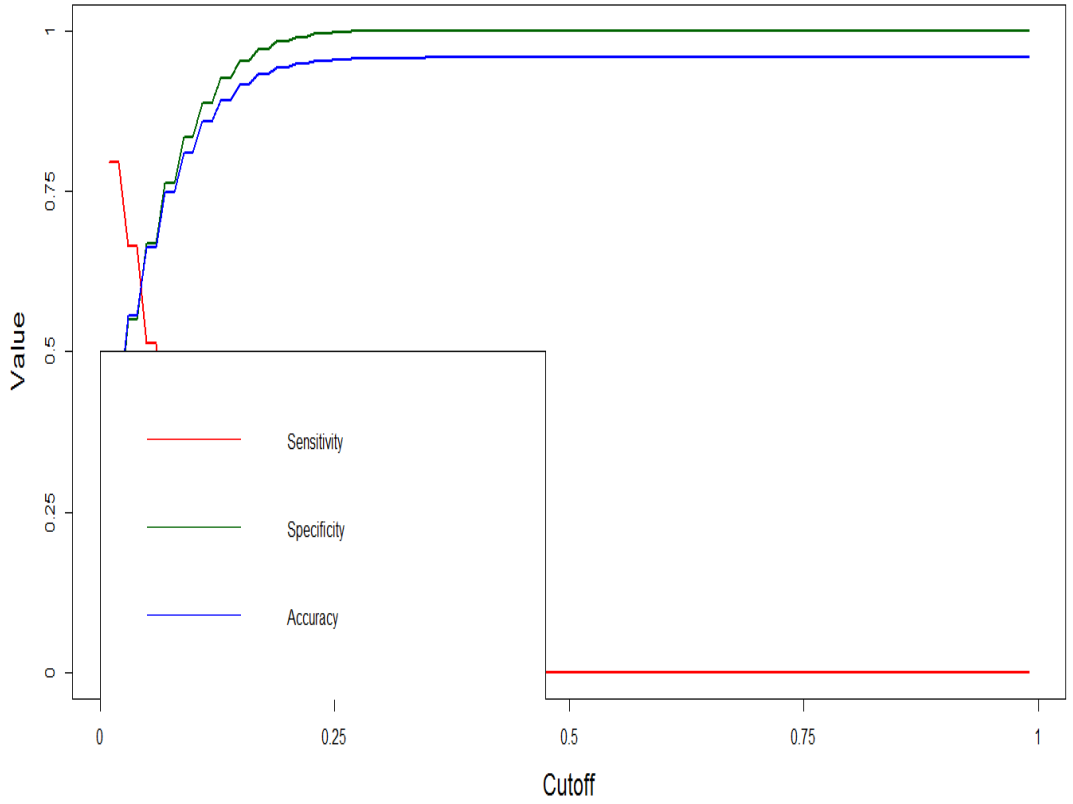
Synthetic Data Sampling Technique

Sensitivity 0.6266

Specificity 0.6319

Accuracy 0.6317

Cutoff Score for Random Forest

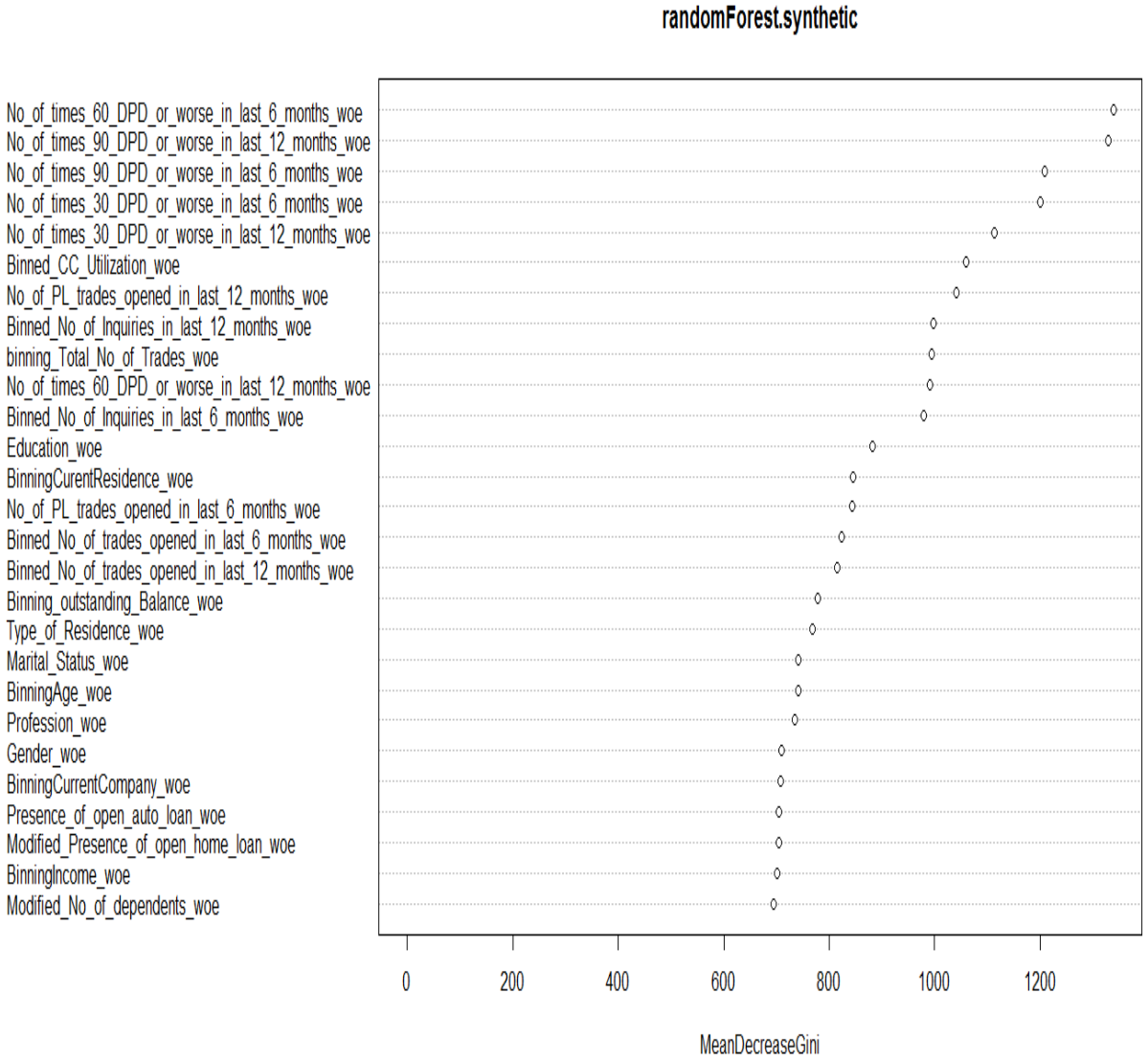


- For synthetic data sampling technique, The confusion matrix for the test data frame generated from random forest model shows the following results.

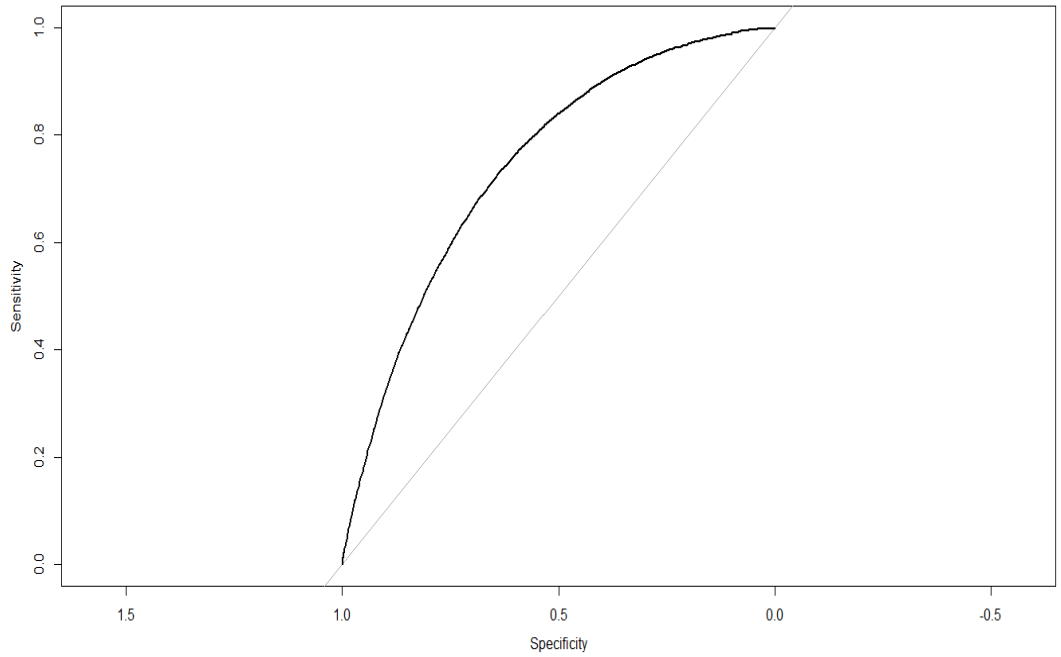
- Prediction no yes
- no 12686 330
- yes 7389 554

The f1 score of the model is 0.125. The precision and recall are also important parameters for evaluation of the model, and in this case we see the values to be 0.0167 and 0.615 consecutively.

Importance plot and the ROC Plot

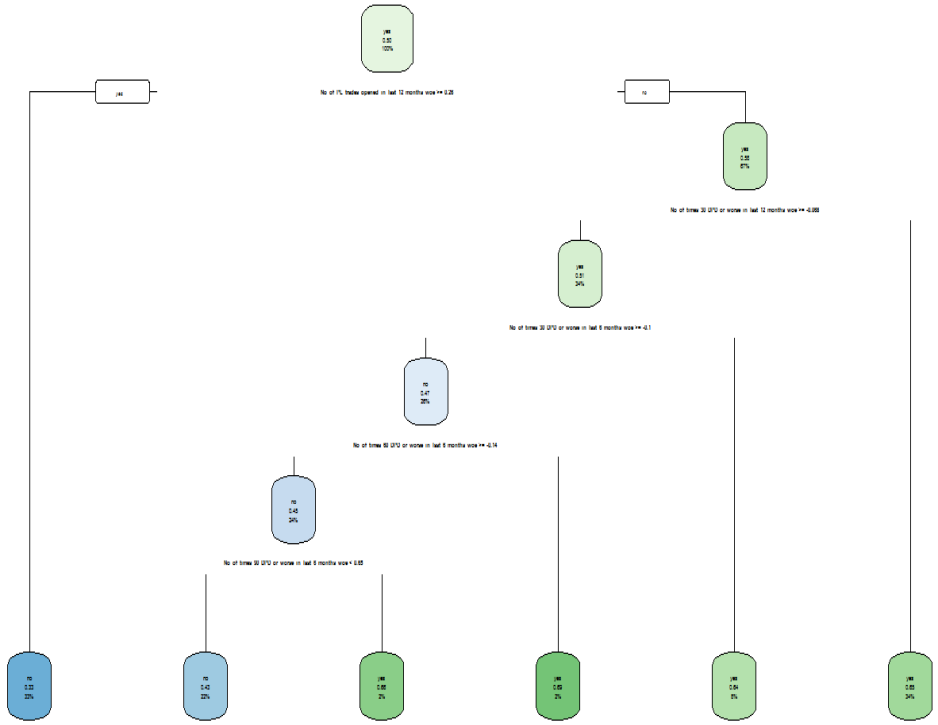


The ROC Plot. The calculated AUC value for this model is 0.7439

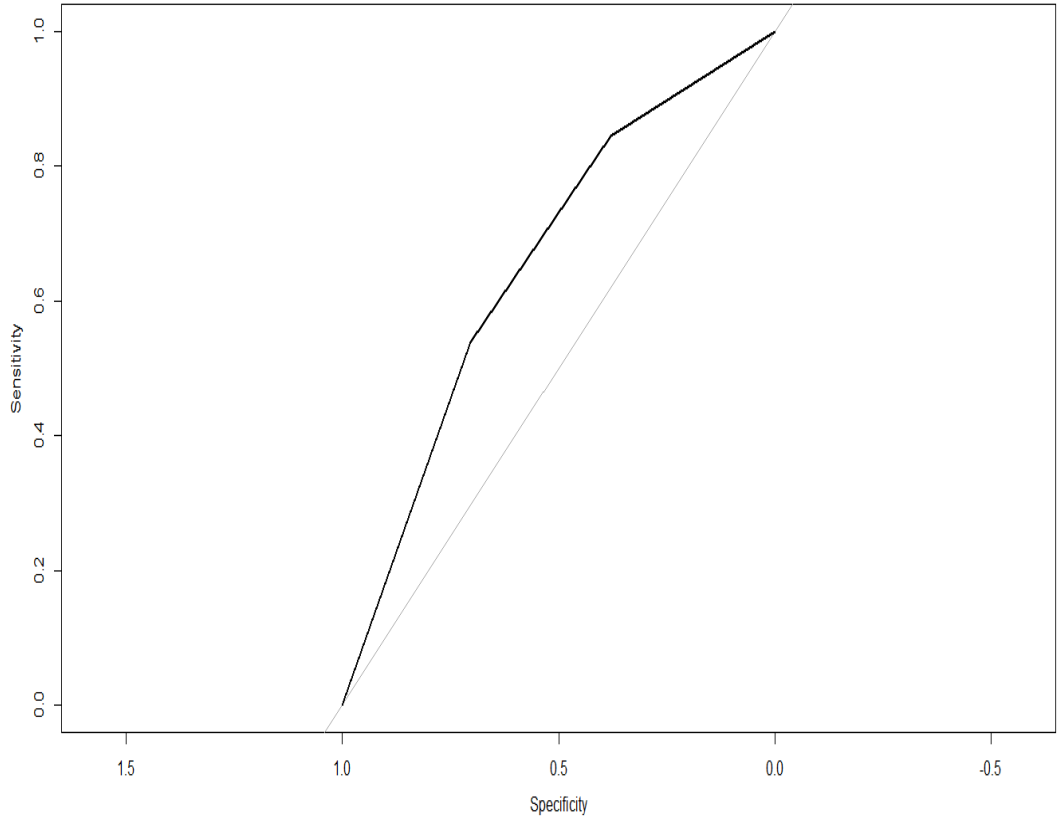


Another Tree Based Algorithm - Decision Trees

Lets have a look at the decision tree itself.



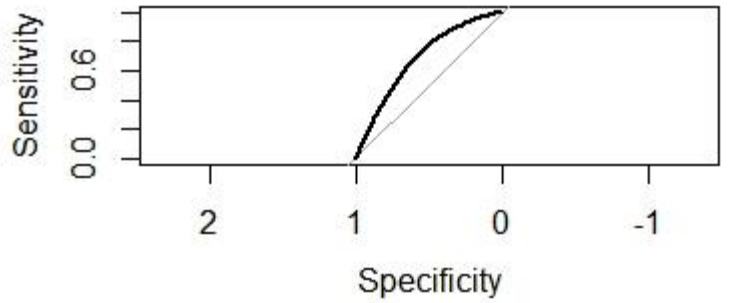
ROC Curve for Decision tree with AUC value as 0.654



Logistic Regression: Both Way Sampling Results

- **Logistic Regression Model on Sampled Data:** We have done both way sampling on the unbalanced dataset and created the Logistic Regression Model again on the sampled version.

- **Result Obtained:**
  - ✓ Accuracy: 65%
  - ✓ Sensitivity: 62.44%
  - ✓ Specificity:- 65.12



So, we see a marginal increase in accuracy and specificity with results obtained after balancing the dataset as compared to unbalanced data.

Choice of Final Model

Based on our analysis and the obtained results Logistic Regression as our Final choice of model for balanced dataset since, it provides a better figure for Accuracy, Sensitivity and specificity on a whole.

Evaluation Parameters	Logistic Regression(Both Way Sampling- Balanced Data)	Random Forests(Synthetic Sampling- Balanced Data)
Accuracy	65%	63.17%
Specificity	65.12%	63.19%
Sensitivity	62.44%	62.66%
F1 Score	0.13	0.125
Precision	0.073	0.017
Recall	0.62	0.615

Analysis on Rejected Applicants

- In our initial analysis, we found that among 71295 customers, a total of 1425 customers do not have a Performance Tag associated with them.
- A customer without a Performance tag illustrates the fact that the candidates applications were rejected, thereby credit card hasn't been provided to those customers and hence,
- However, a comprehensive analysis of the rejected applicants based on their application score will provide us following information:-
  - ✓ **Candidates were rejected by the bank:-**Though based on the credit score they should be provided a credit card which ultimately led to Financial loss of the company.
- Following results for Application score were obtained for rejected applicants:
  - ✓ Minimum : 299.8
  - ✓ 1st Quarter.: 327.3
  - ✓ Median : 334.1
  - ✓ Mean : 333.0
  - ✓ 3rd Quarter.: 339.6
  - ✓ Maximum:- 381
- Based on the values of the Application Score, we have assumed the cut off score = **320** for rejected applicants

Financial Loss due to Rejected Applicants

- Total number of rejected customers by bank : 1425
- Customer should be given Credit card based on credit score 320:

  No : 133

  Yes : 1292
- Revenue Loss for bank :
- Let us assume bank makes $500 per year from 1 credit card customer.
- Bank refused 1292 potential credit card customer.
- Total Loss to the bank = 1292*500 = $646,000 annual loss to the bank

Analysis on Approved Applicants

- To determine cost-benefit analysis for the bank we carried out financial assessment for all credit card customers of the bank using our final model.
- To achieve this, we have found the Potential defaulters among the approved applicants who are the customer who already have the credit card but their credit score is below **320** in this case, This value can be adjusted over a period of time by analyzing default trends.
- Following results for Application score were obtained for rejected applicants:
  - ✓ Minimum : 289.2
  - ✓ 1st Quarter.: 322.5
  - ✓ Median : 338.4
  - ✓ Mean : 338.8
  - ✓ 3rd Quarter.: 356.6
  - ✓ Maximum:- 386.6

Out of all the approved applicants, **Potential defaulters** are stated below:-

Yes : 14980

No : 54887

- **Actual Defaulters:-** The customers who have a Performance Tag = 1 and are potential defaulters, are considered as Actual Defaulters.

Yes : 1194

No : 68673

Financial Assessment on Approved Applicants

Net Defaulter that could be minimized using this model : 1194

Let us assume 1 defaulter results to $10000 to the bank

Net Loss due to default : **$11,194,000, i.e. $11.2 million**

Lost Customer due to model : 14980-1194 = **13,786**

Let us Assume 1 lost customer results in $500 revenue loss to the bank

Net Revenue loss due to lost customers = **$6,893,000 i.e. $6.9 million**

Monetary Benefit for bank using our Model : $11.2 million - $6.9 million = **$4.3 million**

Total Revenue Loss = $ 4.3 million (Approved customer) + $ 646,000 (Rejected customer) = $ 5 Million (approx)

Thank You