BFSI Capstone Project

Final Submission

PRESENTED BY:

ALAY SHAH
RUTUJA MOWADE
PALLAVI KUMARI A
MEGHA GAWDE

Problem Statement

- CredX is a leading credit card provider that gets thousands of credit card applicants every year. But in the past few years, it has experienced an increase in credit loss. The CEO believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’.

Our Goal

- In this project, we will help CredX identify the right customers using predictive models. Using past data of the bank’s applicants, we will determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of our project.

OVERALL APPROACH TO CASE STUDY:

It involved a series of steps, as follows :

1. Business understanding, Data understanding
2. Data Cleaning/Preparation
3. Exploratory Data Analysis
4. Data Modelling
5. Model Evaluation
6. Model Deployment

DATA UNDERSTANDING

- **Demographic/application data**: This is obtained from the information provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status, etc.
- **Credit bureau**: This is taken from the credit bureau and contains variables such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc.
- **performance tag** :- which represents whether the applicant has gone 90 days past due or worse in the past 12-months (i.e. defaulted)

Data Cleaning

- **Deal with ‘NA’ values :**

Total Count : **3180**

However, out of 3180 NA’s, Performace Tag column itself contributed to about half the number of total NA’s i.e. 1425 NA’s in Performance Tag variable. However, there is a method of interpreting NA’s in Performance Tag column, which is termed as “Reject Inference”.

How we treated them :

In this case, NA's in the Performance Tag can be interpreted as customers whose Performance is not recorded because they did not have any i.e. They were just not granted credit card at the first place or their credit applications were rejected. And that is why their data of Performance is "Not Available". Now, getting a credit card application rejected itself signals that the loan granting organization thought this customers would default in the near future, maybe because they might be having some clear red signals in their profile. So, in a way, these customers having NA in their performance Tag column can be assumed no better than the customers who defaulted i.e. having values "1" in the Performance Tag Column

Data Cleaning

After fixing the NA values of Performance Tag Columns, there were only '**1208**' other rows which had one or more NA's which is just (1208 / 71292 = 1.68 %) **1.68**% of total dataset . And **1.69**% is not a significant number. Hence, following that thumb rule, we decided to drop all the remaining rows having one or more NA's, instead of taking time out to interpret it through some other method.

- **Deal with Invalid values :**

Total count : 80

1. Age – 1, Value = -3 2. Income – 79, Value = -0.5

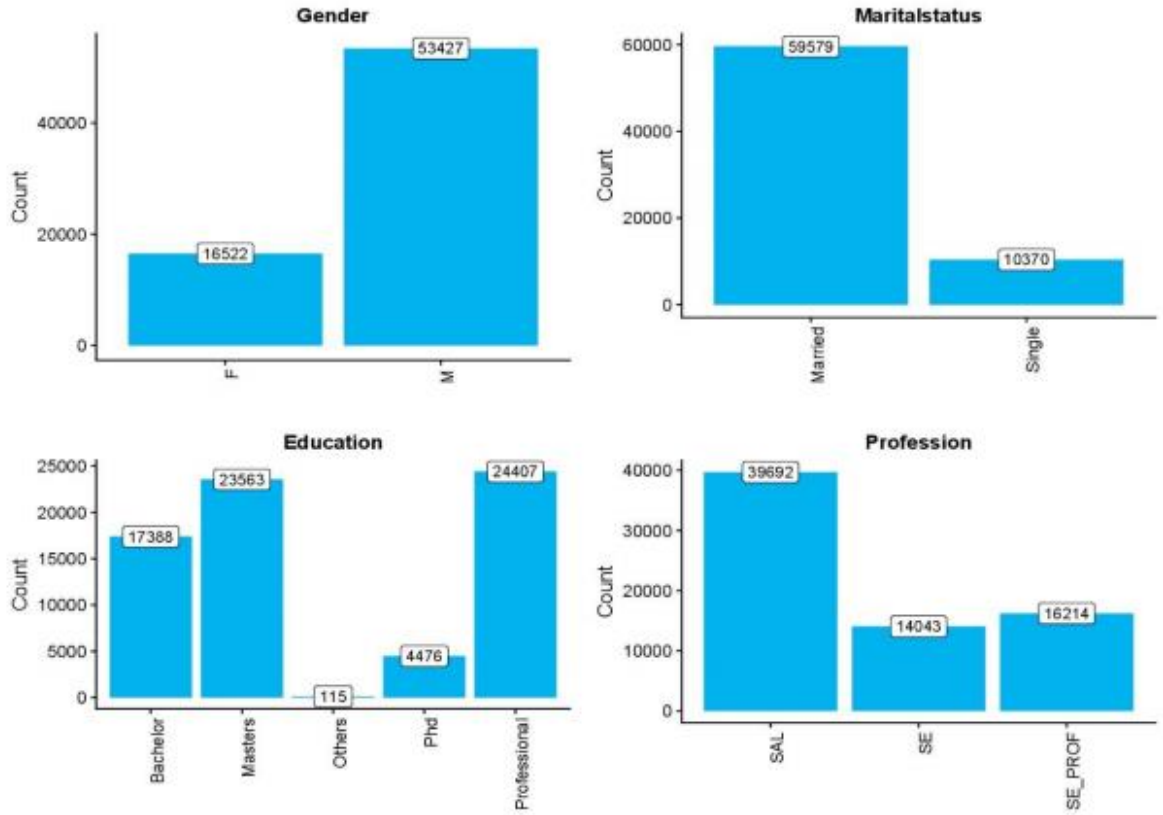
Data Cleaning

How we treated them :

- Since Age cannot be "-3" and Income cannot be "-0.5", Hence clearly that was a data quality issue. Also, again, the data was not significant enough to treat it by taking some mean or median. So we just dropped that bad entries.
- Now, looking from the business perspective, we know that loan cannot be granted to someone whose age is less than 18. However, there were 55 such entries with following unique values: 0,15,16,17. So all entries of customer, whose Age is less than 18 are considered as malicious entries, and we had to drop that.
- Furthermore, after looking more closely at the data, we found out that the max value of the variable "No_of_times_30_DPD_or_worse_in_last_6_months" is 7, whereas it cannot exceed 6 i.e. [30,60,90,120,150,180]. Hence, that is clearly an error. However, instead of dropping that, this error can be tolerated by tweaking it a little i.e. by capping that variable to max possible value -- 6.

Exploratory Data Analysis

1.Overall distribution of data for categorical/DISCRETE variables:

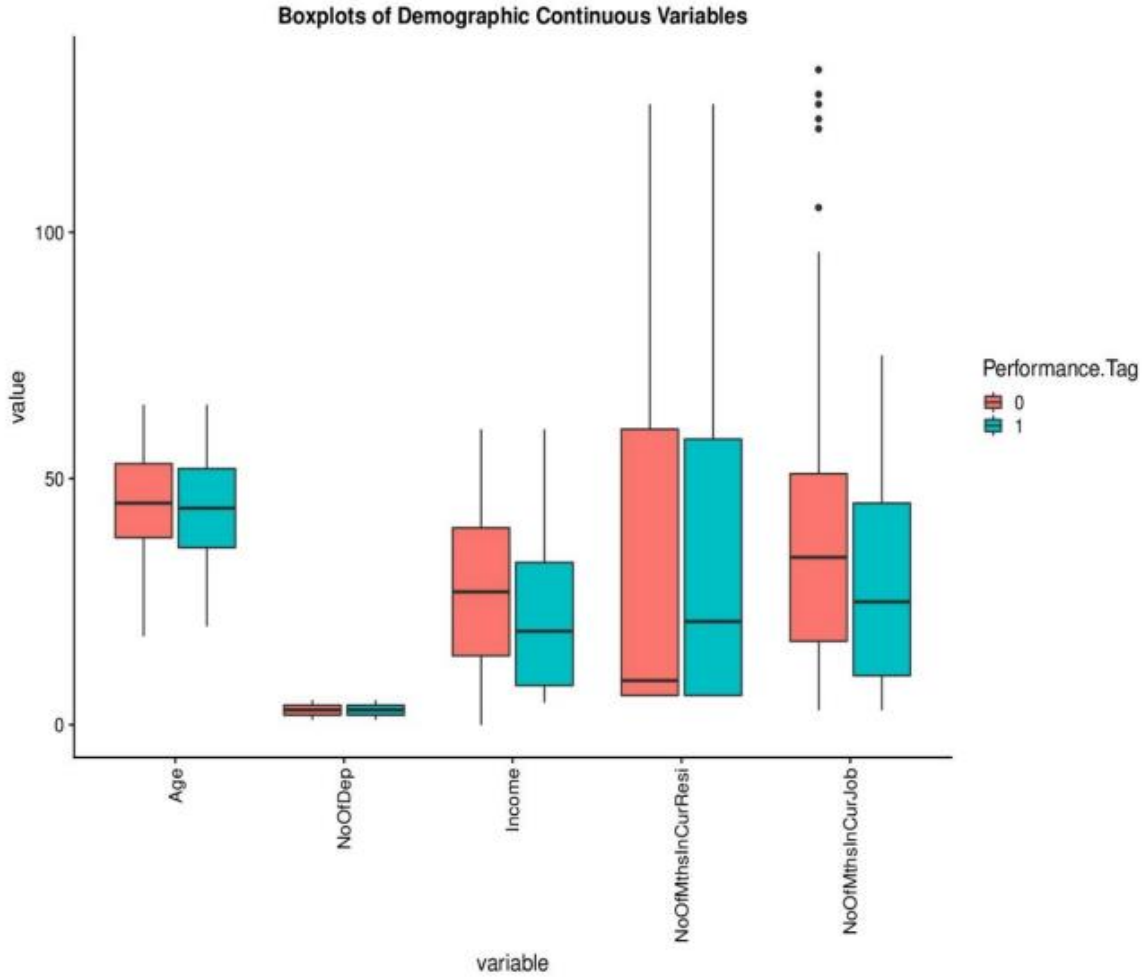


Inferences :

- For Gender, number of loan applications for Males is far larger than Females.
- Married People outnumber the Single People by a large factor in the dataset.
- The Others section in the Education sector had the lowest number of total credit card applicants amongst others in the same sector.
- Number of Salaried Applicants were the highest in the Profession Sector.

Continued…

2.BOX Plots OF CONTINUOUS variables

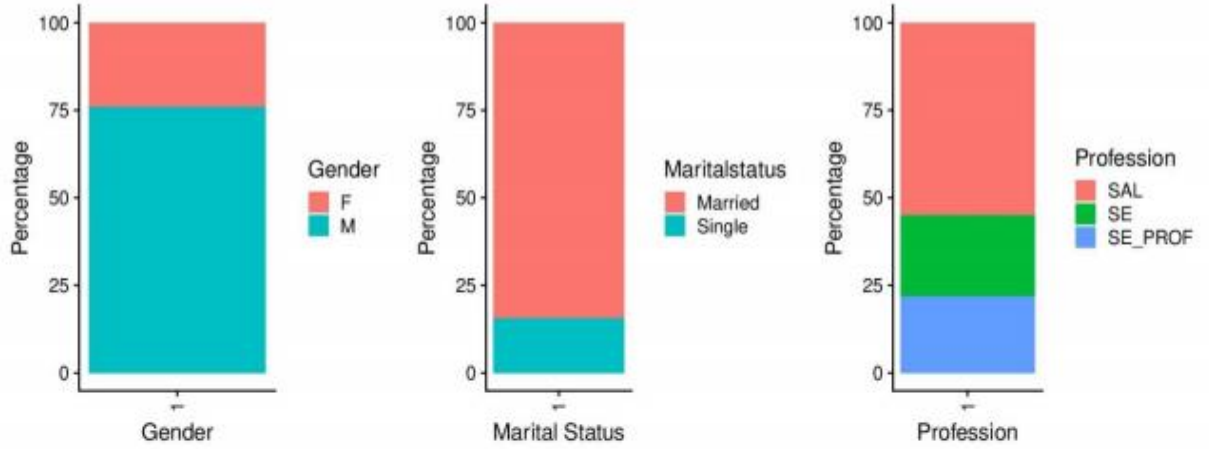


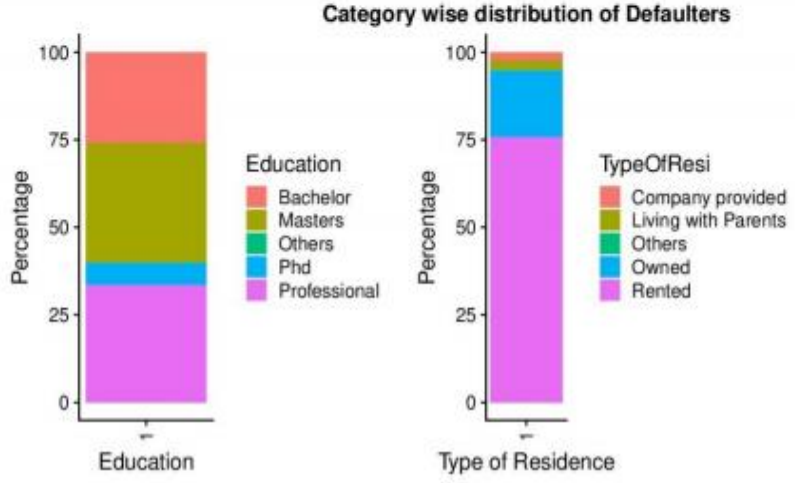
Inferences :

Some outliers exists in the variable “NoofMnthsInCurJob”, which should be treated before model building.

Continued

3.Distribution of DEFAULTERS for CATEGORICAL variables:





Inferences :

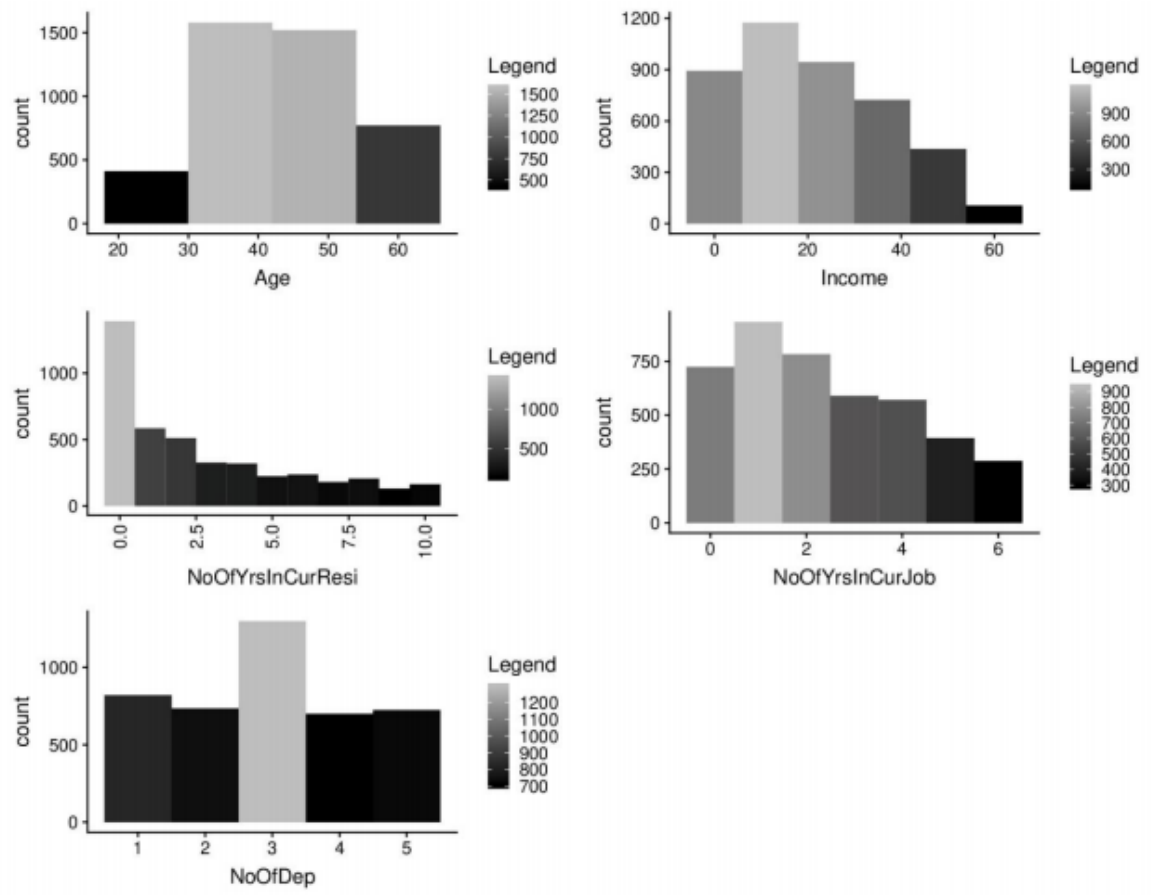
- Male gender has more number of defaulters than Female Gender. o Out of Total Defaulters, Married People accounts to 80% i.e. nearly 4 times as compared to Singles. o The number of defaulters contributed by “SAL” profession is more compared to other two profession "SE" ,"SE_PROF".

- In the Education Sector, Others and PHD contributed to minimum number of defaulters, w.r.t to others in the same sector.

- 75% of total defaulters were living on rent whereas around 20% of the defaulters had their own houses. There were only a few number of defaulters who were “Living with Parents: or in the Company Provided Spaces, or some other other type of residence.

Continued…

- **HISTOGRAM Distribution of DEFAULTERS for Continuous**

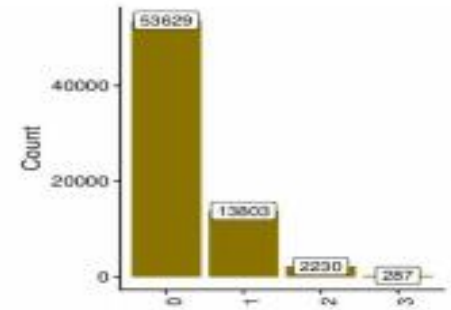


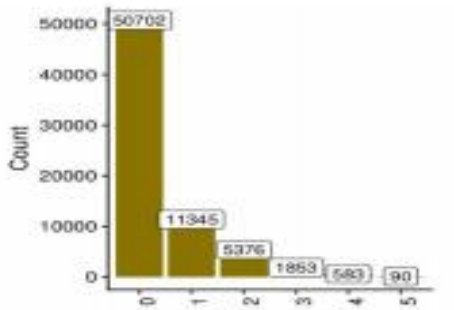
Inferences:

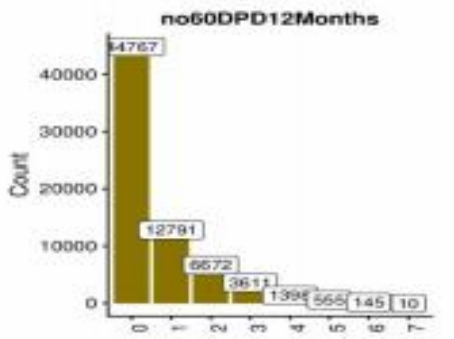
- Most of the defaulters were lying in the age ranging from 30 to 50.
- People having high income tended to default less; quite relatable fact.
- Usually, defaulters were the ones whose “Number of Years in Current Residence” or “Number of Years in Current Job” was less than 3
- People having 3 dependents had the maximum count of defaults, nearly 1.5 times as compared to their counterparts.

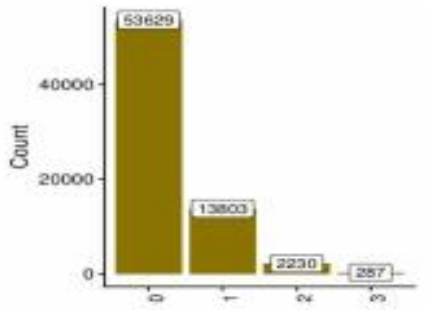
EXPLORATORY DATA ANALYSIS – CREDIT BUREAU

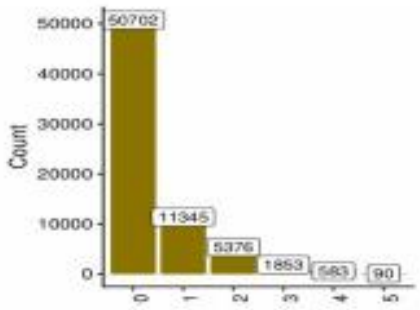
- **Overall distribution of data**

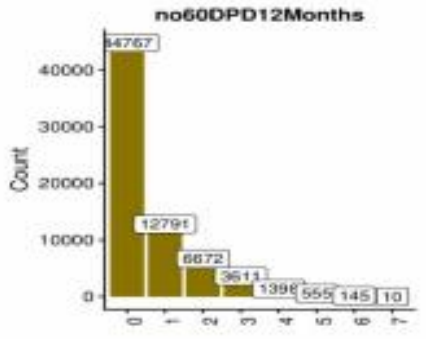


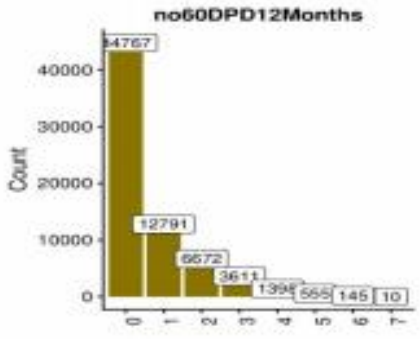


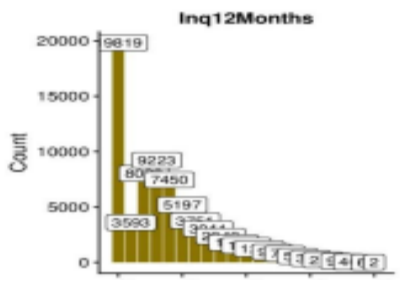


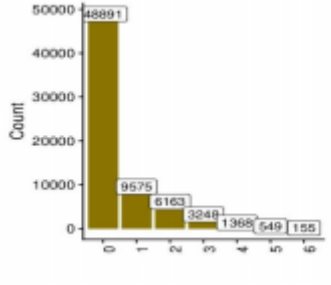


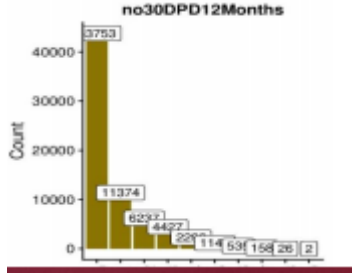




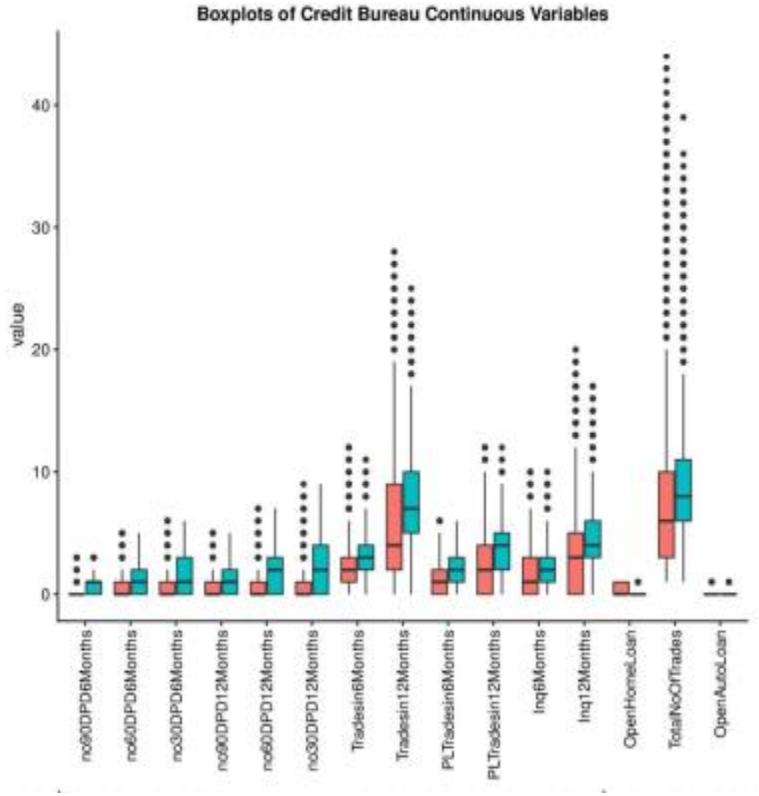


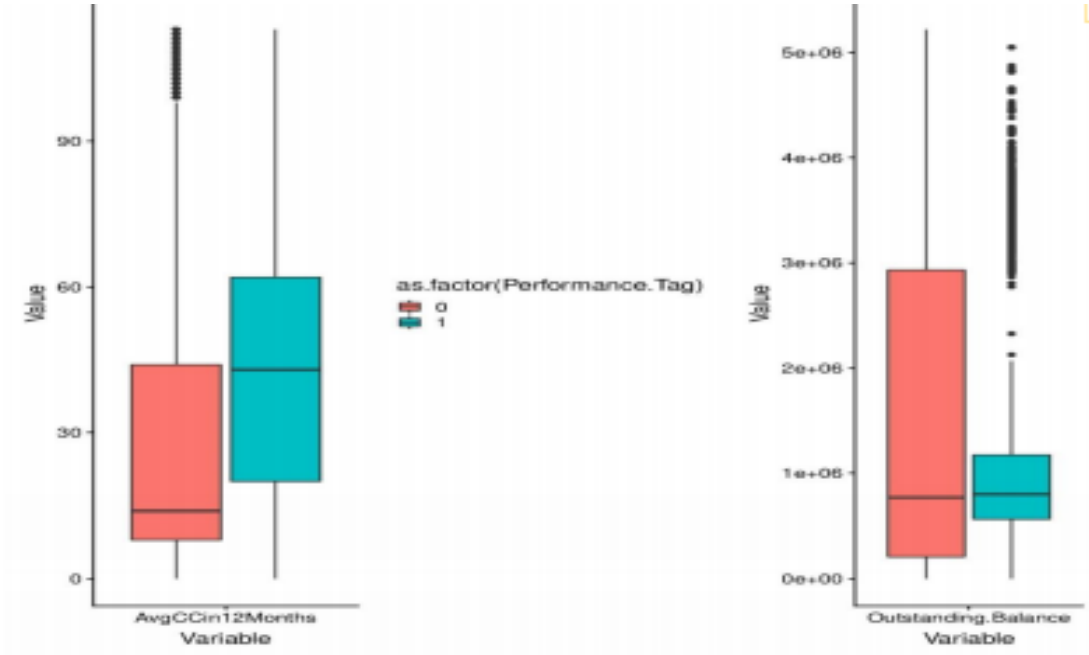






Continued…

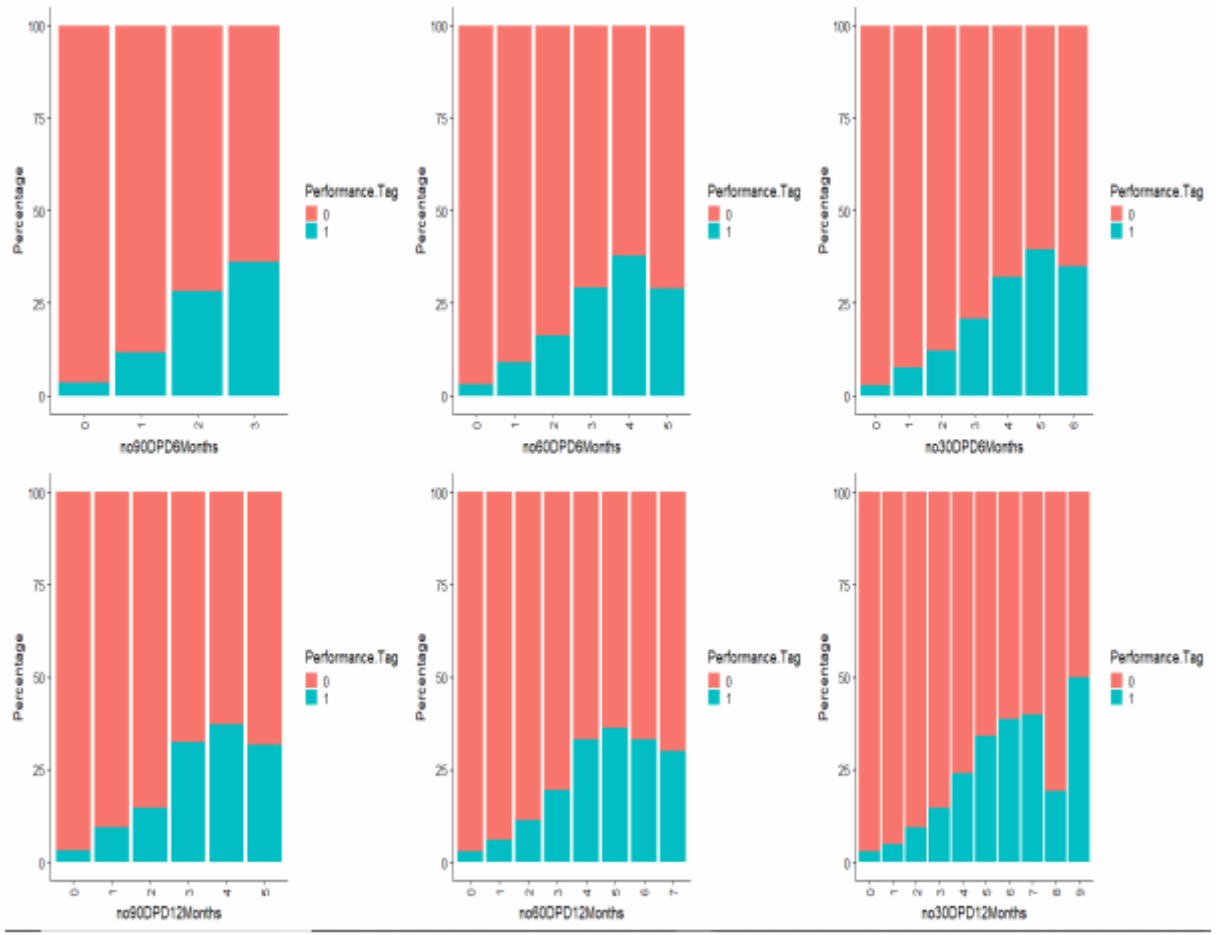




Inferences : The box plots show us that considerable outliers exists in almost all the variables of this dataset, hence we will be treating them or capping them before model building

Continued….

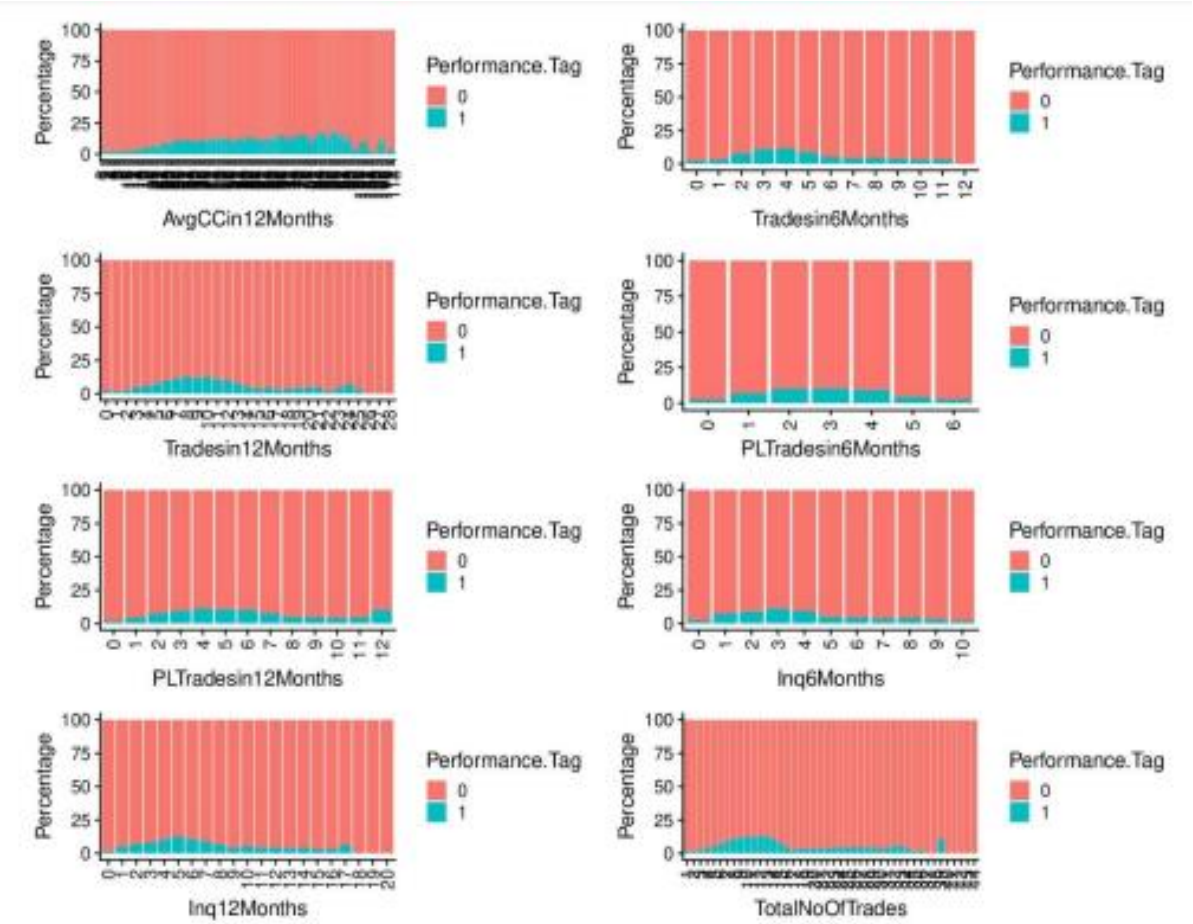
Distribution Of Defaulters



Inferences :

It is clearly apparent from all the (DPD) graphs that as the number of DPD increases, the likelihood of default increases. So all the graphs follow almost a linear increase in default percentage. Hence all the six DPD variables can be of significant importance for predicting the defaulters.

Continued….

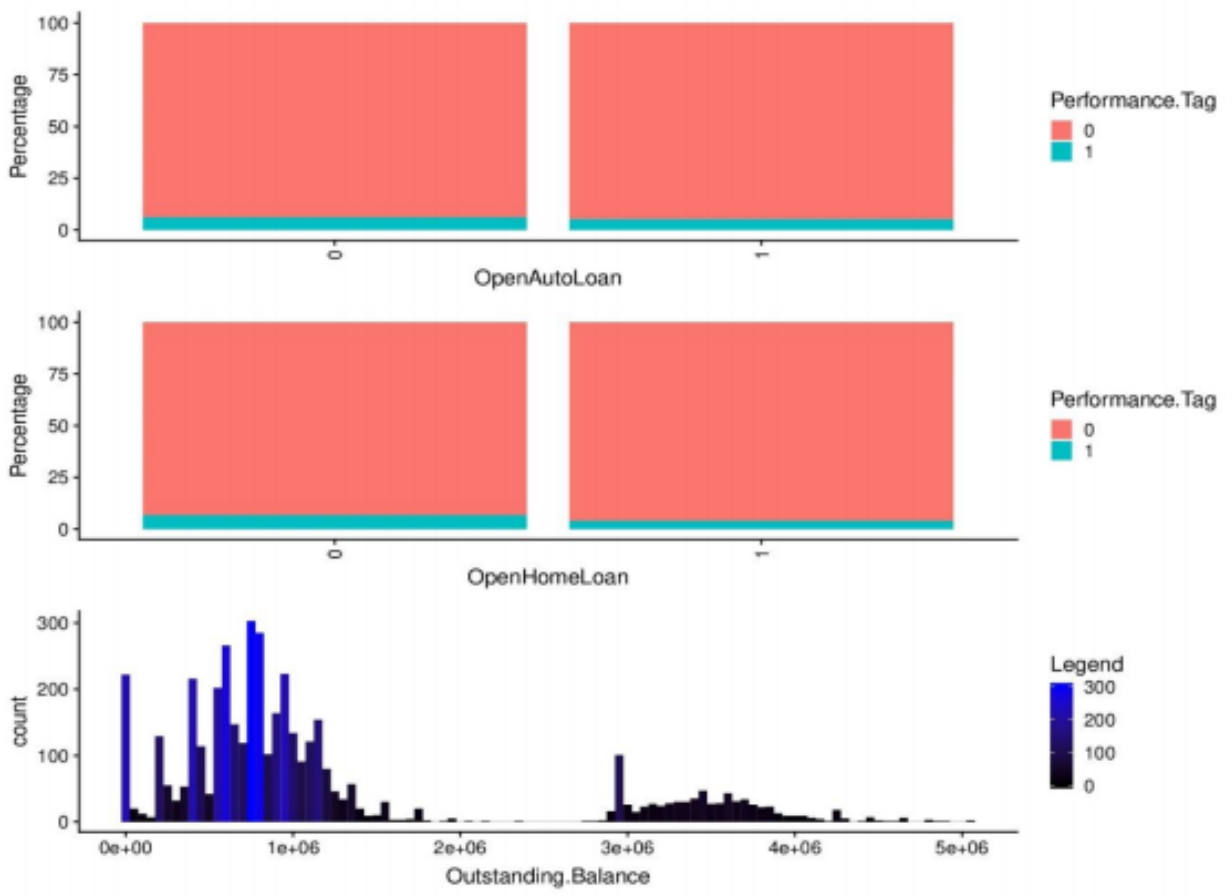


Inferences :

- Three defaulters of almost all the above six variables follow a nearly perfect bell curve i.e. lower percentage of defaulters towards both the ends, and higher percentage in the middle.

- However, the bell curve distribution of defaulters for the “TotalNoOfTrades” variable is slightly skewed towards the right. So we will be fixing them probably with log-transformation method before model building.

Continued….

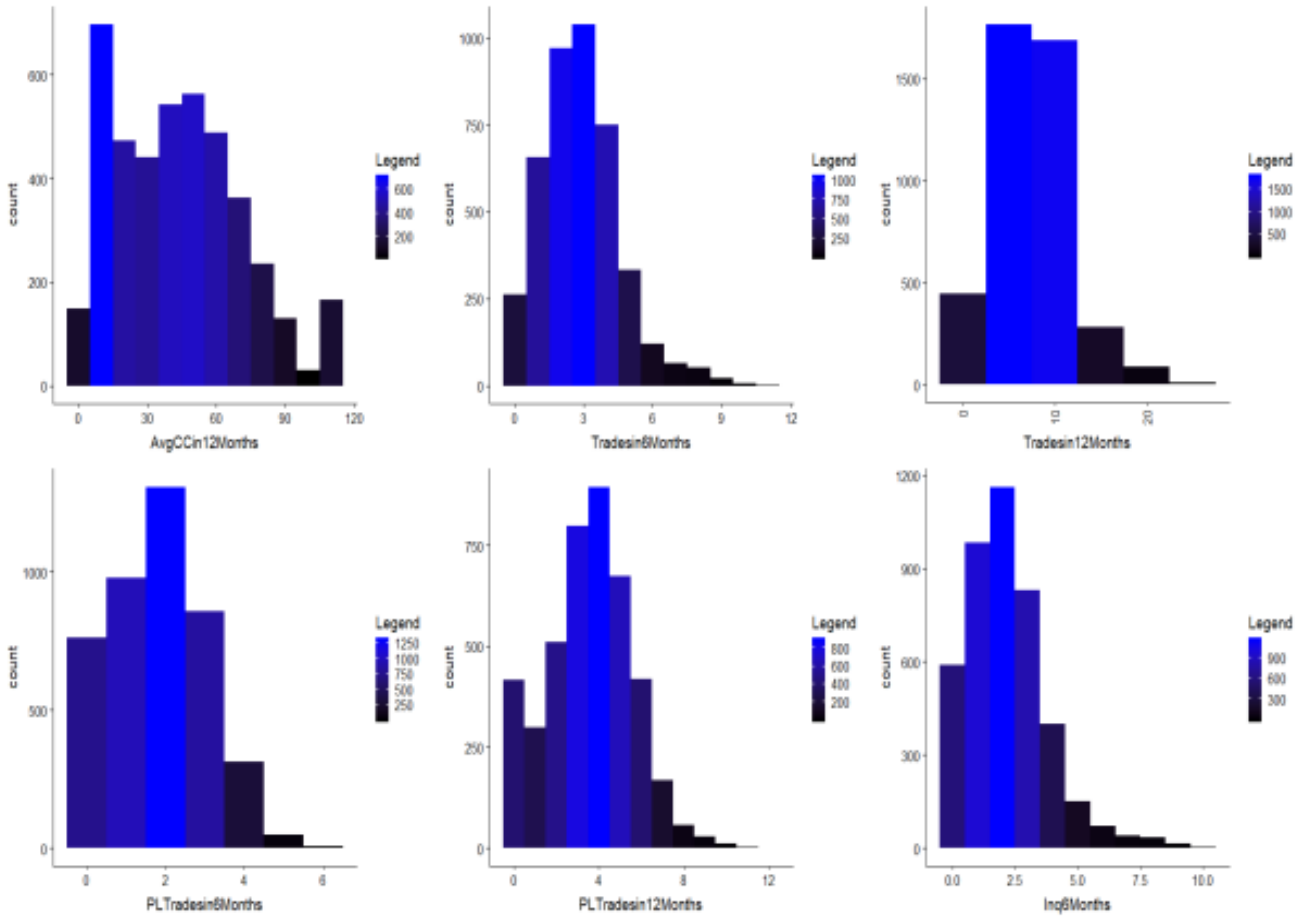


Inferences :

- Having an “Open Home Loan” or not having an “Open Home Loan” does not make much difference in predicting the defaulters. And same goes for the “Open Auto Loan” variable. So these variables does not appear of any significant importance for prediction.

- The Histogram Distribution of Outstanding Balance is almost random. However, the interesting insight is: The number of defaulters having less outstanding balance outnumber the number of defaulters having higher outstanding balance. Hence, this variable can be of moderate importance for us, later.

Continued….



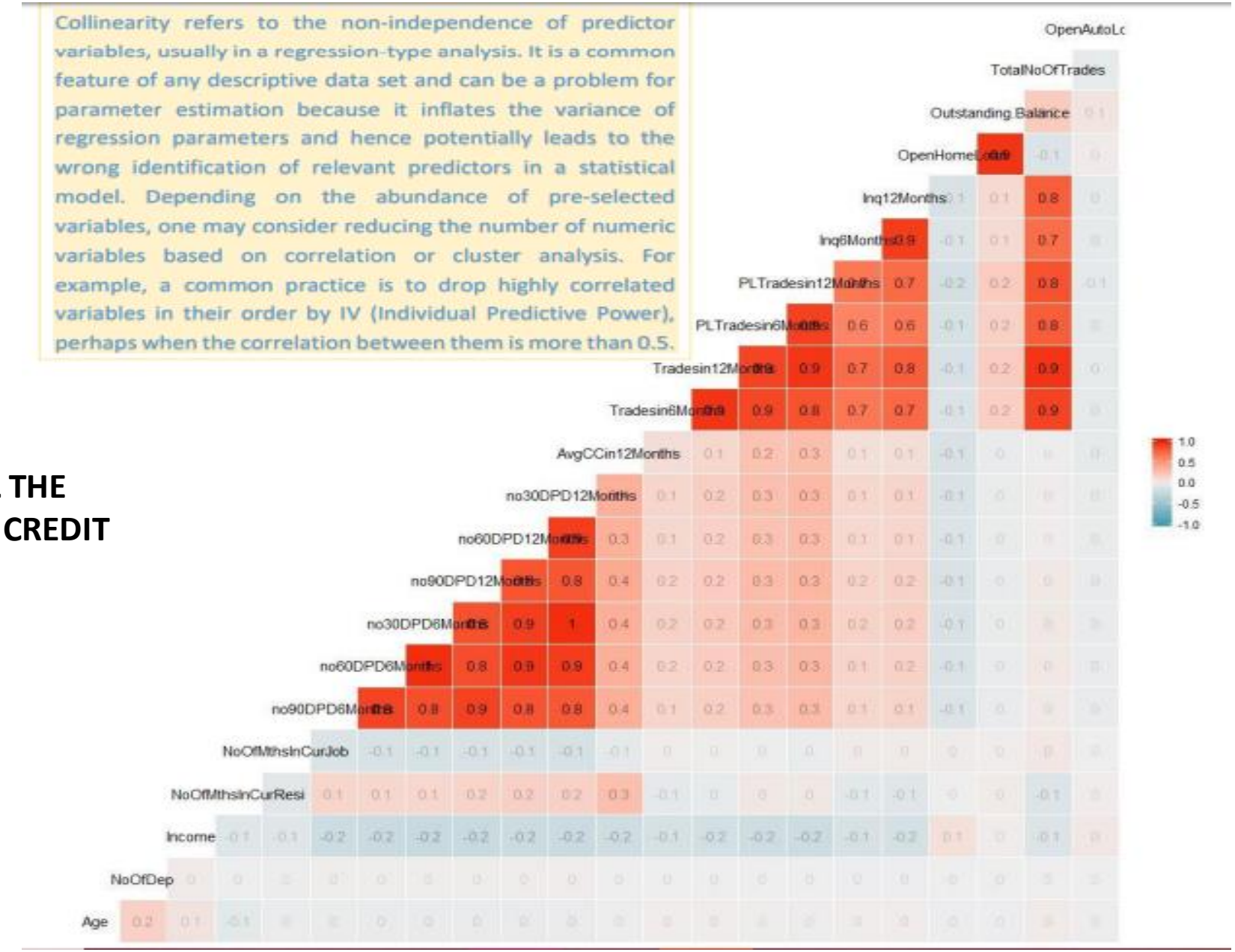
Inferences :

- Again, almost all the above six variables follow a bell curve i.e. lower percentage of defaulters towards both the ends, and higher percentage in the middle.
- However, the histograms of last three variables i.e. “PL Trades in 6 Months”, “PL Trades in 12 months” and “Inq 6 Months” is skewed towards right. So we will be making them more symmetric with log transformation method.

Continued….

Collinearity refers to the non-independence of predictor variables, usually in a regression-type analysis. It is a common feature of any descriptive data set and can be a problem for parameter estimation because it inflates the variance of regression parameters and hence potentially leads to the wrong identification of relevant predictors in a statistical model. Depending on the abundance of pre-selected variables, one may consider reducing the number of numeric variables based on correlation or cluster analysis. For example, a common practice is to drop highly correlated variables in their order by IV (Individual Predictive Power), perhaps when the correlation between them is more than 0.5.

CORRELATION MATRIX OF ALL THE VARIABLES [DEMOGRAPHIC + CREDIT BUREAU]

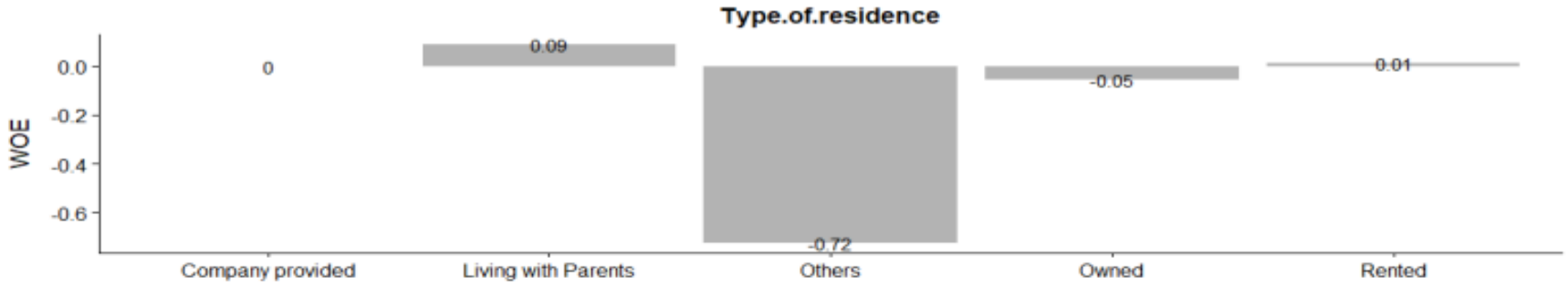


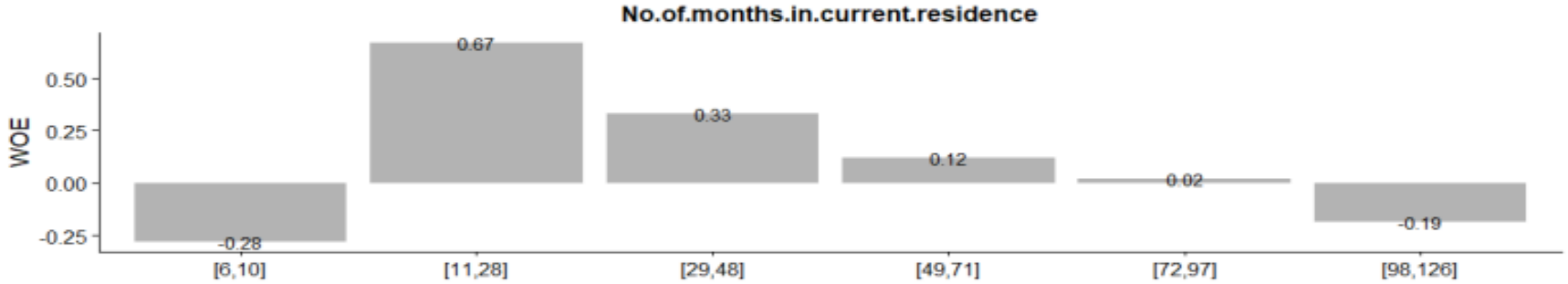
WOE Analysis....

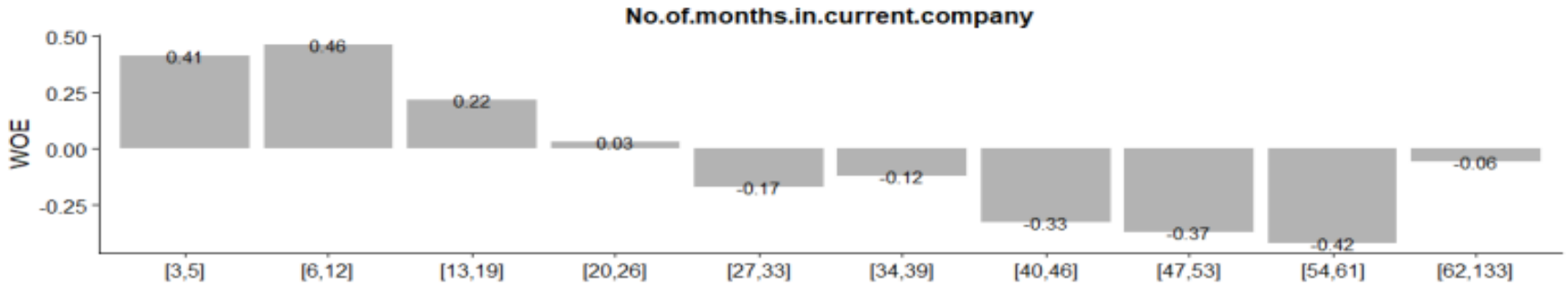
Woe Plots

Note :

WOE describes the relationship between a predictive variable and a binary target variable Y. (Y = "Performance Tag" in our case). Now the greater the value of WOE, the higher the chance of observing Y=1.Hence in our dataset, since Y=1 is the indication of default, the interpretation for WOE specifically for our case will be:
Higher the WOE, higher the chances of default.





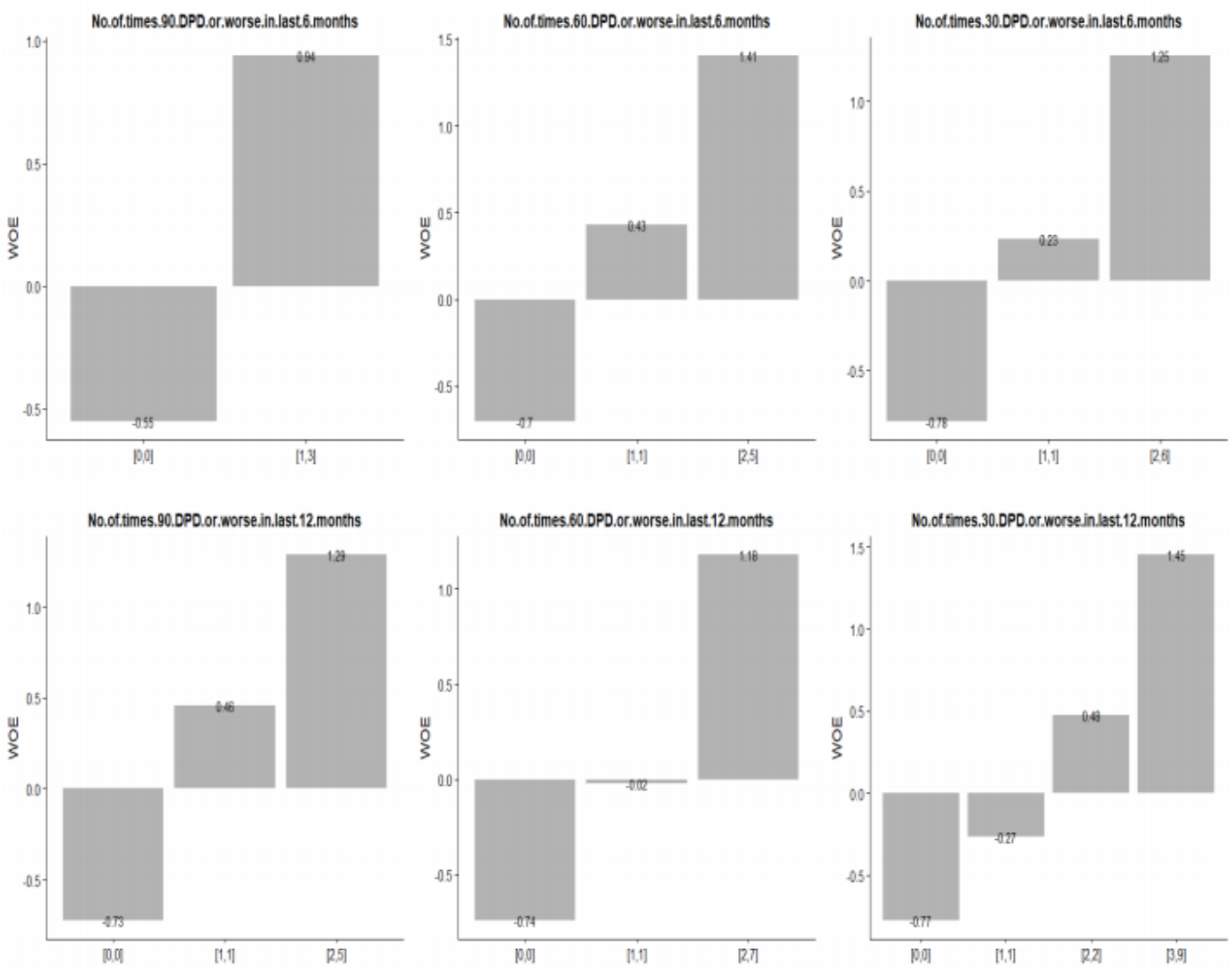


Inferences :

- Applicants who are living with parents tend to default more.
- Applicant whose "No. of months in current residence" is between 11 to 28 has the highest chances of default.
- Applicant whose "No. of months in current company" is between 6 to 12 has the highest chances of default.

Continued….

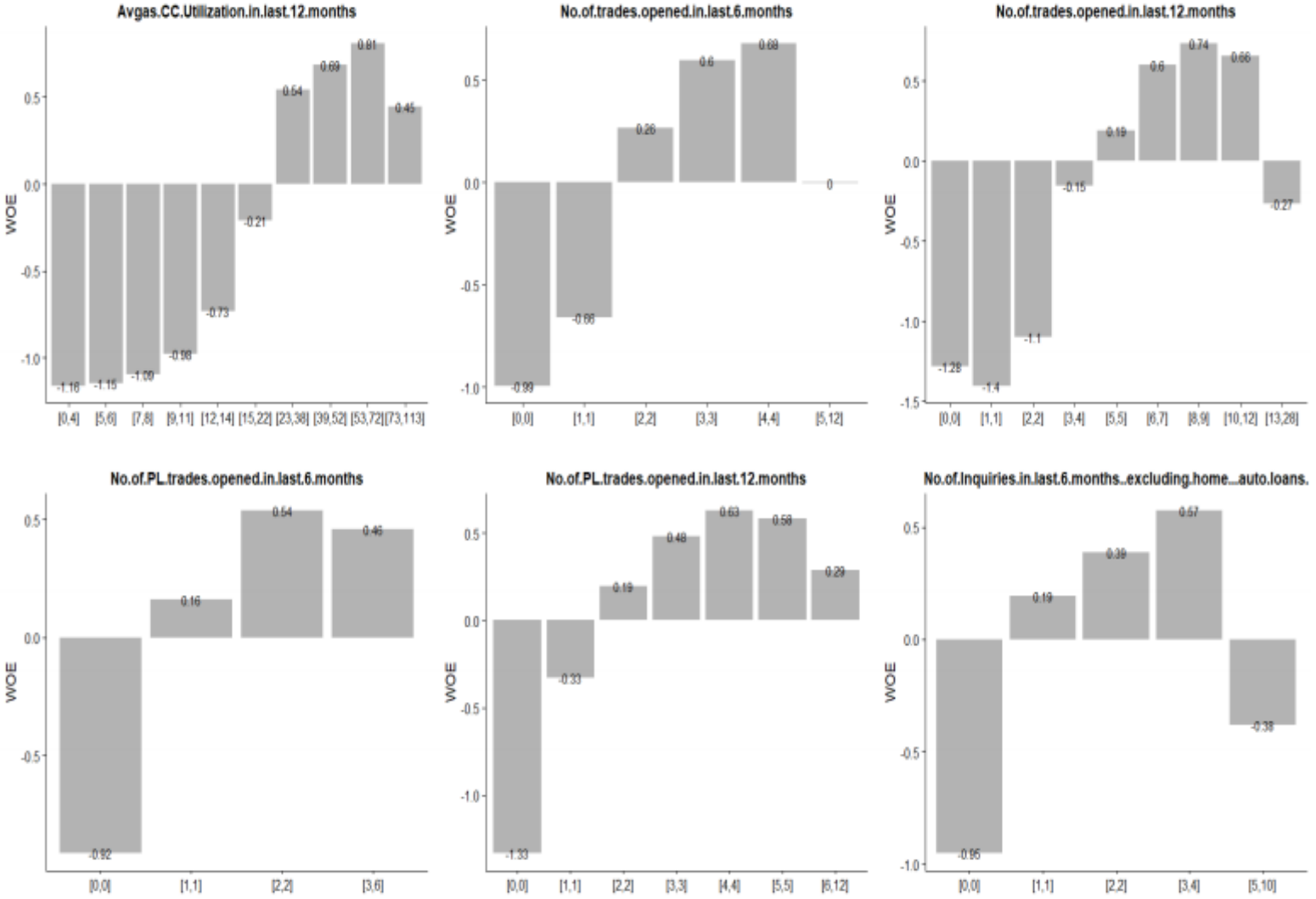
Woe Plots



Inferences :

All the six plots indicate the same insight i.e. as the “No of times of DPD of Applicant” exceeds 2, applicants become more likely to default.

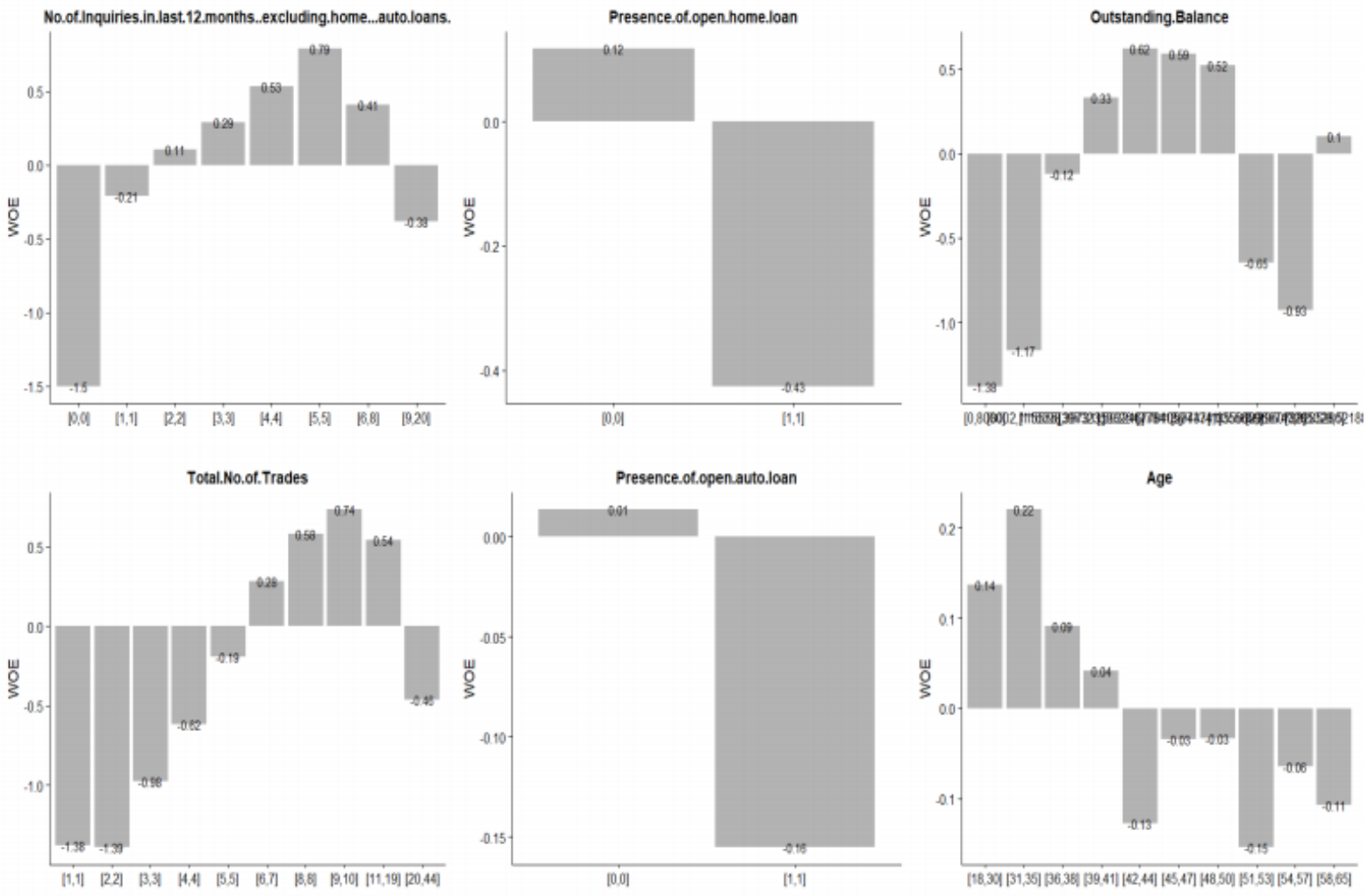
Continued….



Inferences :

- Applicant whose “Average Credit Card Utilization in last 12 months” lies between 53 to 72 has the highest chances of default.
- Applicant whose “No of Trades opened in last 6 months” = 4 has the highest chances of default.
- Applicant whose “No of Trades opened in last 12 months” = 8 or 9 has the highest chances of default. Applicant whose “No of PL Trades opened in last 6 months” = 2 has the highest chances of default.
- Applicant whose “No of PL Trades opened in last 12 months” = 4 has the highest chances of default.
- Applicant whose “No of Inquires in last 6 months” = 3 or 4 has the highest chances of default.

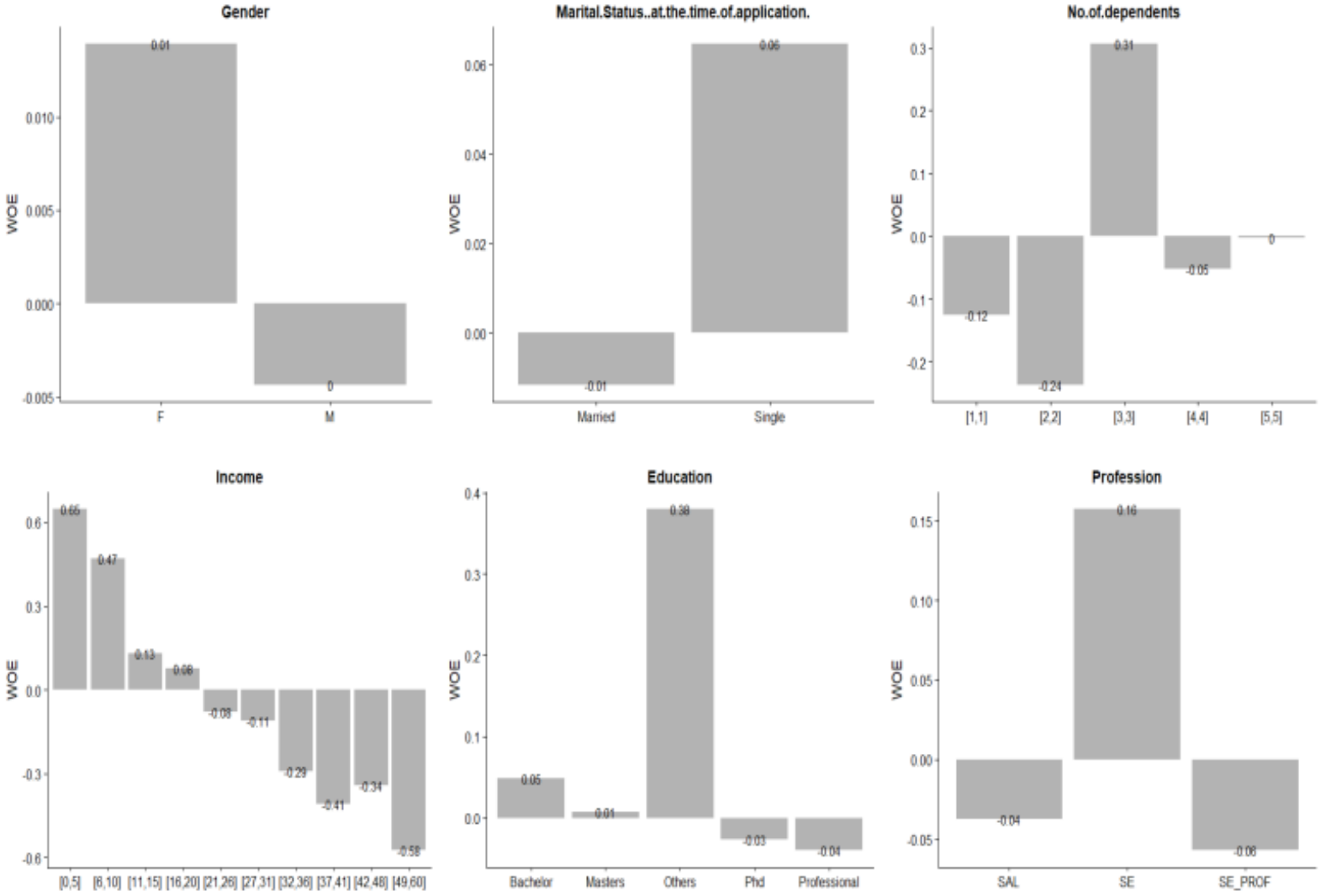
Continued....



Inferences :

- Applicant whose “No of Inquires in last 12 months” = 5 has the highest chances of default.
- Applicant who “Do not have an open home loan” has higher chances of default compared to those who has an open home loan.
- Applicant whose “Outstanding Balance” is between 10 to 20 Lacs has higher chances of default.
- Applicant whose “Total no of Trades” is 9 or 10 has higher chances of default.
- Applicant who “Do not have an open auto loan” has higher chances of default compared to those who has an open auto loan.
- Applicant whose “Age” lies between 31 to 35 has the highest chances of default.

Continued....



Inferences :

- Females tend to default more than males.
- Applicant who are single tend to default more than the married ones.
- Applicant whose “No of dependents” = 3 has higher chances of default.
- Applicant whose “Income” is less than 5K tend to default more.
- Applicant who “Education” is Others has highest chances of default.
- Applicant who are Self Employed tend to default more.

Model Building Approach

In this section, we analysed popular predictive modelling techniques for credit scoring. Each of the methods has its own strengths and weakness, which often vary according to the circumstances and depend on the data quality.

As a part of this Projects we implemented below models for credit scoring on each datasets.

Method	Main technique	Summary
Logistic regression	Maximum likelihood estimation	Determine formula to estimate binary response variable.
Decision trees	Recursive Partitioning Algorithms	Uses tree structure to maximize between group differences.
Neural Networks	Multilayer perceptron	Artificial Intelligence technique, whose results are difficult to interpret.

Logistic Regression :

- Logistic regression(LR) helps in building a reasonable model to describe a relationship between a dependent and one or more independent variables.
- We have applied LR on the required two datasets :

1) Demographic Dataset
2) Demographic and Credit Bureau Dataset (Formed by merging two datasets)

- Dependent variable in both datasets : **Performance Tag**

Implementation : Logistic Regression Model

Demographic data Model

Initial Independent Variables	Independent Variables after Modelling:
Age NoOfDep Income NoOfMthsInCurResi NoOfMthsInCurJob Gender Maritalstatus Education.xMasters Education.xOthers Education.xPhd Education.xProfessional Profession.xSE Profession.xSE_PROF TypeOfResi.xLiving.with.Parents TypeOfResi.xOthers TypeOfResi.xOwned TypeOfResi.xRented	Age ,NoOfDep ,Income ,NoOfMthsInCurJob ,Profession.xSE

Accuracy	Sensitivity	Specificity
58.3%	58.5%	58.2%

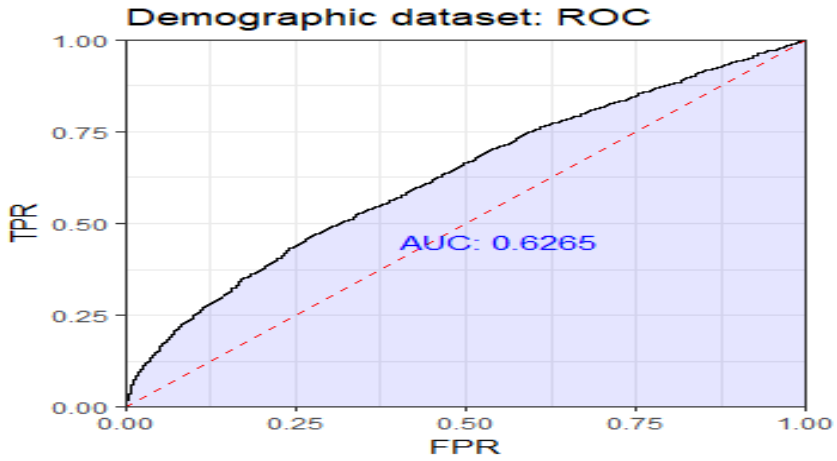
Model using both demographic and credit bureau data

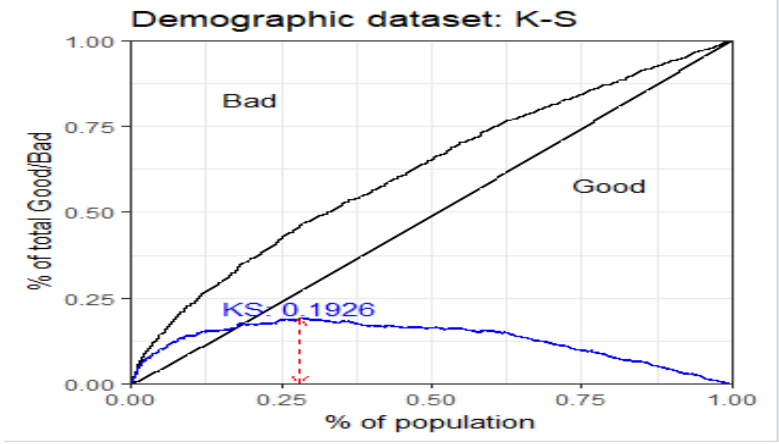
Initial Independent Variables	Independent Variables after Modelling:
"Age" "NoOfDep" "Income" "NoOfMthsInCurResi" "NoOfMthsInCurJob" "no90DPD6Months" "no60DPD6Months" "no30DPD6Months "no90DPD12Months" "no60DPD12Months" "no30DPD12Months" "AvgCCin12Months" "Tradesin6Months" "Tradesin12Months" "PLTradesin6Months" "PLTradesin12Months" "Inq6Months "Inq12Months" "Outstanding.Balance" "TotalNoOfTrades" "Gender" "Maritalstatus" "Education.xMasters" "Education.xOthers" "Education.xPhd" "Education.xProfessional" "Profession.xSE" "Profession.xSE_PROF" "TypeOfResi.xLiving.with.Parents" "TypeOfResi.xOthers" "TypeOfResi.xOwned" "TypeOfResi.xRented" "OpenHomeLoan" "OpenAutoLoan"	Age,NoOfDep ,Income NoOfMthsInCurResi , NoOfMthsInCurJob, no90DPD12Months, AvgCCin12Months, PLTradesin12Months, Outstanding.Balance, Profession.xSE

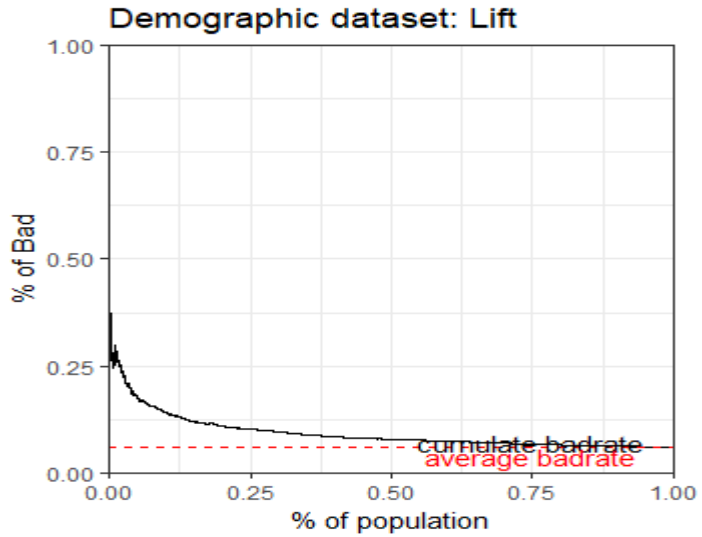
Accuracy	Sensitivity	Specificity
67.1%	69%	67%

In both the cases Confusion matrix created as per the Optimal Cutoff

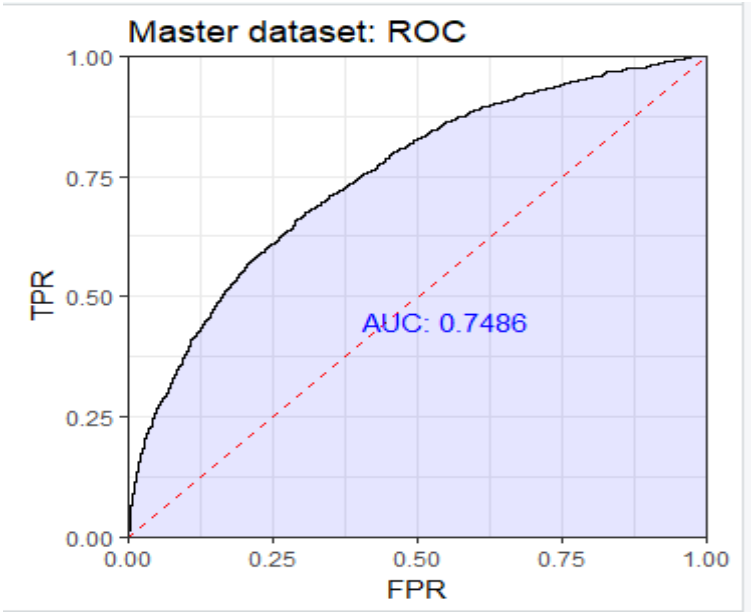
Demographic data Model

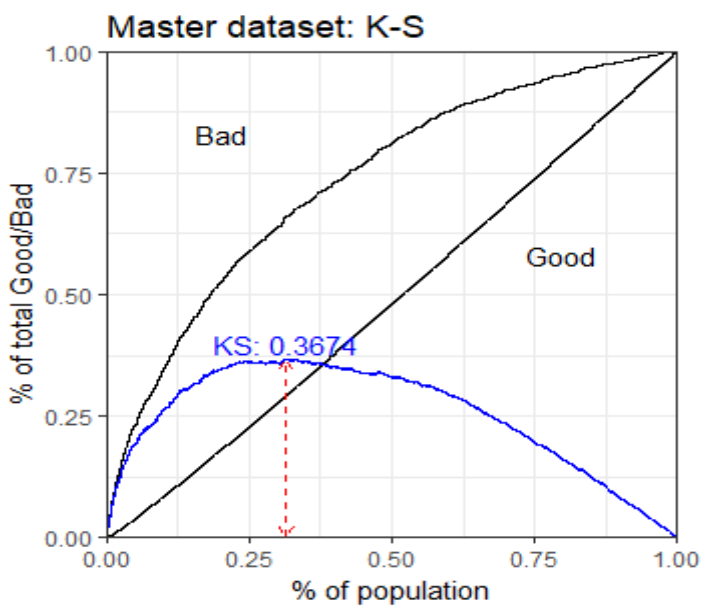


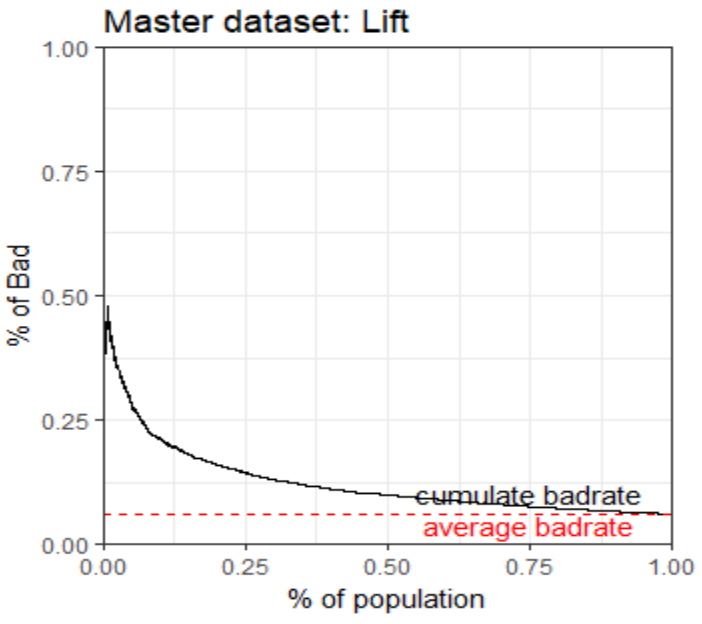




Model using both demographic and credit bureau data







Observations during Analysis :

Unbalanced Data

% Defaults	% Non-Defaults
94%	6%

Balanced Data

% Defaults	% Non-Defaults
61%	39%

- The data is highly biased towards Non-Defaulters in both the datasets. which accounts to nearly 94%. And only 6% of the total
- Hence, to get balanced results, we applied the idea of weighting which is related to sampling in order to get sufficient number of 1's for the maximum likelihood to converge.
- Balancing the dataset helps in boosting variables.
- Yields better results compared to that of unbalanced data.

Comparison between Balanced data and Unbalanced data O/P :

	Simple Logistic Regression	Weighted Logistic Regression
Demographics Data Model	Accuracy : 58.3% Sensitivity : 58.5% Specificity : 58.2%	Accuracy : 60% Sensitivity : 58% Specificity : 60%
Model using both demographic and credit bureau data	Accuracy : 67.1% Sensitivity : 69% Specificity : 67%	Accuracy : 67% Sensitivity : 70% Specificity : 67%

Implementation : Weighted Logistic Regression Model

Weighted Demographic data Model

Initial Independent Variables	Independent Variables after Modelling:
Age NoOfDep Income NoOfMthsInCurResi NoOfMthsInCurJob Gender Maritalstatus Education.xMasters Education.xOthers Education.xPhd Education.xProfessional Profession.xSE Profession.xSE_PROF TypeOfResi.xLiving.with.Parents TypeOfResi.xOthers TypeOfResi.xOwned TypeOfResi.xRented	Age NoOfDep Income NoOfMthsInCurJob Education.xProfessional Profession.xSE

Accuracy	Sensitivity	Specificity
60%	58%	60%

Weighted Model using both demographic and credit bureau data

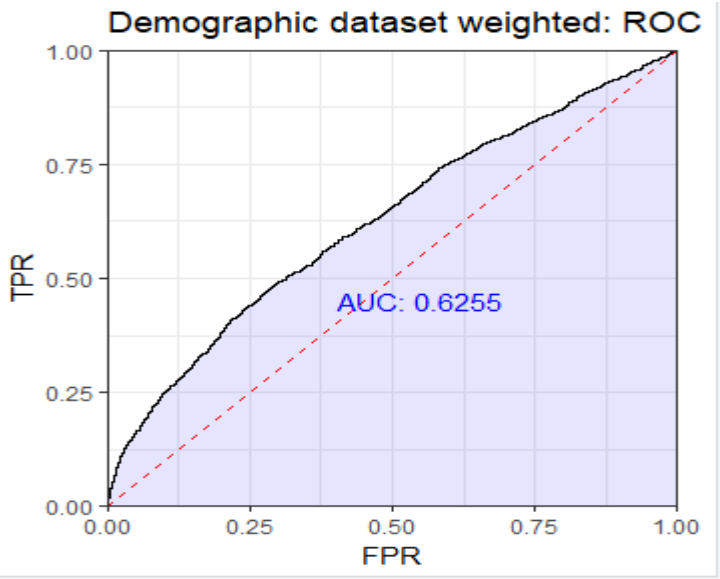
Initial Independent Variables	Independent Variables after Modelling:
"Age" "NoOfDep" "Income" "NoOfMthsInCurResi" "NoOfMthsInCurJob" "no90DPD6Months" "no60DPD6Months" "no30DPD6Months "no90DPD12Months" "no60DPD12Months" "no30DPD12Months" "AvgCCin12Months" "Tradesin6Months" "Tradesin12Months" "PLTradesin6Months" "PLTradesin12Months" "Inq6Months "Inq12Months" "Outstanding.Balance" "TotalNoOfTrades" "Gender" "Maritalstatus" "Education.xMasters" "Education.xOthers" "Education.xPhd" "Education.xProfessional" "Profession.xSE" "Profession.xSE_PROF" "TypeOfResi.xLiving.with.Parents" "TypeOfResi.xOthers" "TypeOfResi.xOwned" "TypeOfResi.xRented" "OpenHomeLoan" "OpenAutoLoan"	Income NoOfMthsInCurResi NoOfMthsInCurJob no90DPD12Months no30DPD12Months AvgCCin12Months PLTradesin12Months Outstanding.Balance Profession.xSE

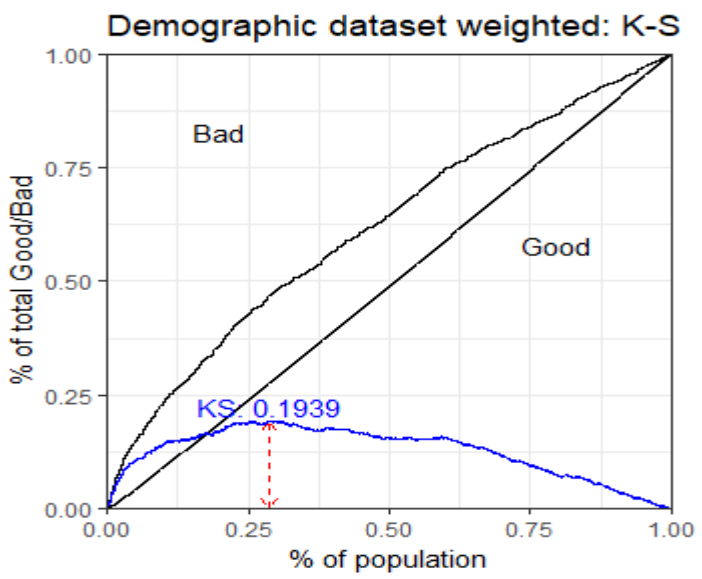
Accuracy	Sensitivity	Specificity
67%	70%	67%

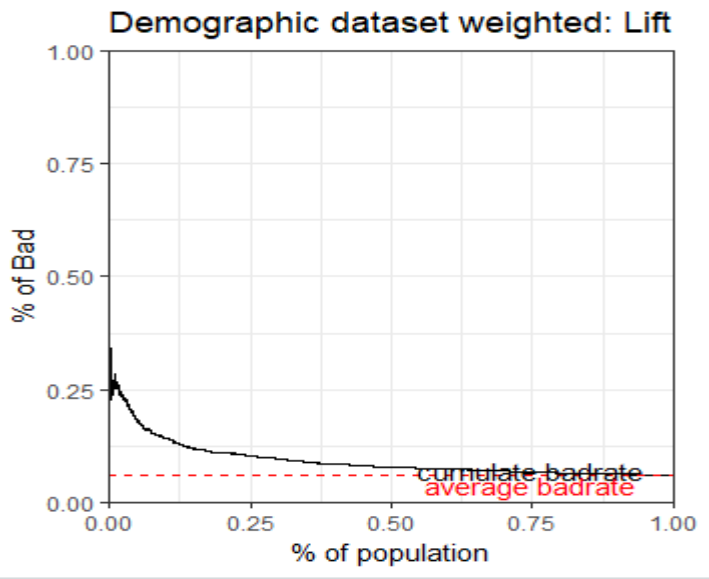
In both the cases Confusion matrix created as per the Optimal Cut off

Observation : Accuracy, sensitivity and specificity values are in better range in case of balanced datasets.

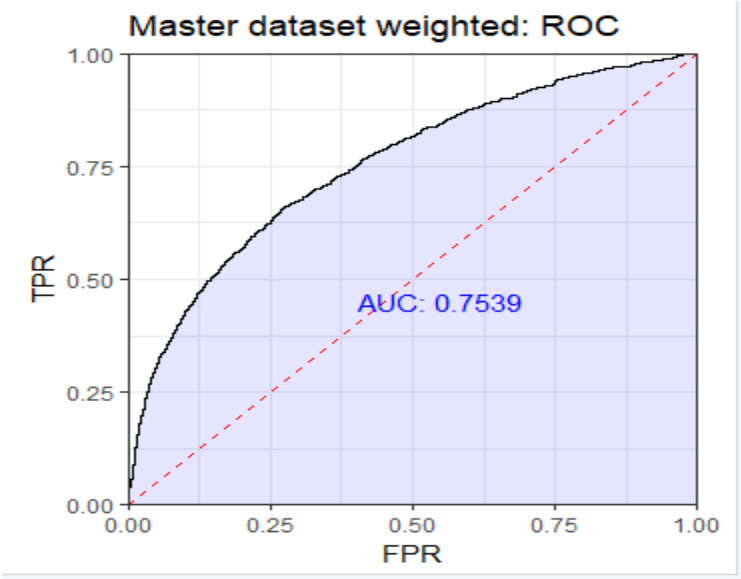
Demographic data Model

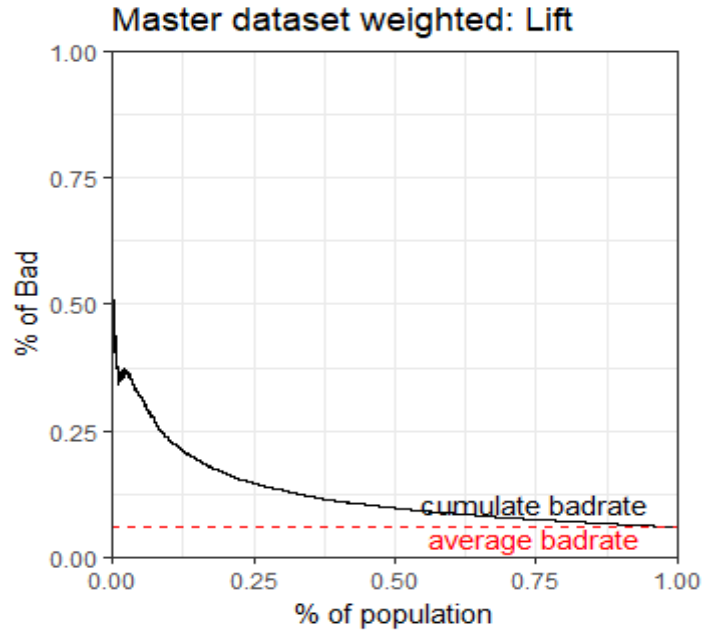






Model using both demographic and credit bureau data





Logistic Regression on WOE data:

- **Weight of evidence (WOE)** is often used to identify the important variables. Apart from assessing variable importance, WOE is also used to impute missing values from the data.
- Some variables contain a significant number of missing values.
- So we have created following two datasets by replacing the actual values of all the variables by the corresponding WOE value :

-Demographics WOE data

-Demographic and credit bureau WOE data

- In both the datasets : Dependent variable : Performance Tag

Implementation : Logistic Regression Model on WOE data

Demographic WOE data Model

Initial Independent Variables	Independent Variables after Modelling:
"Age" "Gender" "Maritalstatus" "NoOfDep" "Income" "Education" "Profession" "TypeOfResi" "NoOfMthsInCurResi" "NoOfMthsInCurJob"	Age NoOfDep Income Profession NoOfMthsInCurResi NoOfMthsInCurJob

Accuracy	Sensitivity	Specificity
63%	61.6%	63%

Model using both demographic and credit bureau WOE data

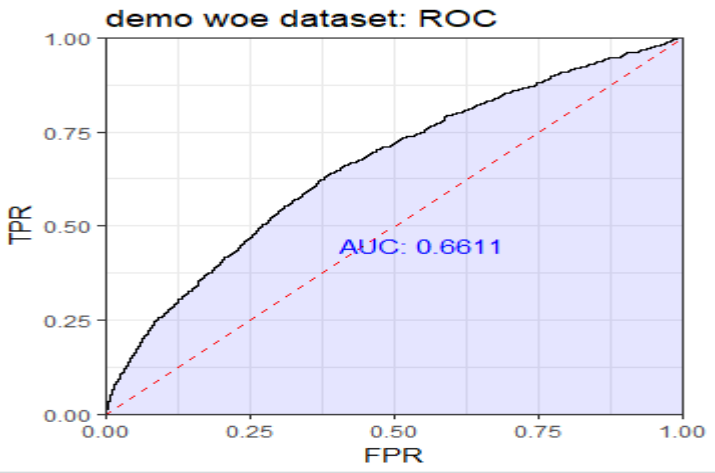
Initial Independent Variables	Independent Variables after Modelling:
"no90DPD6Months" "no60DPD6Months" "no30DPD6Months" "no90DPD12Months" "no60DPD12Months" "no30DPD12Months" "AvgCCin12Months" "Tradesin6Months" "Tradesin12Months" "PLTradesin6Months" "PLTradesin12Months" "Inq6Months" "Inq12Months" "OpenHomeLoan" "OutstandingBal" "TotalTrades" "OpenAutoLoan" "Age" "Gender" "MaritalStatus" "NoofDep" "Income" "Education" "Profession" "TypeofResi" "MonthsCurrResi" "MonthsCurrComp"	no90DPD12Months no30DPD12Months AvgCCin12Months PLTradesin12Months Inq12Months Age NoofDep Income Profession MonthsCurrResi MonthsCurrComp

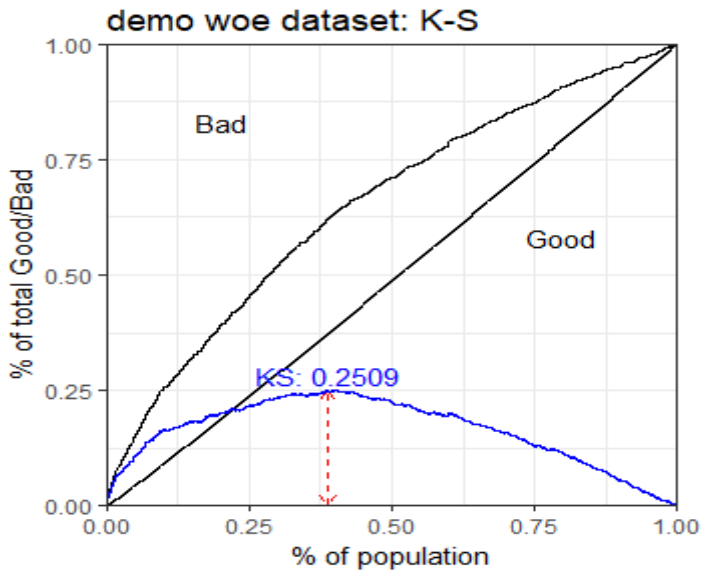
Accuracy	Sensitivity	Specificity
68.12%	69.6%	68%

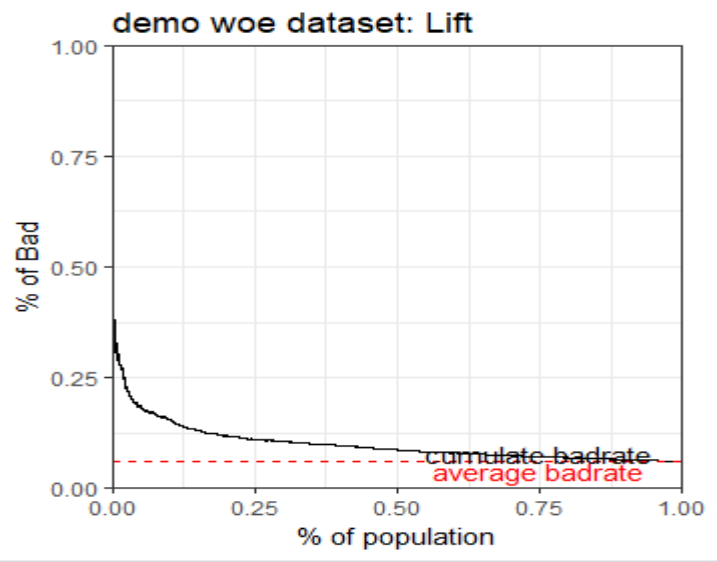
In both the cases Confusion matrix created as per the Optimal Cutoff.

Since the data is unbalanced, after converting into balanced dataset, we have applied logistic regression on the balanced datasets

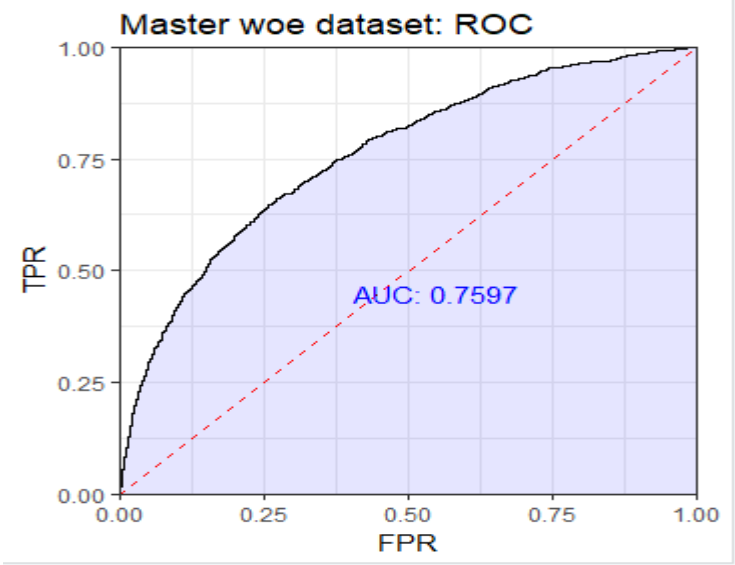
Demographic data Model

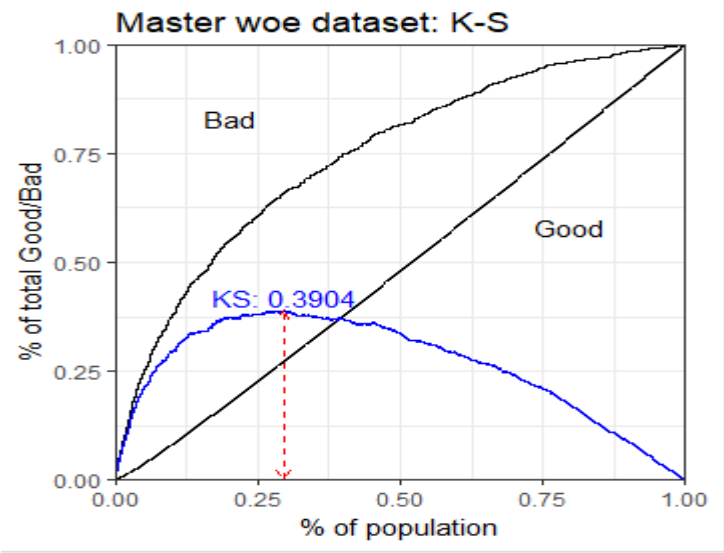


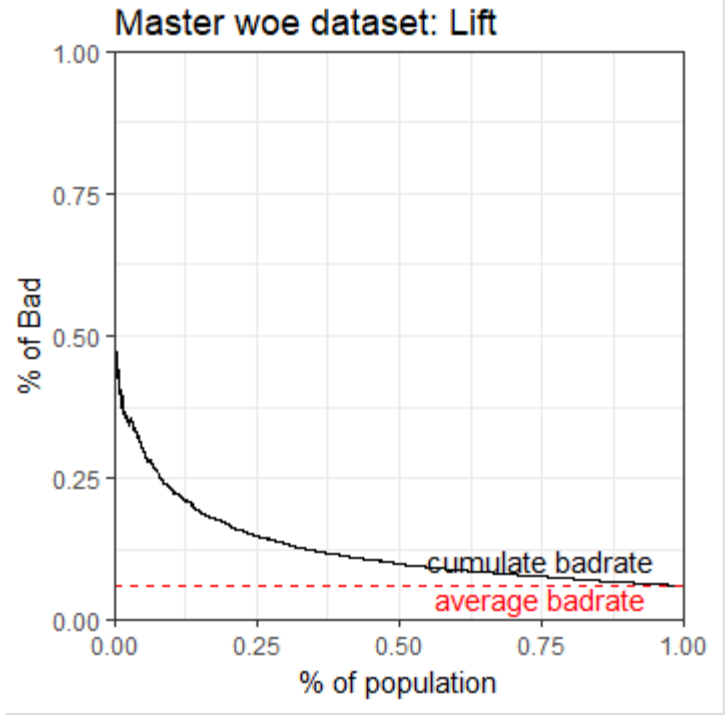




Model using both demographic and credit bureau data







Implementation : Weighted Logistic Regression Model on WOE data

Weighted Demographic WOE data Model

Initial Independent Variables	Independent Variables after Modelling:
"Age" "Gender" "Maritalstatus" "NoOfDep" "Income" "Education" "Profession" "TypeOfResi" "NoOfMthsInCurResi" "NoOfMthsInCurJob" "weights"	Age NoOfDep Income Education Profession NoOfMthsInCurResi NoOfMthsInCurJob

Accuracy	Sensitivity	Specificity
61.2%	62.4%	61.1%

Weighted Model using both demographic and credit bureau WOE data

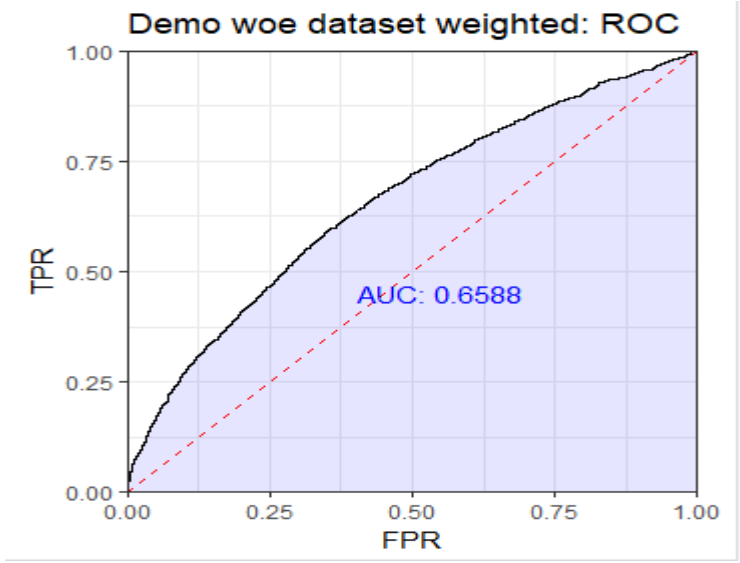
Initial Independent Variables	Independent Variables after Modelling:
"no90DPD6Months" "no60DPD6Months" "no30DPD6Months" "no90DPD12Months" "no60DPD12Months" "no30DPD12Months" "AvgCCin12Months" "Tradesin6Months" "Tradesin12Months" "PLTradesin6Months" "PLTradesin12Months" "Inq6Months" "Inq12Months" "OpenHomeLoan" "OutstandingBal" "TotalTrades" "OpenAutoLoan" "Age" "Gender" "MaritalStatus" "NoofDep" "Income" "Education" "Profession" "TypeofResi" "MonthsCurrResi" "MonthsCurrComp" "weights"	no30DPD12Months AvgCCin12Months PLTradesin6Months Inq12Months OpenHomeLoan NoofDep Income Profession MonthsCurrComp

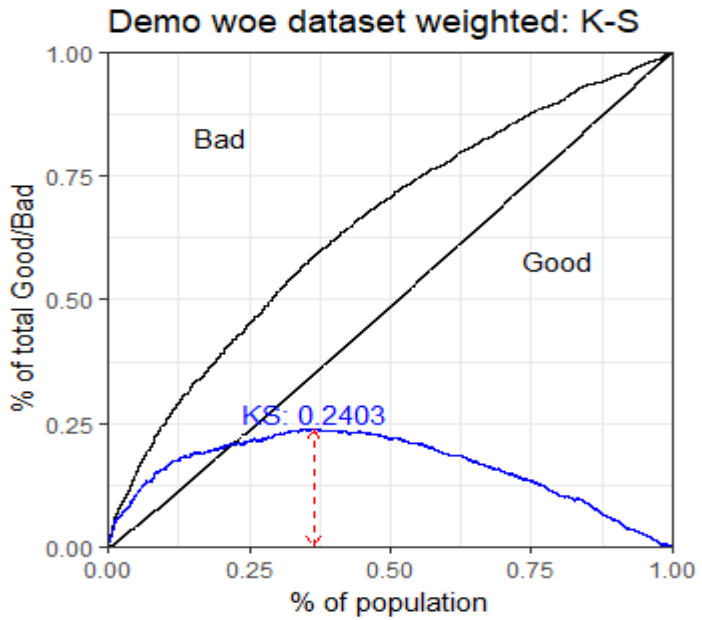
Accuracy	Sensitivity	Specificity
70.13%	68.4%	70.2%

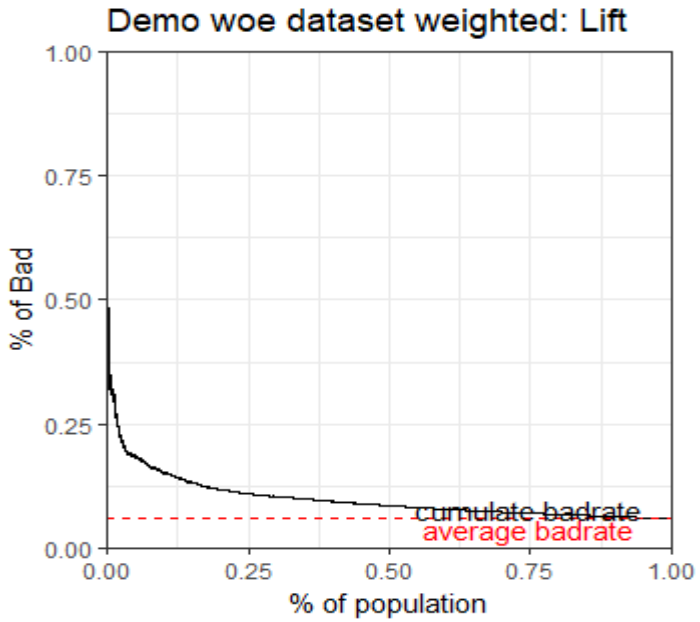
In both the cases Confusion matrix created as per the Optimal Cut off

Observation : Accuracy, sensitivity and specificity values are in better range in case of balanced datasets.

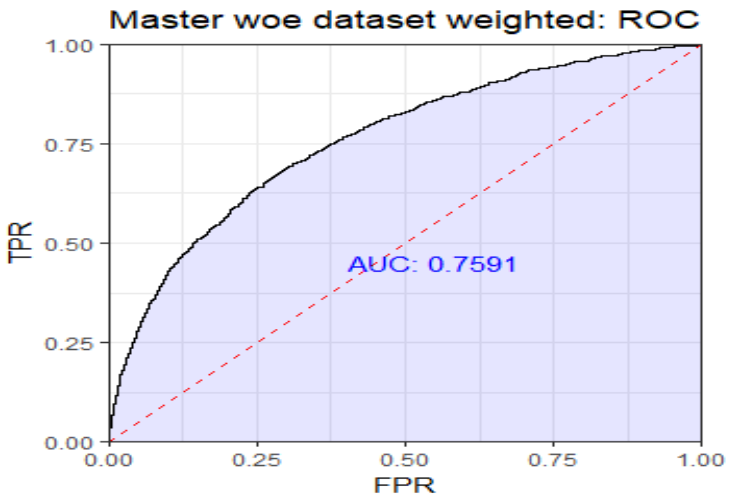
Demographic data Model

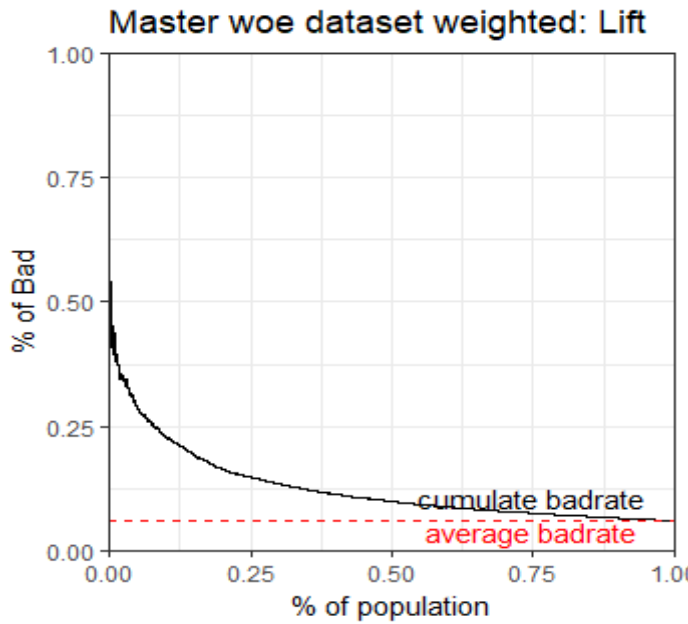






Model using both demographic and credit bureau data





Random Forest :

- Supervised classification algorithm, this algorithm creates the forest with a number of trees.
- In general, the **more trees in the forest** the more robust the forest looks like. In the same way in the random forest classifier, the **higher the number** of trees in the forest gives **the high accuracy** results.
- We have applied RF on the required two datasets :

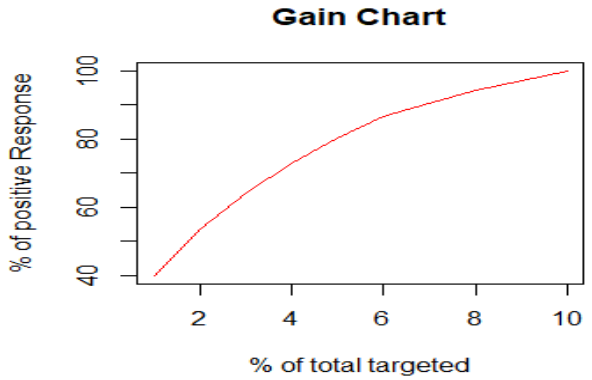
1) Demographic Dataset
2) Demographic and Credit Bureau Dataset (Formed by merging two datasets)

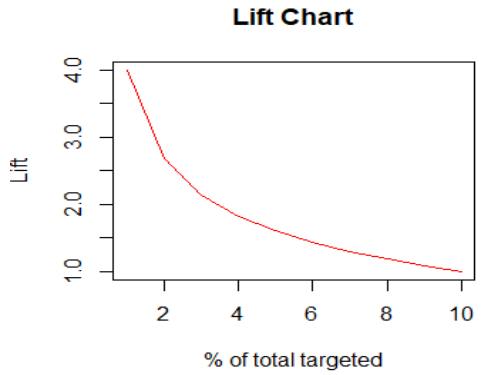
- Dependent variable in both datasets : **Performance Tag**

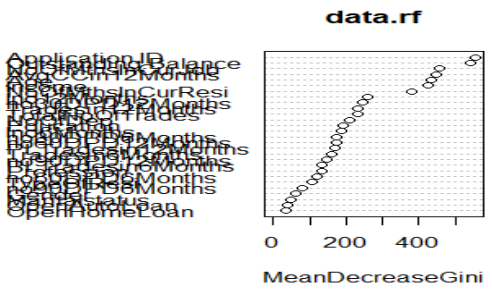
Random Forest Modelling

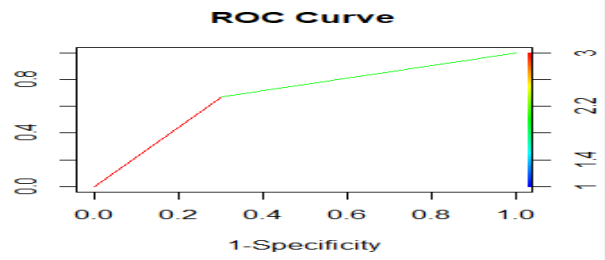
Demographic data Model

Accuracy	Sensitivity	Specificity
58.7%	58.2%	65.8%



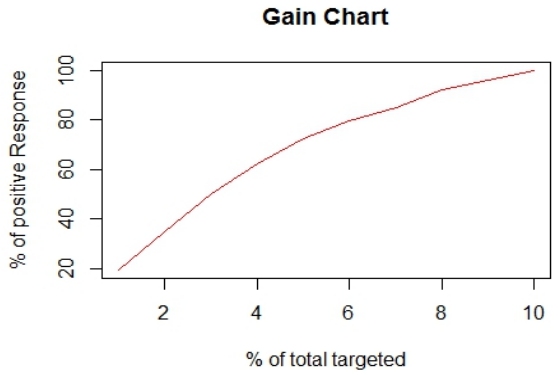


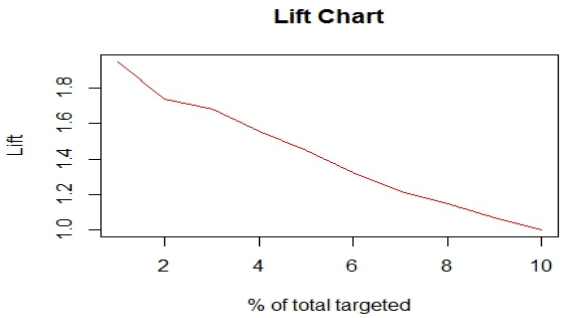


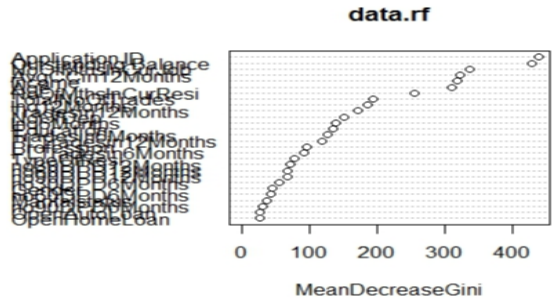


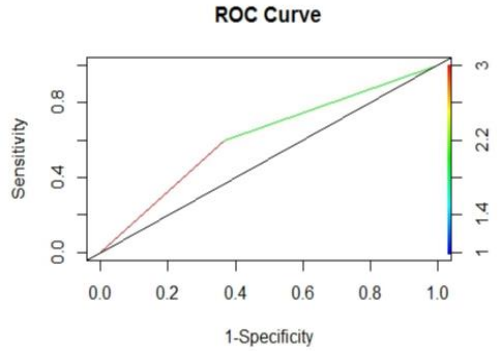
Model using both demographic and credit bureau data

Accuracy	Sensitivity	Specificity
69.3%	69.4%	67.4%









NEURAL NETWORK

- Demographic data Model

Accuracy	Sensitivity	Specificity
67%	53%	68%

- Model using both demographic and credit bureau data

Accuracy	Sensitivity	Specificity
66%	70%	66%

Comparison of Models as per Accuracy, Sensitivity & Specificity

	Demographic Dataset			Demographic & Credit Bureau Dataset		
Model	**Accuracy**	**Sensitivity**	**Specificity**	**Accuracy**	**Sensitivity**	**Specificity**
Simple Logistic Regression Model	58.3%	58.5%	58.2%	67.1%	69%	67%
Weighted Logistic Regression Model	60%	58%	60%	67%	70%	67%
Logistic Regression Model on WOE data	63%	61.6%	63%	68.12%	69.6%	68%
Weighted Logistic Regression model on WOE data	61.2%	62.4%	61.1%	70.13%	68.4%	70.2%
Random Forest model	58.7%	58.2%	65.8%	69.3%	69.4%	67.4%
Neural Network model	67%	53%	68%	66%	70%	66%

Model with Highest Accuracy

Demographic Dataset

- Weighted Logistic Regression model on Demographic WOE data
- Important variables :

Age, NoOfDep, Income, Education, Profession, NoOfMthsInCurResi, NoOfMthsInCurJob

Demographic & Credit Bureau Dataset

- Weighted Logistic Regression model on Demographic & Credit Bureau WOE data
- Important variables :

no30DPD12Months, AvgCCin12Months, PLTradesin6Months, Inq12Months, OpenHomeLoan , NoofDep, Income Profession, MonthsCurrComp

Application Scorecard : Final Model : Final_Model -- Master_WOE_Wt

Cutoffs	Default Percenta	Credi Amount	Potential Customer Loss
No_Cutoffs	6.12	1521436221	0
350+	6.1	1518376821	0
370+	4.96	1224410228	701
390+	3.81	897058300	2423
410+	3.09	656770816	4688
430+	2.22	363925129	9190
450+	1.48	111799758	15161
470+	2.27	8843791	19572

- Clearly as seen in the table, as the Cutoff increases, the Default Percentage and Credit Amount decreases. But at the same time, we are losing some revenue through ignoring Potential Customer.
  Hence, this should be a business call to select which score value as cutoff by taking into account the above stated tradeoffs.
- **However, the optimal cutoff score, as predicted by our model is 417.**
  And the financial benefits offered by following our model are as explained later.
- **Compared Scores of Defaulters with NonDefaulters**
  The mean score of defaulters is **398.49** and the mean score of non defaulters is **427.37.**
  As expected, the mean score of non defaulters is higher than that of defaulters, which further validates the reliability of our model.

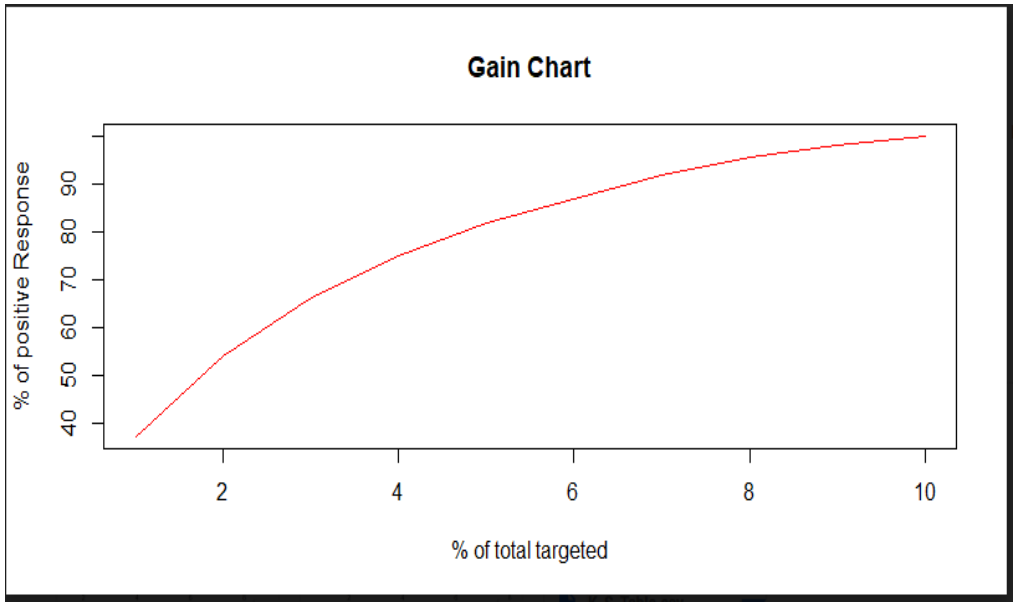
Analyzing the Financial Benefits of using our deployed model- Weighted Master WOE Model

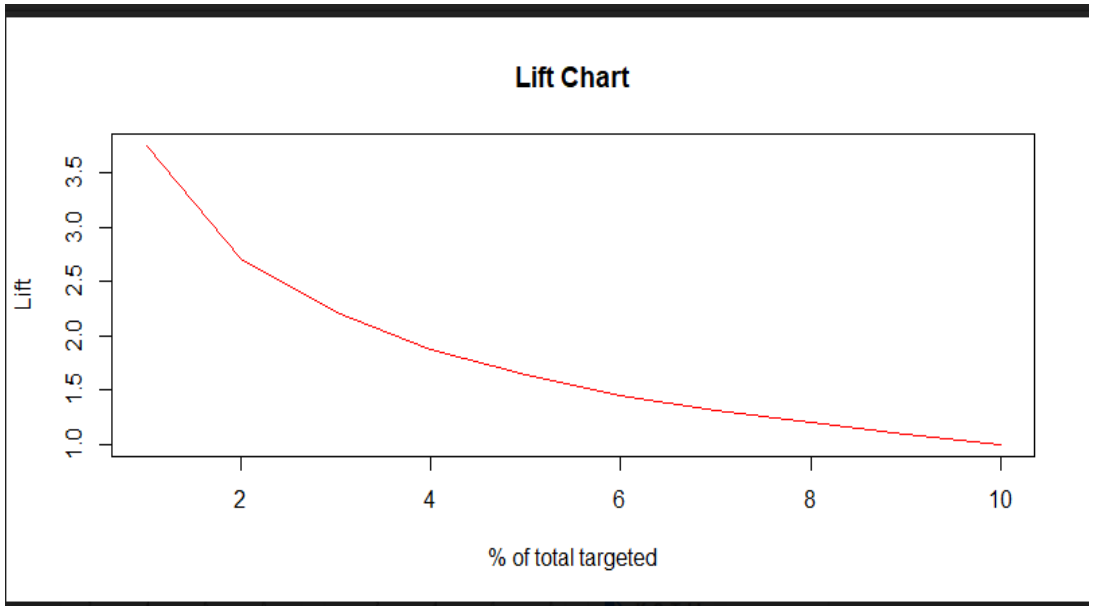
- Using our deployed model, we will reduce our credit_amount by 96,48,29,982
- Business Loss in terms of customers - Though we managed to save a reasonable amount of credit loss, but in return we made a trade off in terms of potential customers, We lost around 30% of the potential customers, which surely affects a little bit on over final collected revenue.
- Using our deployed model, we can reduce the default_percentage by 3.29%

Credit Risk with model vs without model

- Without model, there is 6% credit risk.
- With model, the credit risk is reduced to 3%, i.e. half of that without model
- Thus, using our deployed model, we brought the credit loss down by 3%

Gain chart, Lift Chart for Weighted Demographic & Credit Bureau WOE





We achieved a lift of nearly 2% by the 4th decile

KS table for Weighted Demographic & Credit Bureau WOE

- ***We were successfully able to identify 75% of the defaulters by the 4th decile itself***

bucket	total	totalresp	Cumresp	Gain	Cumlift
1	2099	481	481	37.46105919	3.746105919
2	2098	214	695	54.12772586	2.706386293
3	2099	158	853	66.43302181	2.21443406
4	2098	110	963	75	1.875
5	2099	90	1053	82.00934579	1.640186916
6	2098	65	1118	87.07165109	1.451194185
7	2099	63	1181	91.97819315	1.313974188
8	2098	46	1227	95.56074766	1.194509346
9	2099	35	1262	98.28660436	1.092073382
10	2098	22	1284	100	1

Thank you…!