BFSI Capstone Project

MID-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

1. 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 home & auto loans)	-	-
Presence of open home loan	272 rows doesn't have any value	-
	272 rows doesn't have any value	-

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 for Age Variable

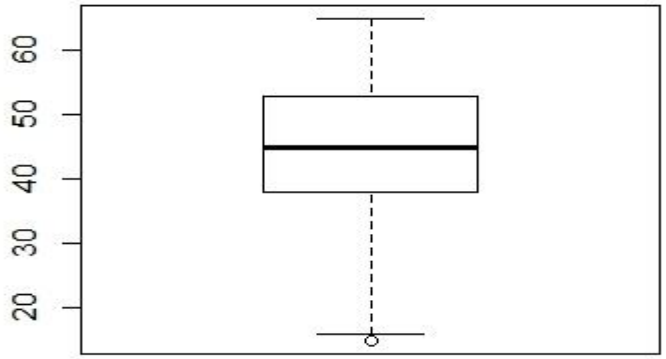


Fig1: Box plot showing distribution of customers based on Age

The Age range is between15 to 65 years. There are around 20 incorrect values which is seen as the outlier here.

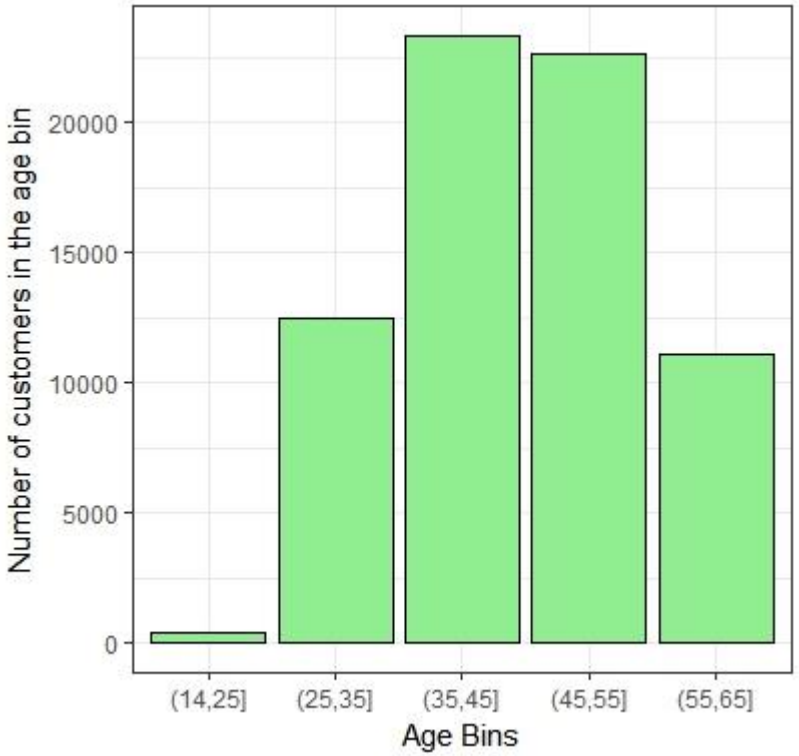


Fig2: Distribution of customer based on bins of age.

Since age is a continuous variable, we have binned the attribute into 5 buckets.

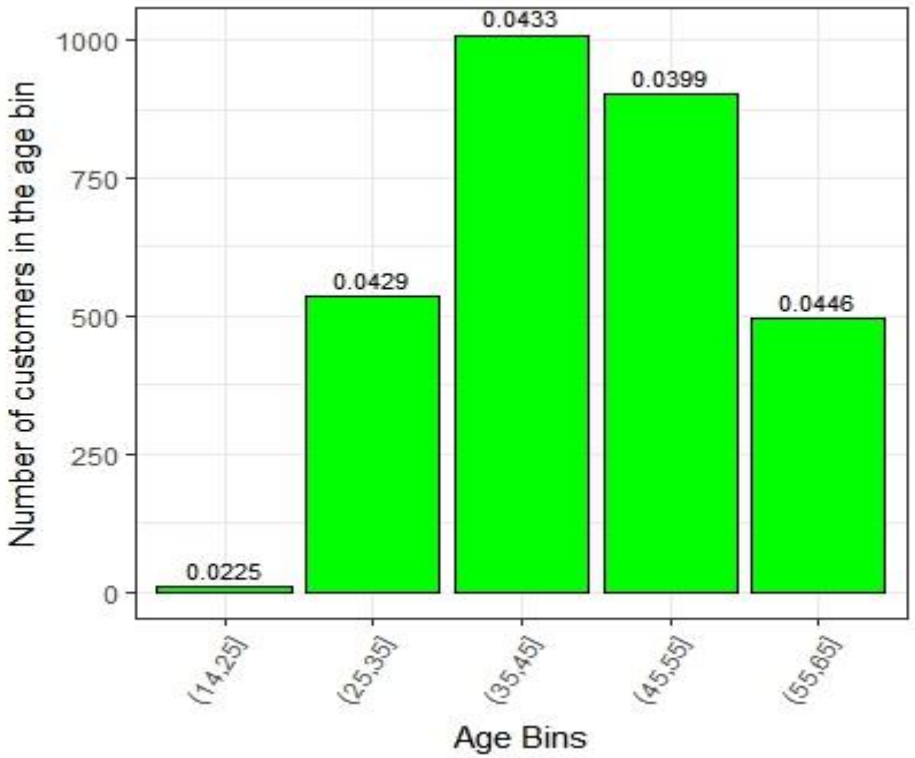


Fig3: It shows for each age bin what is the count of customers who have defaulted.

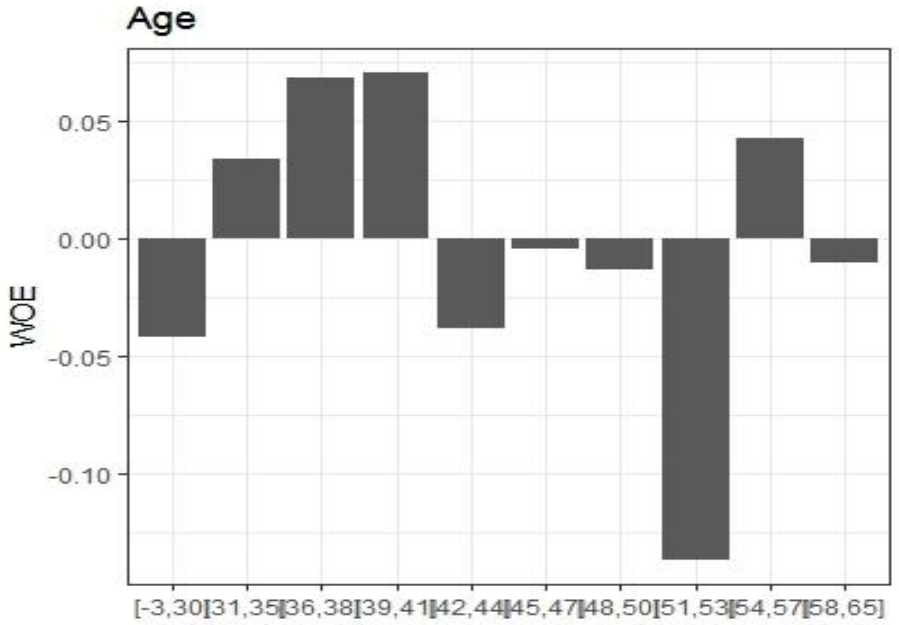


Fig4: Shows the WOE Pattern based on the age bins.

Its seen from the plot that the WOE for the range of age from 36-38 and 39-41 has high WOE value which indicates that it can be significant in predicting the behavior of the customers.

Exploring Gender Variable:

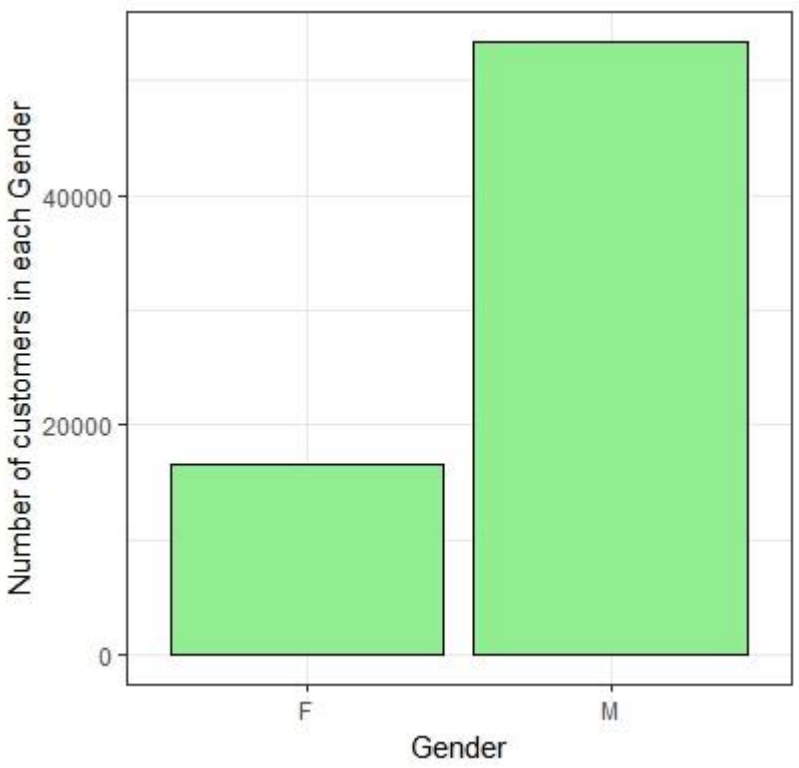


Fig5: Distribution of customer based on Gender

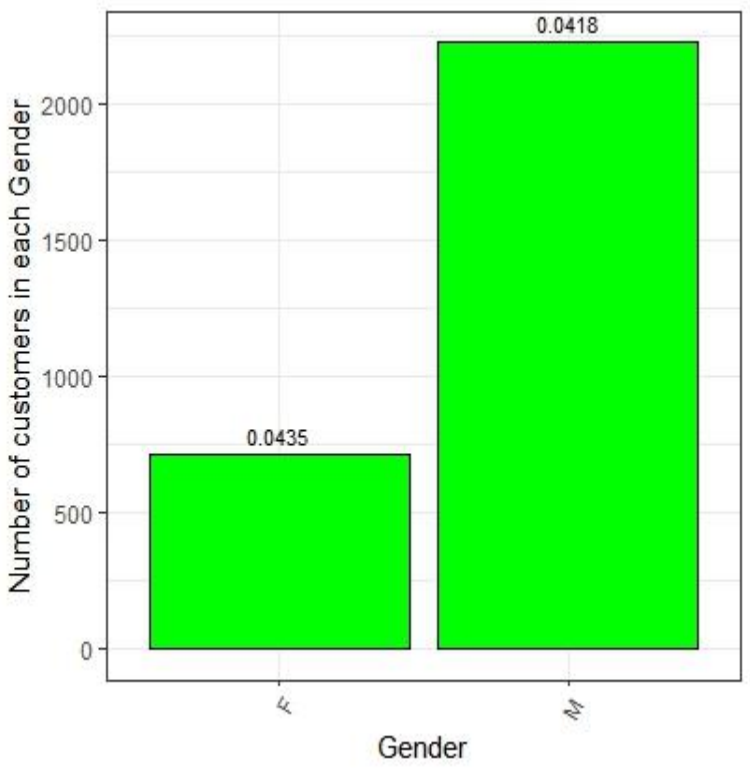


Fig6: Count of customer who defaulted based on gender

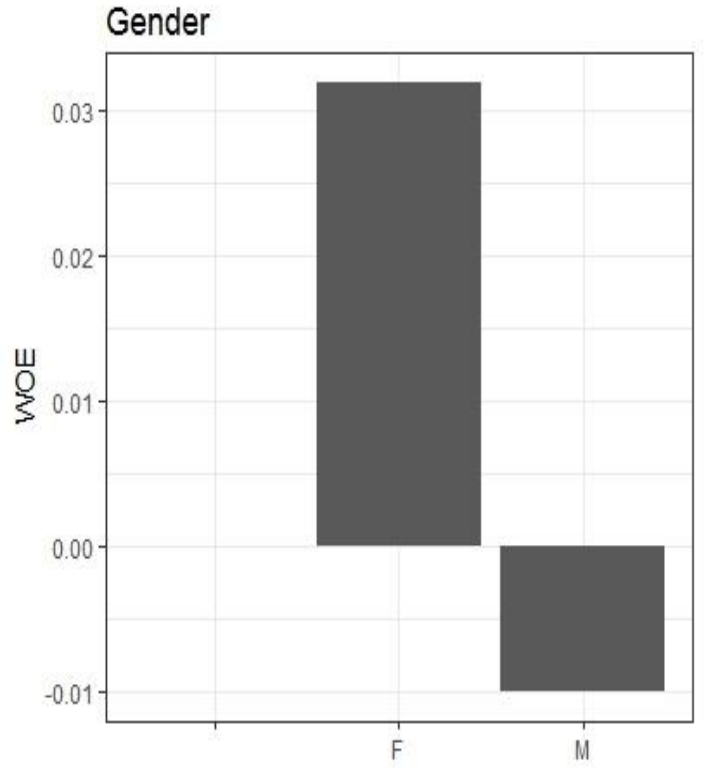


Fig7: WOE pattern based on Gender

Exploring Marital Status Variable:

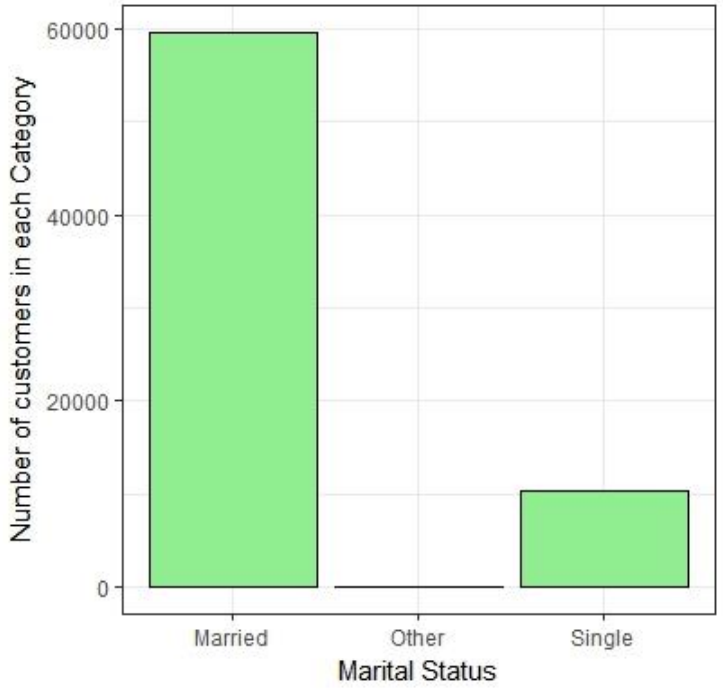


Fig8: Distribution of customer based on Marital Status

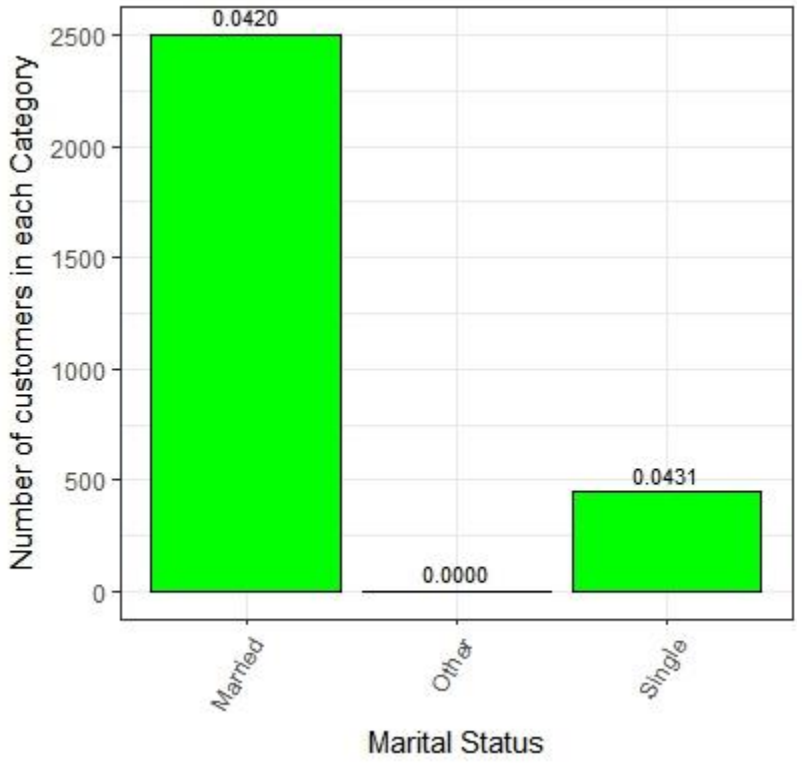


Fig9: Count of customer who defaulted based on marital status

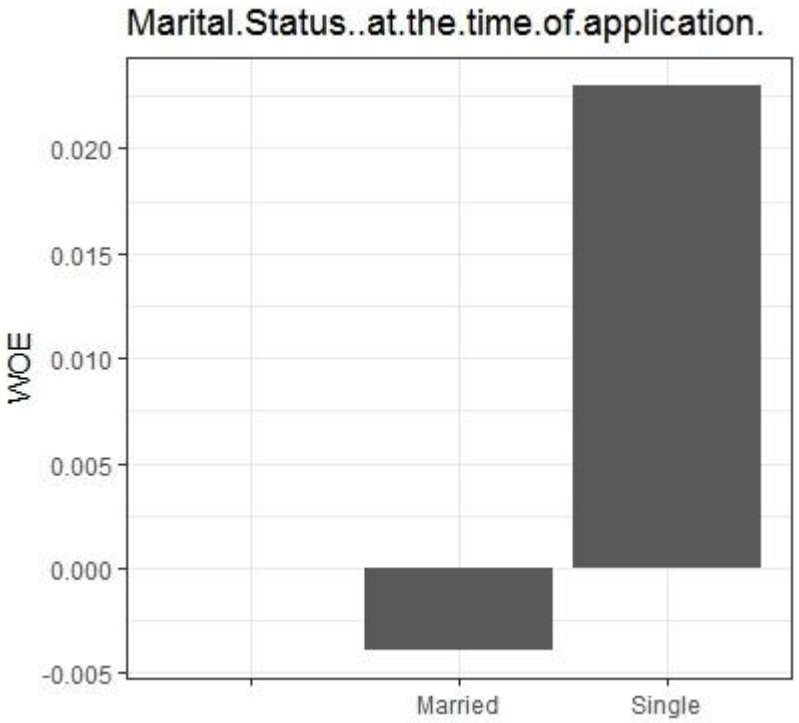


Fig10: WOE Pattern for Gender Variable

Exploring Number of Dependents Variable

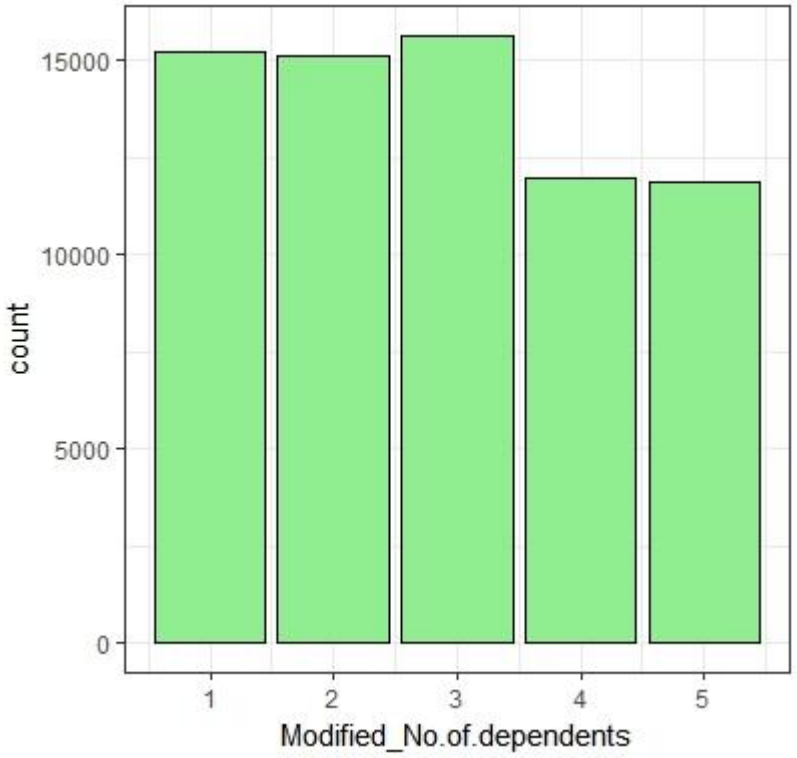


Fig11: Distribution of customers based on number of dependents

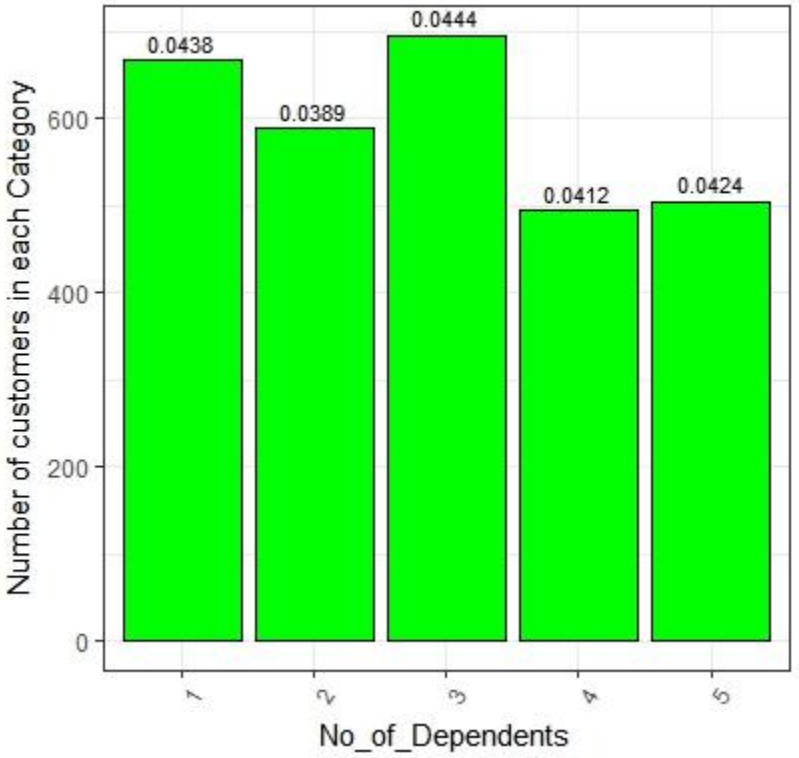


Fig12: Response plot based on number of dependents

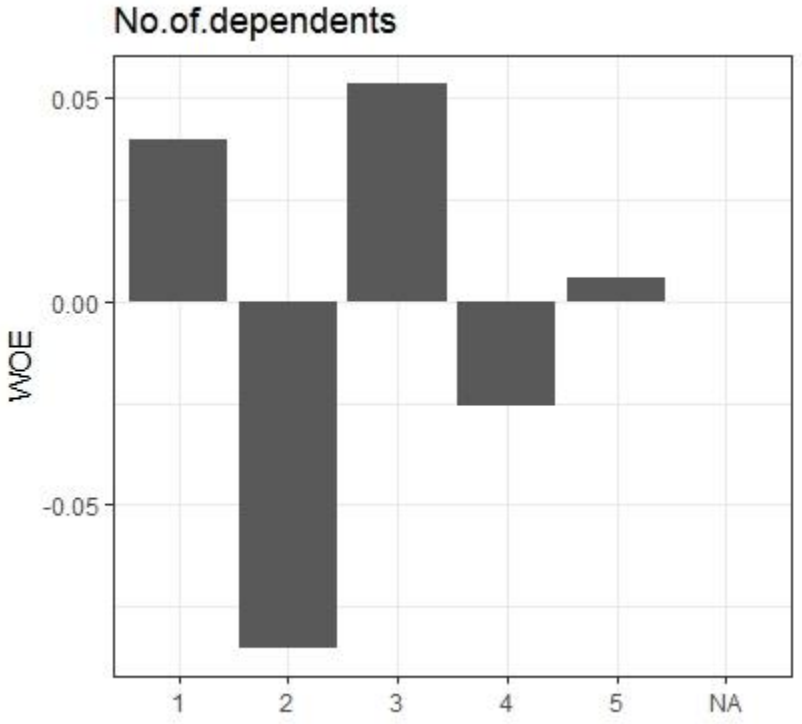


Fig13: WOE Pattern based on number of dependents

Exploring Income Variable

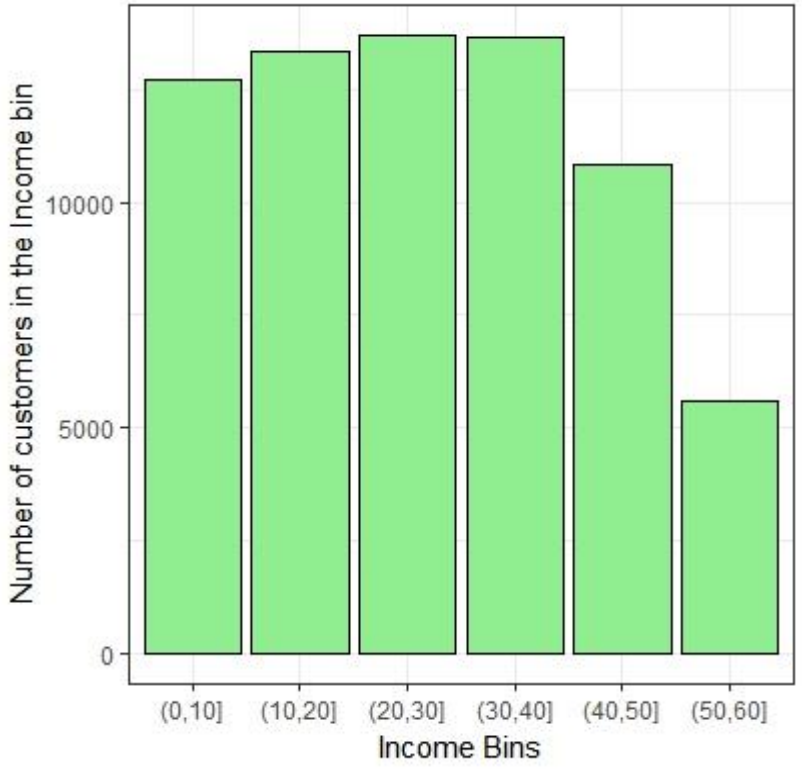


Fig14: Customer distribution based on bins of income

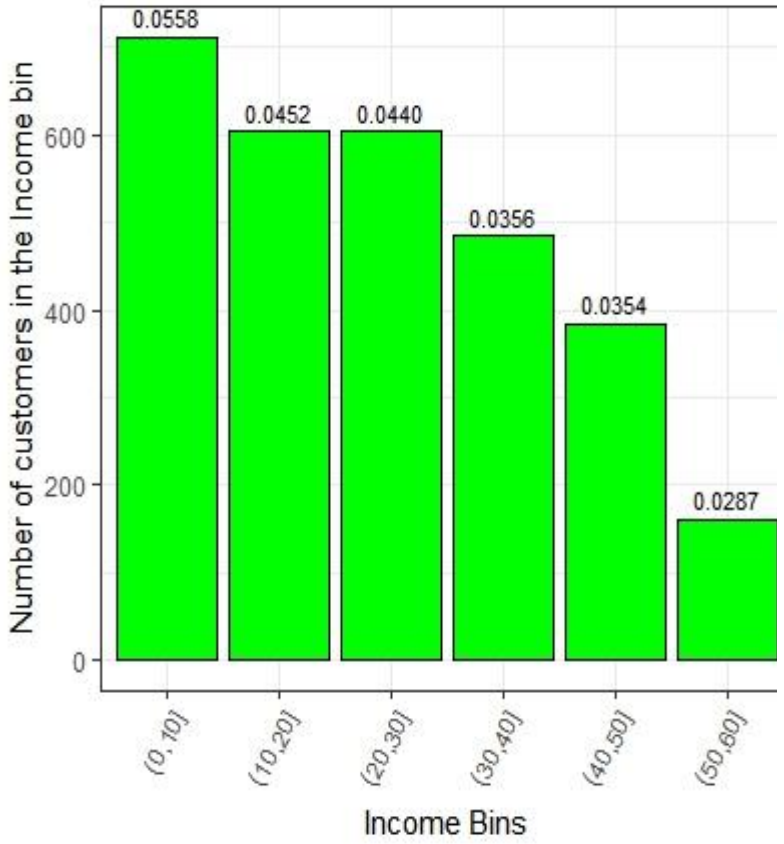


Fig15: Response plot based on income bins

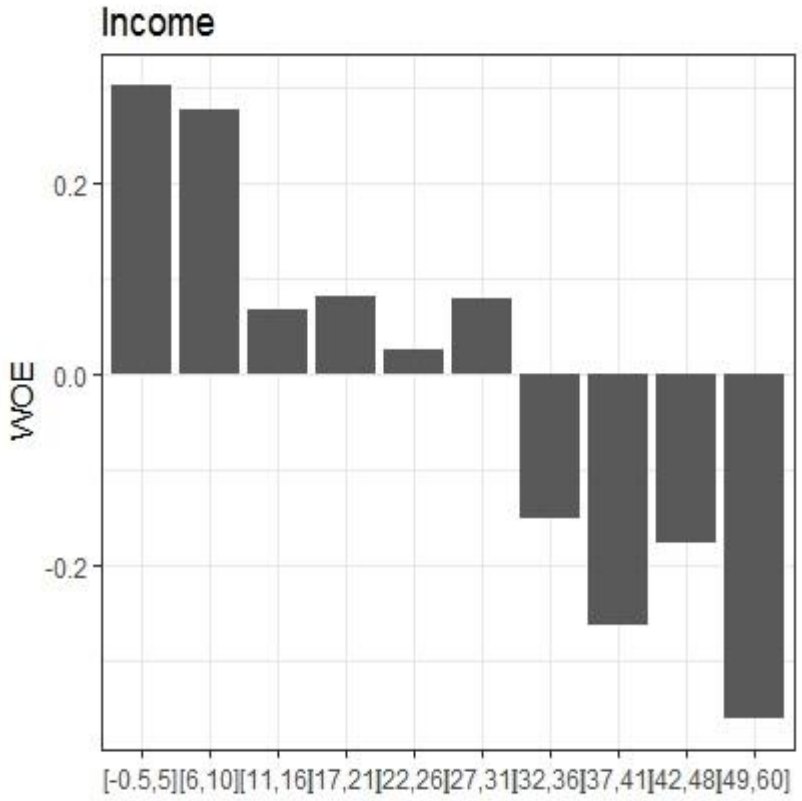


Fig16: WOE Plot based on income bins

Exploring Education Variable

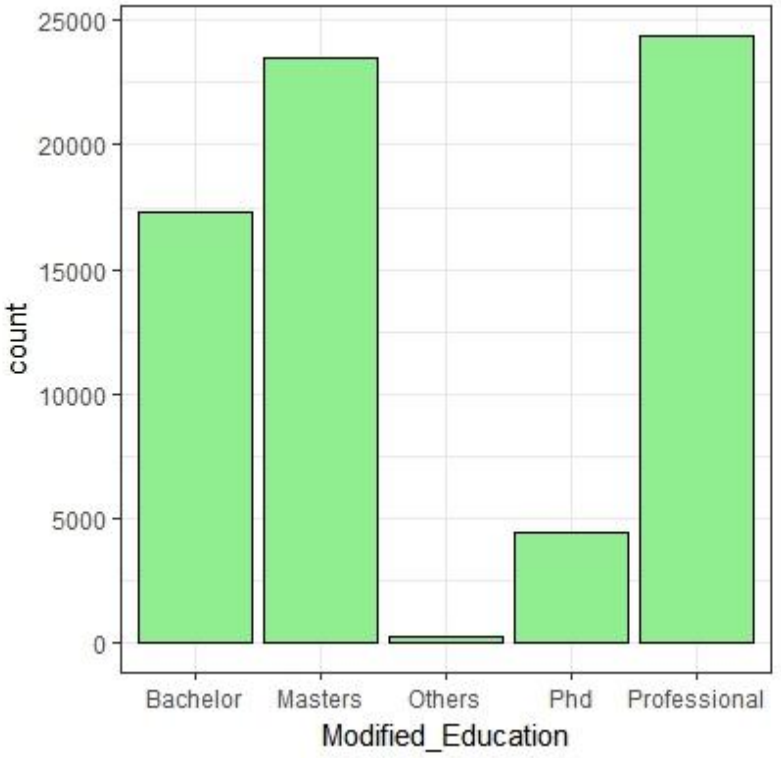


Fig17: Distribution of customer based on Education

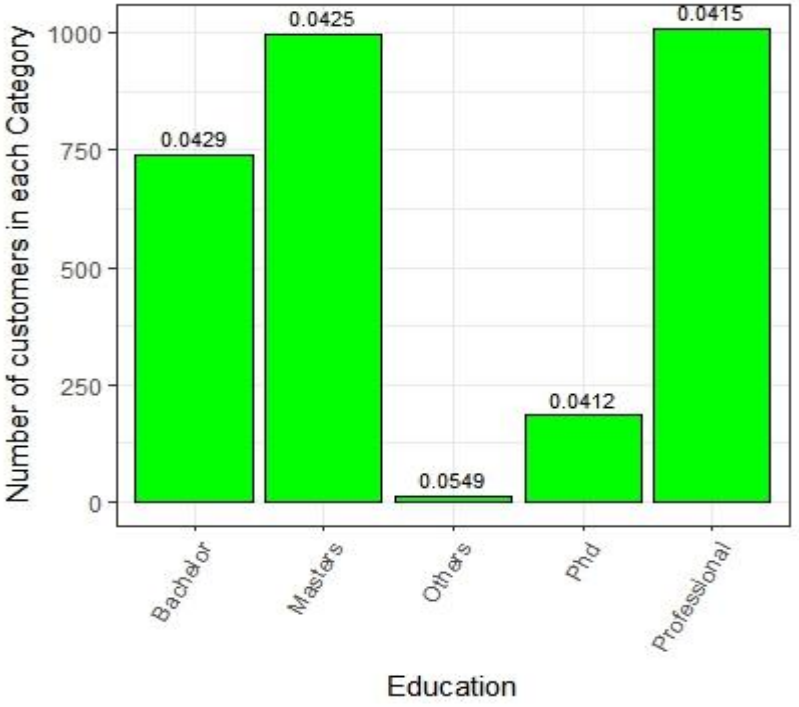


Fig18: Count of customers who defaulted based on Education.
Education qualification, Others have a significant effect on default behavior which is evident from the WOE plot as well.

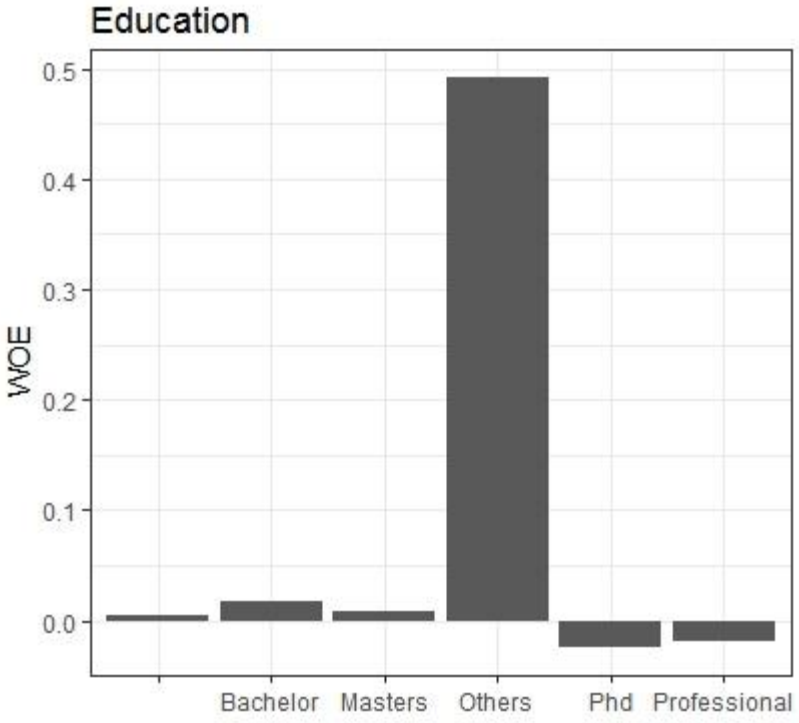


Fig19: WOE pattern for Education

Exploring Profession Variable

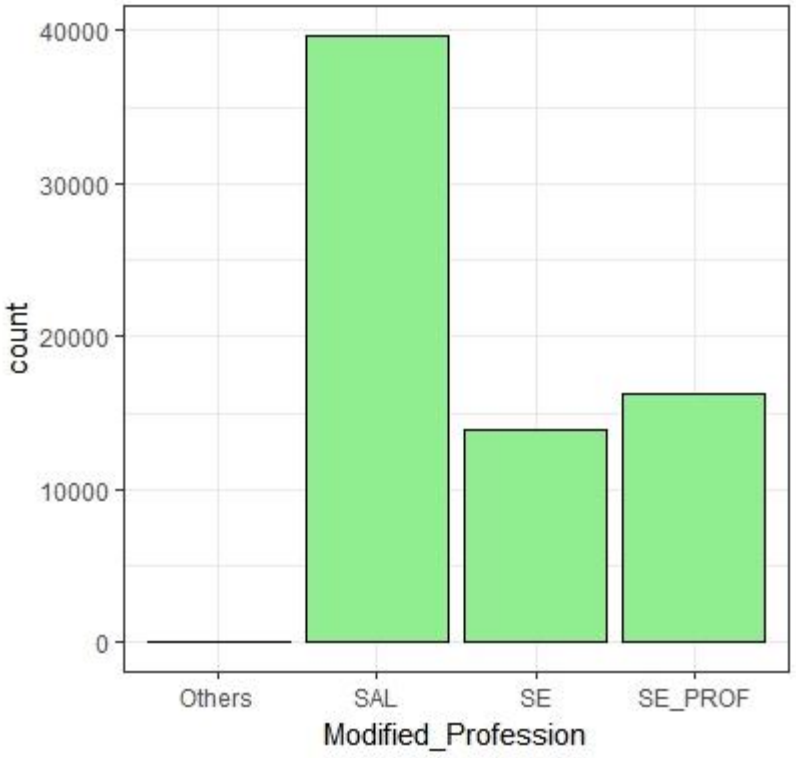


Fig20: Distribution of customer based on profession

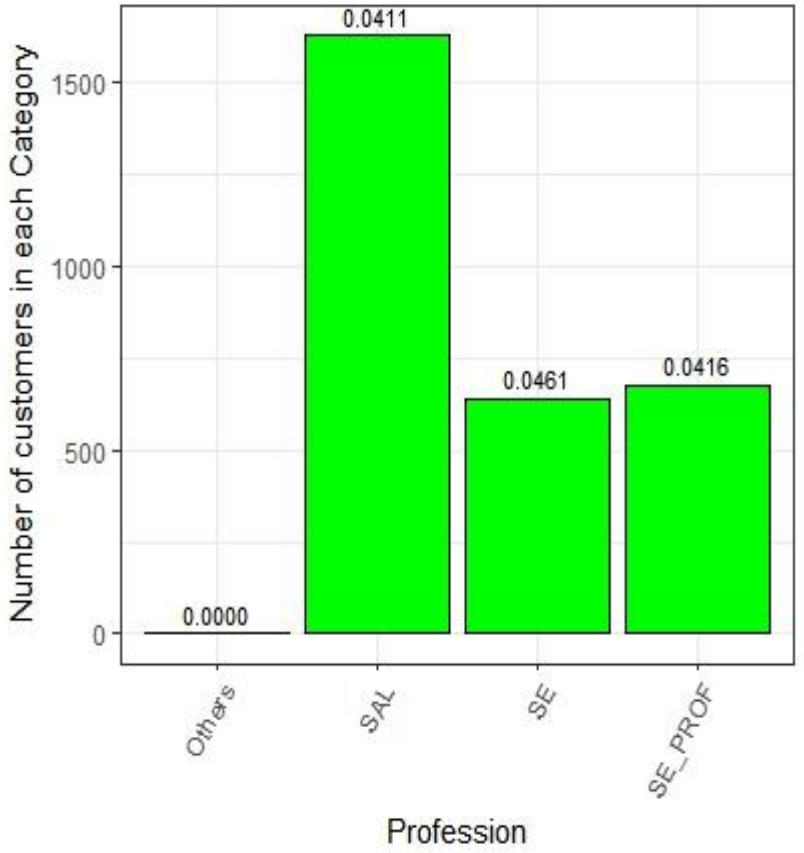


FIg21: Distribution of defaulted customer based on profession

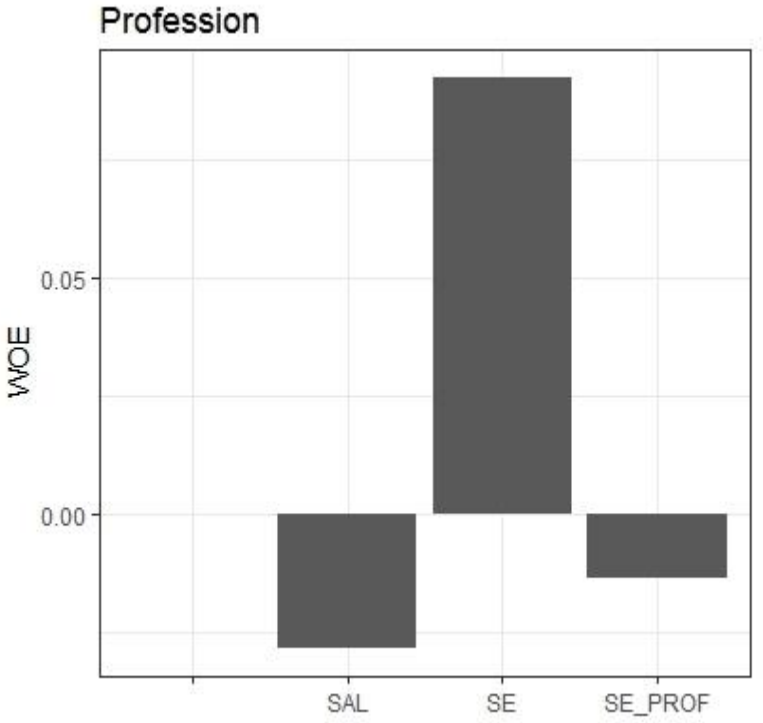


Fig22: WOE pattern for Profession variable

Exploring Type of Residence Variable

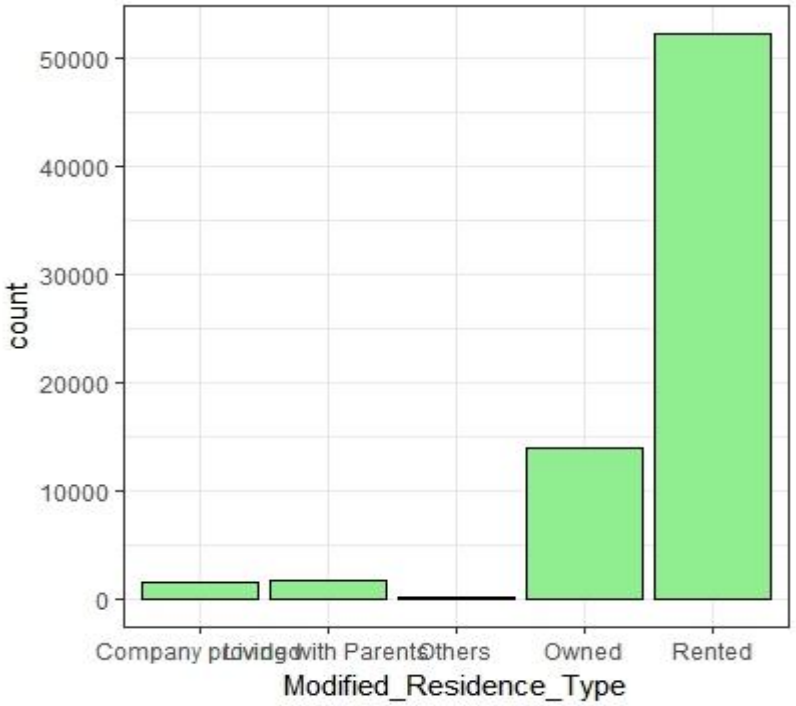


Fig23: Count of customer based on residence type

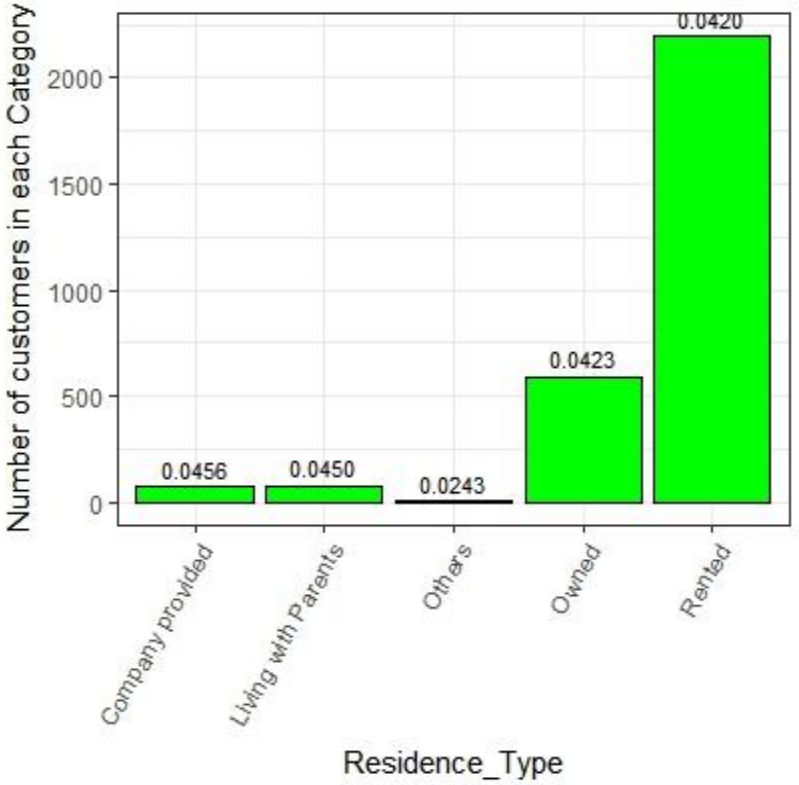


Fig24: Response plot based on residence type

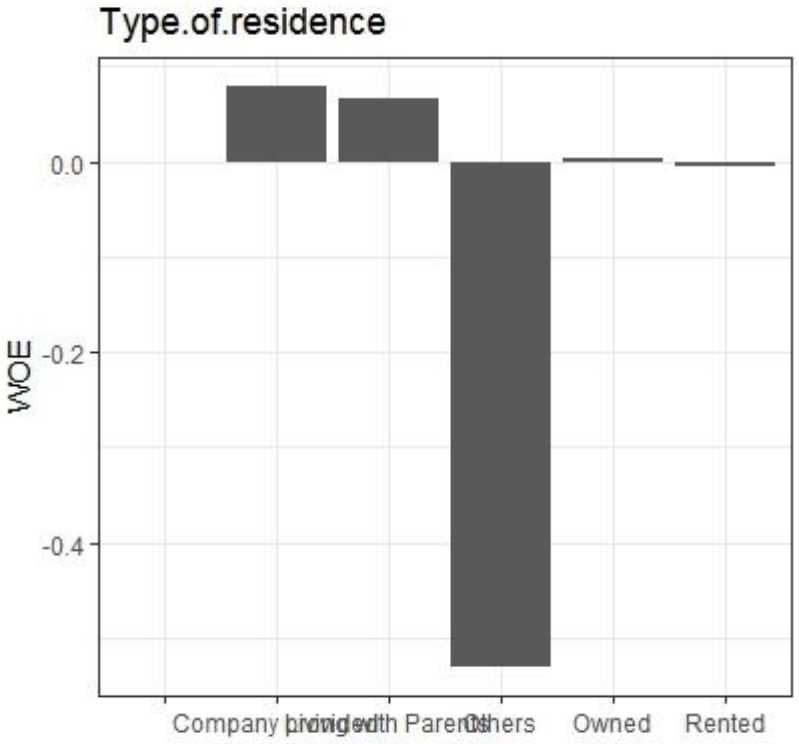


Fig25: WOE Pattern based on residence type. This WOE pattern doesn't show much significant influence in customer default behavior.

Exploring Number of months in current residence

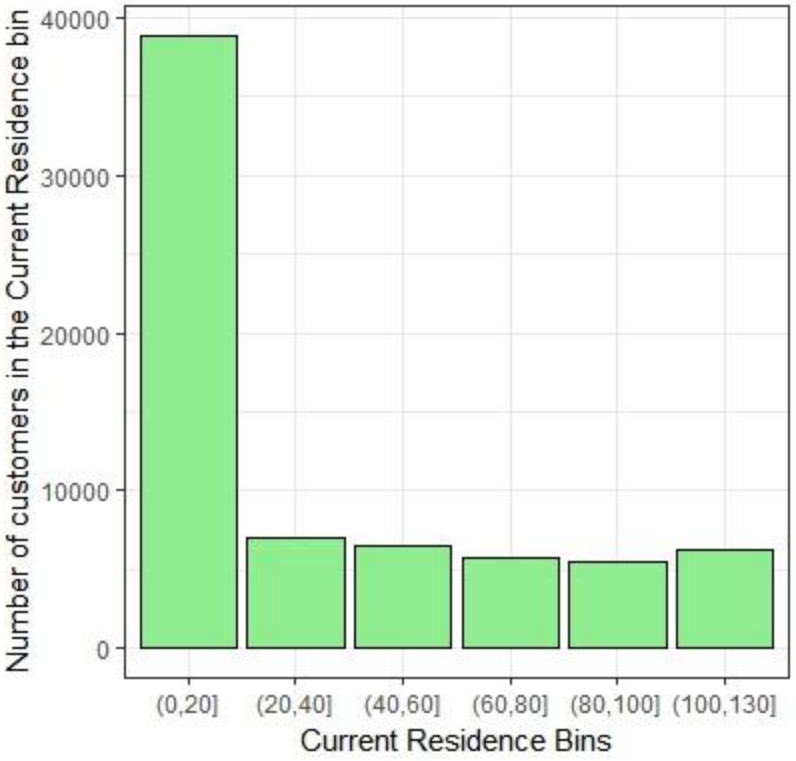


Fig26: Count of customer based on bins for months in current residence

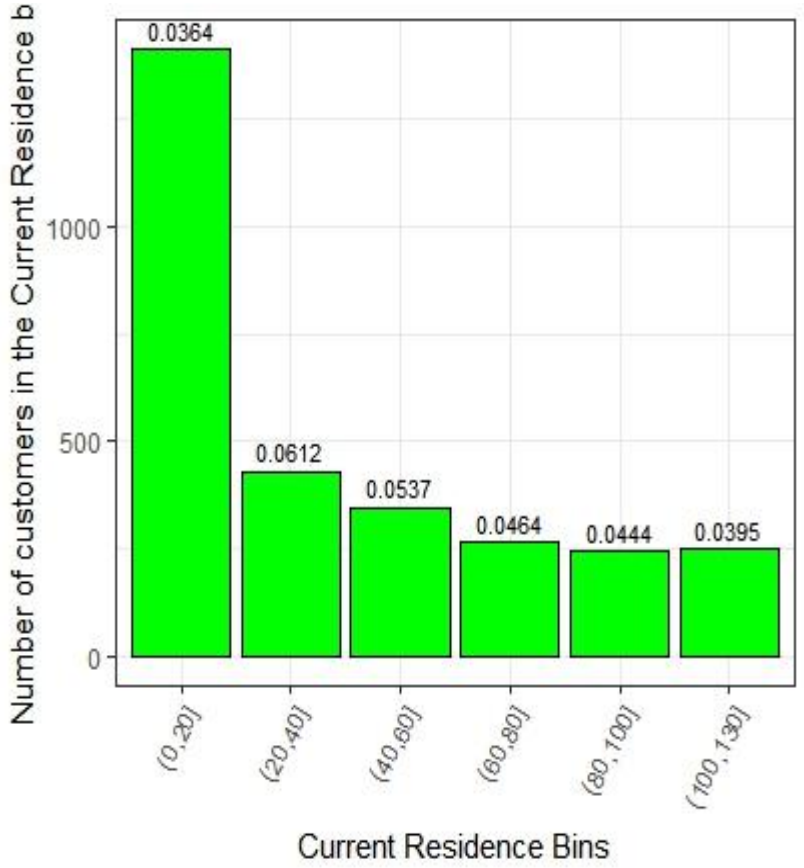


Fig27: Count of customer who defaulted based on bins for months in current residence

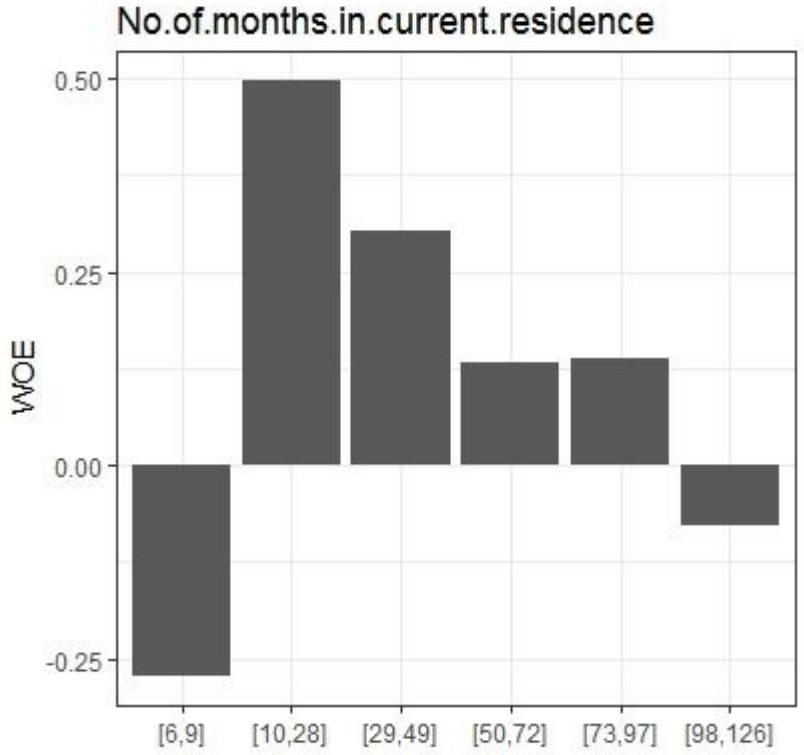


Fig28: WOE Plot for bins based on current residence

Exploring Number of months in current Company

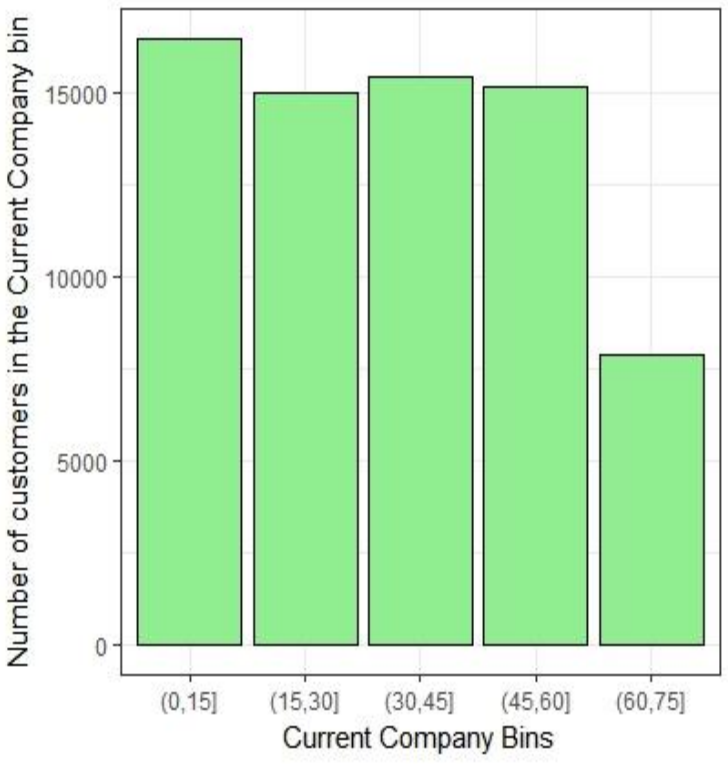


Fig29: Count of customer based on bins for months in current company

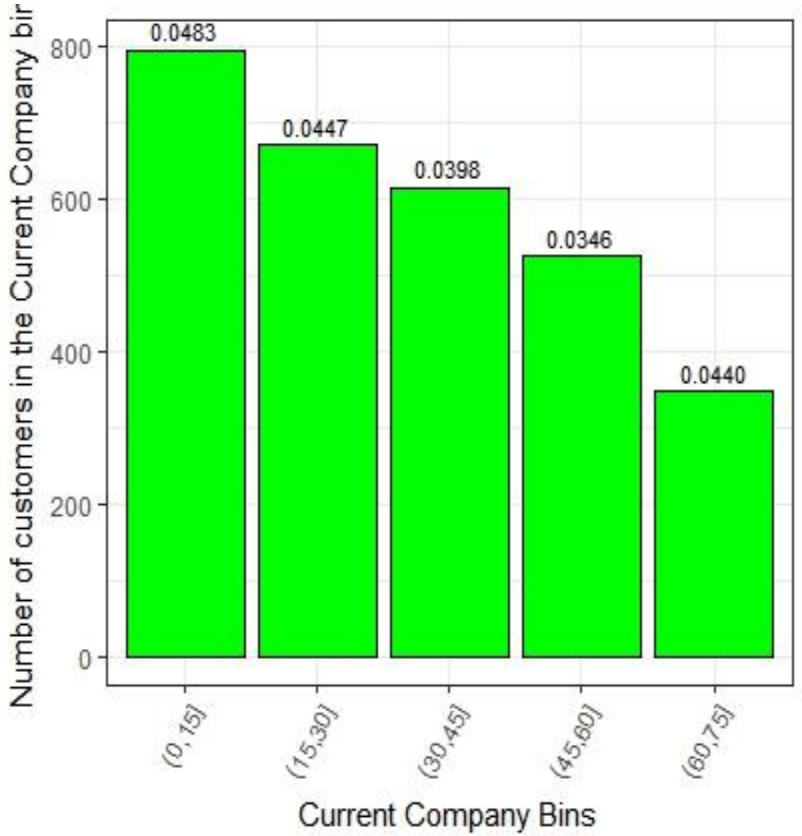


Fig30: Count of customer who defaulted based on bins for months in current company

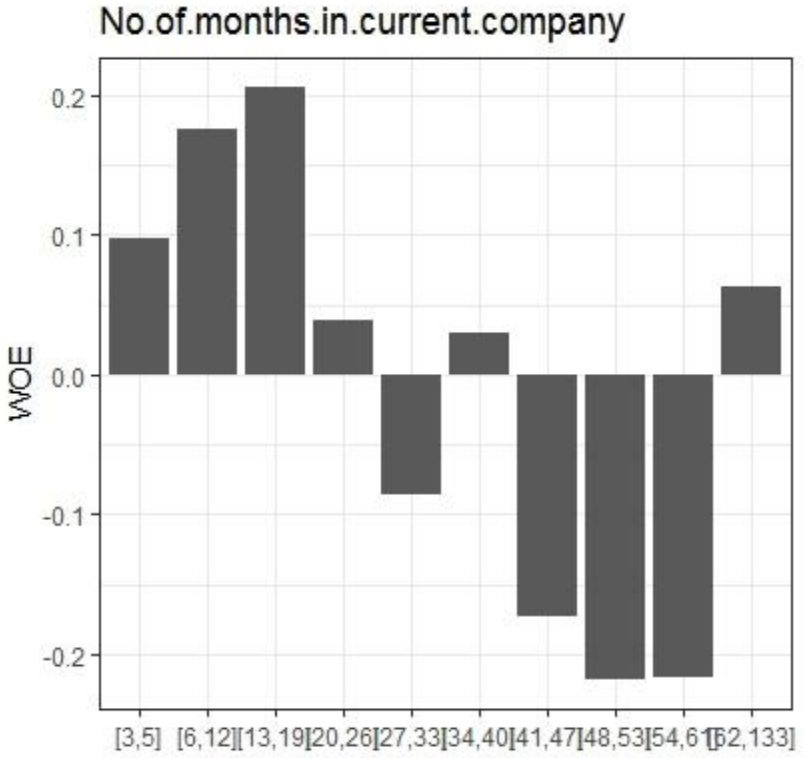


Fig31: WOE Pattern for months in current company

Exploring No. of times 90 DPD or worse in last 6 months

- No of times 90DPD in last 6 months indicates , no of times the specific customer has missed payment for more than 90 days in last 6 months.

- The WOE Chart shows when the count of 90DPD in last 12 months is 1, the chances of the customer being a defaulter is high.

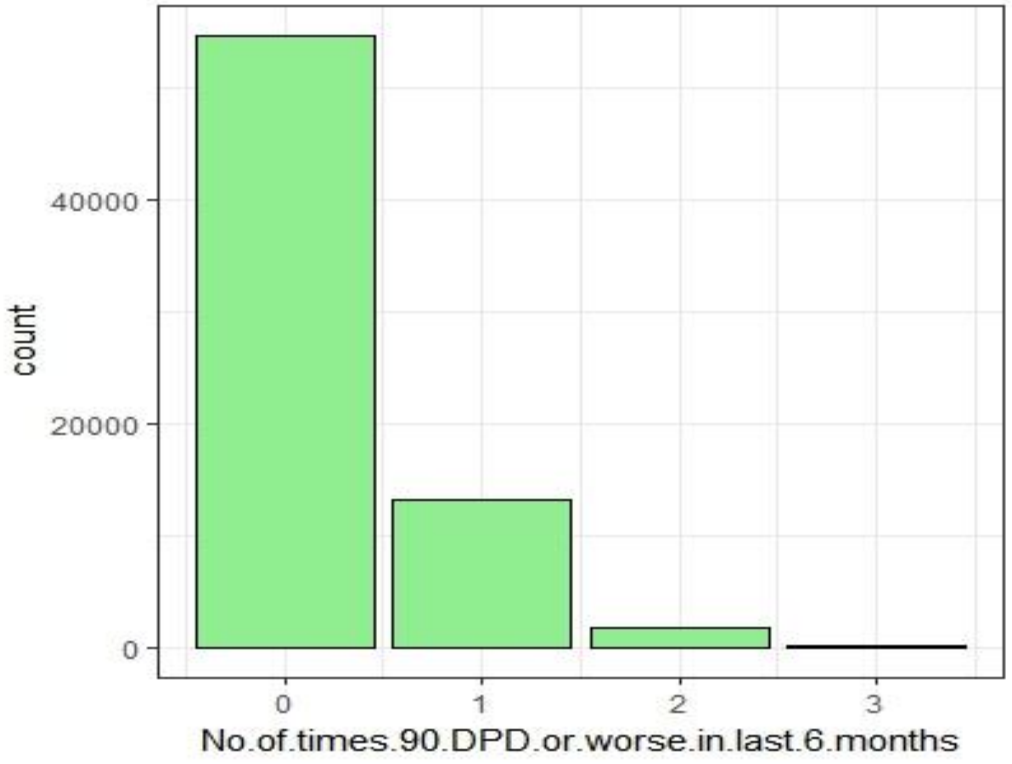


Fig32: Count of customer based on number of times 90 DPD or worse in last 6 months

Exploring No. of times 90 DPD or worse in last 6 months

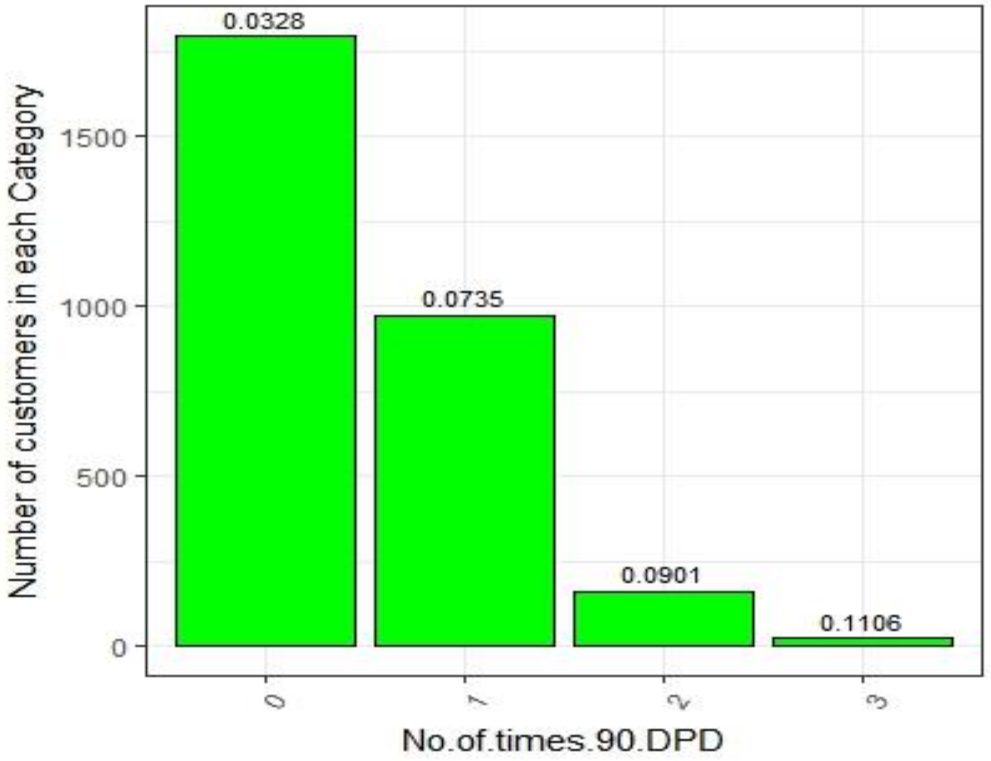


Fig33: Count of defaulted customers based on number of times 90 DPD or worse in last 6 months

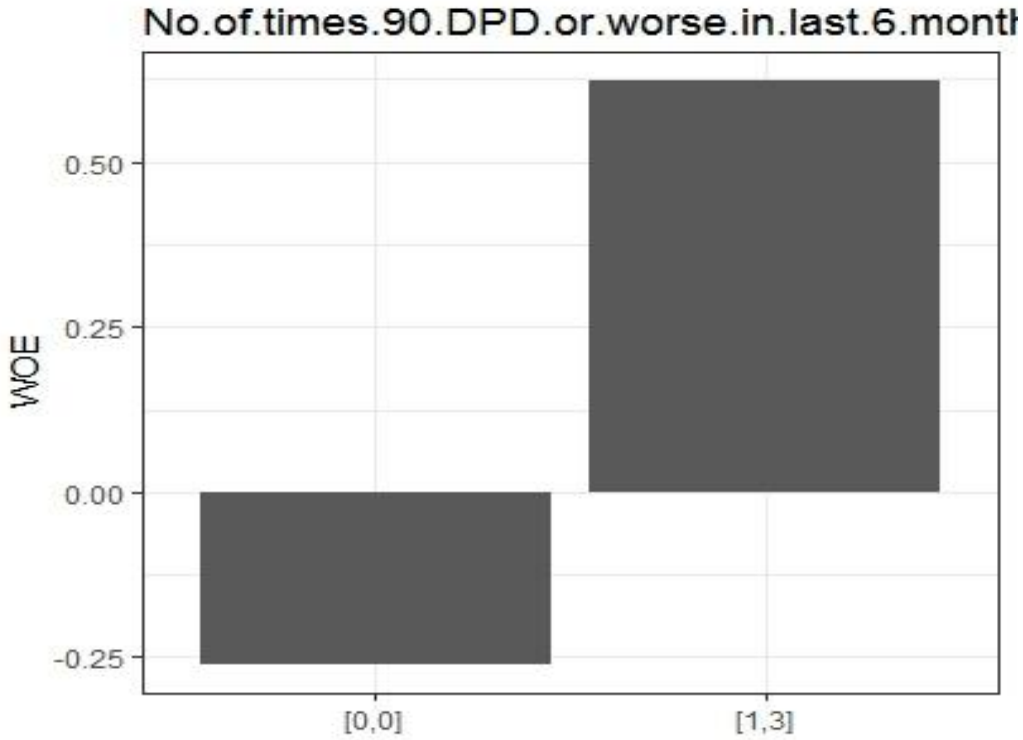
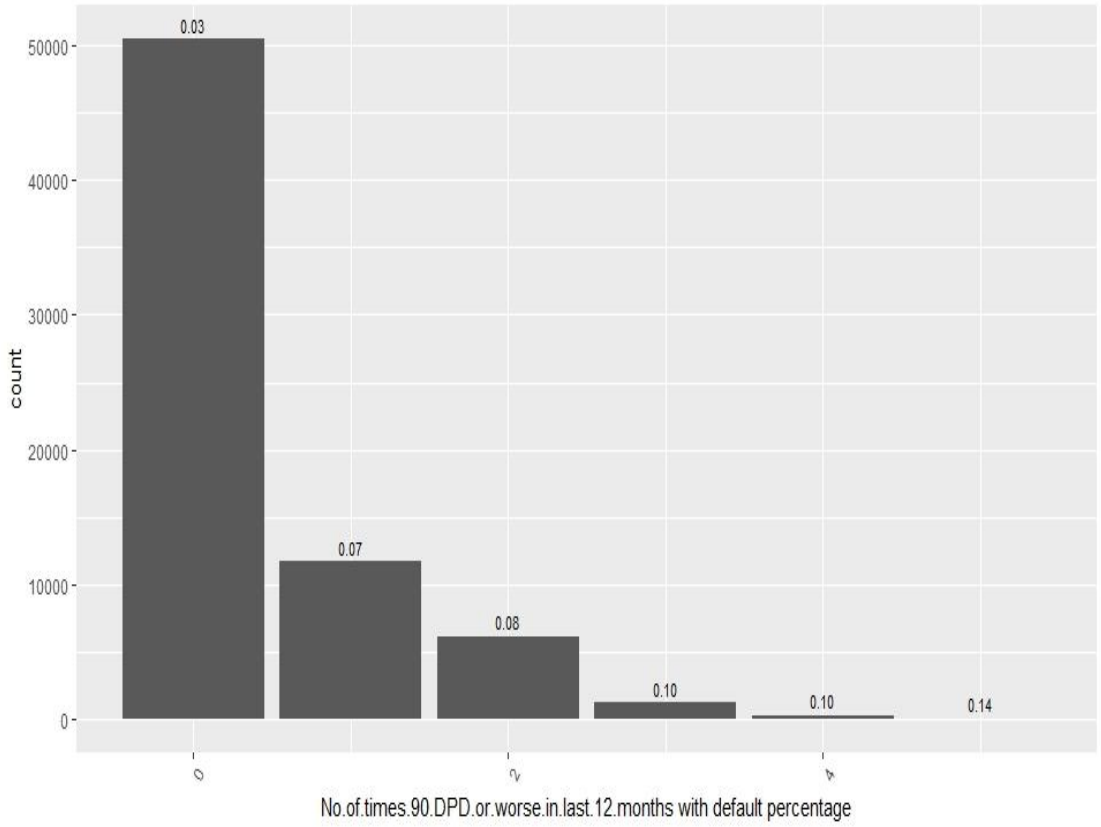
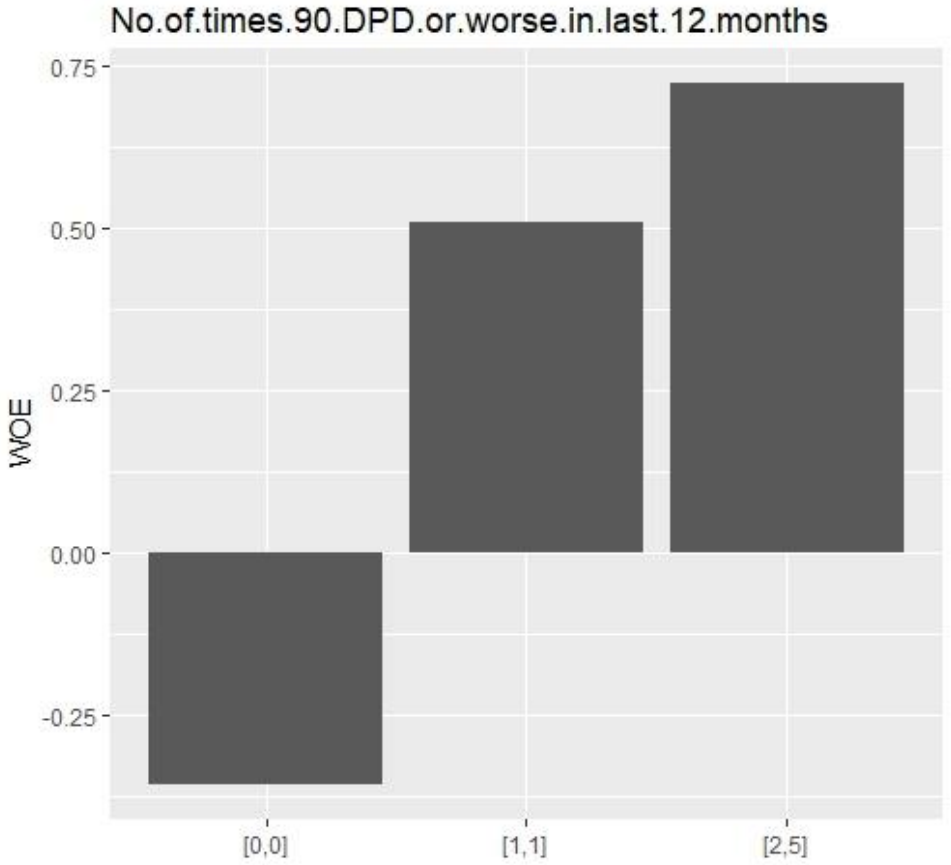


Fig34: WOE Pattern for defaulted customers

Exploring No. of times 90 DPD or worse in last 12 months





Exploring No. of times 60 DPD or worse in last 6 months

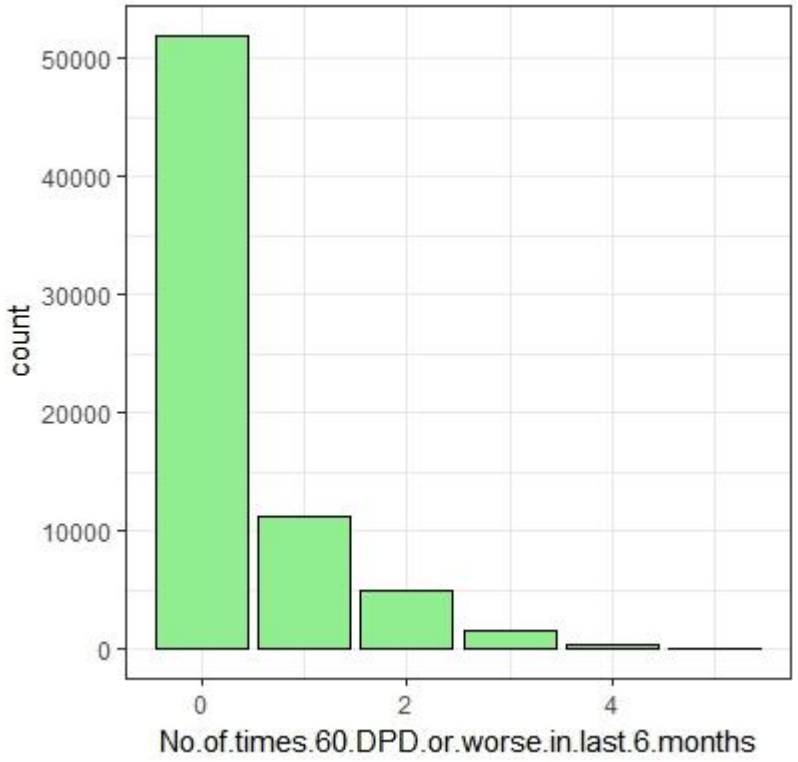


Fig35: Count of customer based on number of times 60 DPD or worse in last 6 months

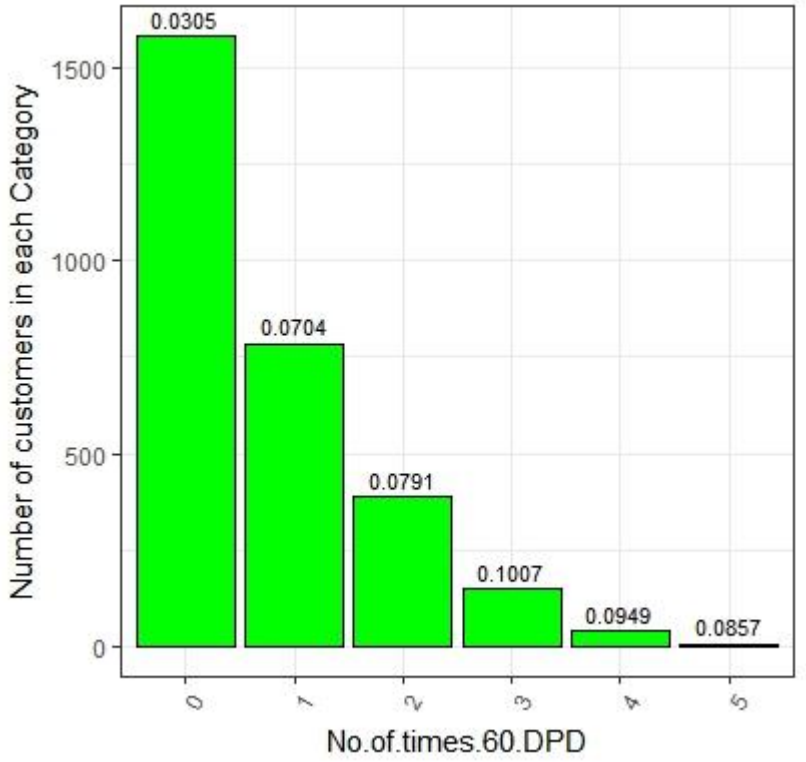


Fig36: Count of customer who defaulted based on number of times 60 DPD or worse in last 6 months

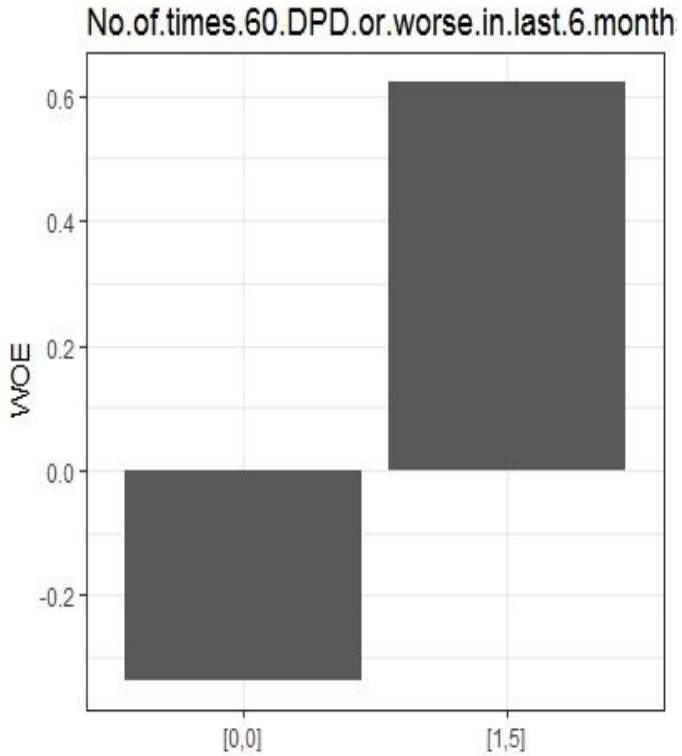


Fig37:WOE Pattern for defaulted customers

Exploring No. of times 30 DPD or worse in last 6 months

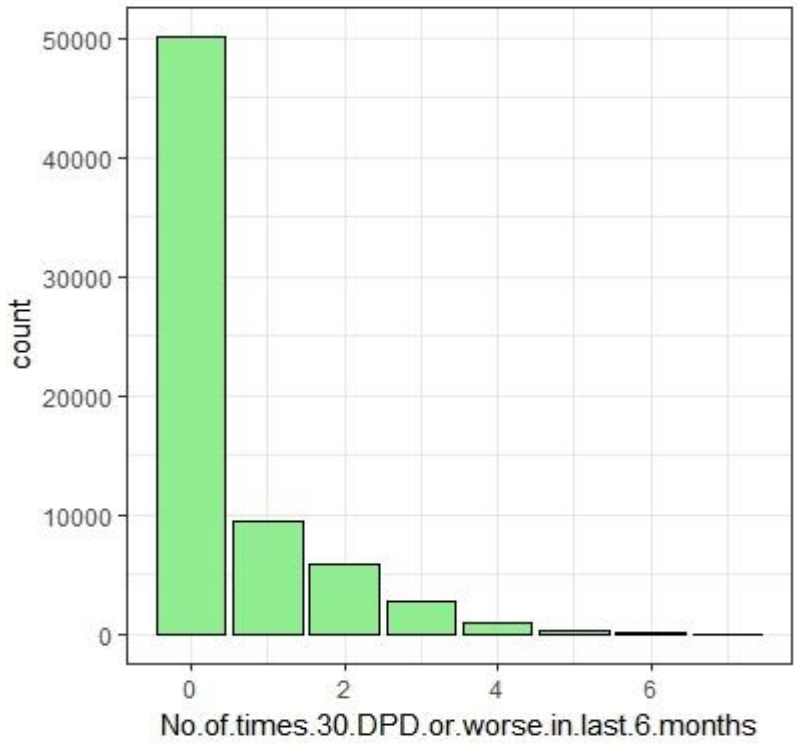


Fig38: Count of customer based on number of times 30 DPD or worse in last 6 months

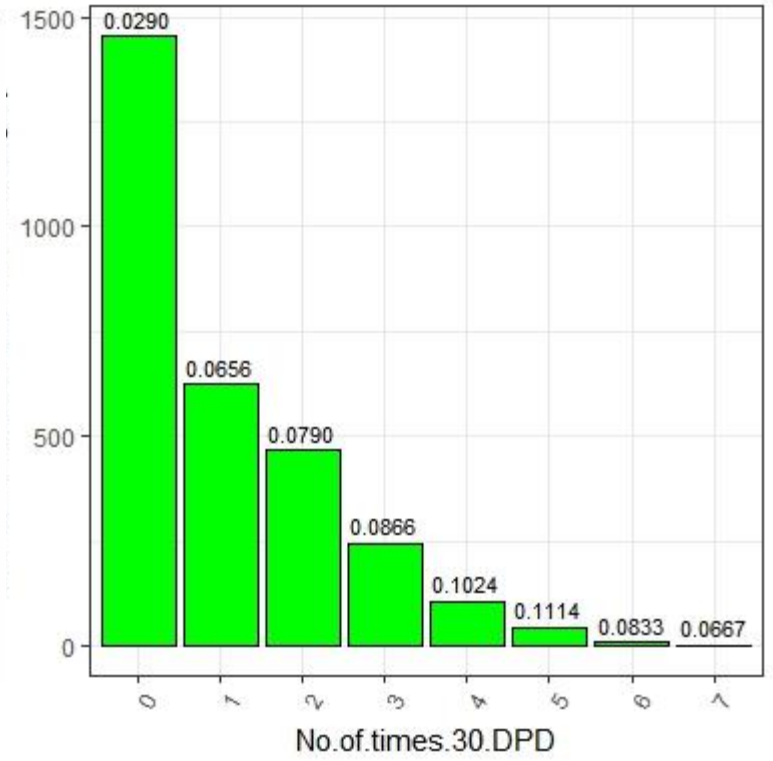


Fig39: Count of customer who defaulted based on number of times 30 DPD or worse in last 6 months

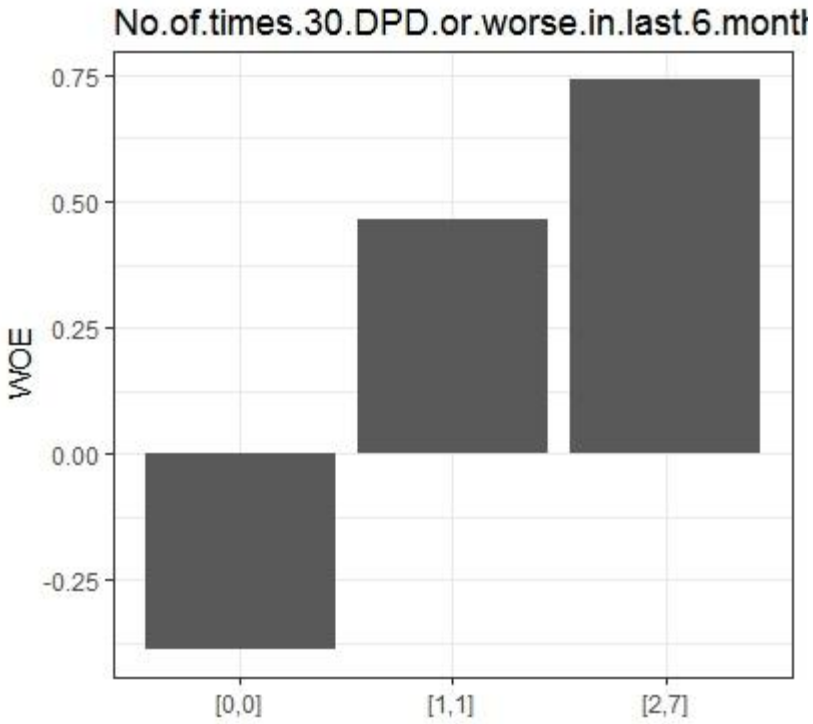
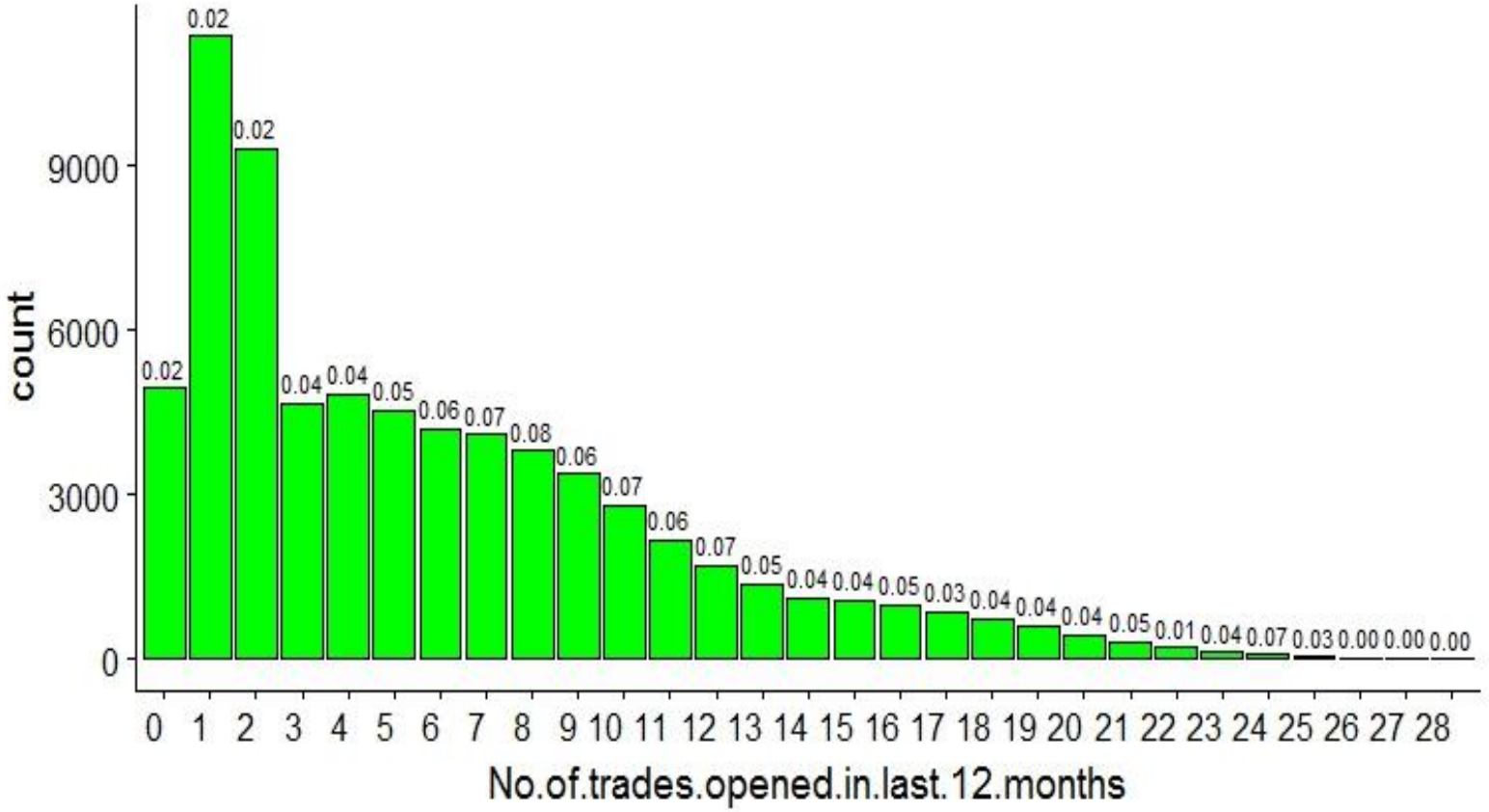


Fig40: Count of customer based on number of times 30 DPD or worse in last 6 months

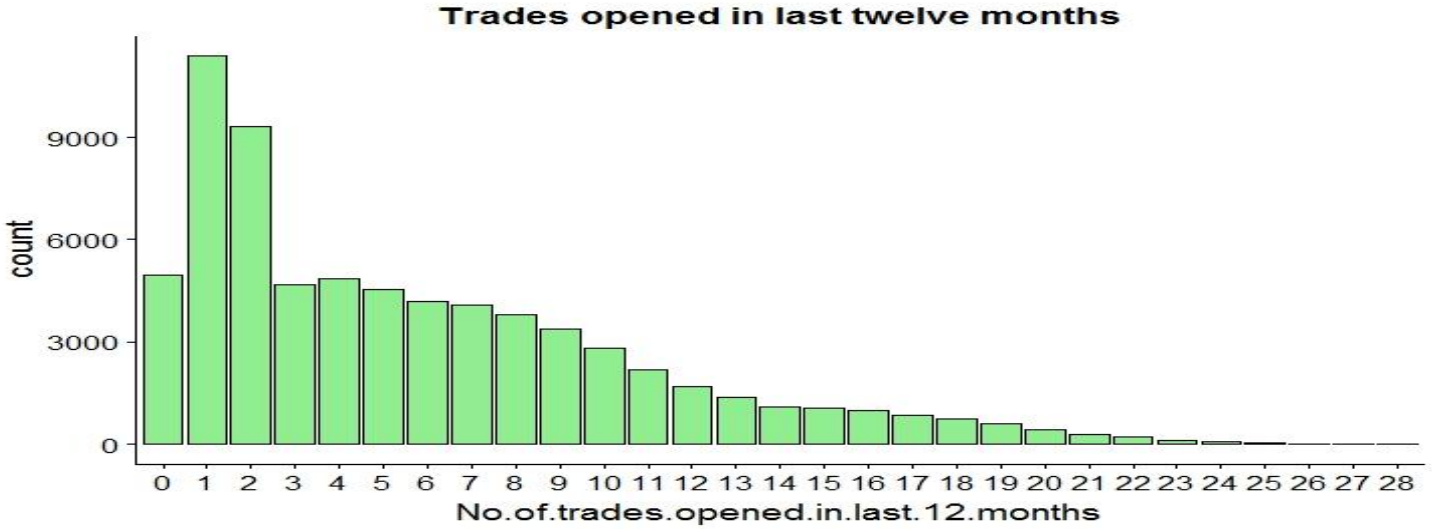
Trade open in last twelve months

It is observed that as the number of trades opened in last 12 months increases , the percentage of defaulters also increase, which indicates that this particular attribute is a contributing factor in deciding if the customer is going to default or not.

Next slides present the count of the customer who have open trade accounts and the corresponding WOE Graph.



Trade Open in last Twelve months

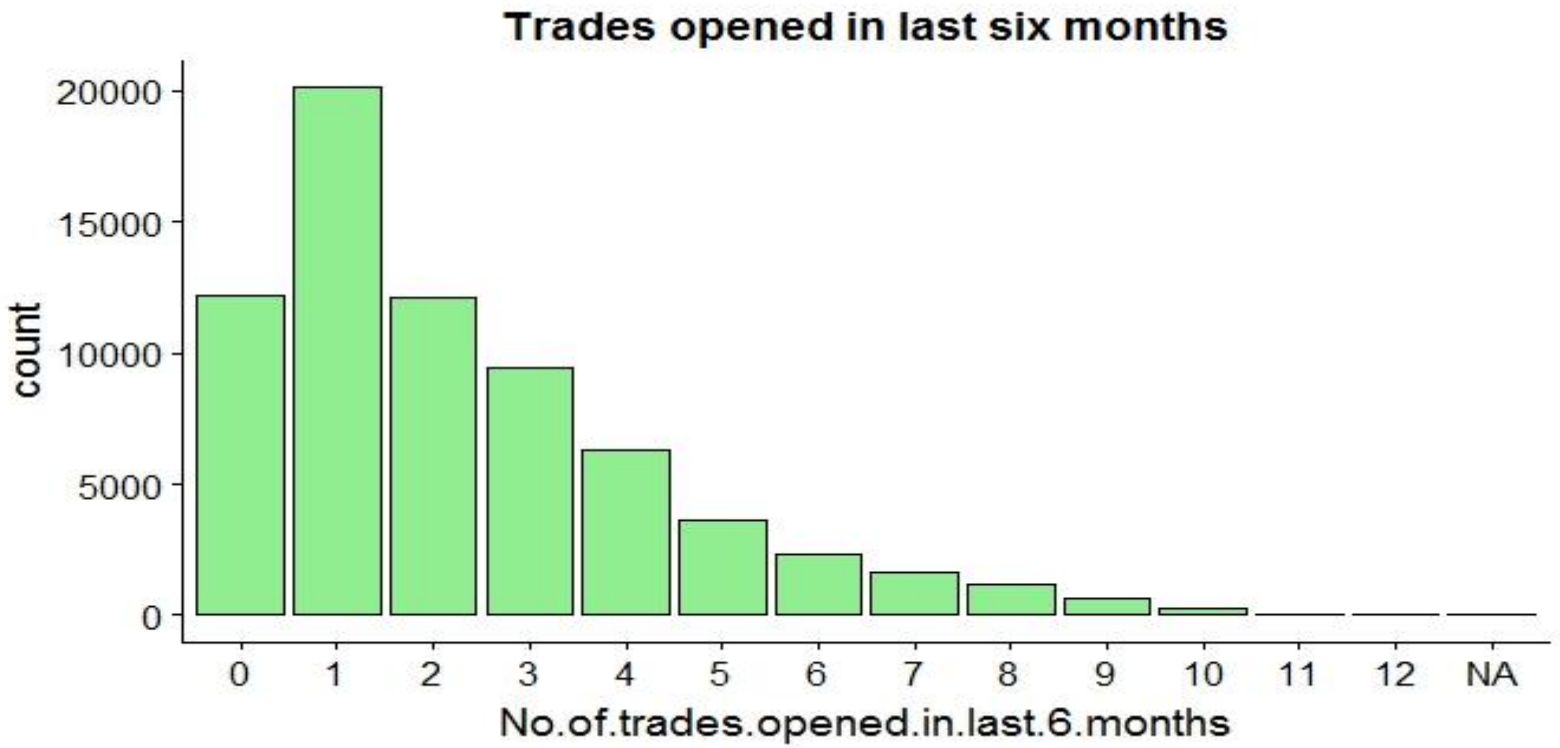


Trade open in last six months

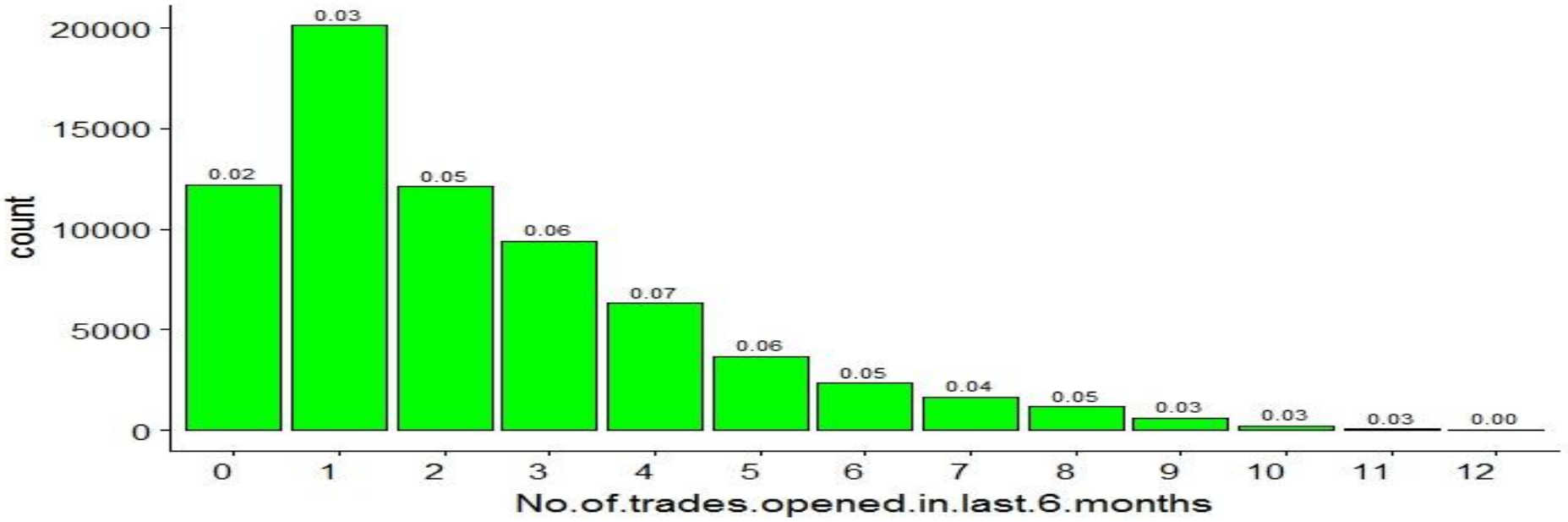
Here we present the count of customers who have open trades in the last six months.

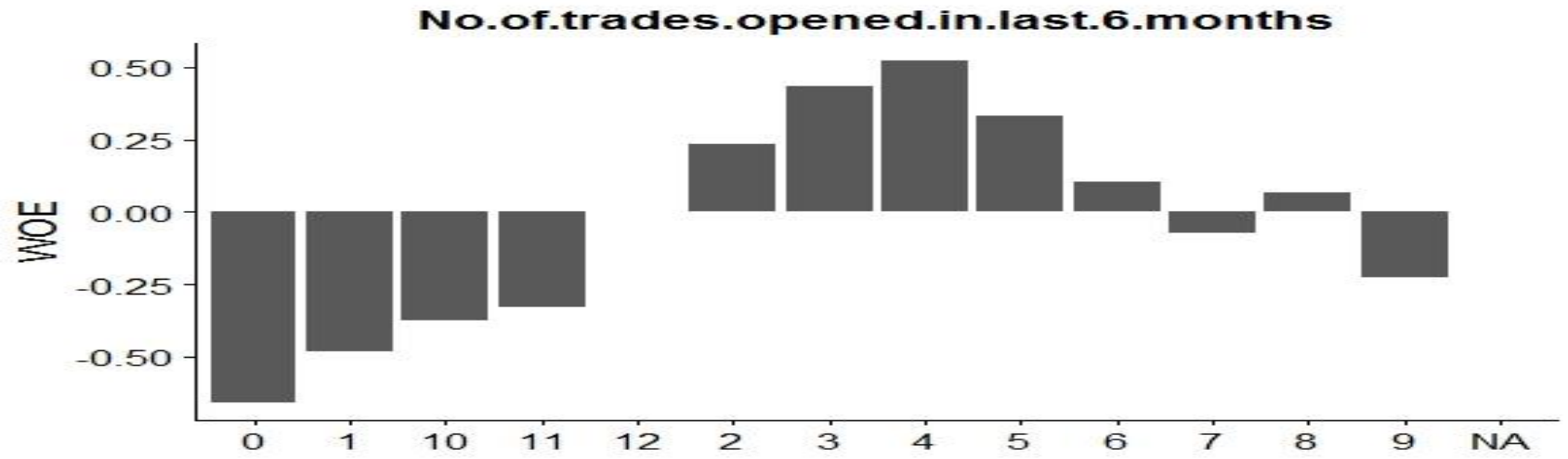
The next slide we have the number of customers who have open trades and the chances of them defaulting.

The observation is the same as the previous one. As the number of open trades increases, the chances of the customer being a defaulter increases. So this variable can be considered as a contributing factor towards deciding for performance.



Trade Open in last Six months



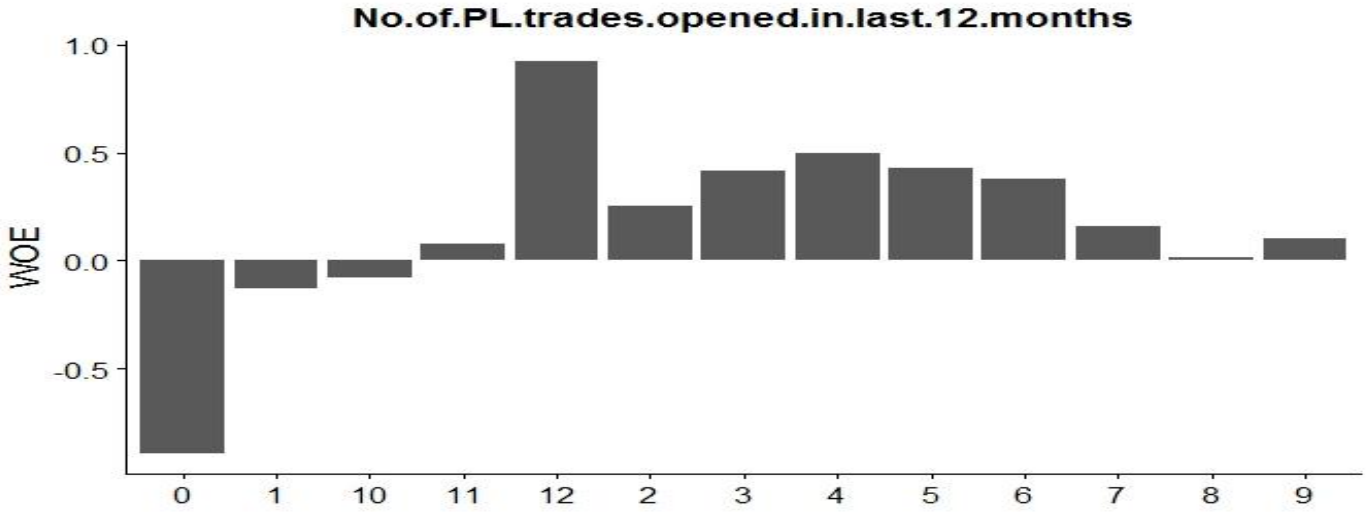


PL Accounts opened in the last 12 months

The WOE Chart shows that the no of PL Account opened in the last 12 months is also a significant variable in determining the performance of the customers.

e.g : The WOE value for ‘No.of PL trades opened in last 12 months’ equals to 3 or 4 is high, which can be a decisive factor towards performance of the customers.

PL Accounts opened in the last 12 months

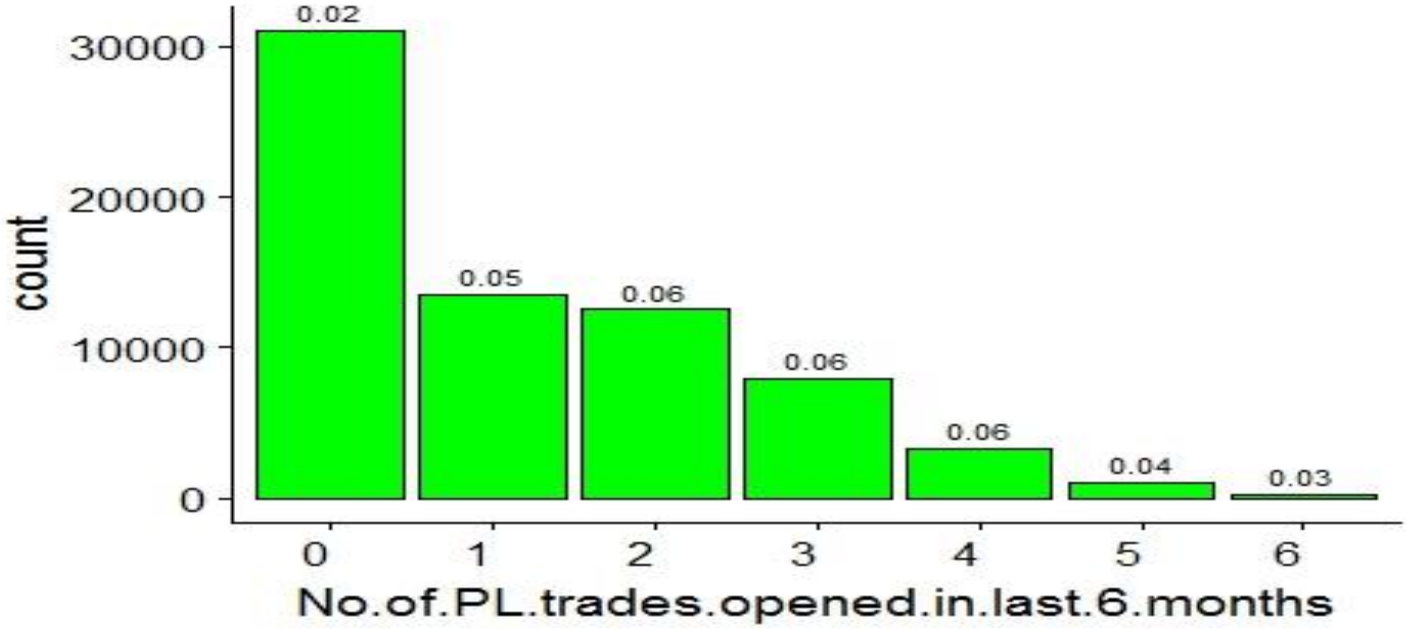


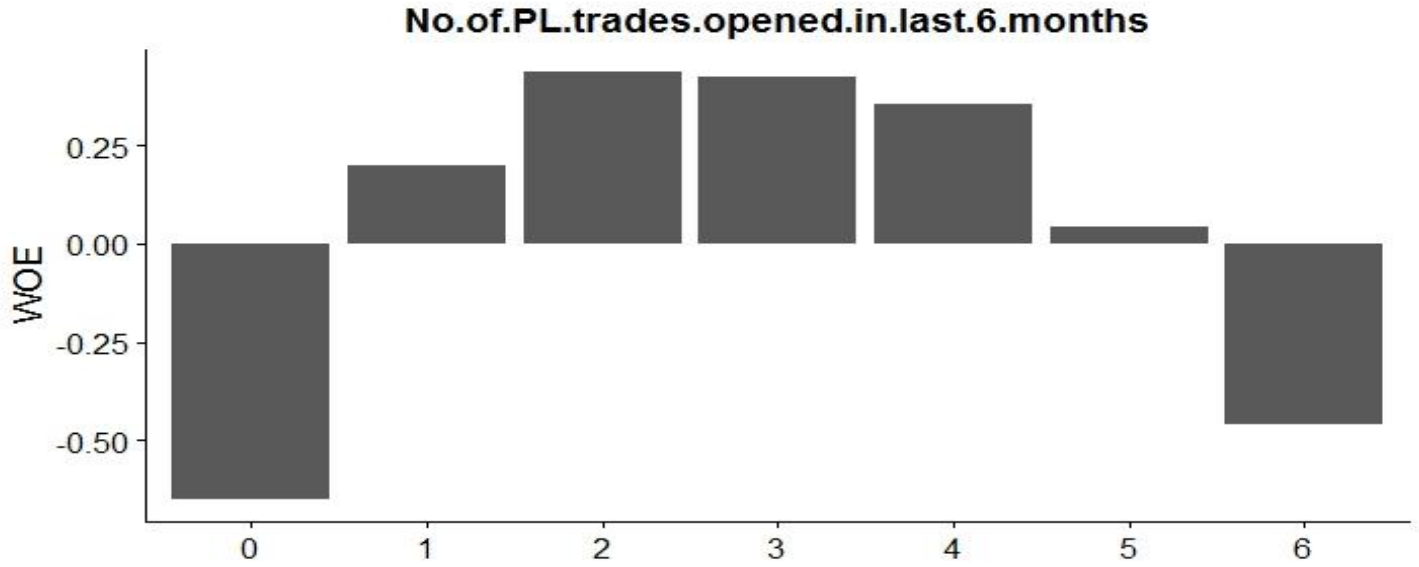
No. of PL Trades opened in last 6 months

No.of.PL.trades.opened.in.last.6.months	WOE
0	-0.64955408
1	0.20042084
2	0.43812623
3	0.42330607
4	0.35235336
5	0.04470458
6	-0.46111815

The WOE Chart shows that the no of PL Account opened in the last 12 months is also a significant variable in determining the performance of the customers.

No. of PL Trades opened in last 6 months

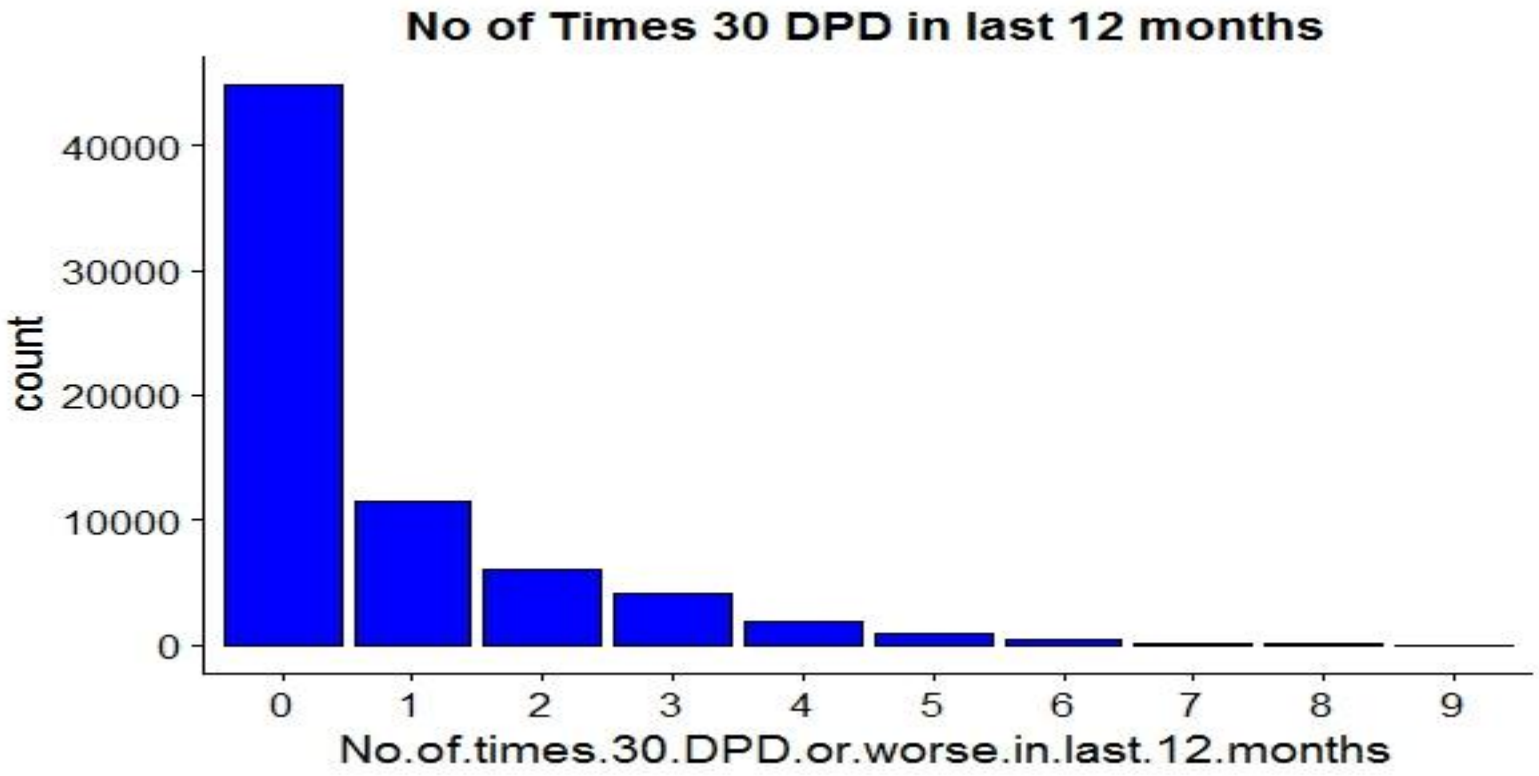




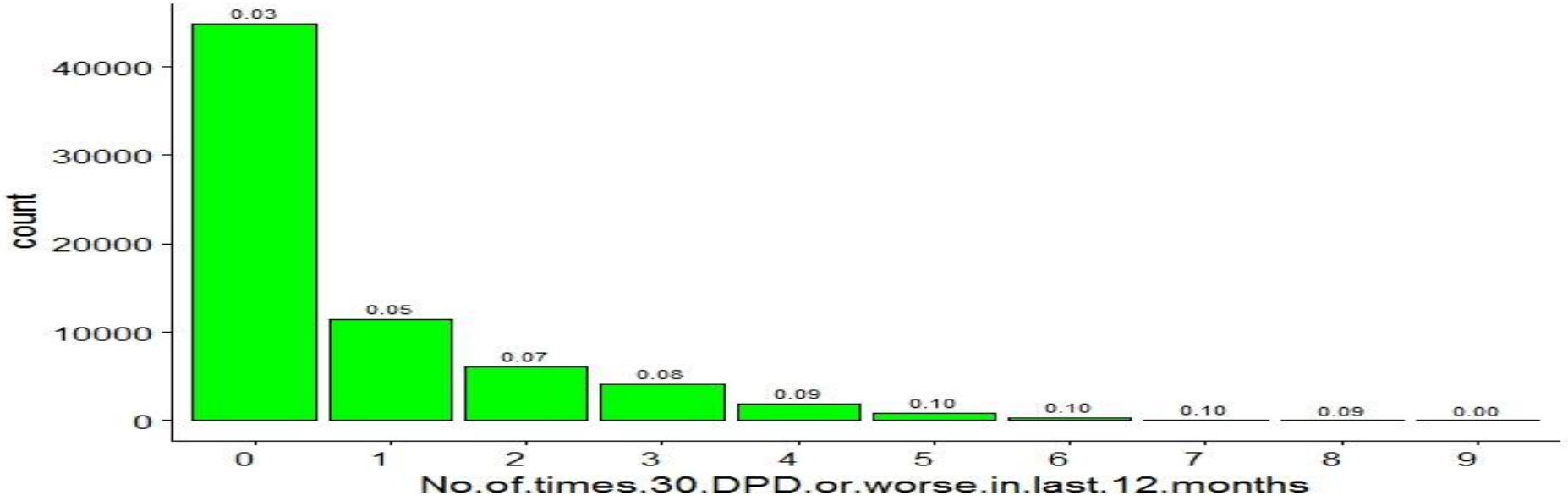
No.of times 30DPD in last 12 months

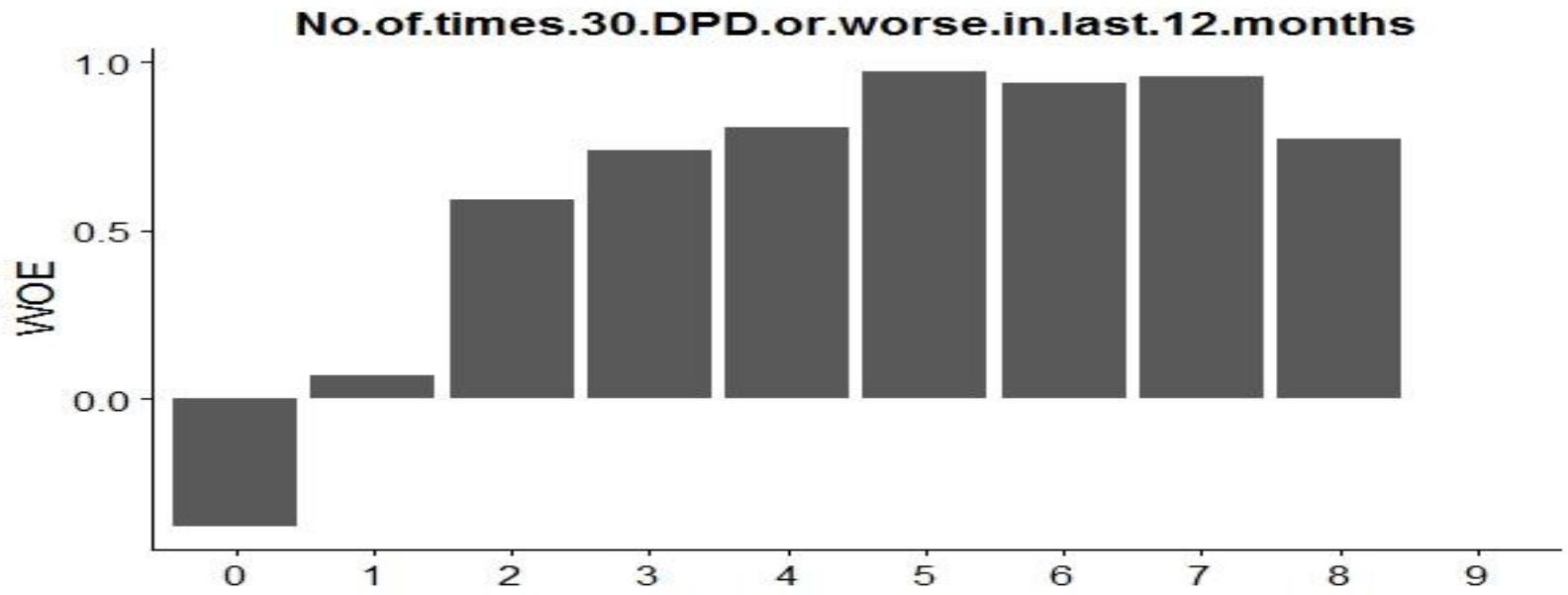
No of times 30DPD in last 12 months indicates , no of times the specific customer has missed payment for more than 30 days in last 12 months.

The WOE Chart shows as the count of 30DPD in last 12 months increases the chances of the customer being a defaulter also increases.



No.of times 30DPD in last 12 months

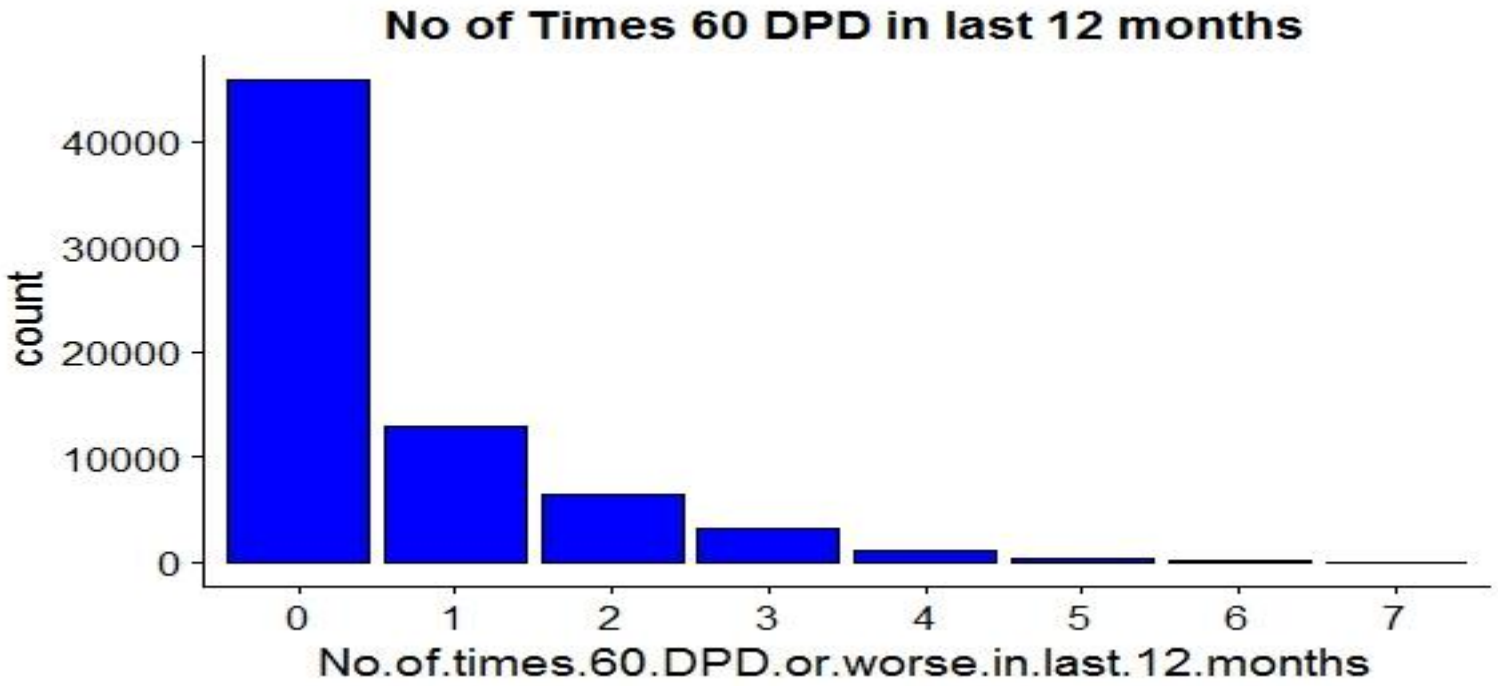




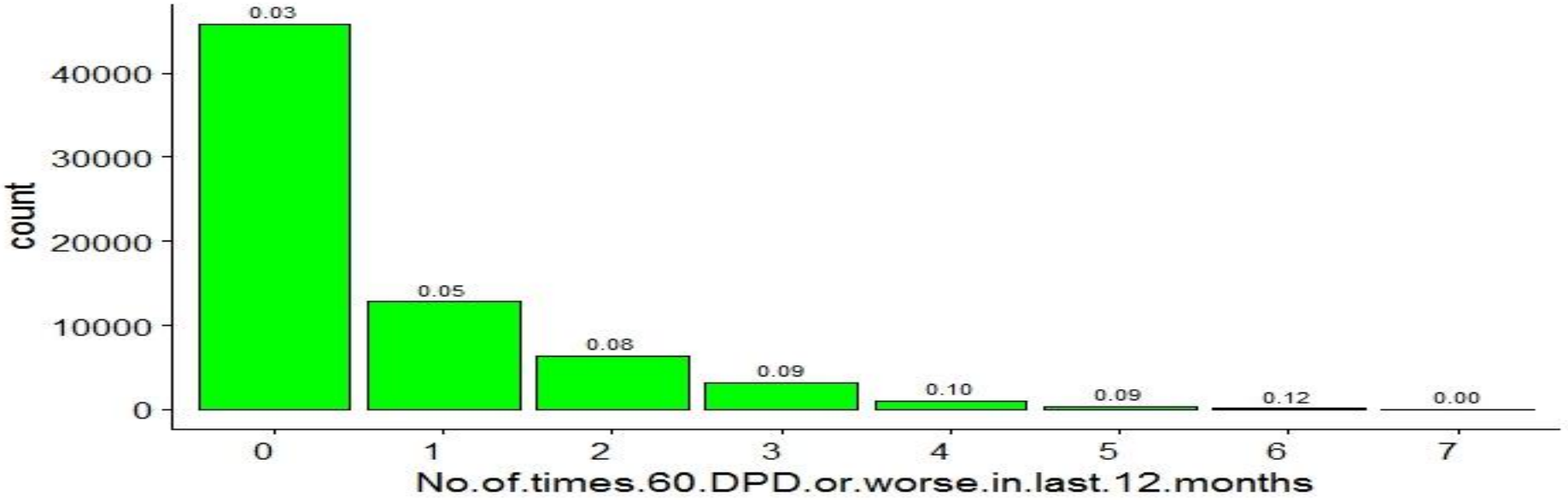
No.of times 60DPD in last 12 months

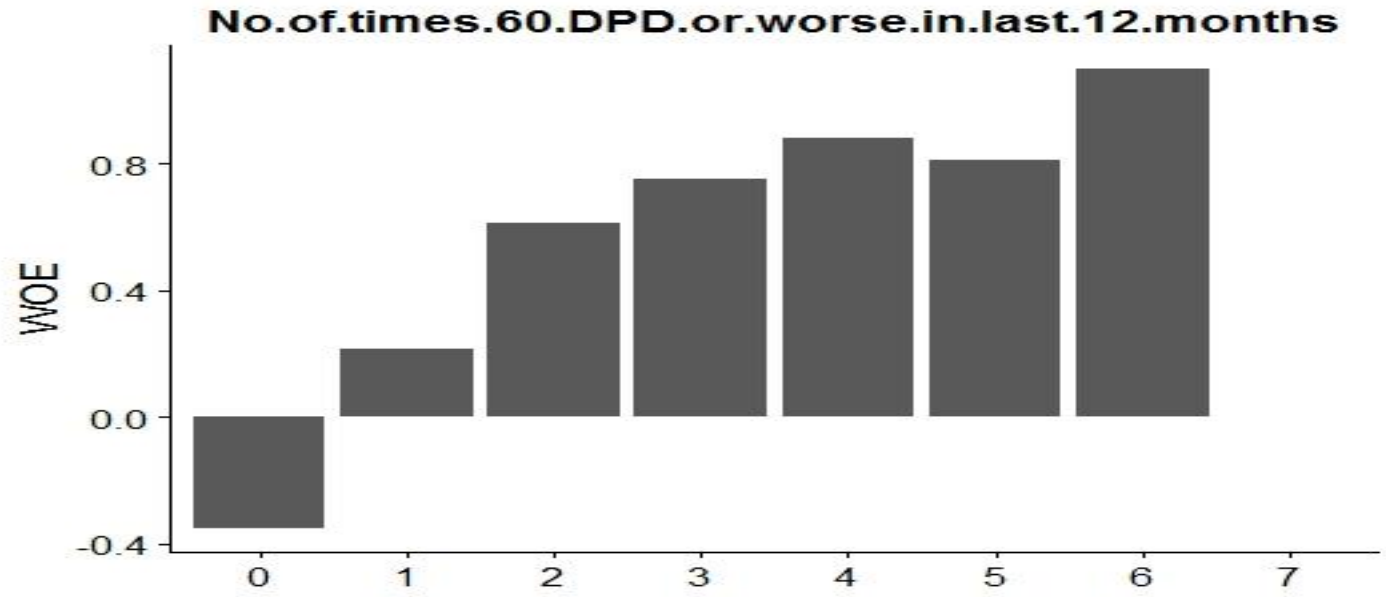
No of times 60DPD in last 12 months indicates , no of times the specific customer has missed payment for more than 60 days in last 12 months.

The WOE Chart shows as the count of 60DPD in last 12 months increases the chances of the customer being a defaulter also increases.



No.of times 60DPD in last 12 months



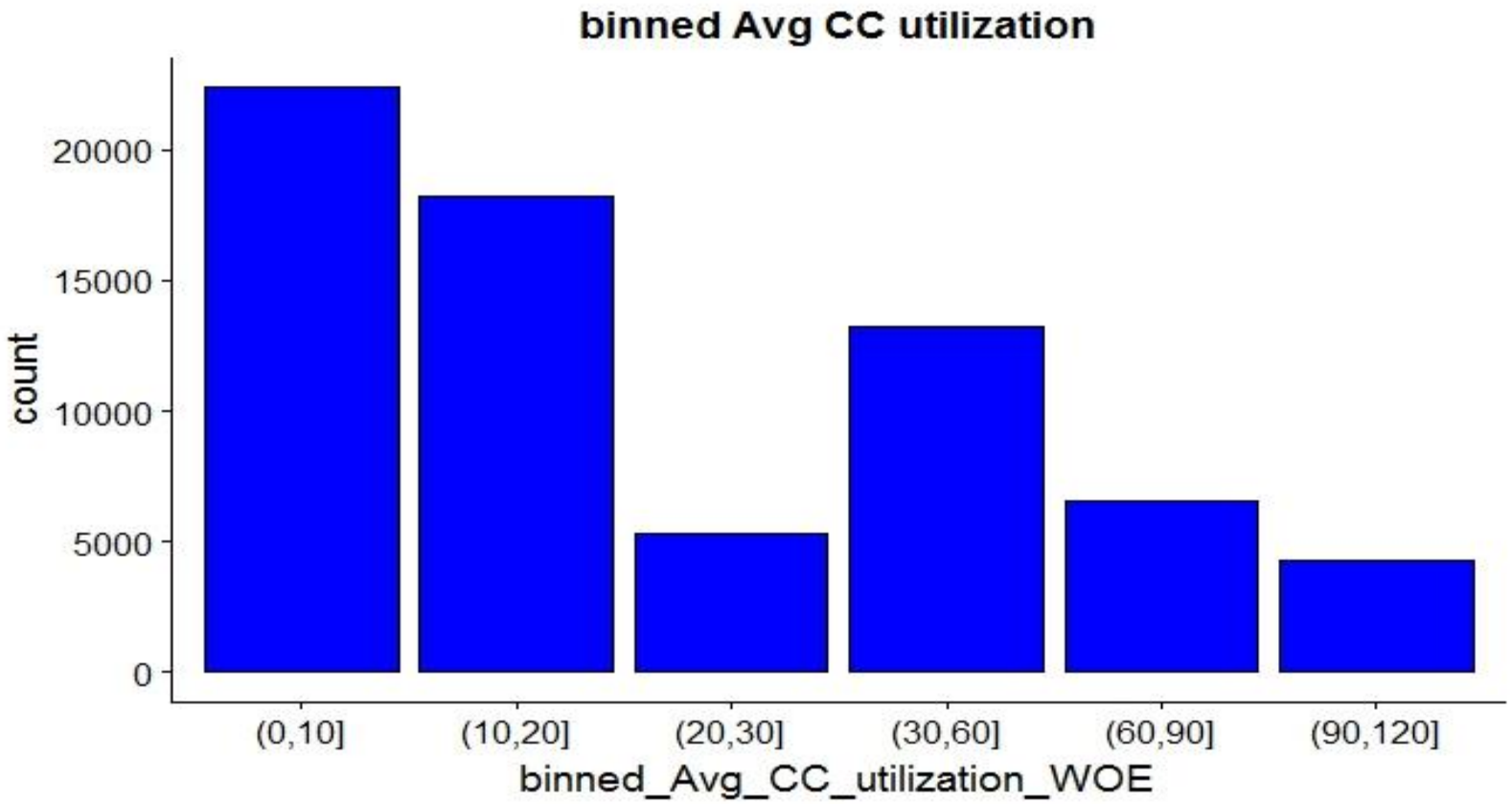


Credit Card Utilization in last 12 months

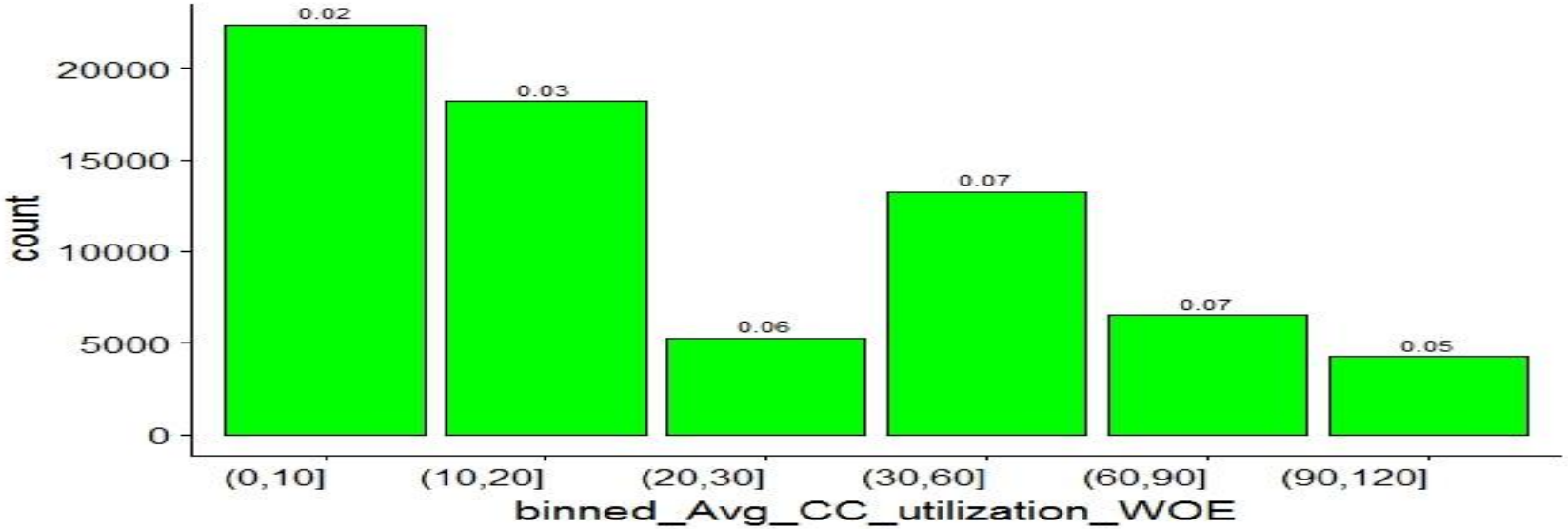
The average Credit card utilization in the last 12 months is a continuous variable, so we have binned the data , and have tried to find the significance of the variable in predicting the default customers.

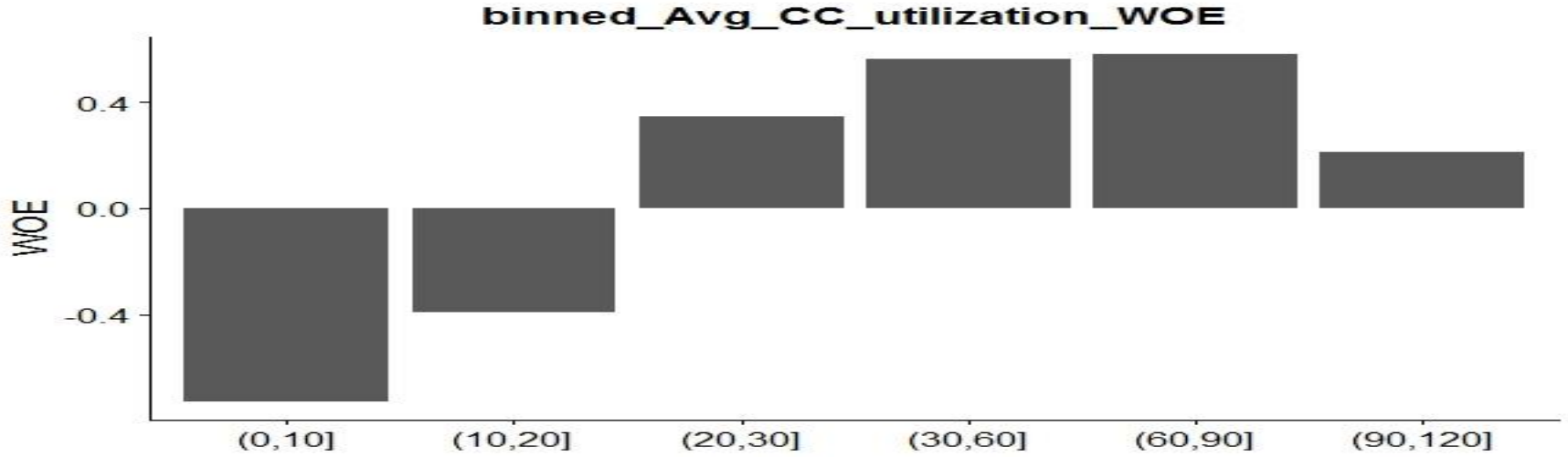
WOE of bins 30-60 and 60-90 has a value of 0.6 indicating that it is a major factor in deciding the defaulters.

binned_Avg_CC_utilization	WOE
(0,10]	0.3201517
(10,20]	-0.3922045
(20,30]	0.3461779
(30,60]	0.5652222
(60,90]	0.5805011
(90,120]	0.2106956



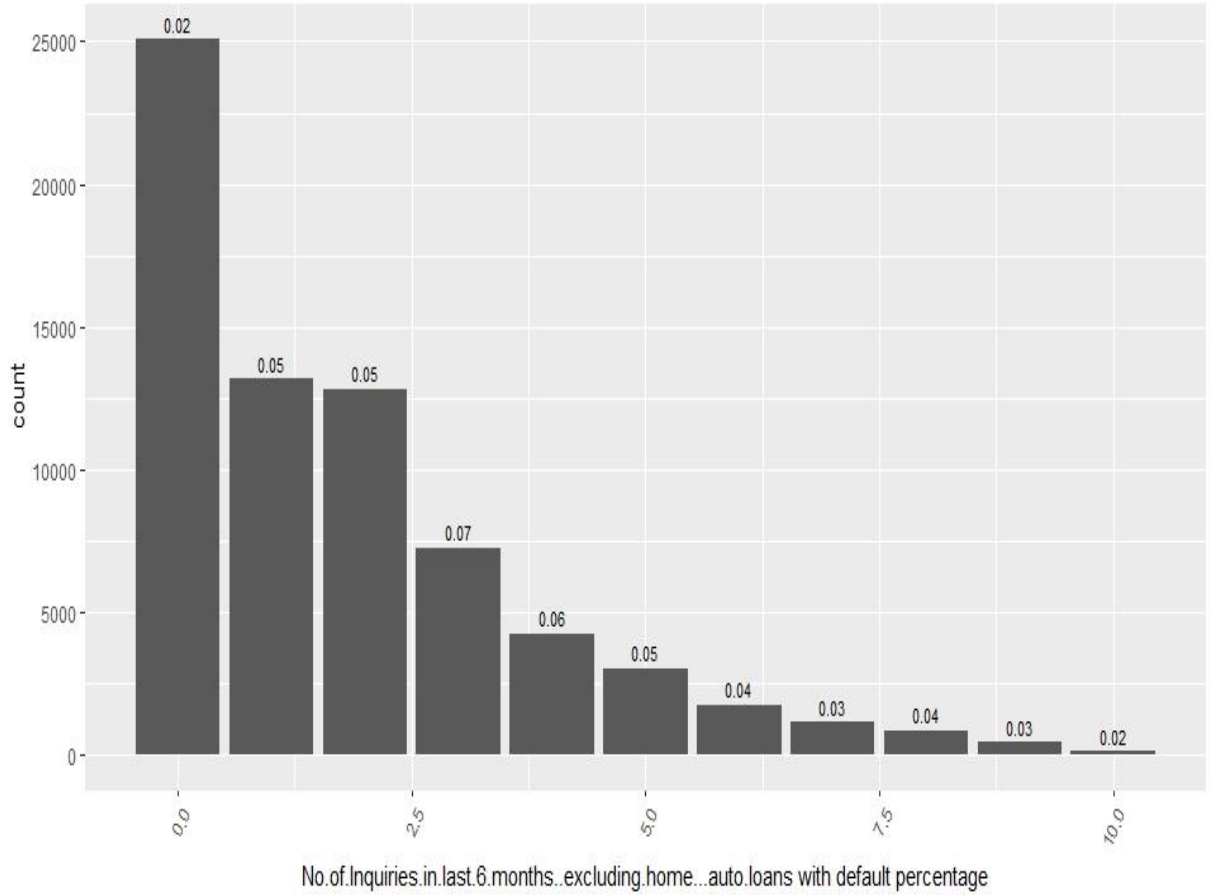
Credit Card Utilization in last 12 months





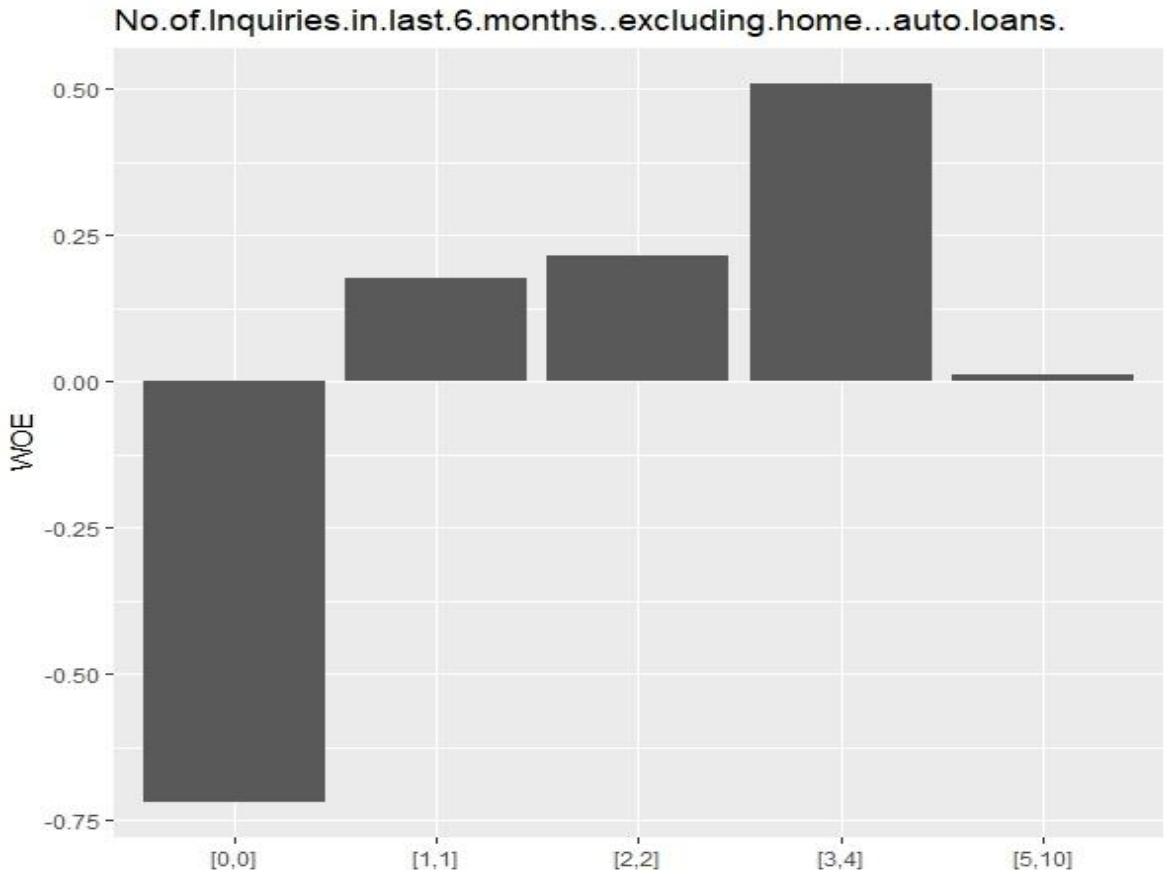
No. of Inquiries in last 6 months, excluding Home and Auto Loans

Number of queries in last 6 months vs percentage of Defaulter shows that the chances of an applicant defaulting is higher if the no of queries is between 2- 5

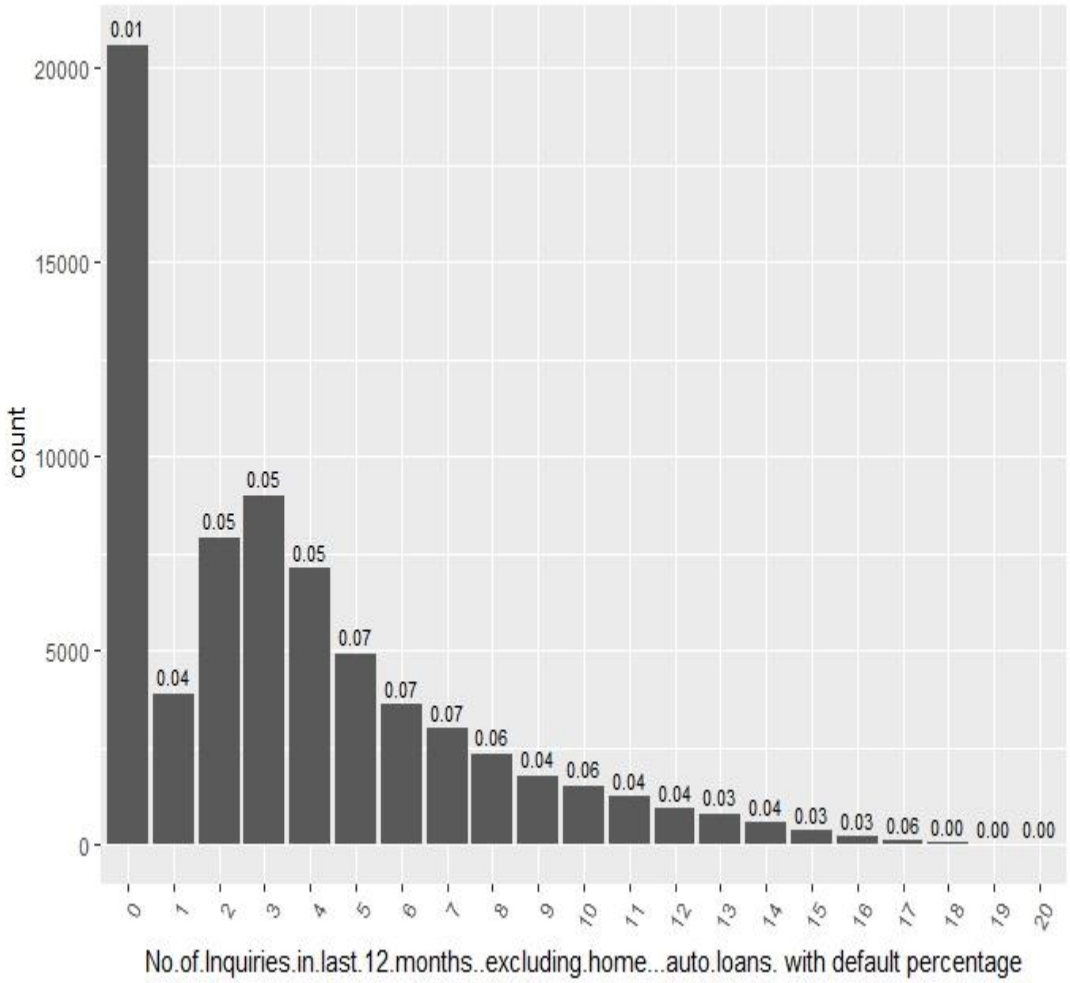


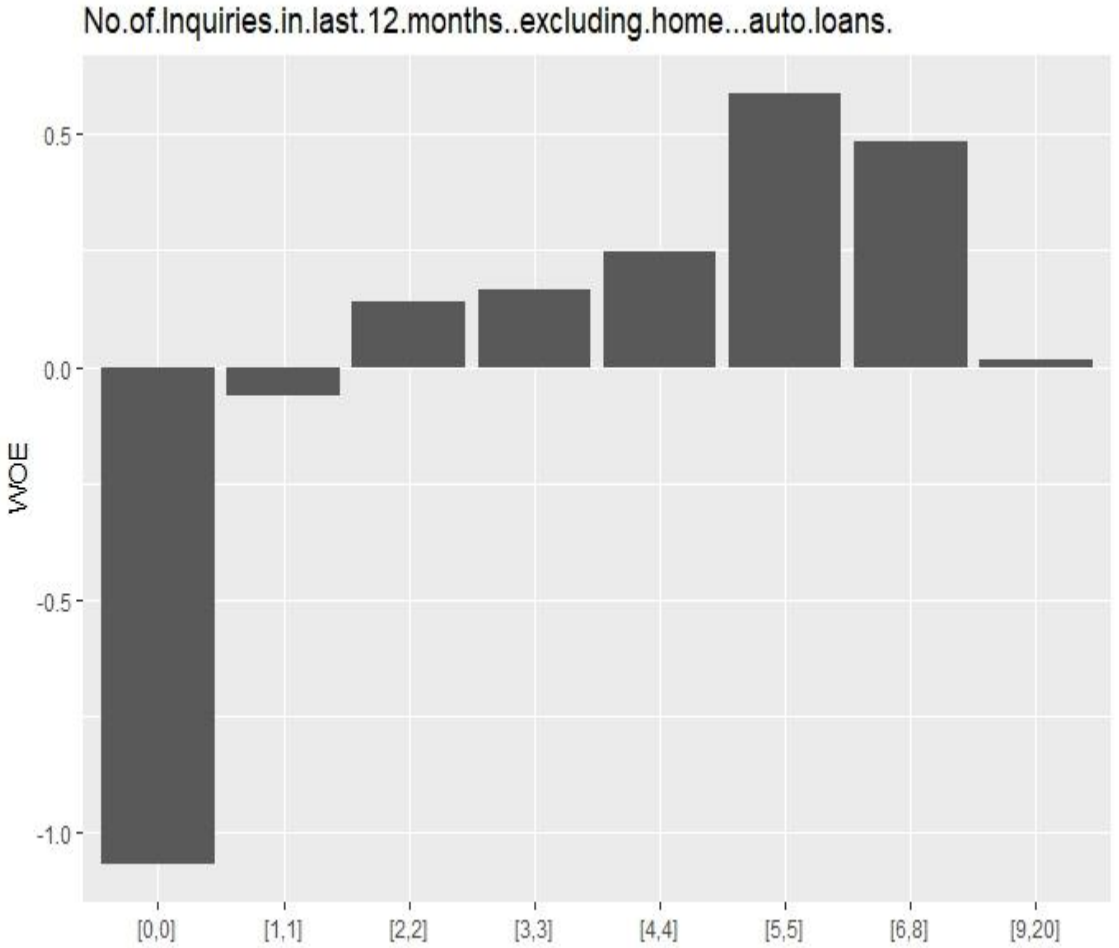
No. of Inquiries in last 6 months, excluding Home and Auto Loans

- From the plot , for loan counts of the range 3-4, WOE is high, which can be a significant factor in deciding the default behavior of customers.

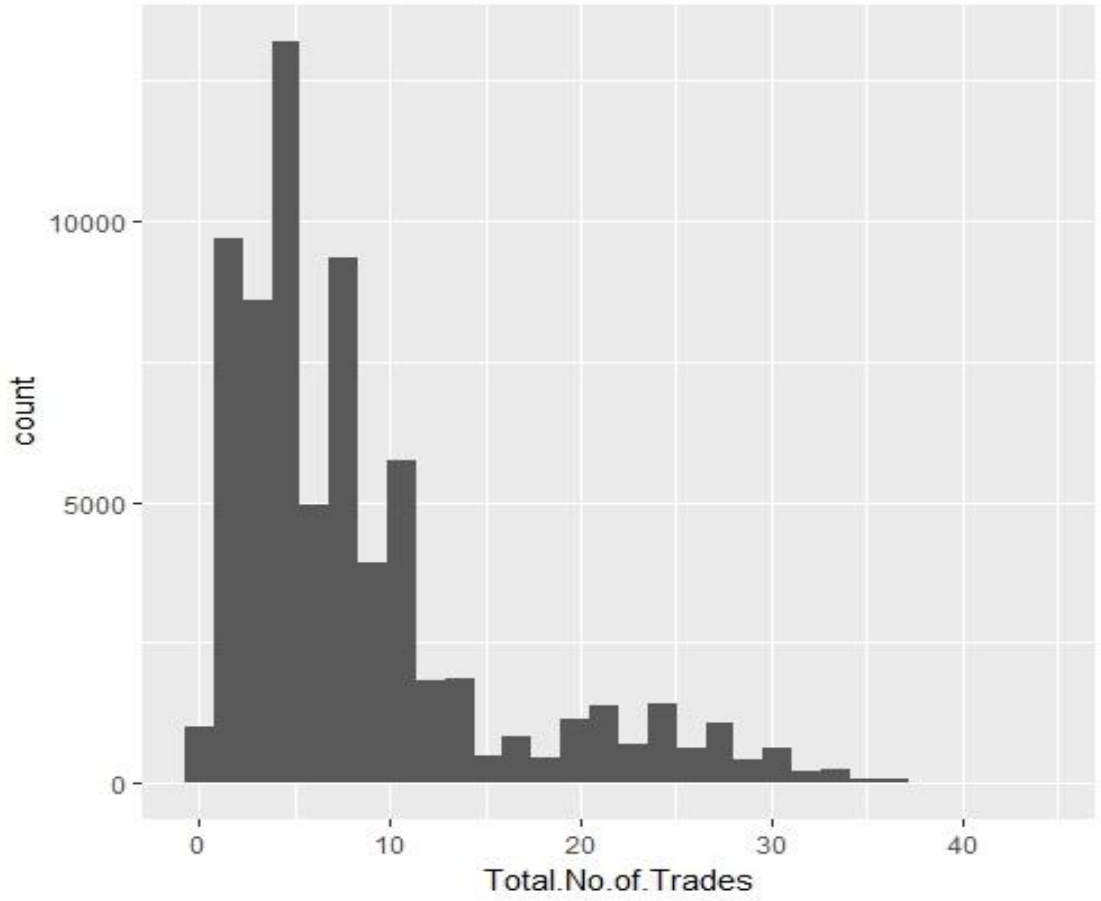


No. of Inquiries in last 12 months, excluding Home and Auto Loans

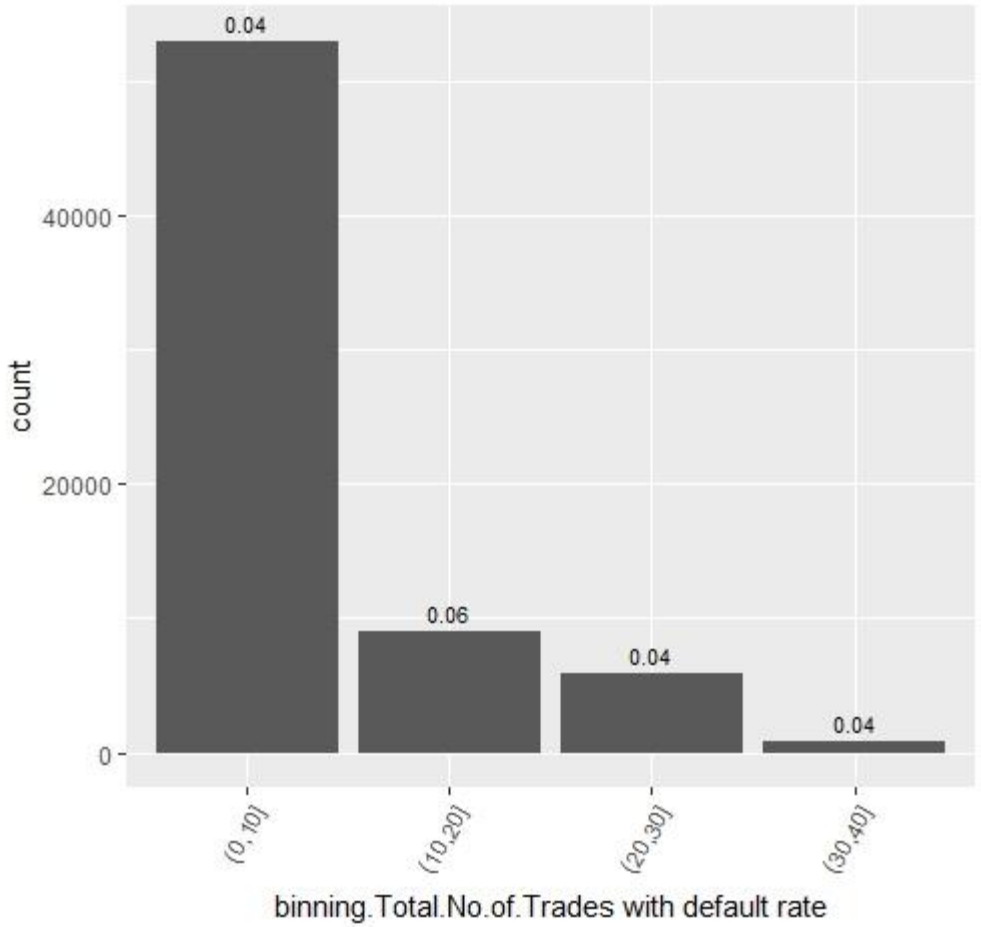


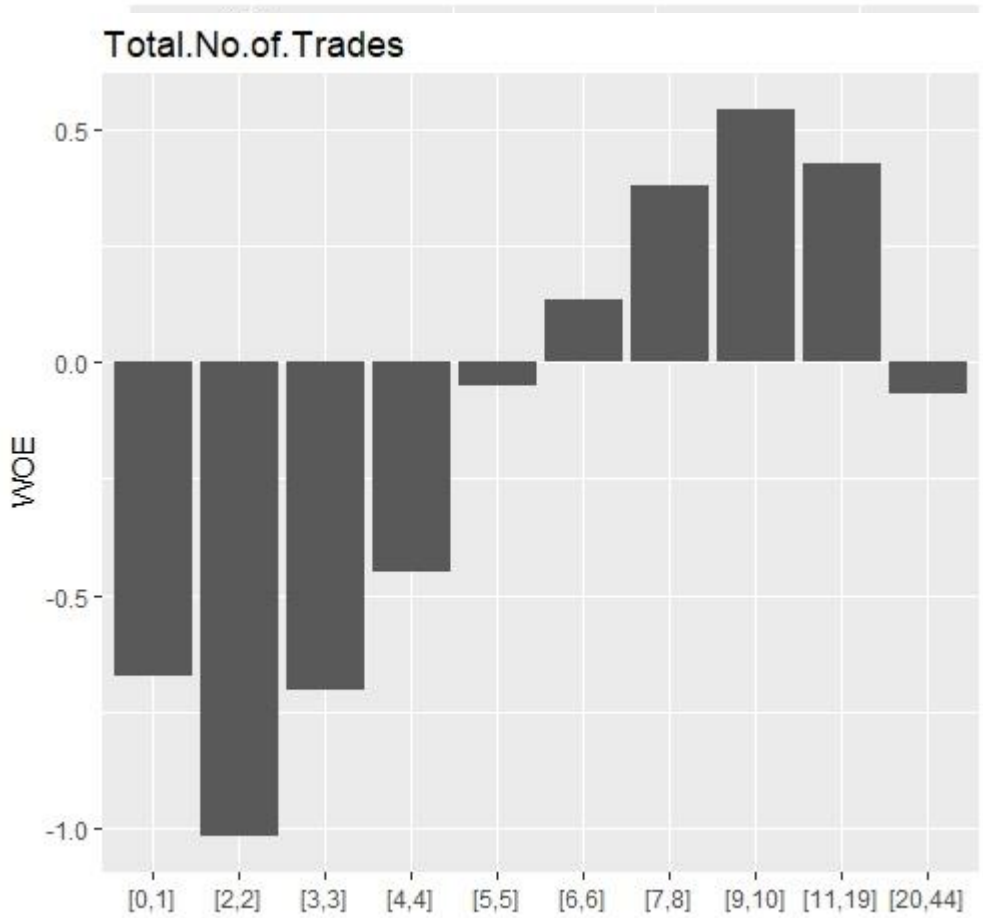


Total Number of Trades

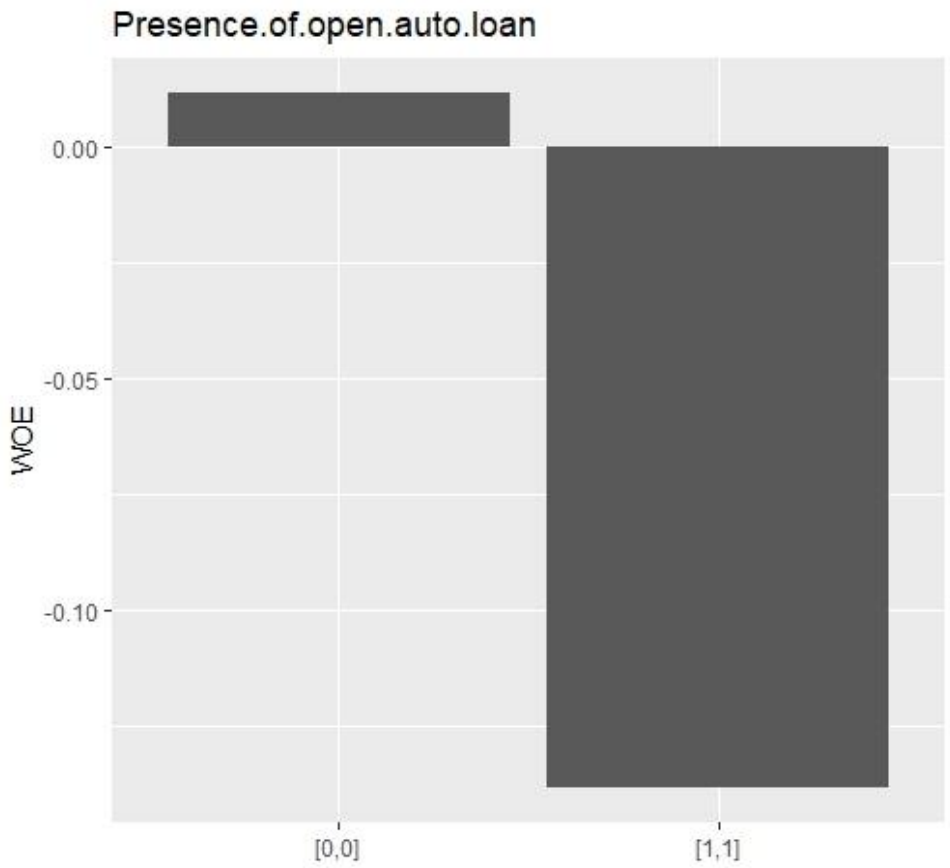


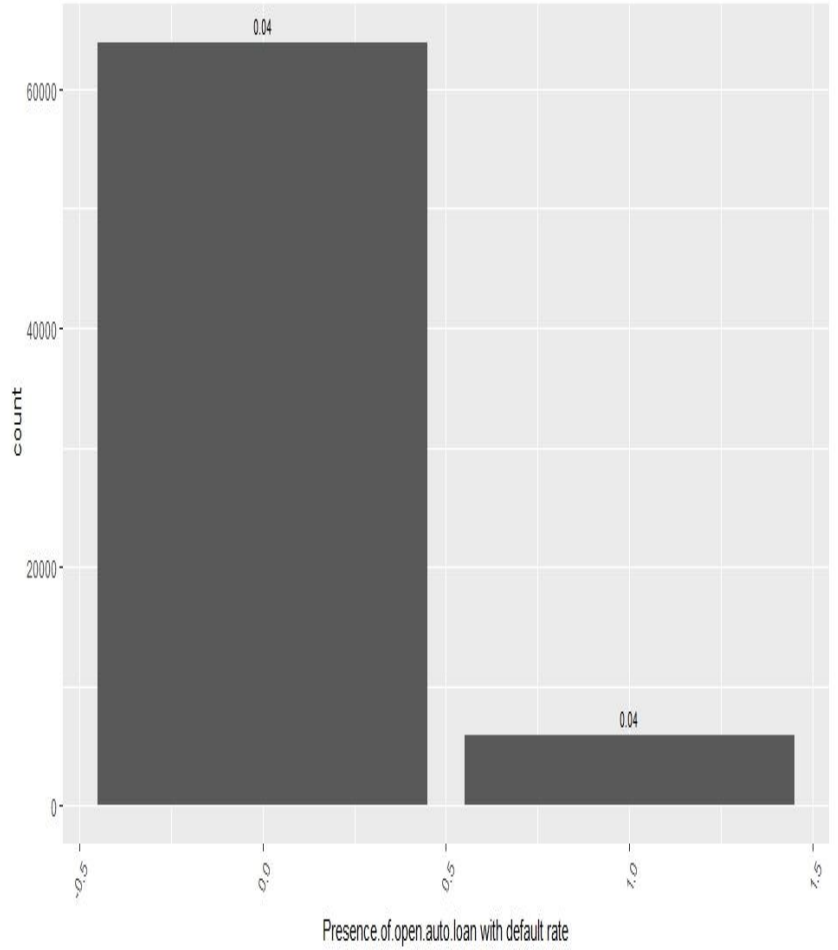
Total Number of Trades





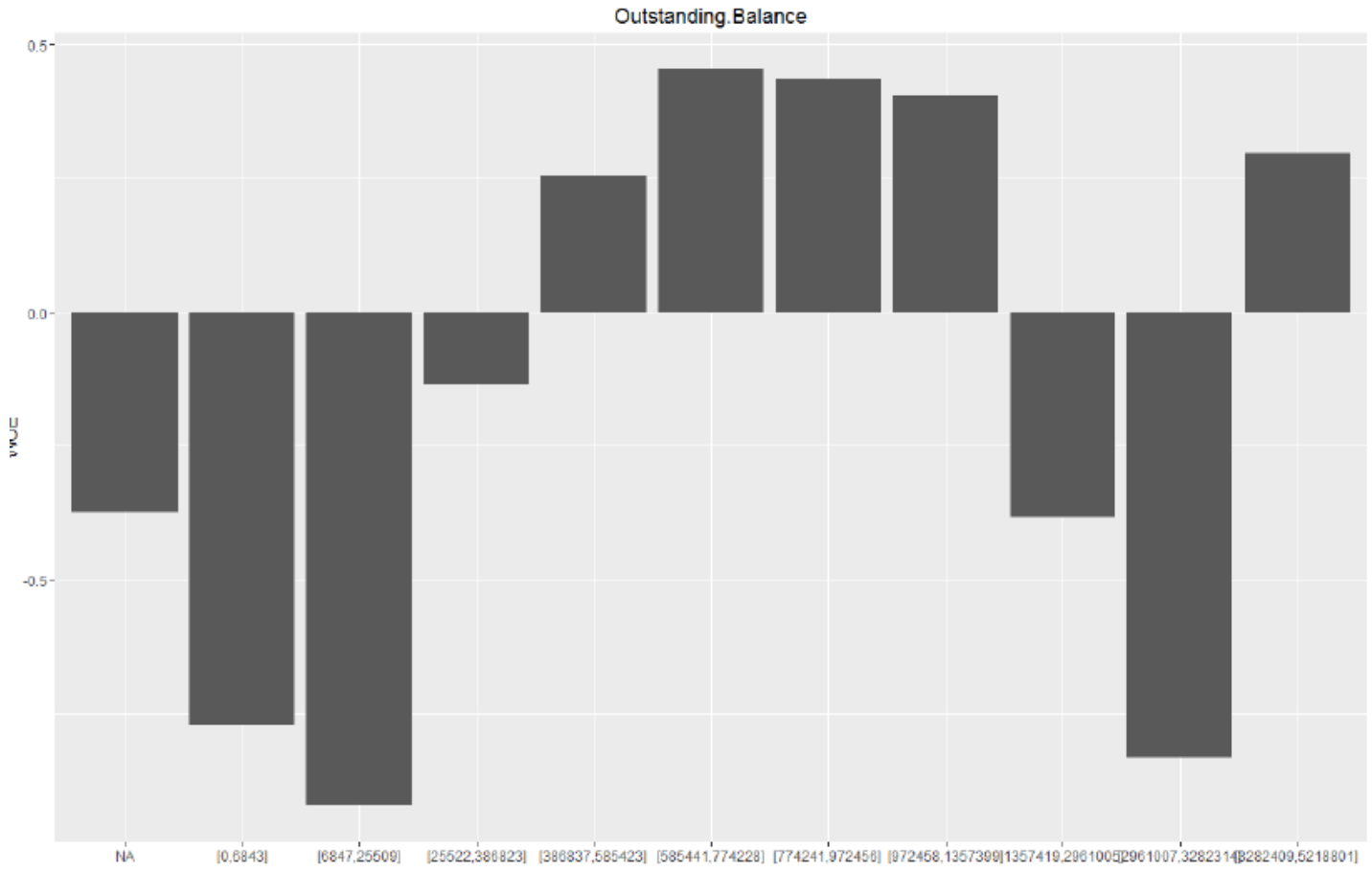
Presence of Open Auto Loan



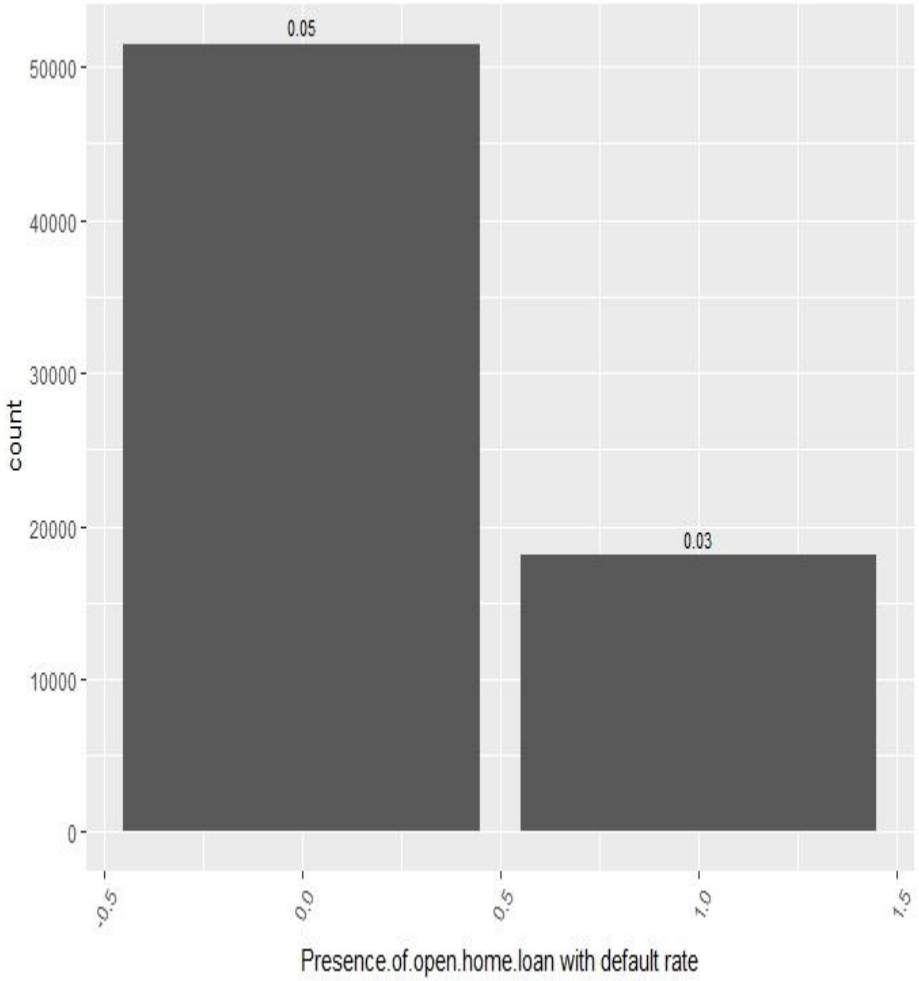


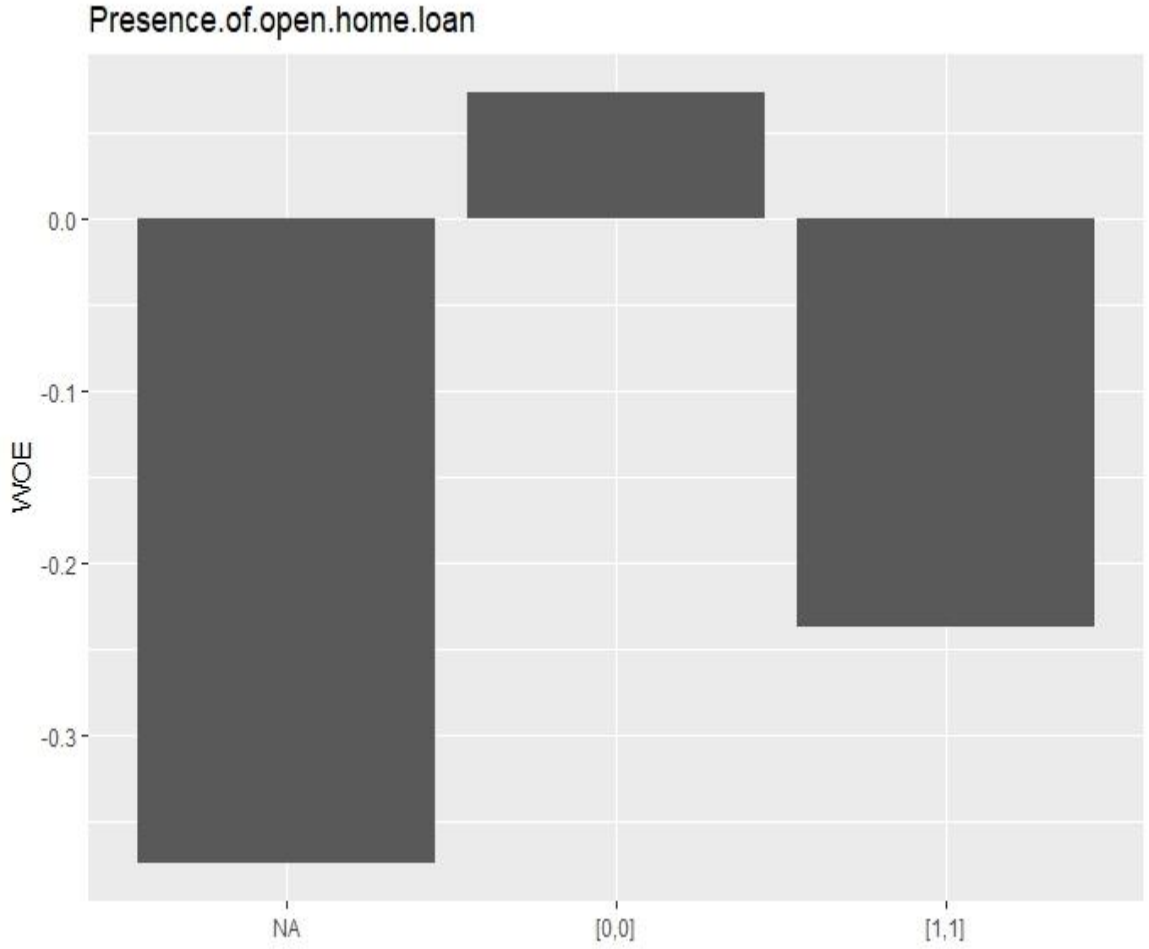
Outstanding Balance

There are 272 missing data present in Outstanding Balance column. We have replaced the data with the corresponding WOE values.



Presence of Open Home Loan





There are 272 missing data present in Open Home Loan column. We have replaced the data with the corresponding WOE values.

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

Next Steps and planned approach

- Once the logistic model is built using the attributes, we will be able to identify the variables which are significant in building the model.
- We propose to evaluate the model using different techniques like accuracy, stability and discriminative power, using stratified k-fold cross validation techniques, and based on that we plan to determine the best model for our case.
- We need to keep in mind that the data is skewed and unbalanced data and we need to decide if we need to use a tree-based or a non-tree based algorithm.
- Next we plan to build the application scorecard and based on that we plan to predict the potential financial benefits for the company.