

CredX Acquisition Analytics

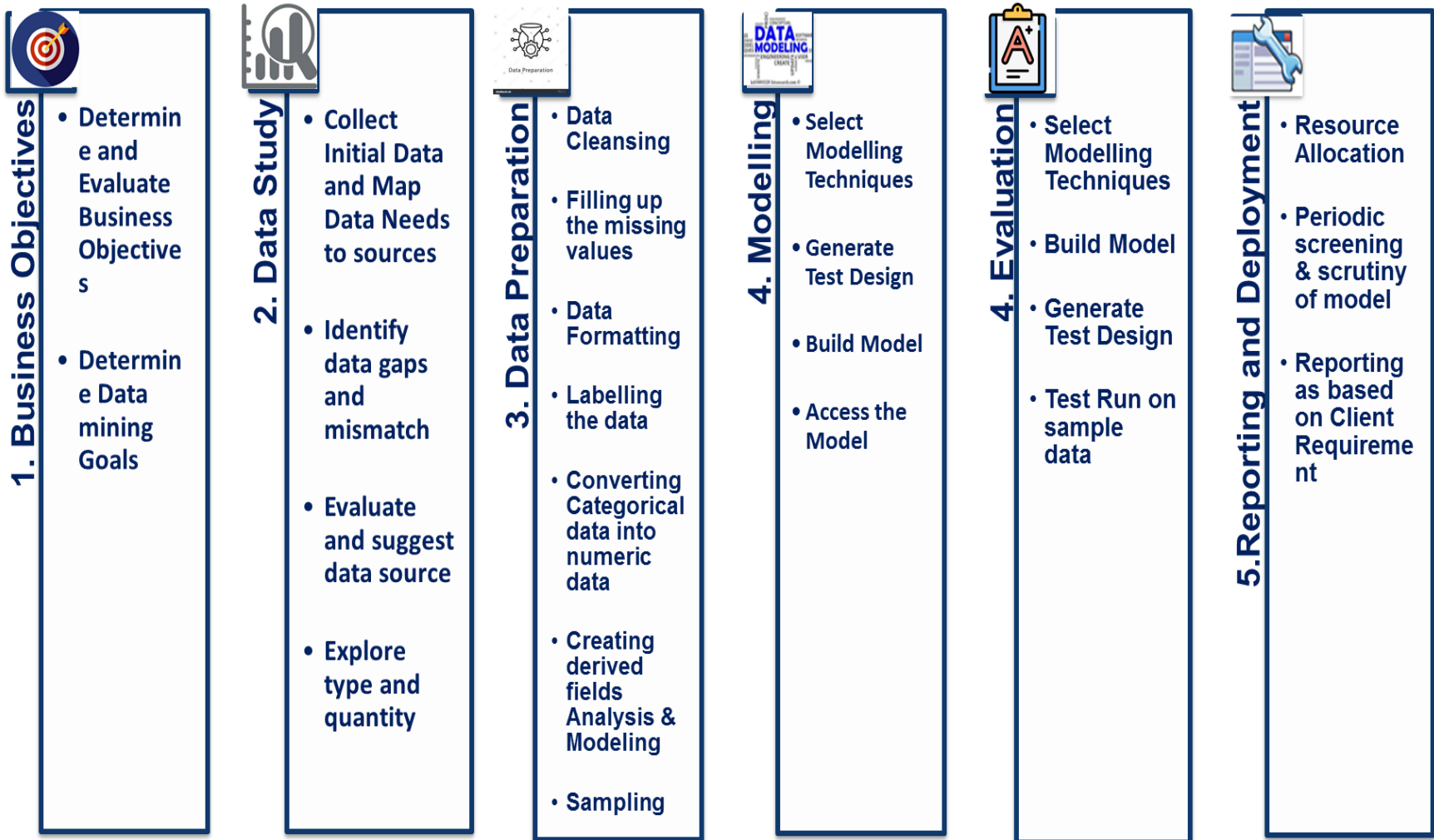
By: Pawan Pandita

Shubham Joshi

Ayush Ranjan

Methodology Used

CRISP-dm (Cross Industry Standard Process For Data Mining) Methodology



1. Business Understanding:-

- 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’.
- Goal:** We need help CredX identify the right customers using predictive models. Using past data of the bank’s applicants, you need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of your project.

2. Data Understanding

There are two data sets in this project — **demographic** and **credit bureau** data.

- 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.

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 children 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")

There are total 71295 in this dataset.

- 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.

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 paid dues since 90days in last 6 months
No of times 60 DPD or worse in last 6 months	Number of times customer has not paid dues since 60 days last 6 months
No of times 30 DPD or worse in last 6 months	Number of times customer has not paid 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 paid 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 paid 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 paid 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
No of Inquiries in last 6 months (excluding home & auto loans)	Number of times the customers has inquired in last 6 months
No of Inquiries in last 12 months (excluding home & 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")

Both data contain a **performance tag** which represents whether the applicant has gone 90 days past due or worse in the past 12-months (i.e. defaulted) after getting a credit card.

In some cases, you will find that all the variables in the credit bureau data are zero and credit card utilization is missing. These represent cases in which there is a no-hit in the credit bureau. You will also find cases with credit card utilization missing. These are the cases in which the applicant does not have any other credit card.

There are total 71295 in this dataset.

Assumptions:-

1. The Income is monthly income and is in 1000.e
2. The records with null **Performance Tag** are those of Non customers and will be used as test data.

Data Preparation

1. **Duplicate Records Treatment:** Duplicate records for the following Application ID were observed in **both demographic and Credit bureau data** and were removed :-

Application ID	Number of Duplicates
653287861	2
671989187	2
765011468	2

2. **Missing Value Imputation:** The following missing values were observed in the data sets :-

Demographic Data				
Column Name	NA and Blank count	%of total	Action	Justification
Number.of.dependents	3	Less than .1%	Imputed using kNN	The assumption behind using KNN for missing values is that a point value can be approximated by the values of the points that are closest to it, based on other variables. Since we have enough records in the population, and missing values are <.16% the aforementioned parameters can be approximated by considering values from records which have similar data in other fields. higher values of k provide smoothing that reduces the risk of over fitting due to noise in the training data, so we assume k=267 here which is square root of total number of observations (71292).
Education	119	.16%	Imputed using kNN	
Profession	14	Less than .1%	Imputed using kNN	
Type.of.residence	8	Less than .1%	Imputed using kNN	
Gender	2	Less than .1%	Imputed using kNN	

Credit Bureau Data				
Column Name	NA and Blank count	%of total	Action	Justification
Presence.of.open.home.loan	272	.3%	Records Deleted	Since Presence.of.open.home.loan is completely on user’s discretion and cannot be derived from the data in hand, as well as data is .3% which is very less hence let’s remove the same.
Outstanding.Balance	119	.3%	Records Deleted	Since Outstanding.Balance is completely on user’s discretion and cannot be derived from the data in hand, as well as data is .3% which is very less hence let’s remove the same.
Avgas.CC.Utilization.in.last.12.months	786	1%	Imputed using kNN	kNN approximates the values of that points that are close to it. Since missing values are 1%, the aforementioned parameters can be approximated by observing behaviors from records which have similar data in other fields and taking average. We assume k=267 here which is square root of total number of observations (71292).

3. **Outlier Treatment:** -

Demographic Data				
Column Name	Percentil	Capped Values	Action	Justification
Age	1%	For Age <18 Set Age =18	Outliers removed using capping	All the values which are capped belong to 1% of the population.
Income	1%	For Income <18 Set Income =18	Outliers removed using capping	
No.of.months.in.current.company	99%	For No.of.months.in.current.company > 74 Set	Outliers removed using capping	

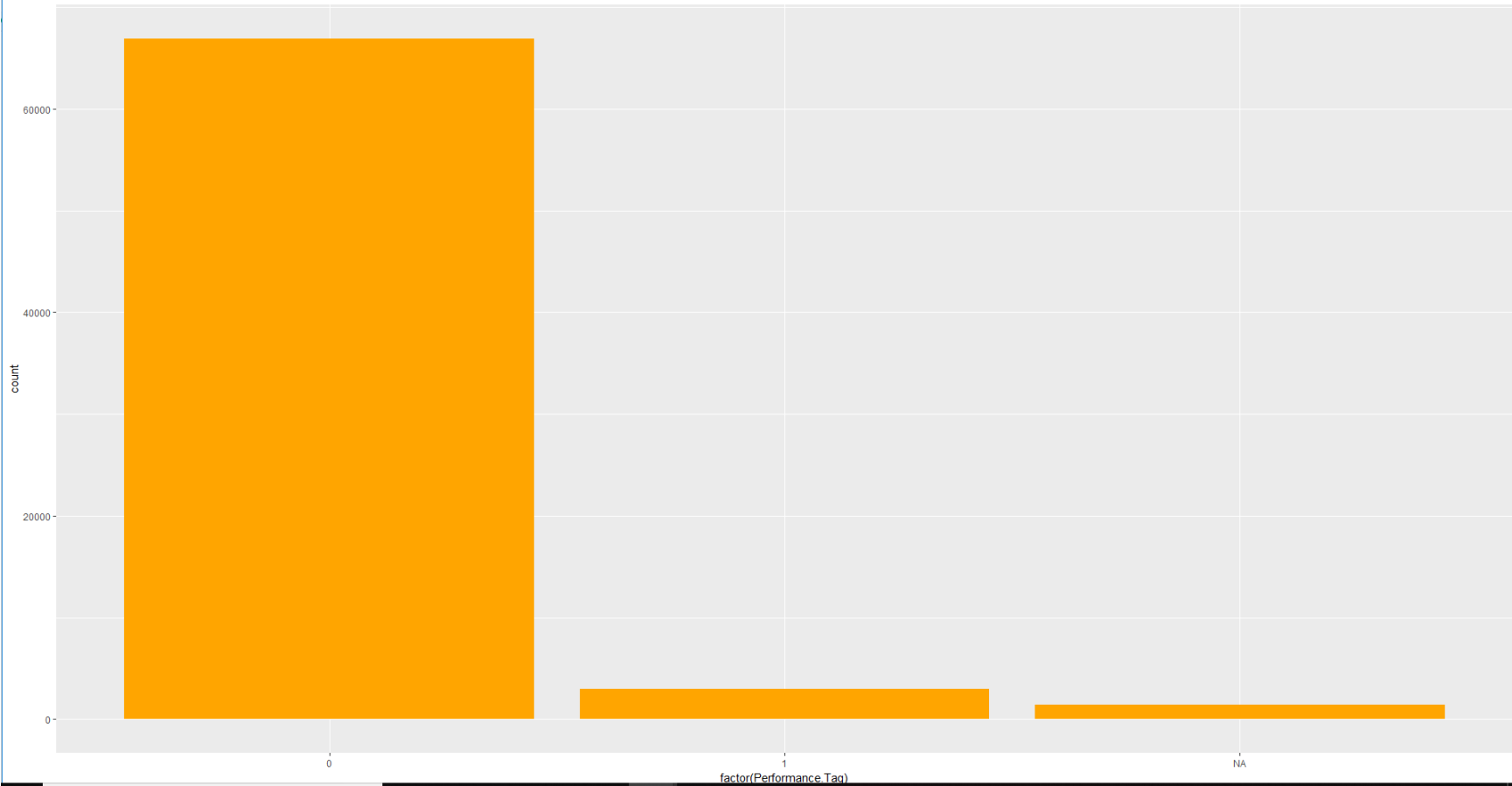
Credit Bureau Data				
Column Name	Percentiles	Capped Values	Action	Justification
No.of.trades.opened.in.last.12.months	99%	For No.of.trades.opened.in.last.12.months > 21 Set No.of.trades.opened.in.last.12.months = 21	Outliers removed using capping	All the values which are capped belong to 1% of the population.
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	99%	For No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.> 15 Set No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.= 15	Outliers removed using capping	
Total.No.of.Trades	99%	For Total.No.of.Trades> 31 Set Total.No.of.Trades= 31	Outliers removed using capping	

4. Derived Metrics

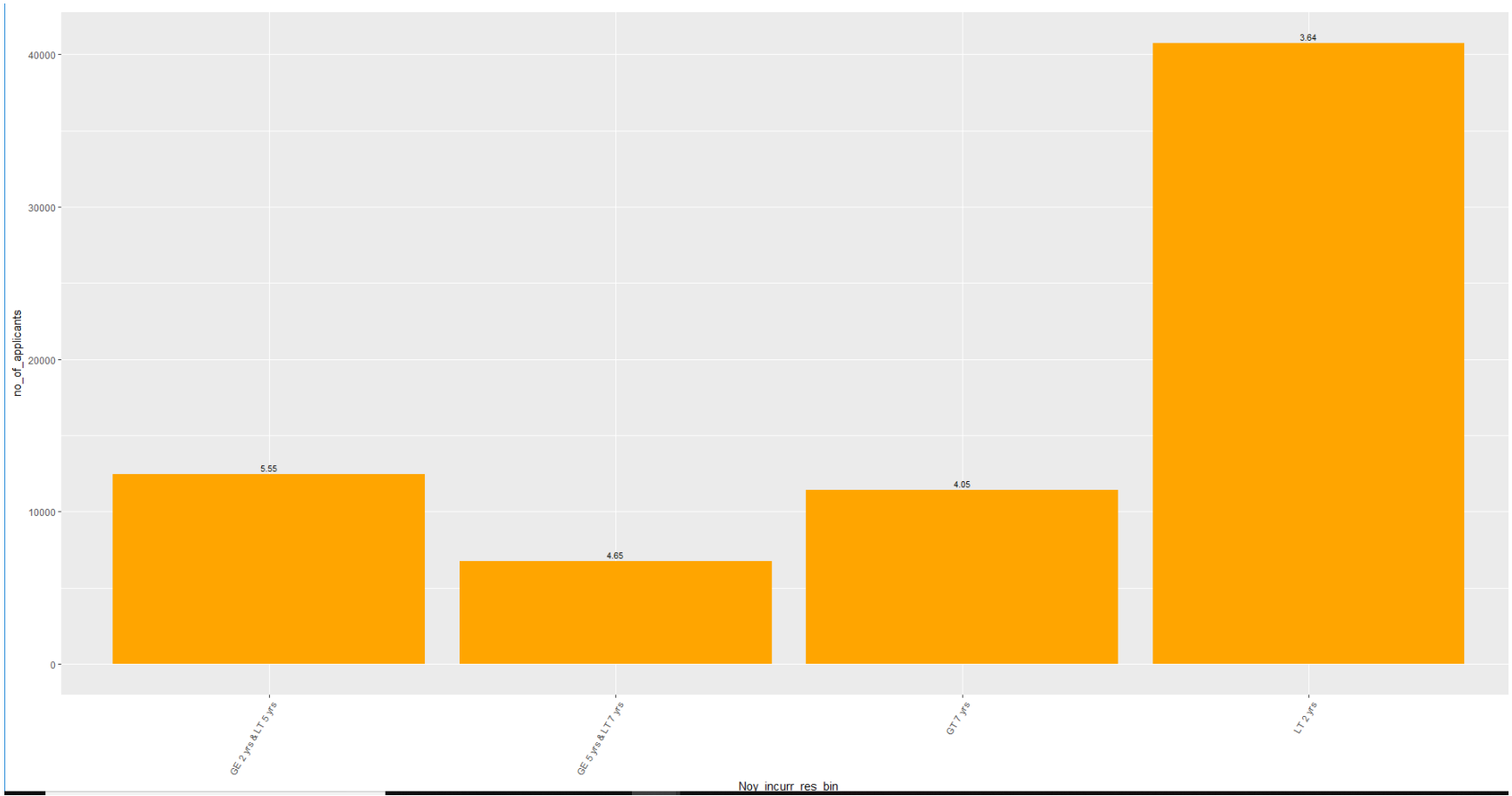
Name	Data Frame column name	Formula
DTIR(Depth to income Ratio)	outbal_inc_rtio	=Outstanding Balance/(Income*1000*12)
Income to dependent ratio	inc_to_noofdep_rtio	=Income/No of dependents
Presence of housing or auto loan	presence.of.ouito.or.home.loan	=1 if auto or home loan is present 0 if no home or auto loan is present

EDA

1) Target variable distribution₁₌

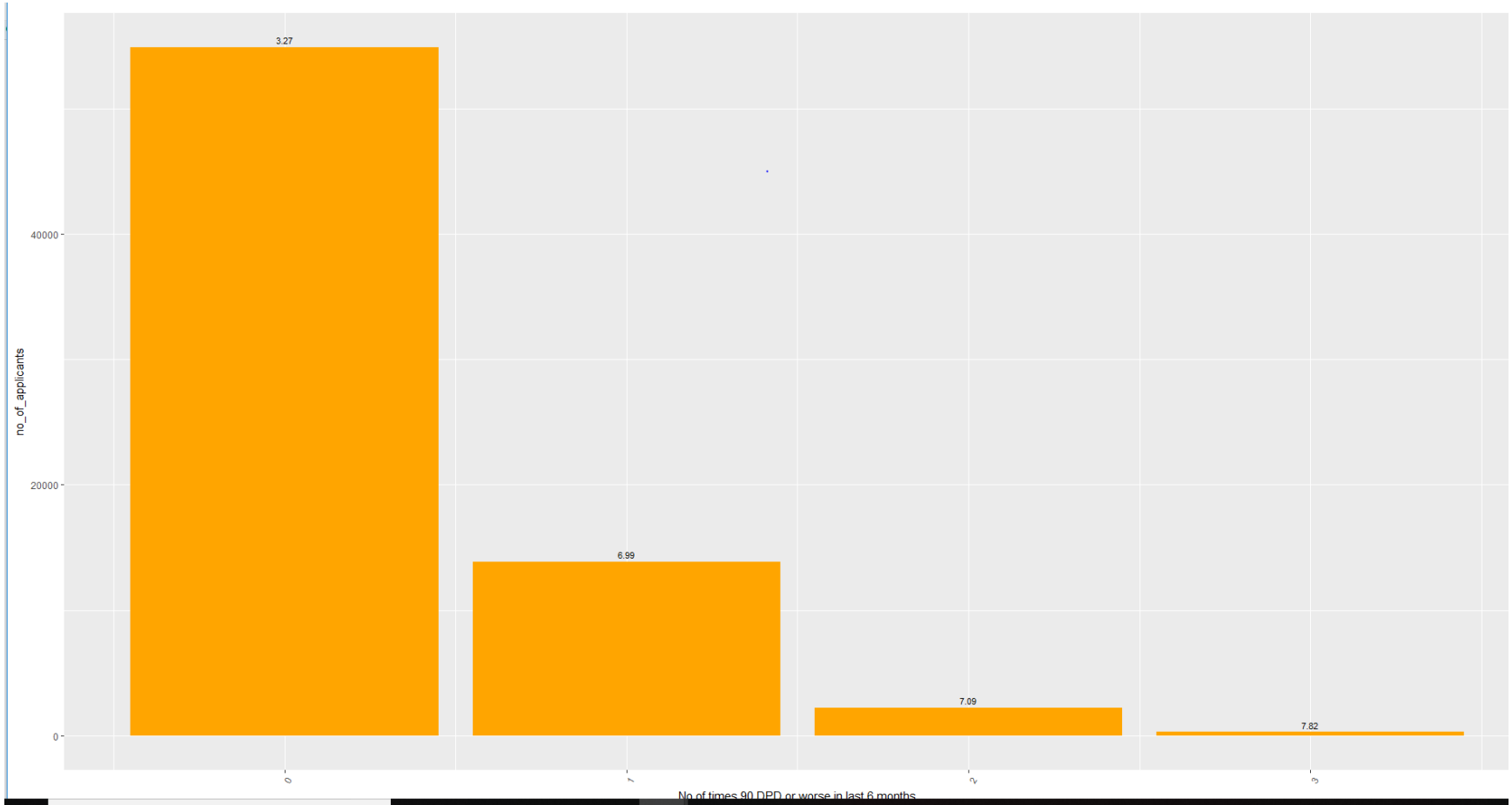


2) No of years in current residence-



If the value is 2-5 then the defaulter rate is quiet high

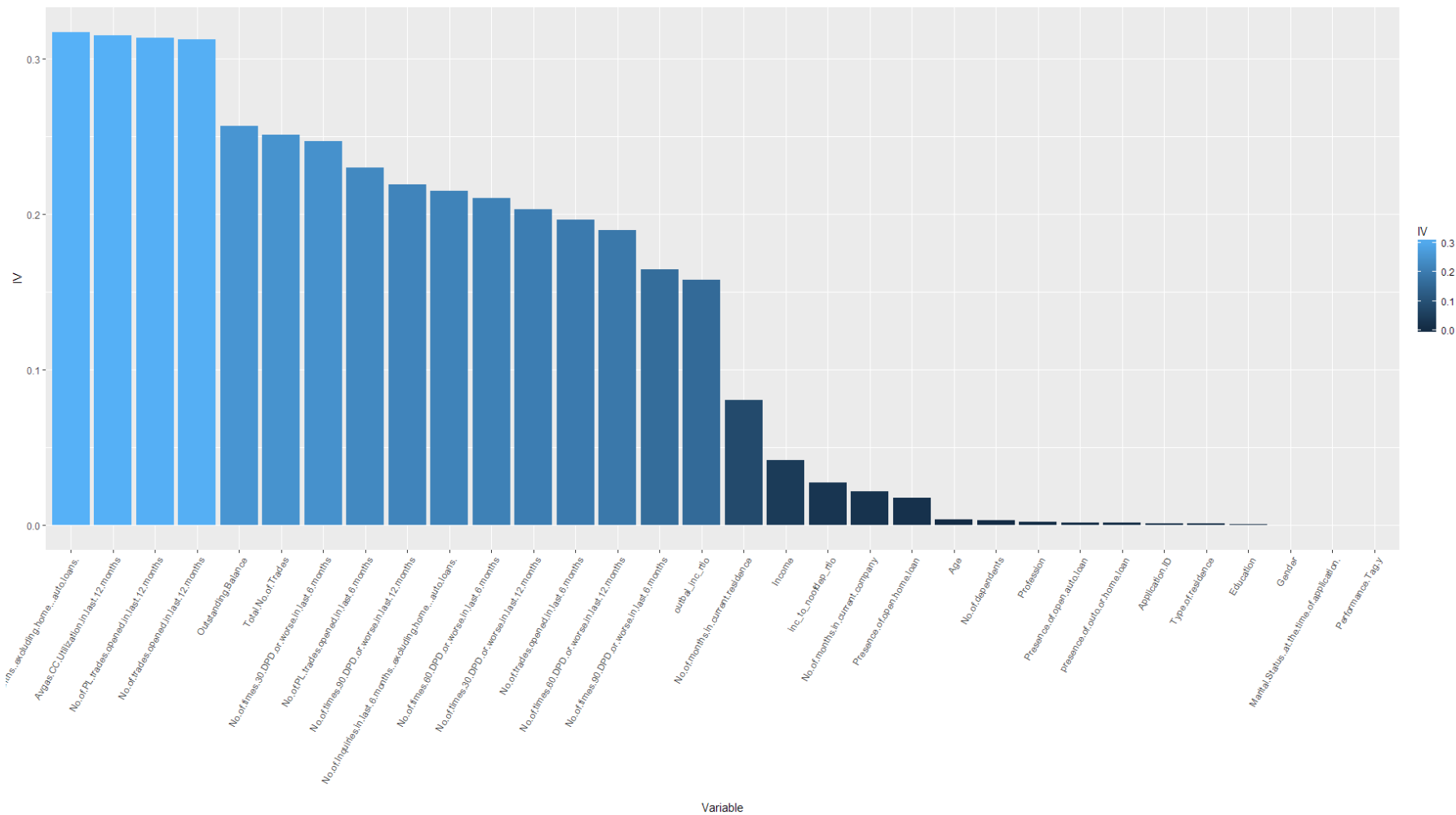
3) No.of.times.90.DPD.or.worse.in.last.6.months



Clearly as the no is increasing the defaulter rate is increasing

1] WOE and IV calculation:-

We calculate WOE and IV for each of the independent variable and created the following bar graph:-



To make inferences from Information Value we use the following rules:-

Information Value	Variable Productiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power

Conclusion:-

Variables with strong predictive power:-

- No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.
- Avgas.CC.Utilization.in.last.12.months
- No.of.PL.trades.opened.in.last.12.months
- No.of.trades.opened.in.last.12.months

Variables with Medium predictive power:-

- Outstanding.Balance
- Total.No.of.Trades
- No.of.times.30.DPD.or.worse.in.last.6.months
- No.of.PL.trades.opened.in.last.6.months
- No.of.times.90.DPD.or.worse.in.last.12.months
- No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.
- No.of.times.60.DPD.or.worse.in.last.6.months
- No.of.times.30.DPD.or.worse.in.last.12.months
- No.of.trades.opened.in.last.6.months
- No.of.times.60.DPD.or.worse.in.last.12.months
- No.of.times.90.DPD.or.worse.in.last.6.months
- outbal_inc_rtio

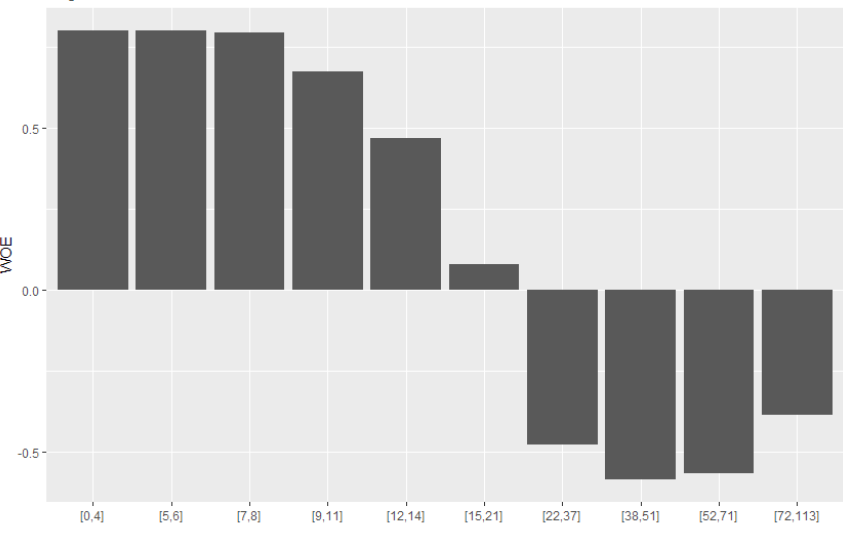
Rest all are of low predictive power.

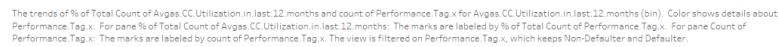
Let’s check WOE plots and bar plots for the above variables:-

1. No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.

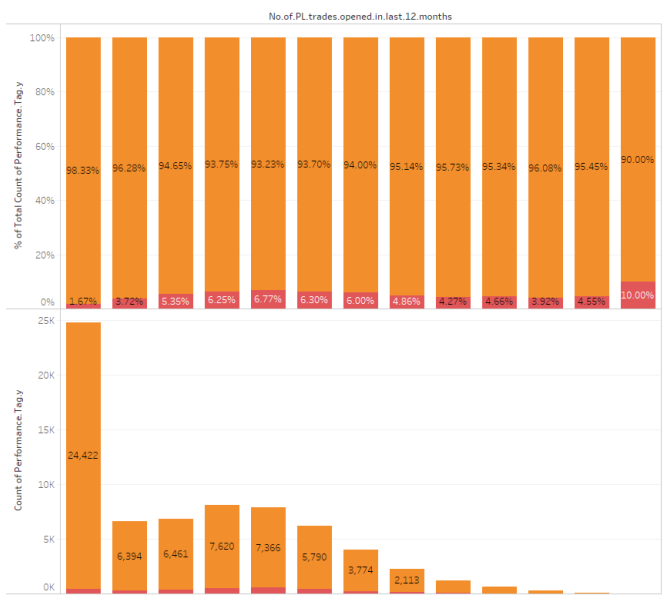
<div><p>No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.</p></div> <div><p>No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.</p><table><tr><th>Bin</th><th>Non-Defaulter (%)</th><th>Defaulter (%)</th></tr><tr><td>0</td><td>98.62%</td><td>1.38%</td></tr><tr><td>1</td><td>95.88%</td><td>4.12%</td></tr><tr><td>2</td><td>95.18%</td><td>4.82%</td></tr><tr><td>3</td><td>95.06%</td><td>4.94%</td></tr><tr><td>4</td><td>94.66%</td><td>5.34%</td></tr><tr><td>5</td><td>92.72%</td><td>7.28%</td></tr><tr><td>6</td><td>93.16%</td><td>6.84%</td></tr><tr><td>7</td><td>93.01%</td><td>6.99%</td></tr><tr><td>8</td><td>93.99%</td><td>6.01%</td></tr><tr><td>9</td><td>96.00%</td><td>4.00%</td></tr><tr><td>10</td><td>94.43%</td><td>5.57%</td></tr><tr><td>11</td><td>95.69%</td><td>4.31%</td></tr><tr><td>12</td><td>95.73%</td><td>4.27%</td></tr><tr><td>13</td><td>96.70%</td><td>3.30%</td></tr><tr><td>14</td><td>95.84%</td><td>4.16%</td></tr><tr><td>15</td><td>96.65%</td><td>3.35%</td></tr></table><p>% of Total Count of Performance.Tag.y and count of Performance.Tag.x for each No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.. Color shows details about Performance.Tag.x. The view is filtered on Performance.Tag.x, which keeps Non-Defaulter and Defaulter.</p></div>	Bin	Non-Defaulter (%)	Defaulter (%)	0	98.62%	1.38%	1	95.88%	4.12%	2	95.18%	4.82%	3	95.06%	4.94%	4	94.66%	5.34%	5	92.72%	7.28%	6	93.16%	6.84%	7	93.01%	6.99%	8	93.99%	6.01%	9	96.00%	4.00%	10	94.43%	5.57%	11	95.69%	4.31%	12	95.73%	4.27%	13	96.70%	3.30%	14	95.84%	4.16%	15	96.65%	3.35%	<div>Conclusions:-</div> <ul style="list-style-type: none">• There are no defaulters for 0 enquires• % of defaulters are more between 5 enquires to 8 enquires, although count of such enquiries are less.• There are very less number of enquiries between 1 and [9 to 15].These may be the genuine customers who enquirer thoroughly before taking the loan.• We need to target the window of [2 to 8] to predict maximum number of defaulters
Bin	Non-Defaulter (%)	Defaulter (%)																																																		
0	98.62%	1.38%																																																		
1	95.88%	4.12%																																																		
2	95.18%	4.82%																																																		
3	95.06%	4.94%																																																		
4	94.66%	5.34%																																																		
5	92.72%	7.28%																																																		
6	93.16%	6.84%																																																		
7	93.01%	6.99%																																																		
8	93.99%	6.01%																																																		
9	96.00%	4.00%																																																		
10	94.43%	5.57%																																																		
11	95.69%	4.31%																																																		
12	95.73%	4.27%																																																		
13	96.70%	3.30%																																																		
14	95.84%	4.16%																																																		
15	96.65%	3.35%																																																		

2. Avgas.CC.Utilization.in.last.12.months

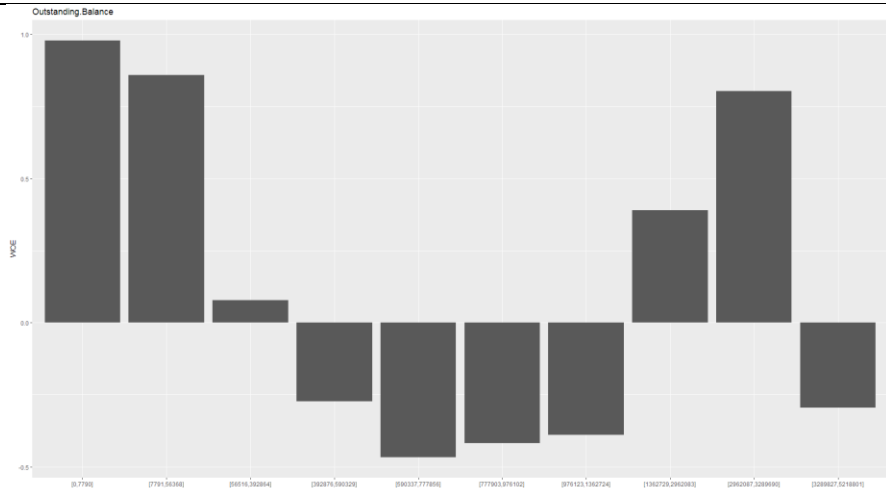
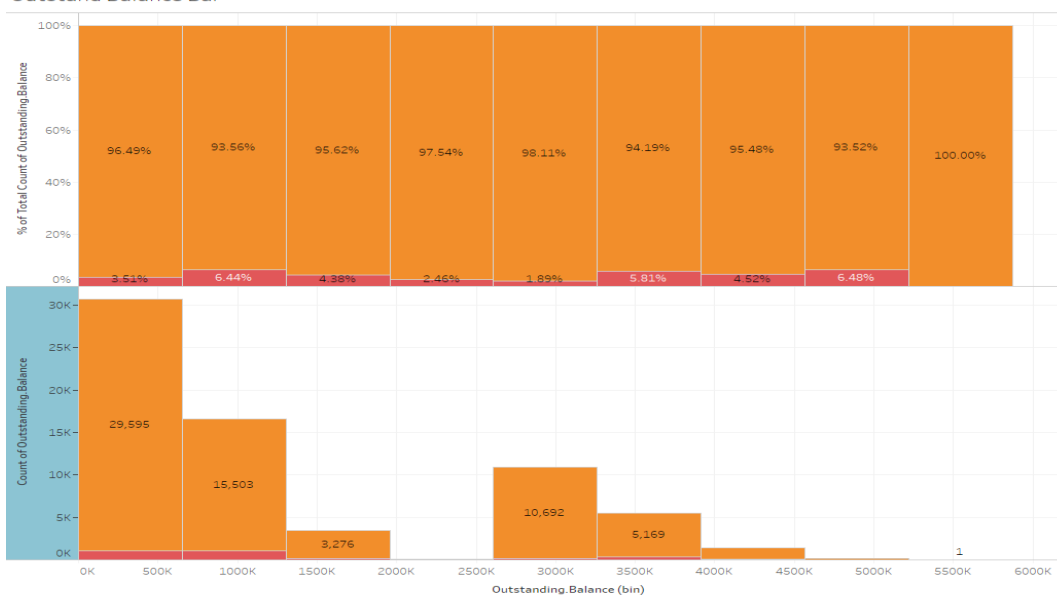
Plots	Conclusions:-
<div><p>Avgas.CC.Utilization.in.last.12.months</p></div>	<ul style="list-style-type: none">Average utilization in range [22 to 133] have high likely hood of being default.Whereas Average utilization below 22 has positive woe value and likelihood of default is less.

[illegible]

4. No.of.trades.opened.in.last.12.months

Plots	Conclusions:-
<div><p>No.of.trades.opened.in.last.12.months</p><p>No.of.PL.trades.opened.in.last.12.months</p><p>% of Total Count of Performance.Tag.y and count of Performance.Tag.y for each No of PL trades opened in last 12 months. Color shows details about Performance.Tag.x. For pane % of Total Count of Performance.Tag.y: The marks are labeled by % of Total Count of Performance.Tag.x. For pane Count of Performance.Tag.y: The marks are labeled by count of Performance.Tag.x. The view is filtered on Performance.Tag.x, which keeps Non-Defaulter and Defaulter.</p></div>	<ul style="list-style-type: none">• All records with No of trades greater > 4 have high likelihood of getting default• The highest likelihood of getting default is highest at 12 , but the count of such records is very less• Most of the defaulters lie in range of [4 to 6] where we have maximum number of records for persons who have Opened Number of trades.

5. Outstanding.Balance

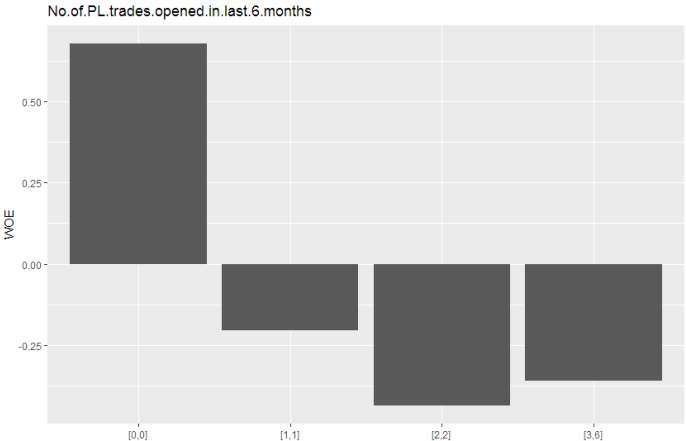
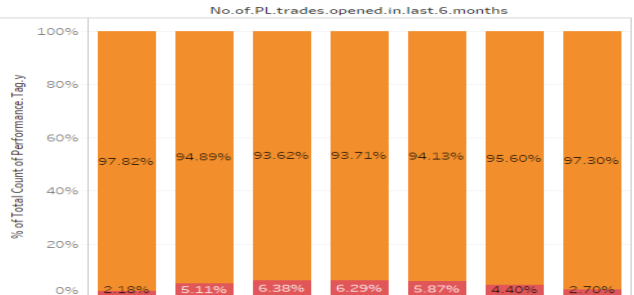
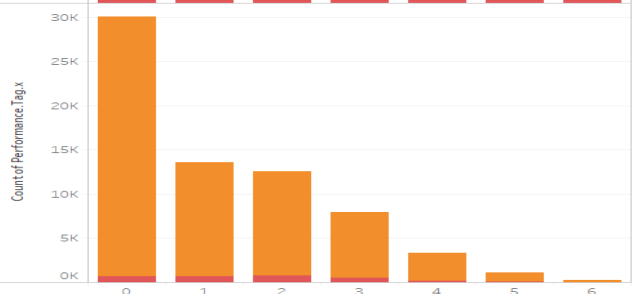
Plots	Conclusions:-
<div><p>Outstanding Balance</p><p>Outstand Balance Bar</p><p>The trends of % of Total Count of Outstanding.Balance and count of Outstanding.Balance for Outstanding.Balance (bin). Color shows details about Performance.Tag.x. For pane % of Total Count of Outstanding.Balance: The marks are labeled by % of Total Count of Performance.Tag.x. For pane Count of Outstanding.Balance: The marks are labeled by count of Performance.Tag.x. The view is filtered on Performance.Tag.x and % of Total Count of Outstanding.Balance. The Performance.Tag.x filter keeps Non-Defaulter and Defaulter. The % of Total Count of Outstanding.Balance filter keeps non-Null values only.</p></div>	<ul style="list-style-type: none">• Most of the defaulters lies between range [392876 to 1362724]• The likelihood of getting default is very less for outstanding balance between [1362729 to 3289690], this seems a little unusual. We observe that there are no records present in range [200000 to 250000] and very less people default between ranges (250000 to 320000).

6. Total.No.of.Trades

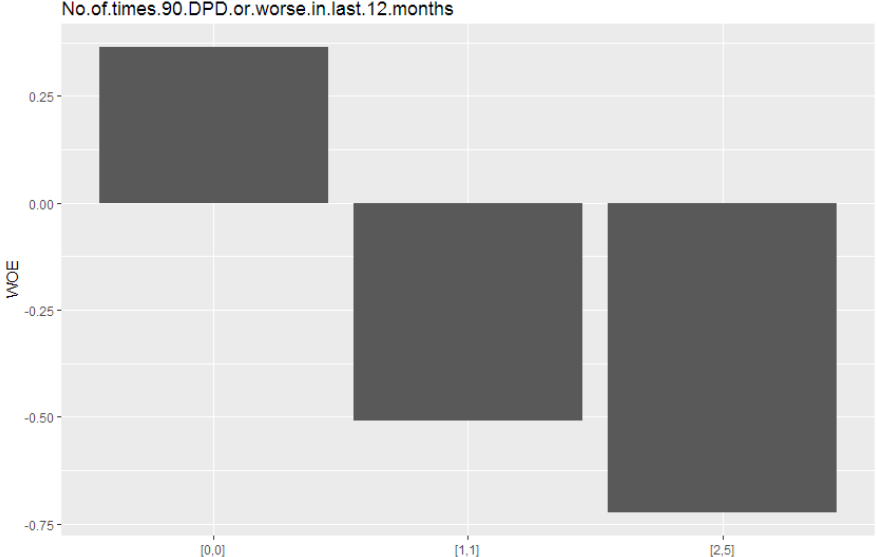
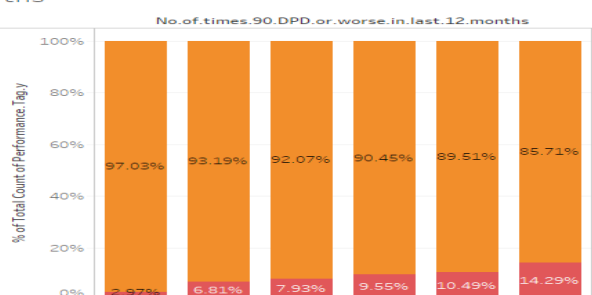
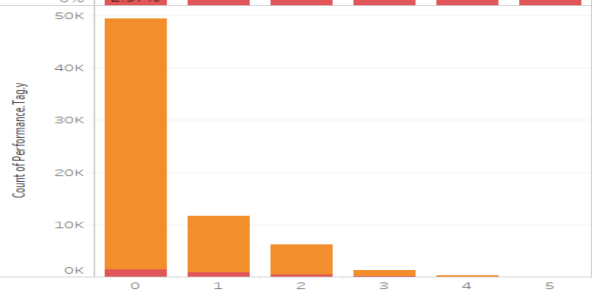


7. No.of.times.30.DPD.or.worse.in.last.6.months

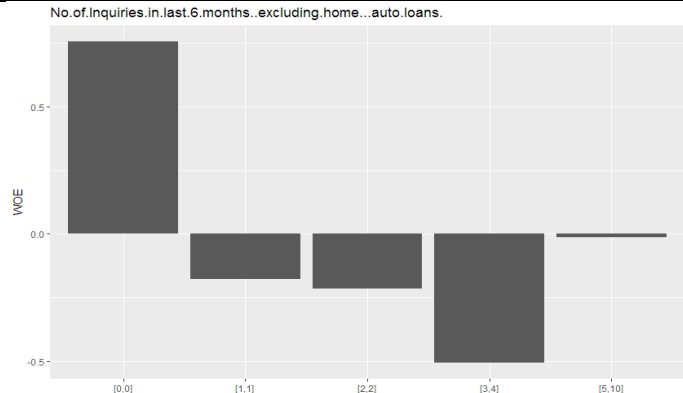
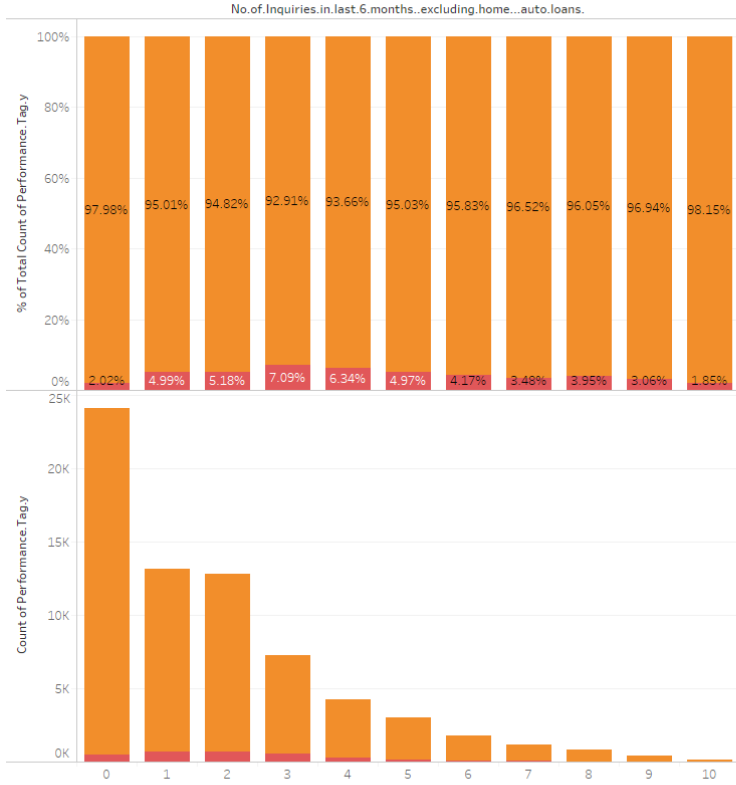


Plots	Conclusions:-
<div><div>No.of.PL.trades.opened.in.last.6.months</div><div>No.of.PL.trades.opened.in.last.6.months</div><div><div>% of Total Count of Performance.Tag.y</div><div>Count of Performance.Tag.x</div><div>% of Total Count of Performance.Tag.y and count of Performance.Tag.x for each No.of.PL.trades.opened.in.last.6.months. Color shows details about Performance.Tag.x. For pane % of Total Count of Performance.Tag.y: The marks are labeled by % of Total Count of Performance.Tag.x. The view is filtered on Performance.Tag.x, which keeps Non-Defaulter and Defaulter.</div></div></div>	

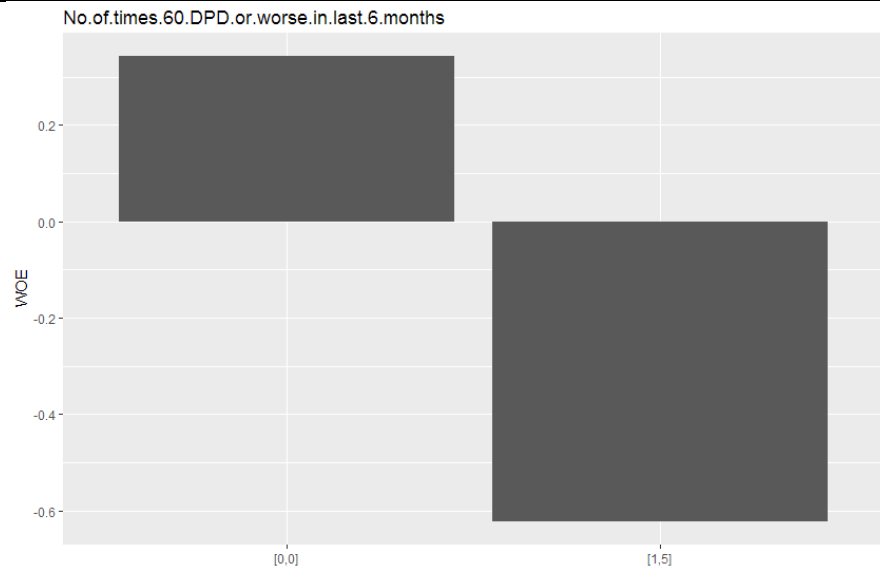

9. No.of.times.90.DPD.or.worse.in.last.12.months

Plots	Conclusions:-
<div><div>No.of.times.90.DPD.or.worse.in.last.12.months</div><div>No.of.times.90.DPD.or.worse.in.last.12.mon ths</div><div><div>% of Total Count of Performance.Tag.y</div><div>Count of Performance.Tag.y</div><div>% of Total Count of Performance.Tag.y and count of Performance.Tag.y for each No.of.times.90.DPD.or.worse.in.last.12.months. Color shows details about Performance.Tag.x. For pane % of Total Count of Performance.Tag.y: The marks are labeled by % of Total Count of Performance.Tag.x. The data is filtered on Performance.Tag.y, which ranges from 0 to 1.</div></div></div>	<ul style="list-style-type: none">• All records having 90 DPD or worst in 12 have high likelihood of getting default• Most of the records where people have gone 90 DPD or worst in 12 lies between rage of [1 to 2]• People with '0' 90 DPD or worst have least likelihood of getting default.

10. No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.

Plots	Conclusions:-																																				
<div><p>No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.</p></div> <div><p>No.of.Inquiries.in.last.6.months..excluding.home...auto.loans. Bar</p><table><tr><th>No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.</th><th>Non-Defaulter (%)</th><th>Defaulter (%)</th></tr><tr><td>0</td><td>97.98%</td><td>2.02%</td></tr><tr><td>1</td><td>95.01%</td><td>4.99%</td></tr><tr><td>2</td><td>94.82%</td><td>5.18%</td></tr><tr><td>3</td><td>92.91%</td><td>7.09%</td></tr><tr><td>4</td><td>93.66%</td><td>6.34%</td></tr><tr><td>5</td><td>95.03%</td><td>4.97%</td></tr><tr><td>6</td><td>95.83%</td><td>4.17%</td></tr><tr><td>7</td><td>96.52%</td><td>3.48%</td></tr><tr><td>8</td><td>96.05%</td><td>3.95%</td></tr><tr><td>9</td><td>96.94%</td><td>3.06%</td></tr><tr><td>10</td><td>98.15%</td><td>1.85%</td></tr></table><p>% of Total Count of Performance.Tag.y and count of Performance.Tag.y for each No.of.Inquiries.in.last.6.months..excluding.home...auto.loans. Color shows details about Performance.Tag.x. For pane % of Total Count of Performance.Tag.y: The marks are labeled by % of Total Count of Performance.Tag.x. The view is filtered on Performance.Tag.x, which keeps Non-Defaulter and Defaulter.</p></div>	No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	Non-Defaulter (%)	Defaulter (%)	0	97.98%	2.02%	1	95.01%	4.99%	2	94.82%	5.18%	3	92.91%	7.09%	4	93.66%	6.34%	5	95.03%	4.97%	6	95.83%	4.17%	7	96.52%	3.48%	8	96.05%	3.95%	9	96.94%	3.06%	10	98.15%	1.85%	<ul style="list-style-type: none">People making enquiries between [3 to 4] have high likelihood of getting default
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	Non-Defaulter (%)	Defaulter (%)																																			
0	97.98%	2.02%																																			
1	95.01%	4.99%																																			
2	94.82%	5.18%																																			
3	92.91%	7.09%																																			
4	93.66%	6.34%																																			
5	95.03%	4.97%																																			
6	95.83%	4.17%																																			
7	96.52%	3.48%																																			
8	96.05%	3.95%																																			
9	96.94%	3.06%																																			
10	98.15%	1.85%																																			

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

Plots	Conclusions:-															
<div><p>No.of.times.60.DPD.or.worse.in.last.6.months</p></div>	12. All people with No.of.times.60.DPD.or.worse.in.last.6.months Value greater than 1 have high likely hood of getting default															
<div><p>No.of.times.90.DPD.or.worse.in .last.6.months</p><table><tr><th>No.of.times.90.DPD.or.worse.in .last.6.months</th><th>Non-Defaulter (%)</th><th>Defaulter (%)</th></tr><tr><td>0</td><td>96.74%</td><td>3.26%</td></tr><tr><td>1</td><td>92.65%</td><td>7.35%</td></tr><tr><td>2</td><td>91.05%</td><td>8.95%</td></tr><tr><td>3</td><td>88.89%</td><td>11.11%</td></tr></table><p>% of Total Count of Performance.Tag.y and count of Performance.Tag.y for each No.of.times.90.DPD.or.worse.in .last.6.months. Color shows details about Performance.Tag.x. For pane % of Total Count of Performance.Tag.y: The marks are labeled by % of Total Count of Performance.Tag.x. The view is filtered on Performance.Tag.x, which keeps Non-Defaulter and Defaulter.</p></div>	No.of.times.90.DPD.or.worse.in .last.6.months	Non-Defaulter (%)	Defaulter (%)	0	96.74%	3.26%	1	92.65%	7.35%	2	91.05%	8.95%	3	88.89%	11.11%	
No.of.times.90.DPD.or.worse.in .last.6.months	Non-Defaulter (%)	Defaulter (%)														
0	96.74%	3.26%														
1	92.65%	7.35%														
2	91.05%	8.95%														
3	88.89%	11.11%														

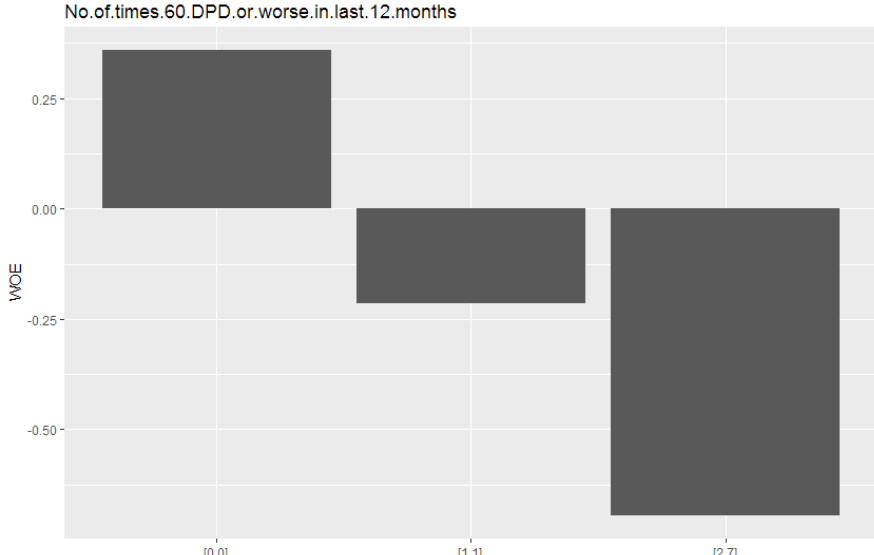
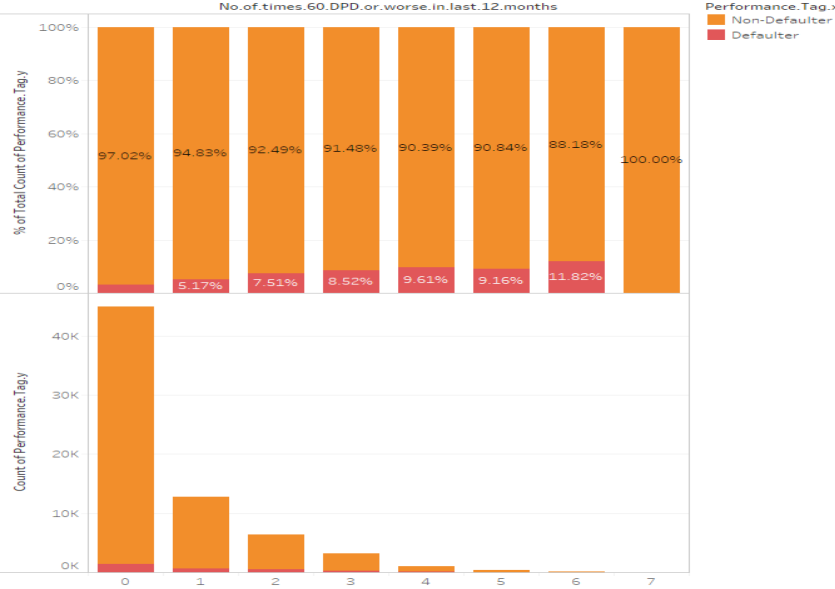
13. No.of.times.30.DPD.or.worse.in.last.12.months

Plots	Conclusions:-																																	
<div><p>No.of.times.30.DPD.or.worse.in.last.12.months</p></div> <div><p>No.of.times.30.DPD.or.worse.in.last.12.months</p><table><tr><th>No.of.times.30.DPD.or.worse.in.last.12.months</th><th>Non-Defaulter (%)</th><th>Defaulter (%)</th></tr><tr><td>0</td><td>97.10%</td><td>2.90%</td></tr><tr><td>1</td><td>95.48%</td><td>4.52%</td></tr><tr><td>2</td><td>92.62%</td><td>7.38%</td></tr><tr><td>3</td><td>91.58%</td><td>8.42%</td></tr><tr><td>4</td><td>91.03%</td><td>8.97%</td></tr><tr><td>5</td><td>89.72%</td><td>10.28%</td></tr><tr><td>6</td><td>89.70%</td><td>10.30%</td></tr><tr><td>7</td><td>89.62%</td><td>10.38%</td></tr><tr><td>8</td><td>91.30%</td><td>8.70%</td></tr><tr><td>9</td><td>100.00%</td><td>0.00%</td></tr></table><p>% of Total Count of Performance.Tag.y and count of Performance.Tag.y for each No.of.times.30.DPD.or.worse.in.last.12.months. Color shows details about Performance.Tag.x. For pane % of Total Count of Performance.Tag.y: The marks are labeled by % of Total Count of Performance.Tag.x. The view is filtered on Performance.Tag.x, which keeps Non-Defaulter and Defaulter.</p></div>	No.of.times.30.DPD.or.worse.in.last.12.months	Non-Defaulter (%)	Defaulter (%)	0	97.10%	2.90%	1	95.48%	4.52%	2	92.62%	7.38%	3	91.58%	8.42%	4	91.03%	8.97%	5	89.72%	10.28%	6	89.70%	10.30%	7	89.62%	10.38%	8	91.30%	8.70%	9	100.00%	0.00%	<ul style="list-style-type: none">• Values between [3 to 9] have high likelihood of getting default• We have very less number of records with values 9 so the same can be capped with category 8, if we are generating our model using original data• As the number of time 30 DPD increases the likelihood of being default increases.
No.of.times.30.DPD.or.worse.in.last.12.months	Non-Defaulter (%)	Defaulter (%)																																
0	97.10%	2.90%																																
1	95.48%	4.52%																																
2	92.62%	7.38%																																
3	91.58%	8.42%																																
4	91.03%	8.97%																																
5	89.72%	10.28%																																
6	89.70%	10.30%																																
7	89.62%	10.38%																																
8	91.30%	8.70%																																
9	100.00%	0.00%																																

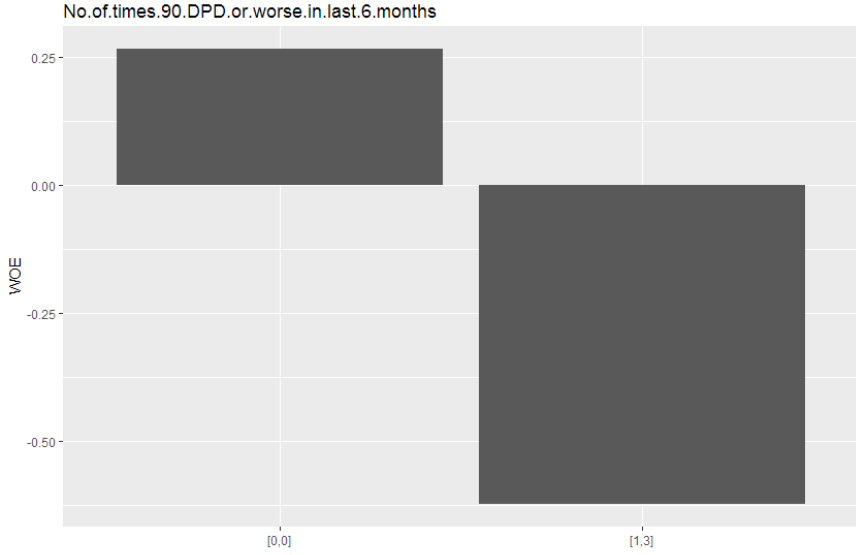
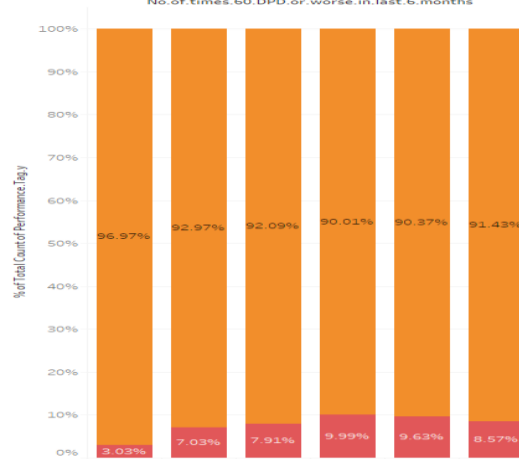
14. No.of.trades.opened.in.last.6.months

Plots	Conclusions:-																																										
<div><p>No.of.trades.opened.in.last.6.months</p></div> <div><p>No.of.trades.opened.in.last.6.months</p><table><tr><th>No.of.trades.opened.in.last.6.months</th><th>Non-Defaulter (%)</th><th>Defaulter (%)</th></tr><tr><td>0</td><td>97.95%</td><td>2.05%</td></tr><tr><td>1</td><td>97.35%</td><td>2.65%</td></tr><tr><td>2</td><td>94.72%</td><td>5.28%</td></tr><tr><td>3</td><td>93.64%</td><td>6.36%</td></tr><tr><td>4</td><td>93.10%</td><td>6.90%</td></tr><tr><td>5</td><td>94.26%</td><td>5.74%</td></tr><tr><td>6</td><td>95.33%</td><td>4.67%</td></tr><tr><td>7</td><td>96.06%</td><td>3.94%</td></tr><tr><td>8</td><td>95.49%</td><td>4.51%</td></tr><tr><td>9</td><td>96.60%</td><td>3.40%</td></tr><tr><td>10</td><td>97.06%</td><td>2.94%</td></tr><tr><td>11</td><td>96.92%</td><td>3.08%</td></tr><tr><td>12</td><td>100.00%</td><td>0.00%</td></tr></table><p>% of Total Count of Performance.Tag.x and count of Performance.Tag.y for each No.of.trades.opened.in.last.6.months. Color shows details about Performance.Tag.x. For pane % of Total Count of Performance.Tag.x: The marks are labeled by % of Total Count of Performance.Tag.x. The data is filtered on Performance.Tag.y, which ranges from 0 to 1.</p></div>	No.of.trades.opened.in.last.6.months	Non-Defaulter (%)	Defaulter (%)	0	97.95%	2.05%	1	97.35%	2.65%	2	94.72%	5.28%	3	93.64%	6.36%	4	93.10%	6.90%	5	94.26%	5.74%	6	95.33%	4.67%	7	96.06%	3.94%	8	95.49%	4.51%	9	96.60%	3.40%	10	97.06%	2.94%	11	96.92%	3.08%	12	100.00%	0.00%	<ul style="list-style-type: none">• Records having no of trades opened in last 6 months 3 and 4 have high likely hood of being default.• If we are using original data for modeling than we can cap the data at 6 , as the number of records above it are very less.• Records having value greater than '2' have more likelihood of being default compared to values having value less than '2'.
No.of.trades.opened.in.last.6.months	Non-Defaulter (%)	Defaulter (%)																																									
0	97.95%	2.05%																																									
1	97.35%	2.65%																																									
2	94.72%	5.28%																																									
3	93.64%	6.36%																																									
4	93.10%	6.90%																																									
5	94.26%	5.74%																																									
6	95.33%	4.67%																																									
7	96.06%	3.94%																																									
8	95.49%	4.51%																																									
9	96.60%	3.40%																																									
10	97.06%	2.94%																																									
11	96.92%	3.08%																																									
12	100.00%	0.00%																																									

15. No.of.times.60.DPD.or.worse.in.last.12.months

Plots	Conclusions:-																											
<div><p>No.of.times.60.DPD.or.worse.in.last.12.months</p></div> <div><p>No.of.times.60.DPD.or.worse.in.last.12.months</p><table><tr><th>No.of.times.60.DPD.or.worse.in.last.12.months</th><th>Non-Defaulter (%)</th><th>Defaulter (%)</th></tr><tr><td>0</td><td>97.02%</td><td>3.17%</td></tr><tr><td>1</td><td>94.83%</td><td>5.17%</td></tr><tr><td>2</td><td>92.49%</td><td>7.51%</td></tr><tr><td>3</td><td>91.46%</td><td>8.52%</td></tr><tr><td>4</td><td>90.39%</td><td>9.61%</td></tr><tr><td>5</td><td>90.84%</td><td>9.16%</td></tr><tr><td>6</td><td>88.18%</td><td>11.82%</td></tr><tr><td>7</td><td>100.00%</td><td>0.00%</td></tr></table><p>% of Total Count of Performance.Tag.y and count of Performance.Tag.y for each No.of.times.60.DPD.or.worse.in.last.12.months. Color shows details about Performance.Tag.x. For pane % of Total Count of Performance.Tag.y: The marks are labeled by % of Total Count of Performance.Tag.x. The view is filtered on Performance.Tag.x, which keeps Non-Defaulter and Defaulter.</p></div>	No.of.times.60.DPD.or.worse.in.last.12.months	Non-Defaulter (%)	Defaulter (%)	0	97.02%	3.17%	1	94.83%	5.17%	2	92.49%	7.51%	3	91.46%	8.52%	4	90.39%	9.61%	5	90.84%	9.16%	6	88.18%	11.82%	7	100.00%	0.00%	<ul style="list-style-type: none">• Values between [3 to 9] have high likelihood of getting default• We have very less number of records with values 9 so the same can be capped with category 8, if we are generating our model using original data.• As the number of time 30 DPD increases the likelihood of being default increases.
No.of.times.60.DPD.or.worse.in.last.12.months	Non-Defaulter (%)	Defaulter (%)																										
0	97.02%	3.17%																										
1	94.83%	5.17%																										
2	92.49%	7.51%																										
3	91.46%	8.52%																										
4	90.39%	9.61%																										
5	90.84%	9.16%																										
6	88.18%	11.82%																										
7	100.00%	0.00%																										

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

Plots	Conclusions:-																					
<div><p>No.of.times.90.DPD.or.worse.in.last.6.months</p></div> <div><p>No.of.times.60.DPD.or.worse.in.last.6.months</p><table><tr><th>No.of.times.60.DPD.or.worse.in.last.6.months</th><th>Non-Defaulter (%)</th><th>Defaulter (%)</th></tr><tr><td>0</td><td>96.97%</td><td>3.03%</td></tr><tr><td>1</td><td>92.97%</td><td>7.03%</td></tr><tr><td>2</td><td>92.09%</td><td>7.91%</td></tr><tr><td>3</td><td>90.01%</td><td>9.99%</td></tr><tr><td>4</td><td>90.37%</td><td>9.63%</td></tr><tr><td>5</td><td>91.43%</td><td>8.57%</td></tr></table><p>% of Total Count of Performance.Tag.y for each No.of.times.60.DPD.or.worse.in.last.6.months. Color shows details about Performance.Tag.x. The marks are labeled by % of Total Count of Performance.Tag.x. The view is filtered on Performance.Tag.x, which keeps Non-Defaulter and Defaulter.</p></div>	No.of.times.60.DPD.or.worse.in.last.6.months	Non-Defaulter (%)	Defaulter (%)	0	96.97%	3.03%	1	92.97%	7.03%	2	92.09%	7.91%	3	90.01%	9.99%	4	90.37%	9.63%	5	91.43%	8.57%	<ul style="list-style-type: none">• All records with greater than '0' DPD have high likelihood of being default
No.of.times.60.DPD.or.worse.in.last.6.months	Non-Defaulter (%)	Defaulter (%)																				
0	96.97%	3.03%																				
1	92.97%	7.03%																				
2	92.09%	7.91%																				
3	90.01%	9.99%																				
4	90.37%	9.63%																				
5	91.43%	8.57%																				

Plots	Conclusions:-
<div><div>outbal_inc_rtio</div></div> <div><div>Sheet 36</div><p>The trends of % of Total Count of Performance.Tag.y and count of Performance.Tag.y for outbal_inc_rtio (bin). Color shows details about Performance.Tag.x. The marks are labeled by % of Total Count of Performance.Tag.x. The view is filtered on Performance.Tag.x, which keeps Non-Defaulter and Defaulter.</p></div>	

Modelling

1. Model was built on data frame contains WOE values for all the independent variables.
2. As we observed above, there are only 2% of defaulters in the dataset. This imbalanced classification needs to be handled before going into modelling, as model will not be able to learn appropriately. We used synthetic minority oversampling technique (SMOTE) for imbalanced classification.
3. N fold cross validation was done for model stability.
4. Starting with logistic regression cleaning variables based on Aic, Vif and p values will give us a refined list of variables to go ahead.
5. To improve the performance of the model we tried more complex models. Decision Tree and Random forest with variable importance plot will give us the idea of important variables as well.
6. The best results were achieved from logistic regression model.

Model Evaluation criteria:

1. **Discriminatory power:** KS statistic, ROC curve, sensitivity and specificity, rank-ordering, etc. can be used to check the discriminatory power of the model.
2. **Accuracy:** Accuracy can be measured using the confusion matrix. However, Since the cost of miss-classification of defaulter is more than the miss-classification of non-defaulters accuracy alone can't be a major factor in deciding the performance of the model. As one of the class has higher importance (defaulter) so considering sensitivity or specificity will help.
3. **Stability:** Check the population across each variable and the fraction of high/med/low-risk customers
The predictive pattern
4. **Reject Inferencing:** We will run the final model on rejected application dataset and see the performance of our model.

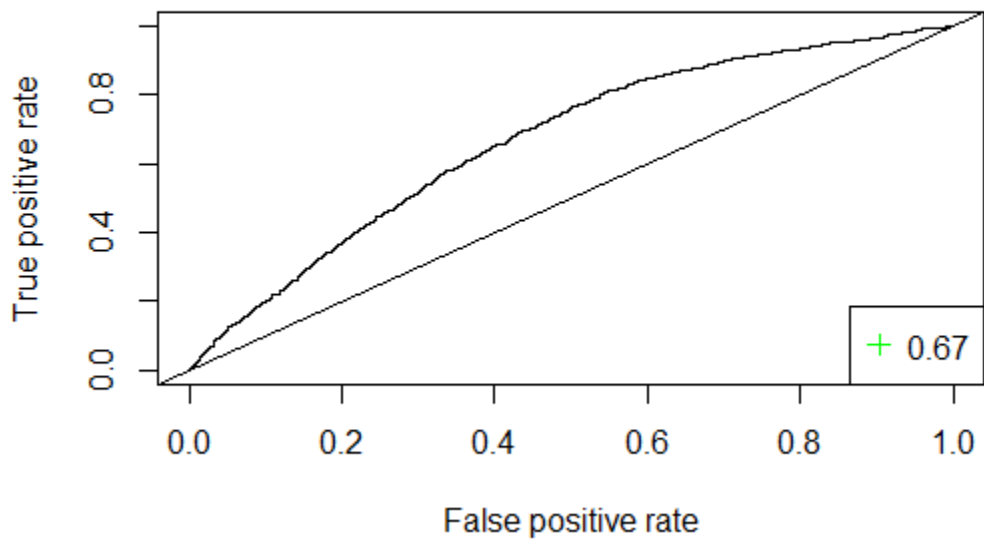
Summary of the total Models generated:-

		Cross validation	Test Data Set				Validatin Data Set
Models		AUC	AUC	Acuracy	Specificity	Sensitivity	Specficity of Validation set
Models on original data	Logistic Regression	NA	0.66	0.6471	0.65097	0.55977	0.9661871
	Decision Tree using Information Value as split Criteria	0.829797	0.65	0.6017	0.59962	0.64943	0.9906475
	Decision Tree using gini Index Value as split Criteria	0.8145131	0.66	0.6106	0.60922	0.64253	0.9697842
	Random Forest	0.9164863	0.65	0.553	0.68736	0.54705	0.9179856
	Gradient Boosting	0.883344	0.66	0.5481	0.54149	0.6977	0.9438849
Models on woe data	Logistic Regression	NA	0.67	0.534	0.52519	0.7331	0.9978134
	Decision Tree using Information Value as split Criteria	0.7424283	0.65	0.669	0.60113	0.669	0.8994169
	Decision Tree using gini Index Value as split Criteria	0.7363733	0.5	0.9578	1	0	0
	Random Forest	0.9	0.62	0.4703	0.45998	0.70396	0.9395044
	Gradient Boosting	0.6748559	0.68	0.5507	0.54274	0.73193	0.9970845

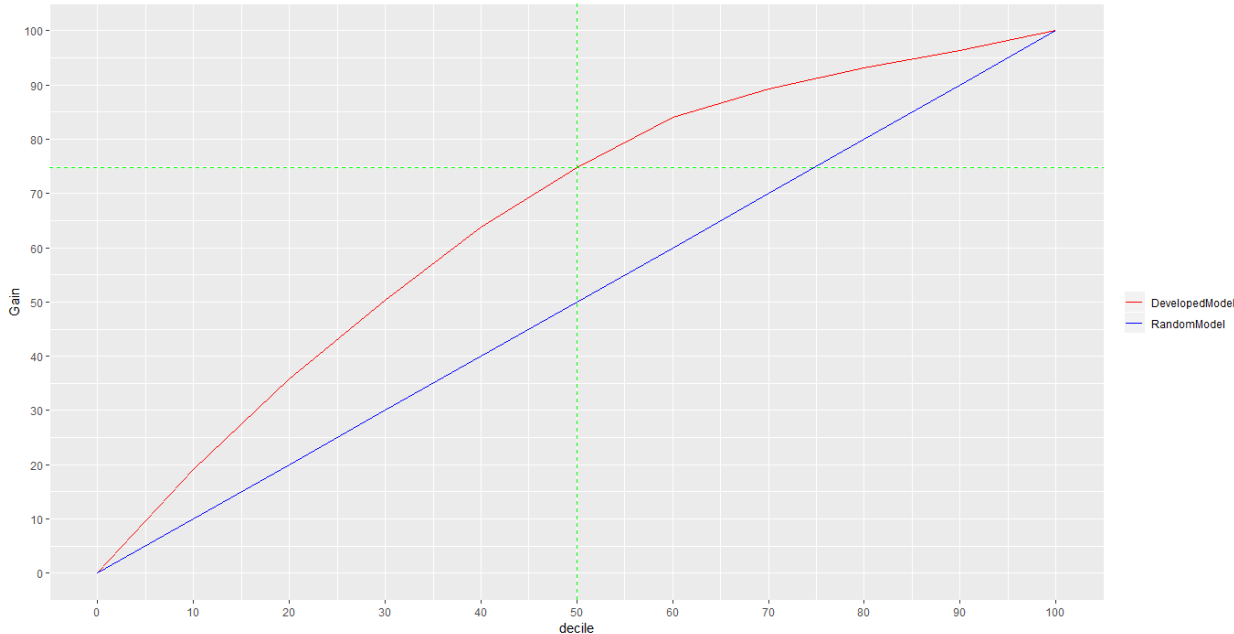
Looking at the above chart it is clear that the best model is Logistic Regression using WOE data.

Best model results

1. AUC of the final model:-



2. Gain Chart-



3. KS- Statistics-

	decile	defaulters	nondefaulters	cum_defaulters	cum_nondefaulters	Observations	percentcum_defaulters	percentcum_nondefaulters	ks_value
1	1	164	1871	164	1871	2035	19.1142191142191	9.59979476654695	9.51442434767217
2	2	143	1892	307	3763	2035	35.7808857808858	19.3073370959466	16.4735486849391
3	3	125	1910	432	5673	2035	50.3496503496504	29.1072344792201	21.2424158704302
4	4	116	1919	548	7592	2035	63.8694638694639	38.9533093894305	24.9161544800334
5	5	93	1942	641	9534	2035	74.7086247086247	48.9173935351462	25.7912311734785
6	6	80	1955	721	11489	2035	84.032634032634	58.9481785531042	25.0844554795299
7	7	45	1990	766	13479	2035	89.2773892773893	69.1585428424833	20.1188464349059
8	8	33	2002	799	15481	2035	93.1235431235431	79.4304771677783	13.6930659557648
9	9	28	2007	827	17488	2035	96.3869463869464	89.728065674705	6.6588807122414
10	10	31	2002	858	19490	2033	100	100	0

From the above table it is clear that the 74% of the defaulters are preset in first 5 deciles, hence KS-value comes to be 50%

Conclusion-

Important variable which can help to identify the defaulter correctly-

- Avgas.CC.Utilization.in.last.12.months
- Age
- No of dependents
- Income
- Profession
- No.of.months.in.current.residence
- No.of.months.in.current.company
- No.of.times.60.DPD.or.worse.in.last.12.months
- Avgas.CC.Utilization.in.last.12.months
- Outstanding.Balance
- Total.No.of.Trades
- Presence.of.open.auto.loan,

Application score card:

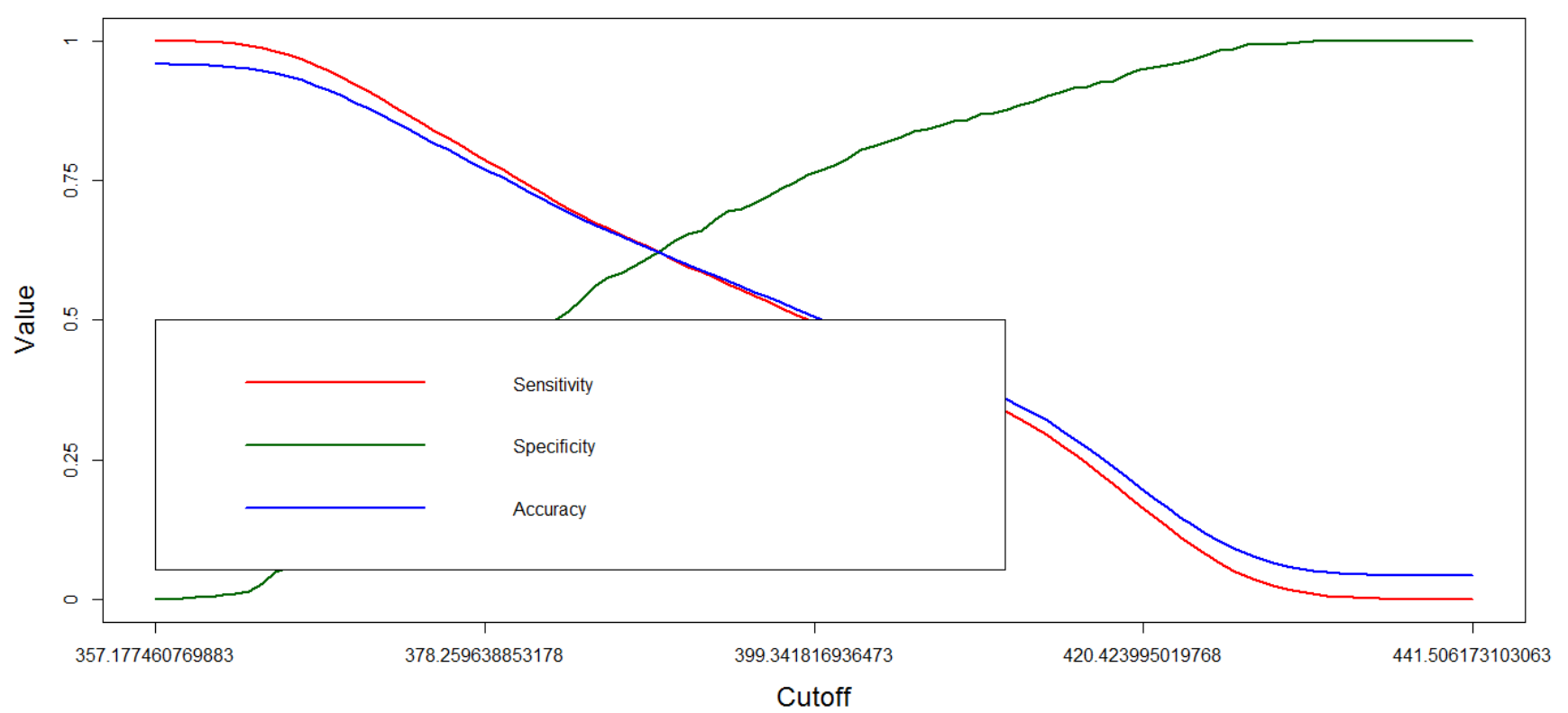
The score is the basis to decide whether to grant credit or not. Based on the final model we will create an application score card. Below formula will be used to calculate the score-
Score = Offset + (Factor * log(odds))
Factor = PDO/ln(2)
Offset = Score-(Factor*log(odds))

- Where PDO = points to double the odds = 20
- Score = scoring value for which you want to receive specific odds of the defaulter = 400
- Odds = odds of the defaulter for specific scoring value = 10
- Factor = scaling parameter calculated on the basis of formula presented above

We have PDO = 20, Base Score=400 & odds = 10

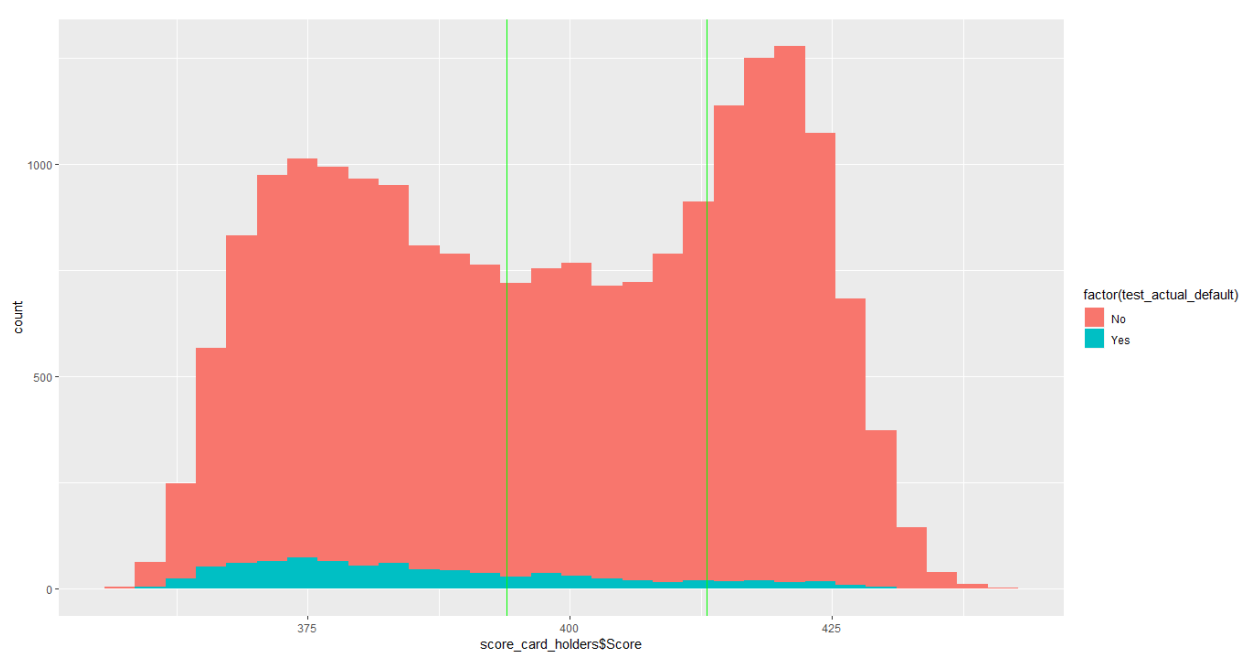
Based on the scores and threshold cutoff, we can categories the applicants as very high risk, high risk, medium risk and low risk. Once the applications are categorized as mentioned above we can measure the revenue that can be saved by not giving credit to very high and high risk customers.

Optimal Score Cutoff: After creating a score card, to identify the optimal cut-off we plot Sensitivity, Specificity and Accuracy against the calculated scores:-





Clearly looking at the above graph we observe that the cutoff score below which there are high likelihood of a customer being a default is : **394**

Rank Ordering:-



By analyzing the above Graph we can rank order the scores as follows:-



Label	Min Score	Max Score
Bad Customers	300	394
Average Customers	>394	413
Good Customers	>413	500

On Reject Inferencing we observe that 99.7% of the rejected population have score < 394 which validates our model.

Credit Loss Avoided:-

68.2% Total credit loss saved by our model = sum of outstanding balance of true positive / (total outstanding balance of true positive + false negative)