BFSI CAPSTONE PROJECT
FINAL-SUBMISSION

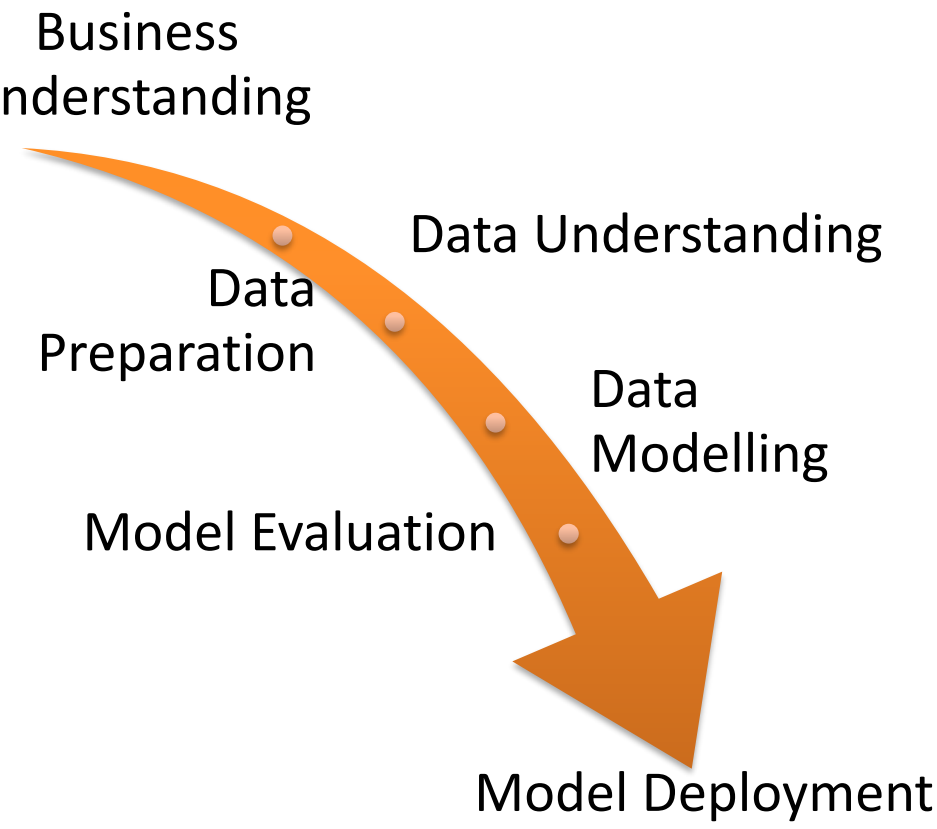
Date: - 03/18/2018

Submission By:- Satya Mishra

We will use the **Cross Industry Standard Process for Data Mining (CRISP–DM) framework** while executing this project.

Workflow (Life Cycle):-

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Data Modelling
5. Model Evaluation
6. Model Deployment



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. So the best strategy to mitigate credit risk is to 'acquire the right customers'.

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

Data Understanding

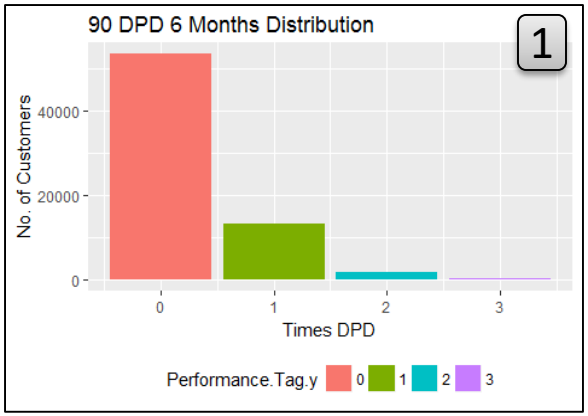
- We have two datasets that we need to merge to start our analysis.
- There are a total of 29 variables.
- Out of these 29 variables Performance Tag is the dependent variable (we need to predict the outcome of this variable).
- We will not consider Application ID in our analysis.
- There are a total of 27 independent variables on which we will build the model/s

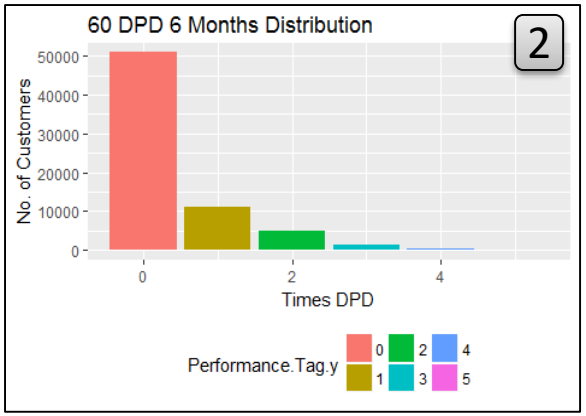
LIST OF VARIABLES	
Variables	**Description**
Application ID	Unique ID of the customers
Age	Age of customer
Gender	Gender of customer
Marital Status	Marital status of customer (at the time of application)
No of dependents	No. of childrens of customers
Income	Income of customers
Education	Education of 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
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
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")

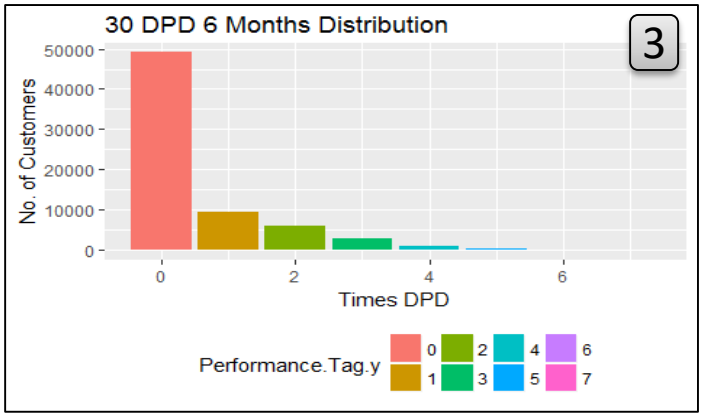
DO NOT USE

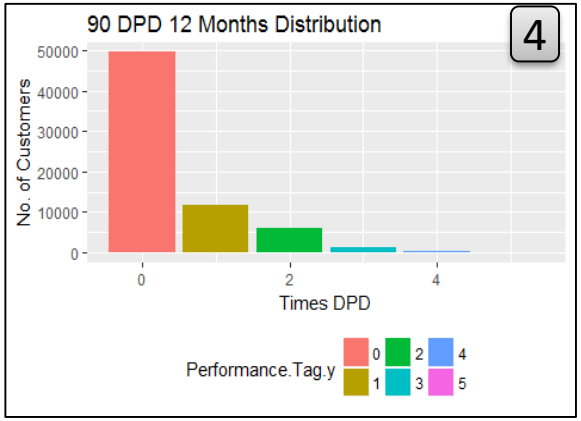
DEPENDENT VARIABLE

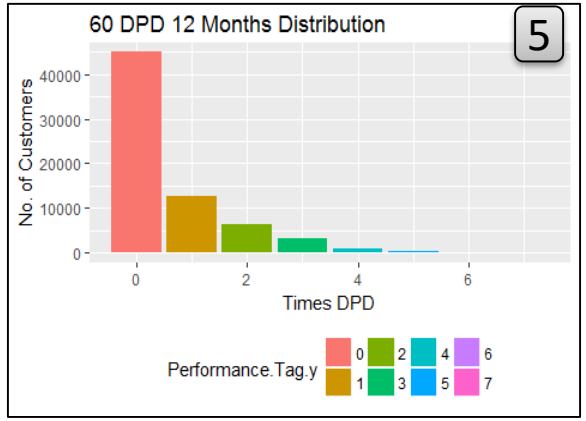
Univariate Analysis

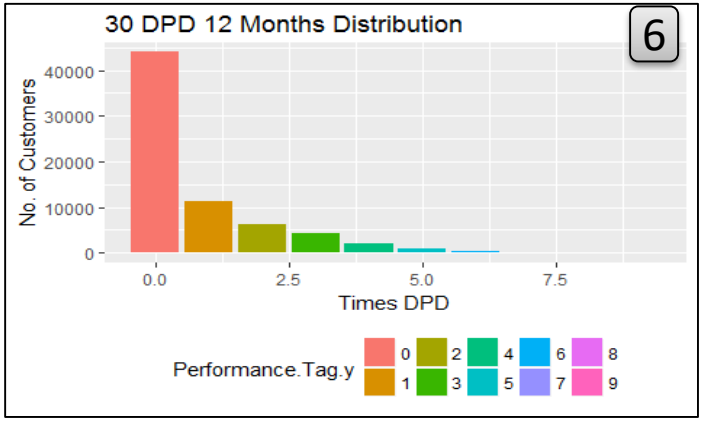


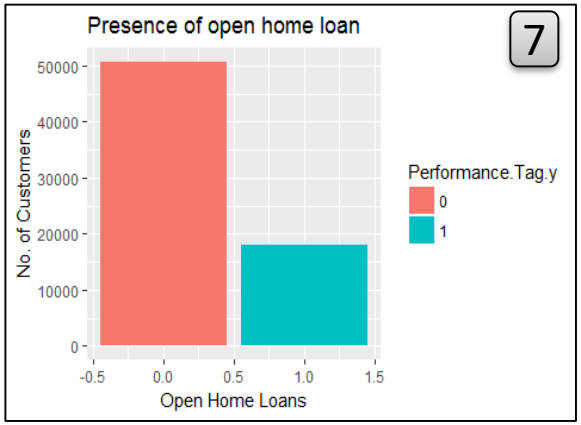


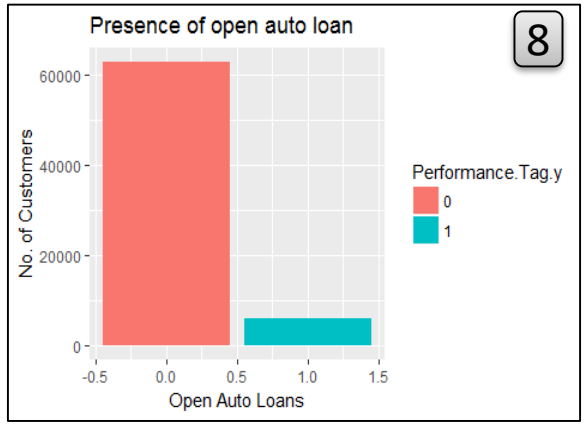


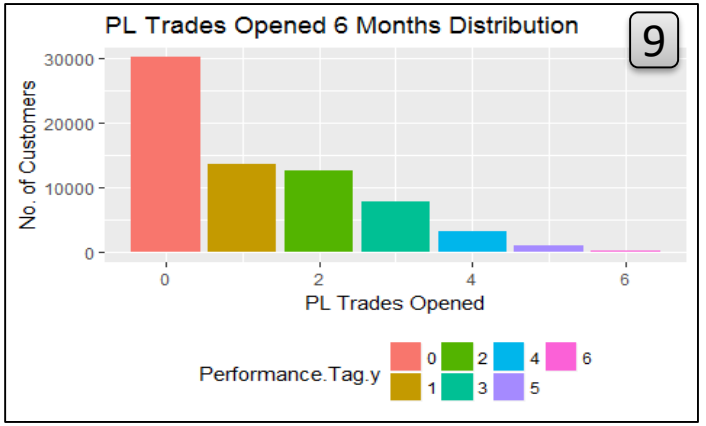












Univariate Analysis

Findings for the variables numbered 1 to 9 and 12

1

Variable 1 - No of times 90 DPD or worse in last 6 months				
90 DPD in last 6 month	0	1	2	3
Count in each bucket	53673	13205	1765	207
% Contribution	77.96%	19.18%	2.56%	0.30%

The number of buckets keep on decreasing with the increase in time frame. Number of people who have paid their dues on time has increased, when we shift the time frame for 30 DPD to 90 DPD .

2

Variable 2 - No of times 60 DPD or worse in last 6 months						
60 DPD in last 6 month	0	1	2	3	4	5
Count in each bucket	50883	11127	4905	1461	404	70
% Contribution	73.90%	16.16%	7.12%	2.12%	0.59%	0.10%

There is a 6.5% jump when we shift the time frame for 30 DPD to 90 DPD.

3

Variable 3 - No of times 30 DPD or worse in last 6 months								
30 DPD in last 6 month	0	1	2	3	4	5	6	7
Count in each bucket	49111	9500	5890	2824	1036	379	95	15
% Contribution	71.33%	13.80%	8.55%	4.10%	1.50%	0.55%	0.14%	0.02%

4

Variable 4 - No of times 90 DPD or worse in last 12 months						
90 DPD in last 12 month	0	1	2	3	4	5
Count in each bucket	49507	11653	6153	1235	267	35
% Contribution	71.91%	16.93%	8.94%	1.79%	0.39%	0.05%

The number of buckets keep on decreasing with the increase in time frame. Number of people who have paid their dues on time has increased, when we shift the time frame for 30 DPD to 90 DPD .

5

Variable 5 - No of times 60 DPD or worse in last 12 months								
60 DPD in last 12 month	0	1	2	3	4	5	6	7
Count in each bucket	44967	12727	6414	3192	1040	393	110	7
% Contribution	65.31%	18.49%	9.32%	4.64%	1.51%	0.57%	0.16%	0.01%

There is a 8.0% jump when we shift the time frame for 30 DPD to 90 DPD.

6

Variable 6 - No of times 30 DPD or worse in last 12 months										
30 DPD in last 12 month	0	1	2	3	4	5	6	7	8	9
Count in each bucket	43958	11386	6112	4132	1918	846	368	106	23	1
% Contribution	63.85%	16.54%	8.88%	6.00%	2.79%	1.23%	0.53%	0.15%	0.03%	0.00%

Univariate Analysis

Findings for the variables numbered 1 to 9 and 12 continued..

7

Variable 7 - Presence of open home loan		
Status	0	1
Count in each bucket	50784	18063
% Contribution	73.76%	26.24%

8

Variable 8 - Presence of open auto loan		
Status	0	1
Count in each bucket	62917	5930
% Contribution	91.38%	8.61%

We see that almost 74% of the applicants who do not have a home loan have not defaulted in payments and 91% of applicants who do not have auto loan have not defaulted in payments. The percentage of applicants who default is more for open home loan category.

9

Variable 9 - No of PL trades opened in last 6 months							
PL Trade opened in last 6 months	0	1	2	3	4	5	6
Count in each bucket	30090	13539	12556	7937	3339	1090	296
% Contribution	43.70%	19.66%	18.24%	11.53%	4.85%	1.58%	0.43%

12

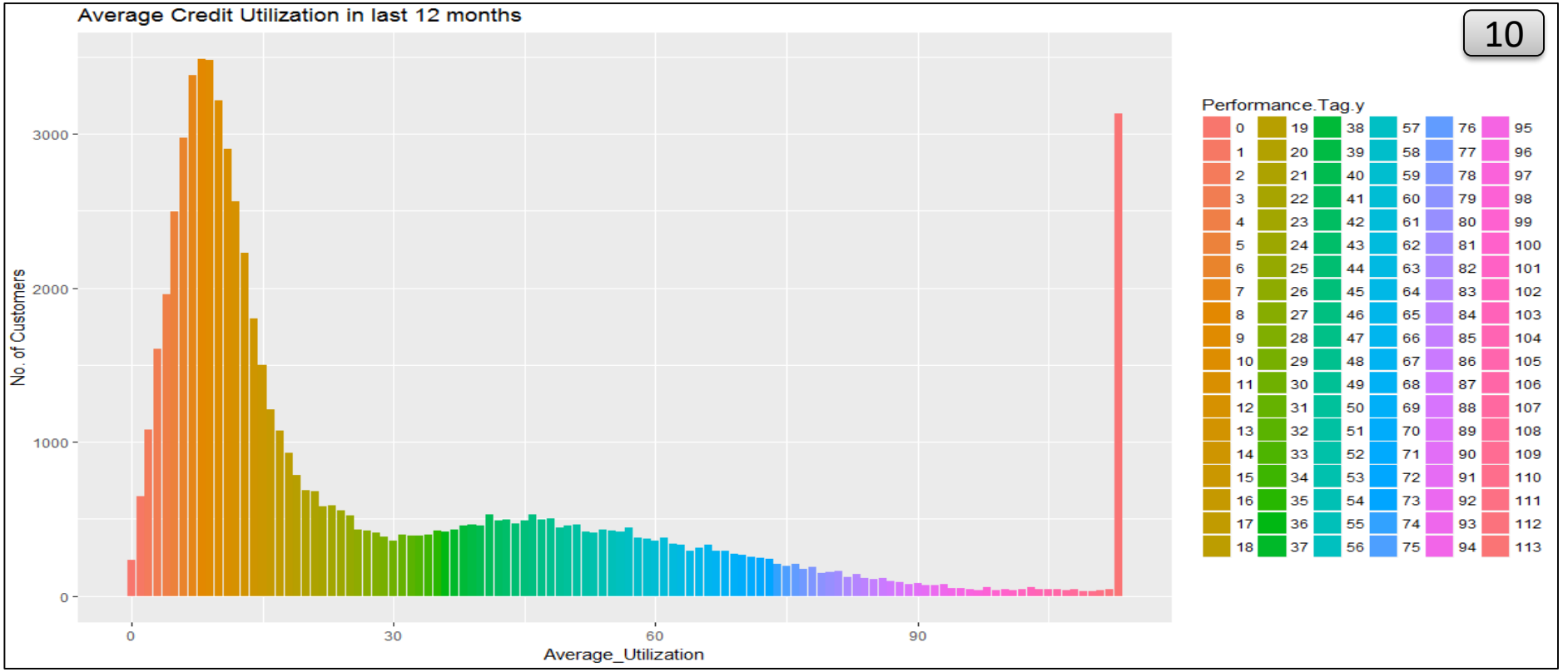
Variable 12 - No of PL trades opened in last 12 months													
PL Trade opened in last 12 months	0	1	2	3	4	5	6	7	8	9	10	11	12
Count in each bucket	24837	6641	6825	8127	7901	6179	4015	2221	1172	601	255	66	10
% Contribution	36.07%	9.65%	9.91%	11.80%	11.48%	8.97%	5.83%	3.23%	1.70%	0.87%	0.37%	0.10%	0.01%

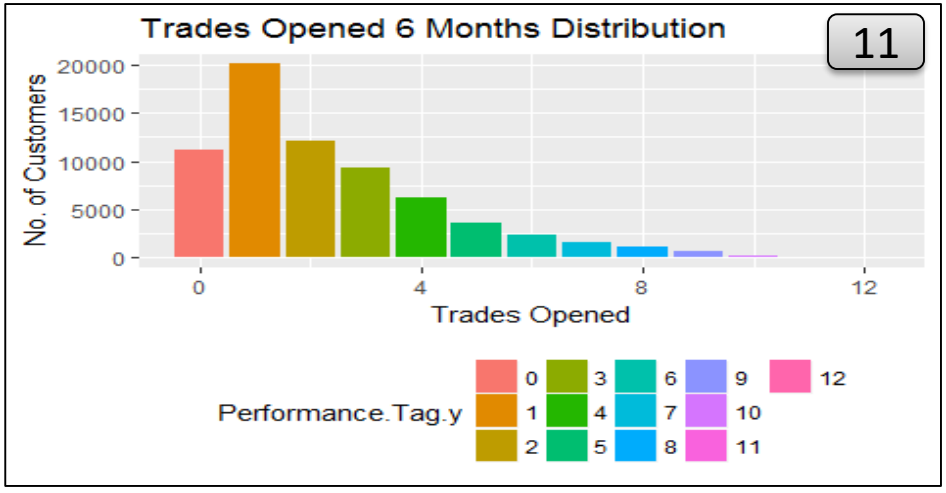
We see that the highest amount of PL trades that were opened in the last 6 months fall under bucket 1 and 2, whereas for PL trades that were opened in the last 12 months fall under bucket 3 and 4. We can see that people have opened less PL trade accounts in the last 6 months, as the percentage of 0 in last 6 months is higher than last 12 months.

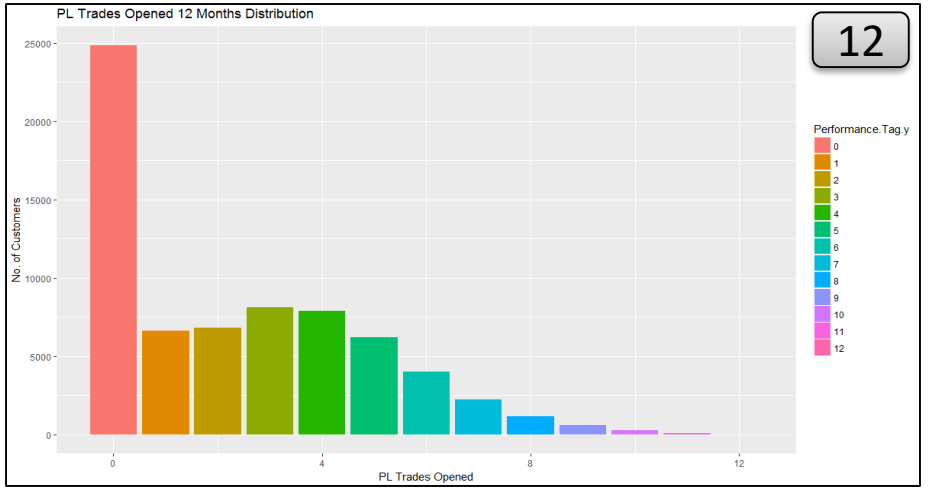
Variable - Outstanding Balance	
Performance_Tag	Balance (In Millions)
0	83487.91
1	3695.06

The loss due to default in payments is more than 3695 millions, we will have to look at how to reduce this value.

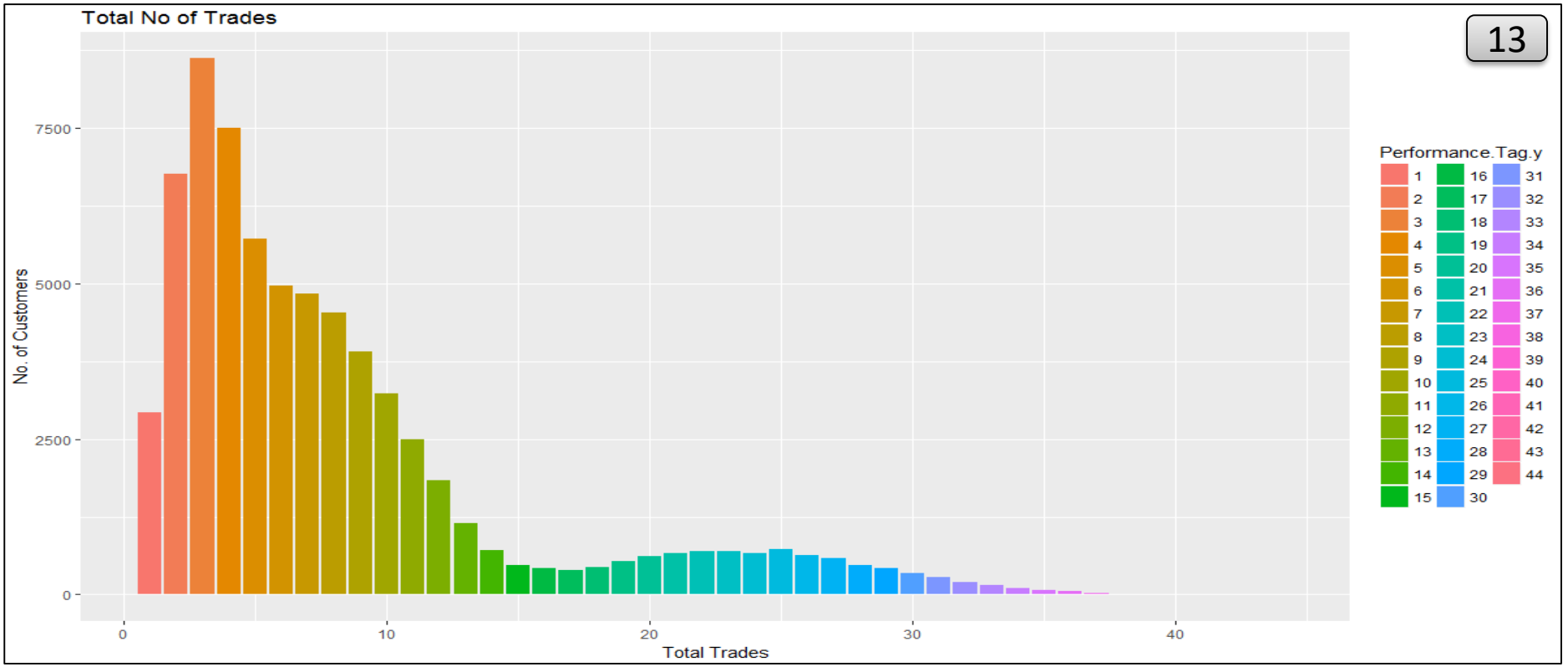
Univariate Analysis (continued)

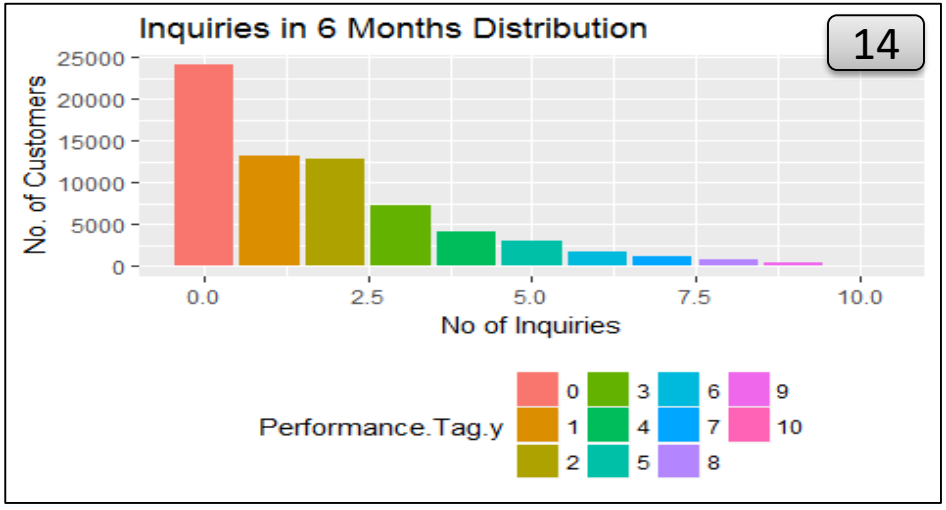


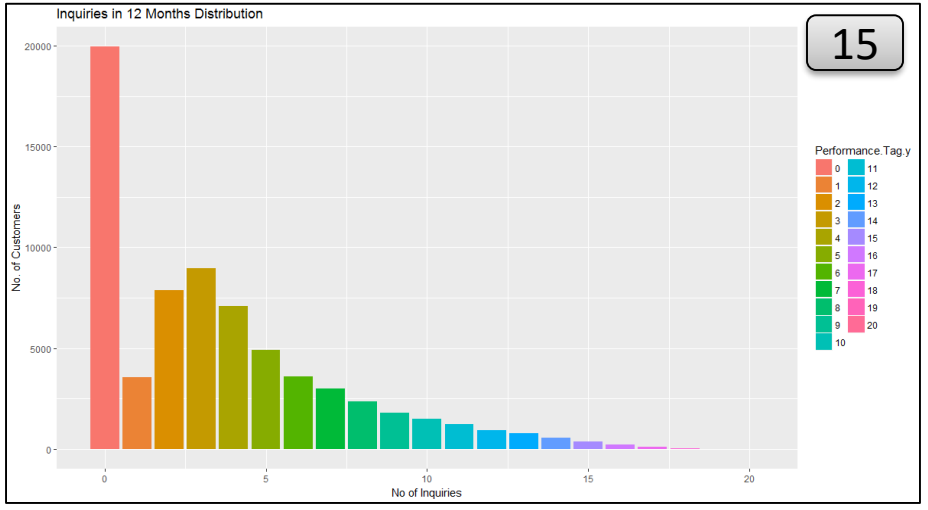




Univariate Analysis (continued)

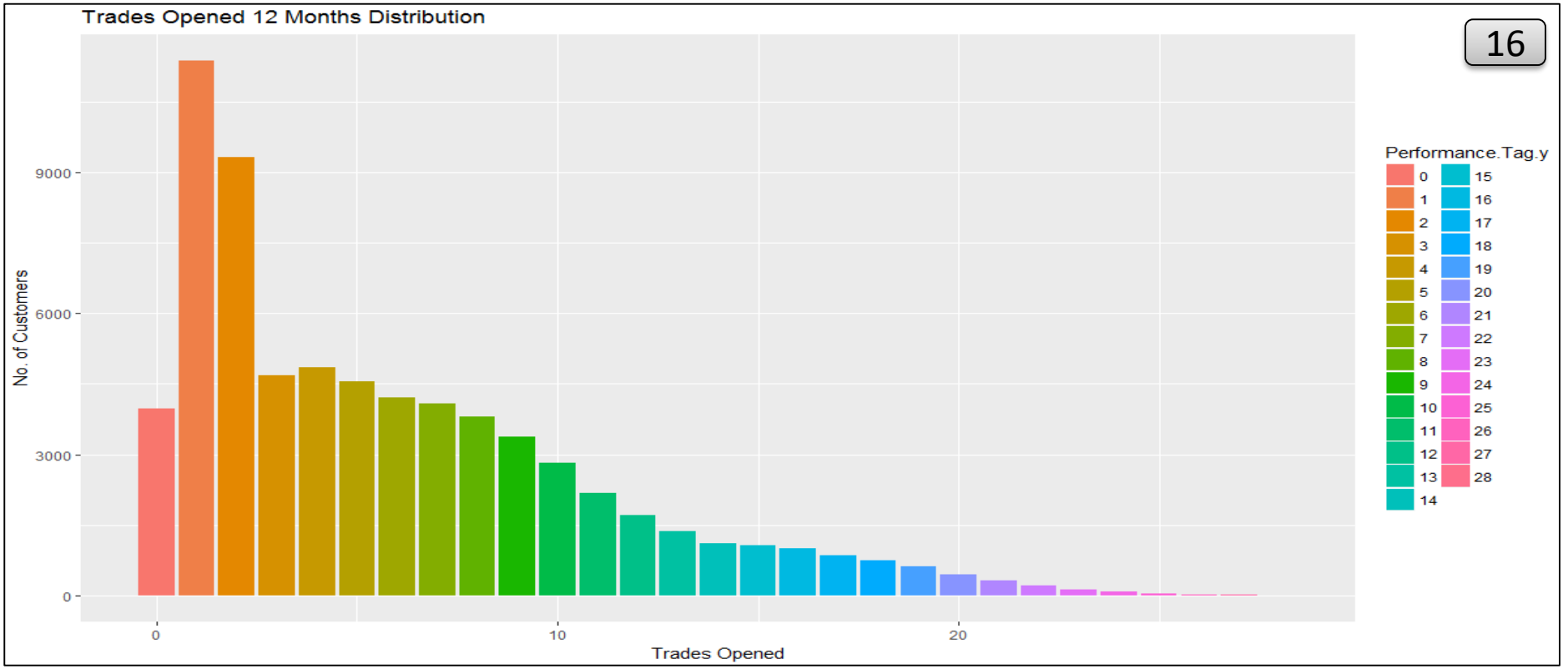




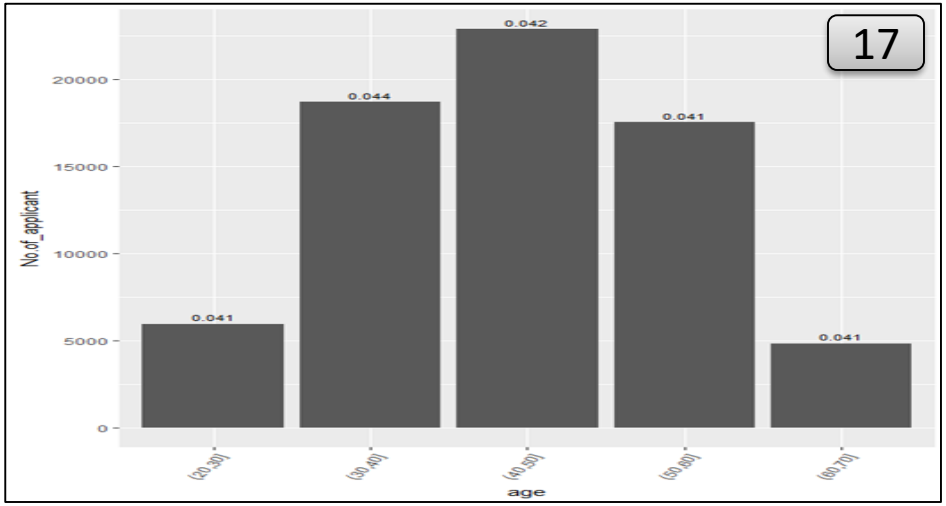


Univariate Analysis (continued)

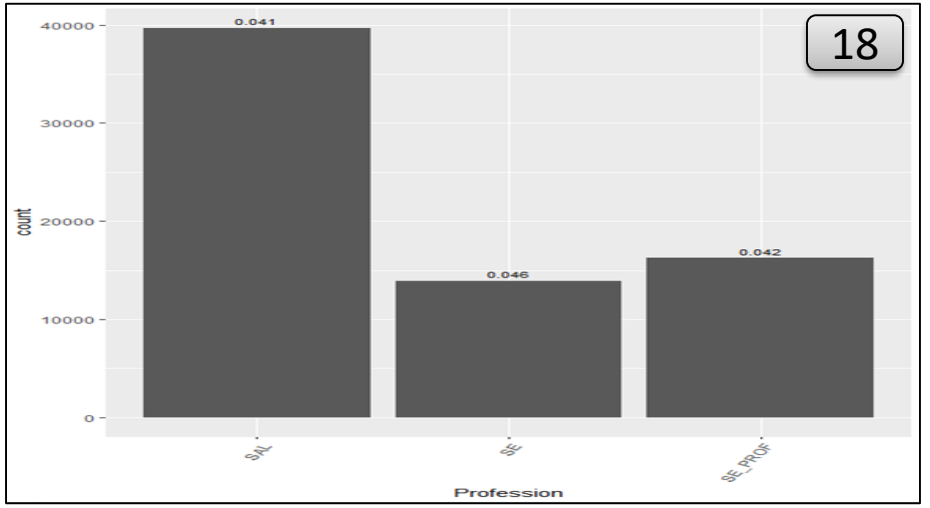
16



17



18



Univariate Analysis
Findings for the variables numbered 10, 11 and 13 to 18

Variable – 10 (Avgas CC Utilization in last 12 months)
This talks about how many times have applicants on an average withdrawn credit for their utilization.

10

Variable 10 - Avgas CC Utilization in last 12 months						
Top 6 buckets	8	9	7	10	113	6
Count in each bucket	3485	3482	3384	3216	3134	2974
% Contribution	5.06%	5.06%	4.92%	4.67%	4.55%	4.32%

25% of the applicant have used their credit lines between 6 to 10 times, we see that 4.5% of applicants have used their credit line 113 times.

11

Variable 11 - No of trades opened in last 6 months													
Trade opened in last 6 months	0	1	2	3	4	5	6	7	8	9	10	11	12
Count in each bucket	11205	20116	12110	9397	6287	3661	2336	1649	1154	618	238	65	11
% Contribution	16.27%	29.22%	17.59%	13.65%	9.13%	5.32%	3.39%	2.40%	1.68%	0.90%	0.35%	0.09%	0.02%

16

Variable 16 - No of trades opened in last 12 months															
Trade opened in last 12 months	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Count in each bucket	3969	11376	9323	4678	4847	4547	4202	4091	3795	3371	2811	2176	1702	1369	1114
% Contribution	5.76%	16.52%	13.54%	6.79%	7.04%	6.60%	6.10%	5.94%	5.51%	4.90%	4.08%	3.16%	2.47%	1.99%	1.62%
Trade opened in last 12 months	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
Count in each bucket	1068	992	860	736	612	434	308	218	121	73	34	11	9	3	
% Contribution	1.55%	1.44%	1.25%	1.07%	0.89%	0.63%	0.45%	0.32%	0.18%	0.11%	0.05%	0.02%	0.01%	0.00%	

While the range of number of trades opened in the last 12 months falls within 0 to 28, the range has substantially reduced in the last 6 months and is now between 0 to 12. This tells us that applicants are opening less trade lines with the bank or the bank has put a restriction on the opening of new trade lines.

Univariate Analysis
Findings for the variables numbered 10, 11 and 13 to 18

14

Variable 14 - No of Inquiries in last 6 months (excluding home & auto loans)											
Inquiries in last 6 months	0	1	2	3	4	5	6	7	8	9	10
Count in each bucket	24118	13145	12809	7248	4244	3019	1750	1149	835	425	108
% Contribution	35.03%	19.09%	18.60%	10.53%	6.16%	4.38%	2.54%	1.67%	1.21%	0.62%	0.16%

15

Variable 15 - No of Inquiries in last 12 months (excluding home & auto loans)											
Inquiries in last 12 months	0	1	2	3	4	5	6	7	8	9	10
Count in each bucket	19961	3572	7886	8961	7094	4916	3612	2992	2345	1777	1508
% Contribution	28.99%	5.19%	11.45%	13.02%	10.30%	7.14%	5.25%	4.35%	3.41%	2.58%	2.19%
Inquiries in last 12 months	11	12	13	14	15	16	17	18	19	20	
Count in each bucket	1231	936	789	553	360	212	97	40	6	2	
% Contribution	1.79%	1.36%	1.15%	0.80%	0.52%	0.31%	0.14%	0.06%	0.01%	0.00%	

We see a decrease in the number of inquiries that were made in the last 6 months when compared to the last 12 months. The decrease is almost 7%. The first 6 months shows that applicants used inquire more number of times in that time frame.

13

Variable 13 - Total No of Trades						
Top 6 buckets	3	4	2	5	6	7
Count in each bucket	8612	7489	6765	5710	4965	4830
% Contribution	12.51%	10.88%	9.83%	8.29%	7.21%	7.02%

3 trades holds 1/8th of all the trades opened.

Variable -17 (Age)

Default rate for age group 30- 40 is more compared to others . But not much significant difference with other age group.

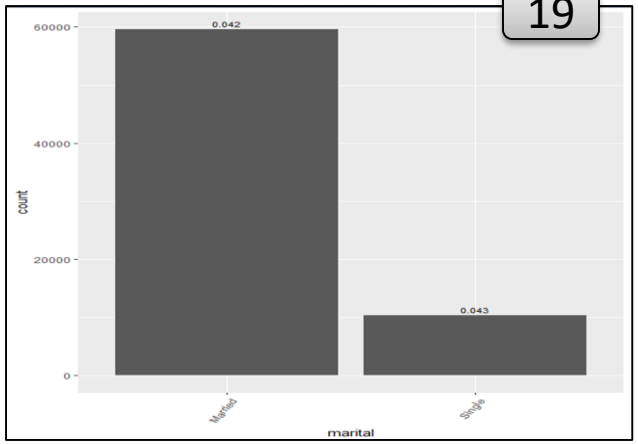
Variable -18 (Profession)

Profession			
Performance_Tag	SAL	SE	SE_PROF
0	38045	13285	15579
1	1629	642	677

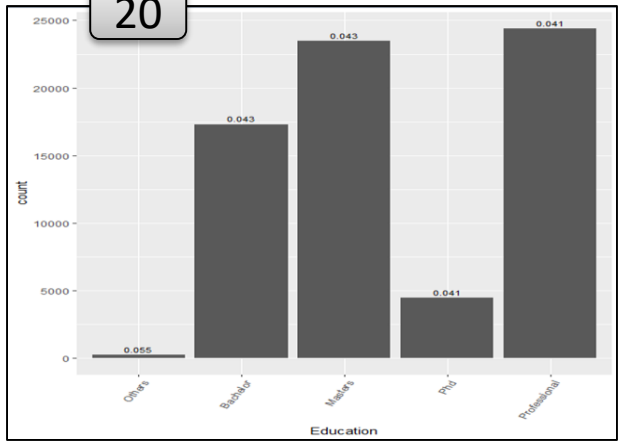
People with SE profession default the most

Univariate Analysis (continued)

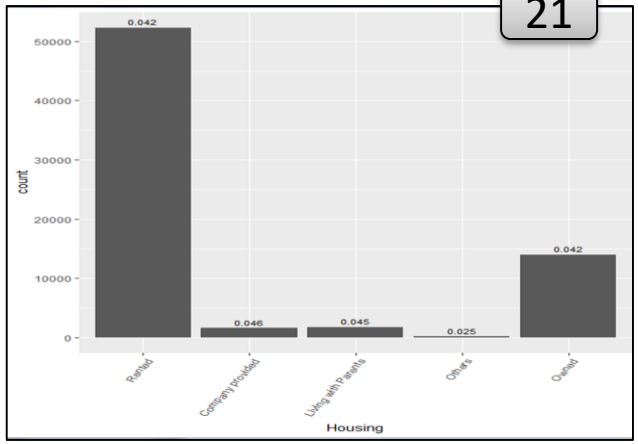
19



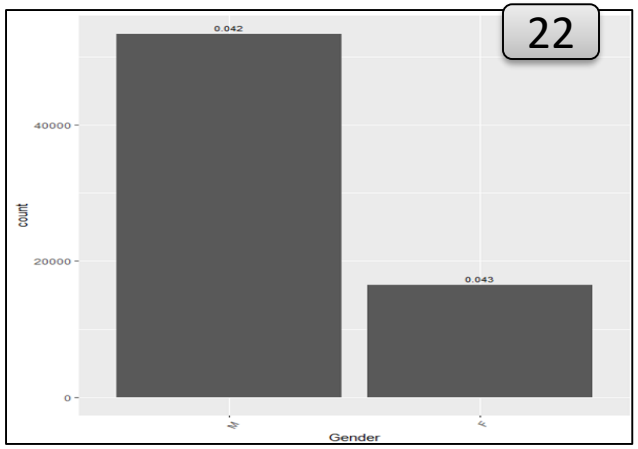
20



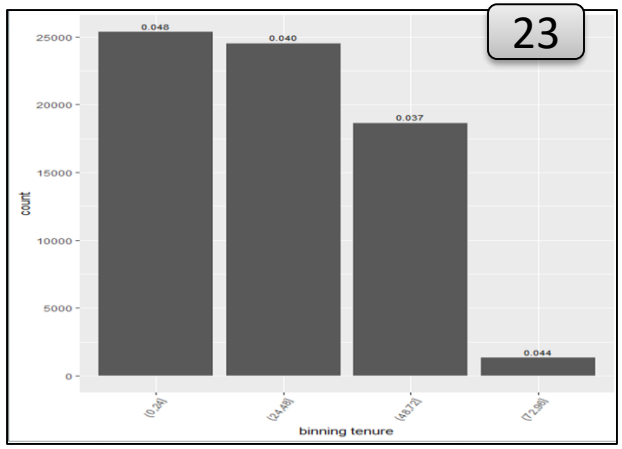
21



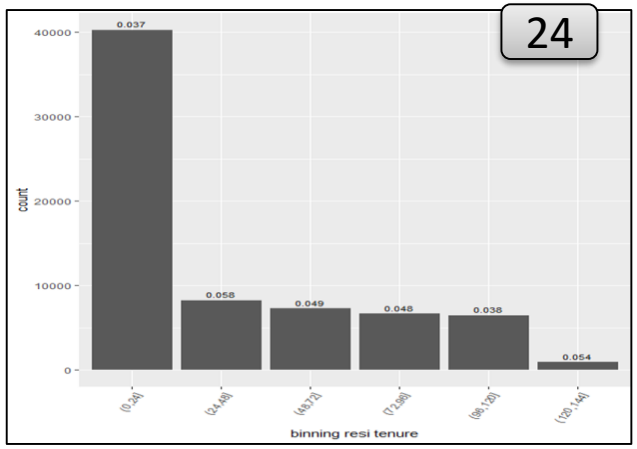
22



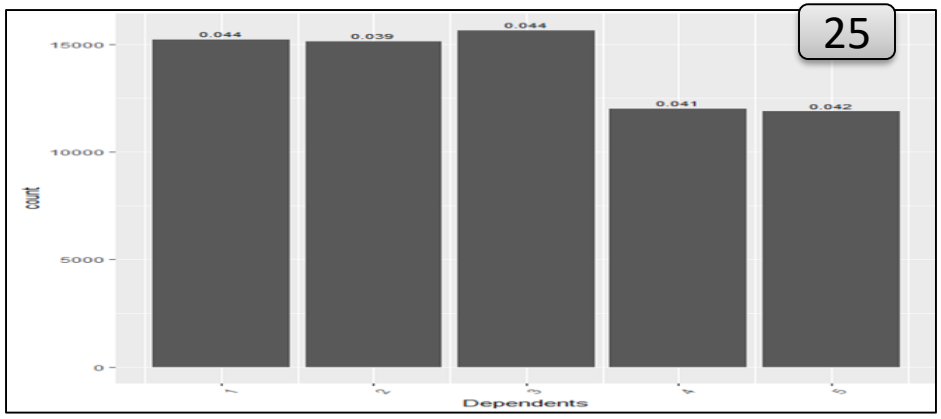
23



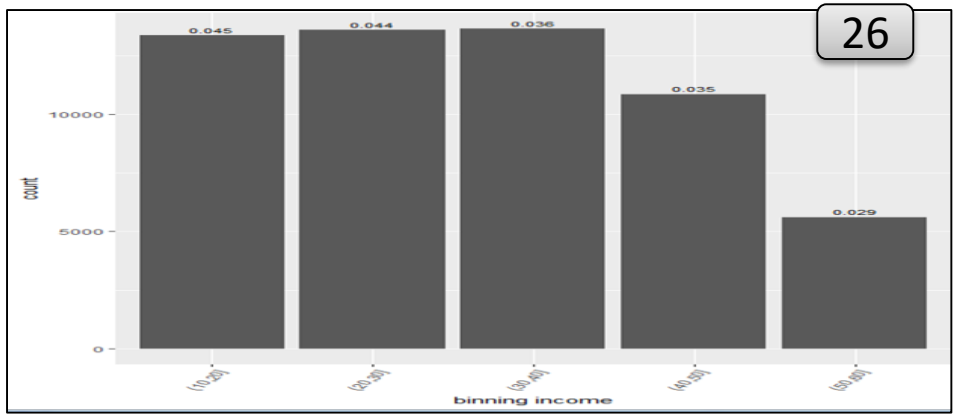
24



25



26



Univariate Analysis
Findings for the variables numbered 19 to 26

19

Marital Status		
Performance_Tag	Married	Single
0	57044	9872
1	2503	445

Single default more than married people

20

Education					
Performance_Tag	Bachelor	Masters	Others	Phd	Professional
0	16560	22483	111	4280	23375
1	742	998	8	184	1011
% Contribution	4.48%	4.44%	7.21%	4.30%	4.33%

Others Contribute the highest, followed by bachelor's when it comes to defaulting

21

Type Of Residence					
Performance_Tag	Company provided	Living with Parents	Others	Owned	Rented
0	1530	1698	193	13410	50083
1	73	80	5	593	2197
% Contribution	4.77%	4.71%	2.59%	4.42%	4.39%

People who stay in company provided accommodation default the most followed by people who stay with parents

22

Gender		
Performance_Tag	F	M
0	15788	51132
1	718	2230
% Contribution	4.55%	4.36%

Females default more than male.

Variable – 23 (No of months in current company) People who have less than 2 years of experience default the most.

Variable – 24 (No of months in current residence) People who have stayed between than 2 to 4 years in their current location default the most.

Univariate Analysis
Findings for the variables numbered 19 to 26

No Of Dependants					
Performance_Tag	1	2	3	4	5
0	14551	14540	14950	11506	11372
1	667	588	695	494	504
% Contribution	4.58%	4.04%	4.65%	4.29%	4.43%

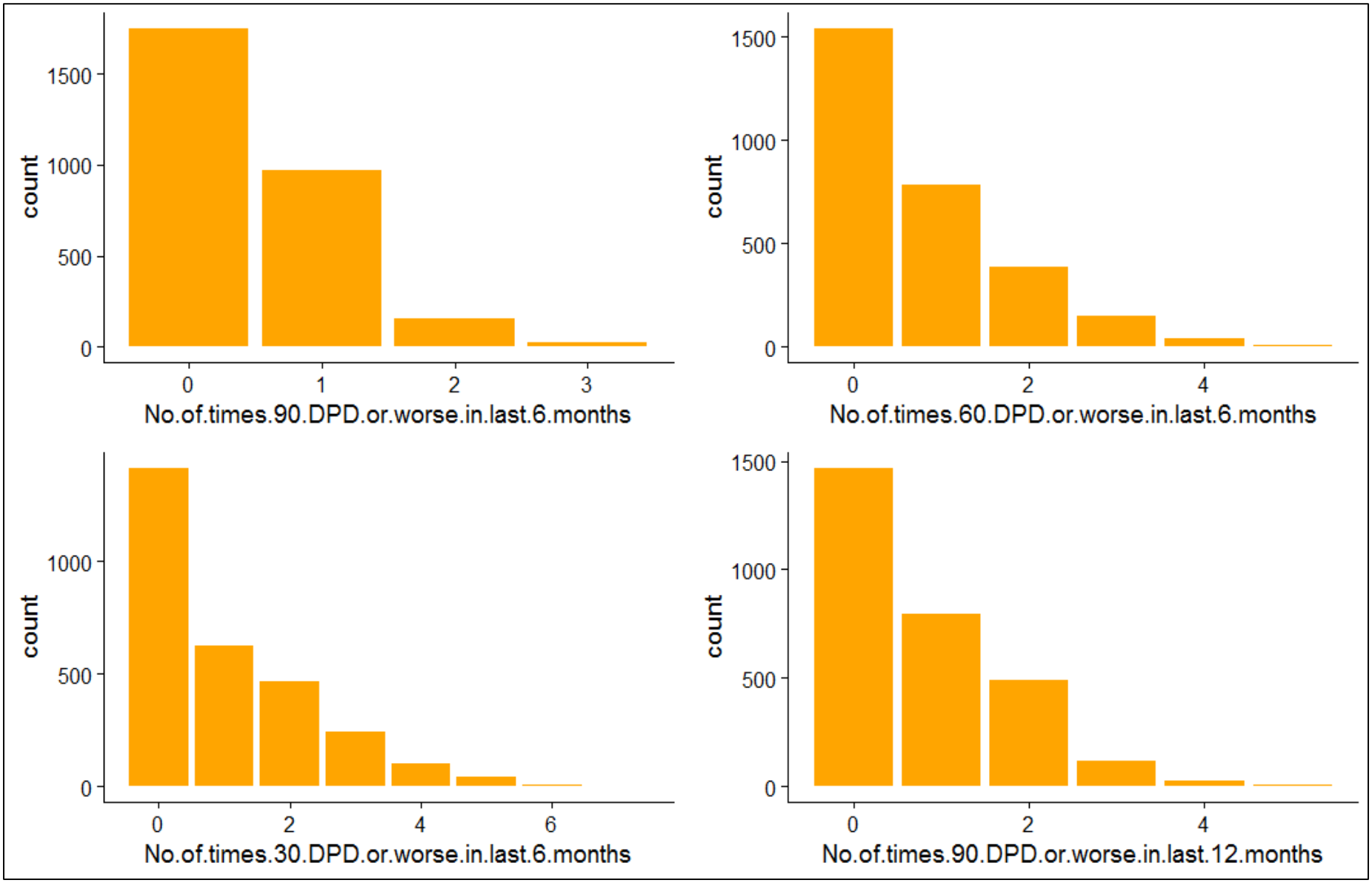
People having 3 dependents default the most

Variable – 26 (Income) Income which is scaled between 10 to 20 has the highest default rate.

Summary Of Univariate Analysis

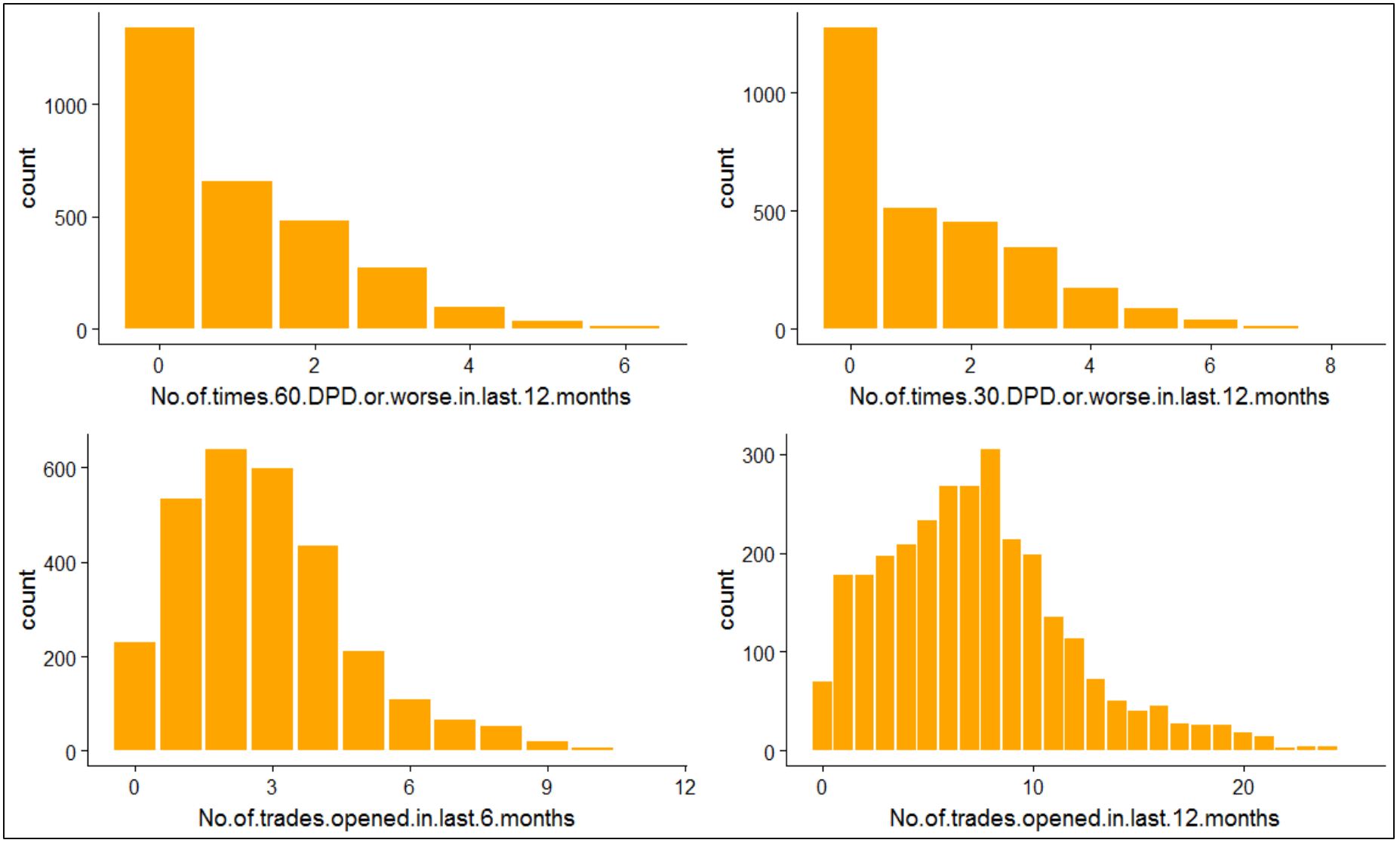
1.We looked into the different variable that compromise of both categorical and quantative types but did not find any unusual patterns while doing a performance tag verse other variable analysis .
2.There are no such variable that single handedly influence the performance tag.
3.We understand that most of the CredX companies customer work as SAL professionals.
4.The age variable has a normally distributed curve with most of the performance tag as 1, happening in the age group of 30 to 40.
5.Most of the customers have a masters degree and stay in a rented place
6.Most of the applicants are married and usually male have a higher usage of credit than females.

Segmented Analysis (Where Performance Tag is “1”)



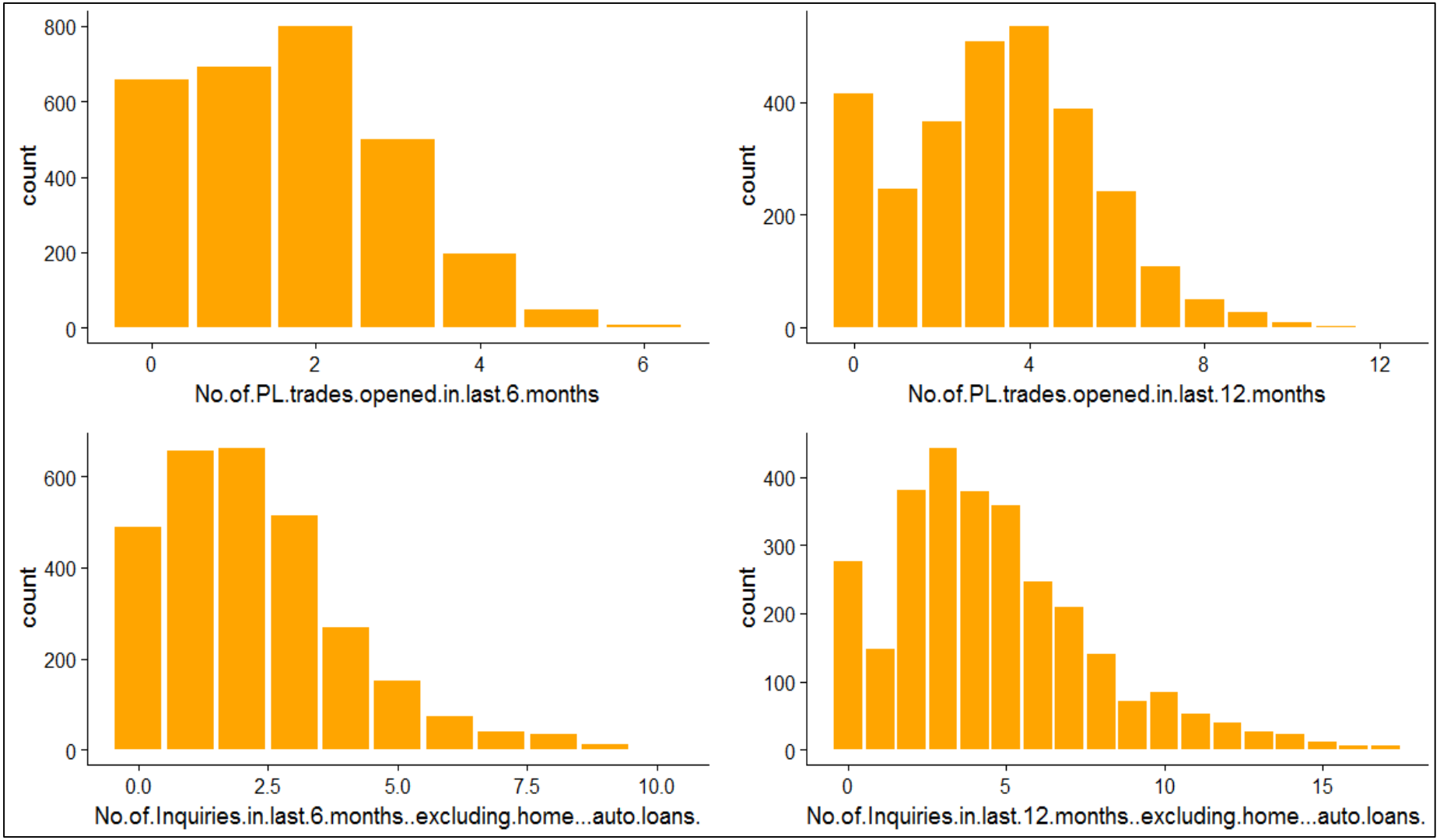
While doing segmented analysis of performance tag (where tag is 1) over other variables we do not see any of the variable having a major impact. The attrition across the variable’s tells the same story .

Segmented Analysis (Where Performance Tag is “1”)



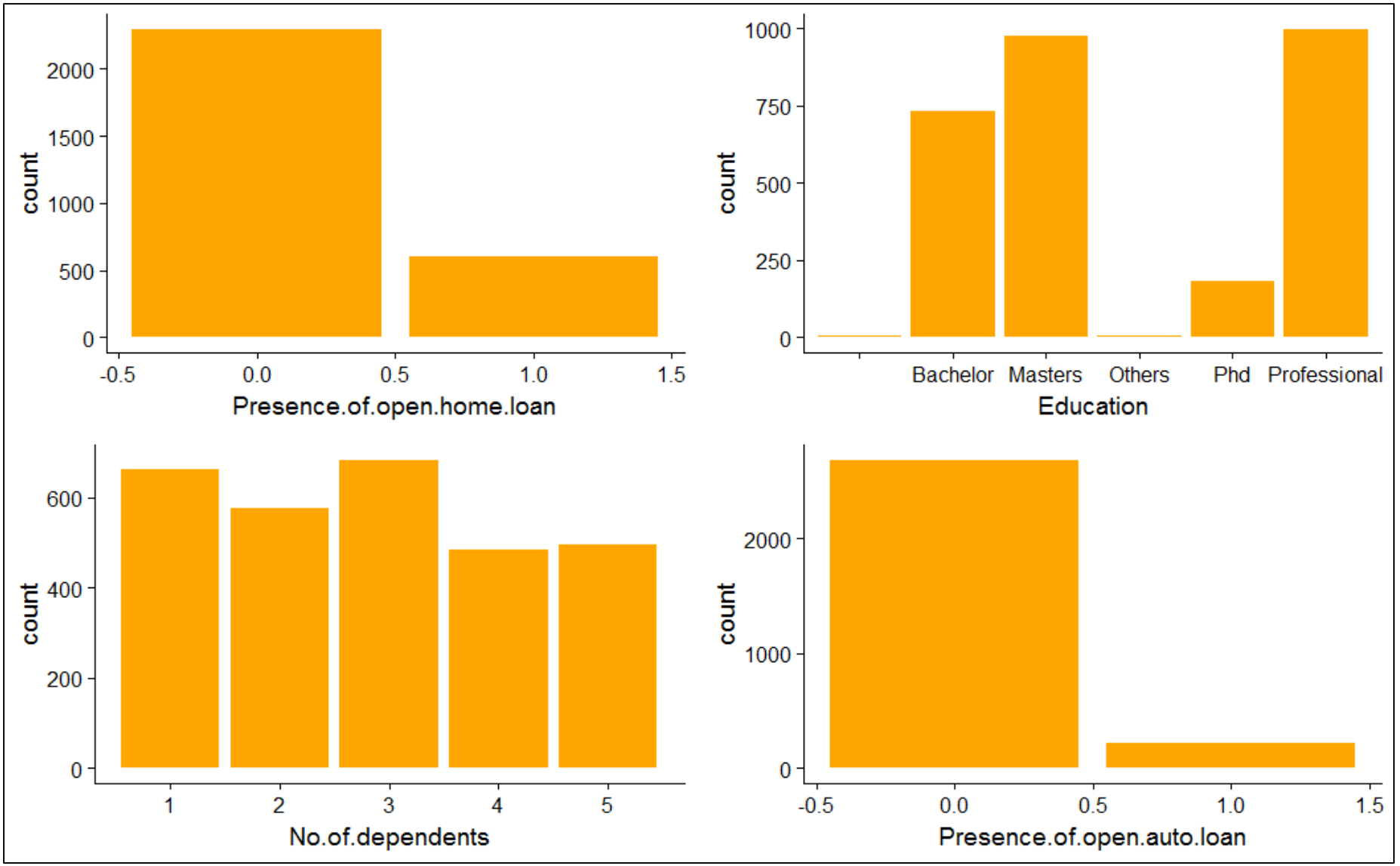
While doing segmented analysis of performance tag (where tag is 1) over other variables we do not see any of the variable having a major impact. The attrition across the variable’s tells the same story .

Segmented Analysis (Where Performance Tag is “1”)



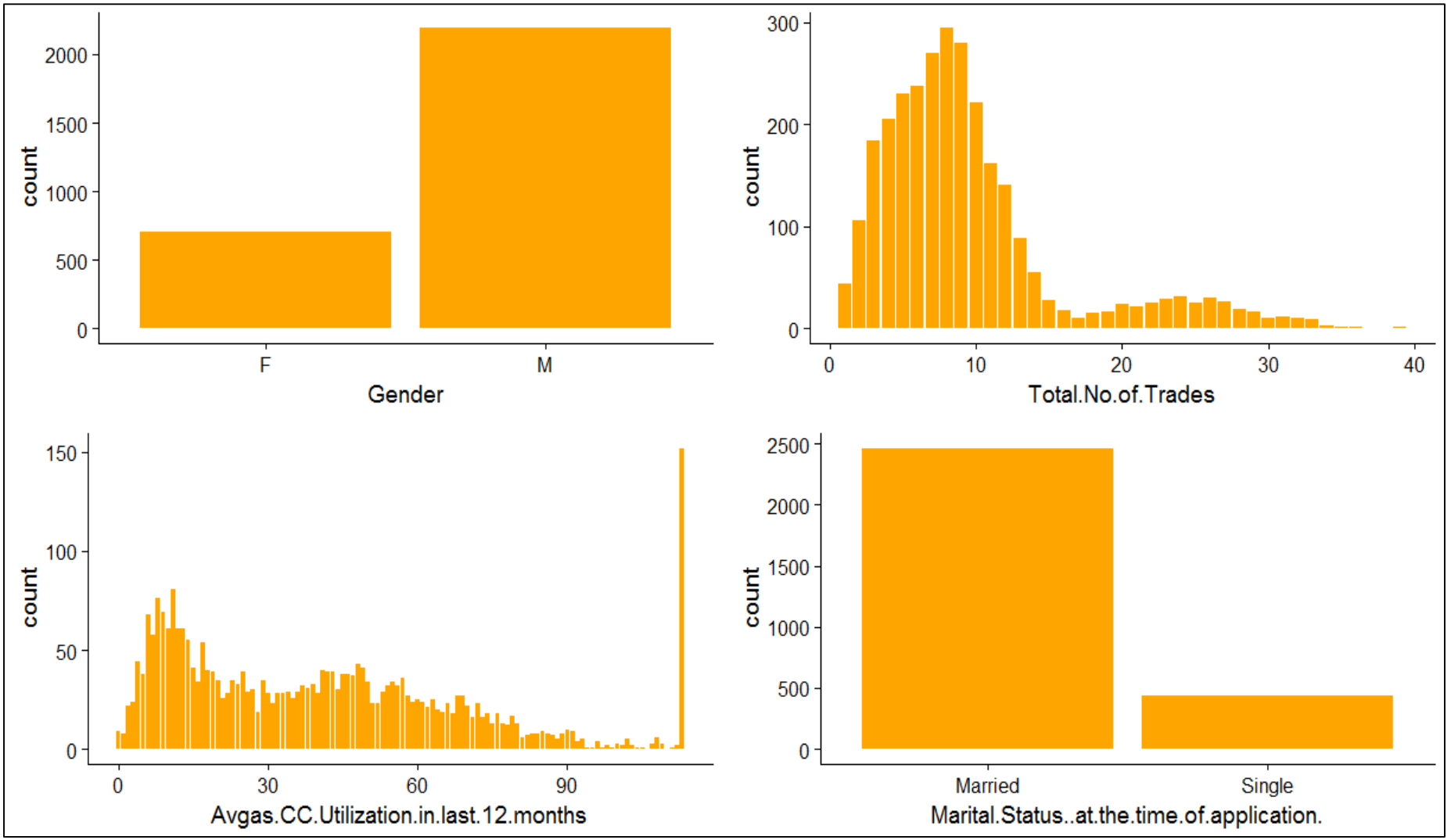
While doing segmented analysis of performance tag (where tag is 1) over other variables we do not see any of the variable having a major impact. The attrition across the variable’s tells the same story .

Segmented Analysis (Where Performance Tag is “1”)



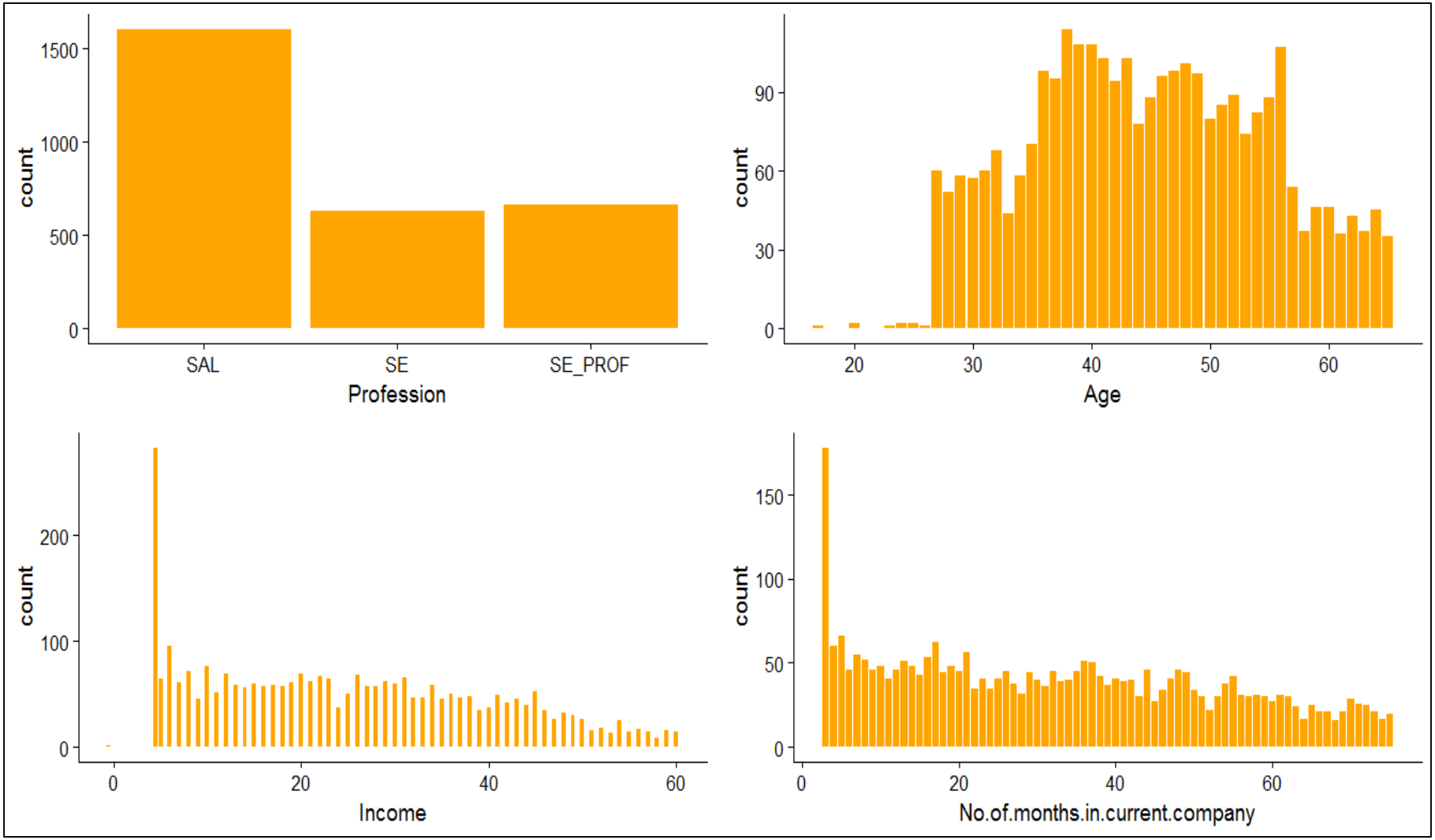
While doing segmented analysis of performance tag (where tag is 1) over other variables we do not see any of the variable having a major impact. The attrition across the variable’s tells the same story .

Segmented Analysis (Where Performance Tag is “1”)



While doing segmented analysis of performance tag (where tag is 1) over other variables we do not see any of the variable having a major impact. The attrition across the variable’s tells the same story .

Segmented Analysis (Where Performance Tag is “1”)



While doing segmented analysis of performance tag (where tag is 1) over other variables we do not see any of the variable having a major impact. The attrition across the variable’s tells the same story .

Methodology

- Application ID was used as the primary key to combine the credit bureau and demographic datasets. This was done after checking the uniqueness of application ID in each dataset.
- After merging the dataset the performance tag variable from demographic dataset was removed. This was done after looking at the difference if any in both the performance tag.
- The number of data points in the combined datasets is 71301.

Treatment of NA values

The below table holds the number of NA's that we got after doing the merging on both the datasets.

Variable	No_Of_NA's
Avgas CC Utilization in last 12 months	1058
No of trades opened in last 6 months	1
Presence of open home loan	272
Outstanding Balance	272
No of dependents	3
Performance Tag y	1425

Data is missing from the dependent variable, this missing data is around 2% of the total data points. We have removed these data points. After removing these NA's we are not left with any missing values in any of the variables.

Dependent Variable

We introduced NA's in Age variable after we converted -3 and 0 as missing values. We check the WOE for this bucket and saw the weight as 0 for this group. We removed these data points from our calculation.

There were 6 blank data points introduced in marital status variable and 1 for gender variable, while doing the merge, we checked the WOE for these and found out it to be 0 for each of the group. We removed these data points from our calculation.

The model building stage will have 66825 variables for model creation.

Binning activity

The below table holds the variables that were binned and the numbers of bins created.

Variable	No_Of_Bins_Created
Age	11
Income	10
No of months in current residence	6
No of months in current company	10
No of times 90 DPD or worse in last 6 months	2
No of times 60 DPD or worse in last 6 months	2
No of times 30 DPD or worse in last 6 months	3
No of times 90 DPD or worse in last 12 months	3
No of times 60 DPD or worse in last 12 months	3
No of times 30 DPD or worse in last 12 months	3
Avgas CC Utilization in last 12 months	10
No of trades opened in last 6 months	6
No of trades opened in last 12 months	9
No of PL trades opened in last 6 months	4
No of PL trades opened in last 12 months	7
No of Inquiries in last 6 months (excluding home & auto loans)	5
No of Inquiries in last 12 months (excluding home & auto loans)	8
Outstanding Balance	10
Total No of Trades	10

The bins were created looking at the WOE value calculated for each variable. The weight of evidence (WOE) tells the predictive power of an independent variable in relation to the dependent variable. In an ideal scenario WOE should always be monotonic, which means either they should increase or decrease with the bins. For some of our variables we were not able to achieve the character of a WOE, so we went ahead and used the original bins.

Scaling variables

We scaled the outstanding balance variable before using it for model building.

WOE Values

VARIABLE NAME	BINS	WOE	IV
No.of.times.90.DPD.or.worse.in.last.6.months	[0,0]	-0.266	0.049
	[1,3]	0.624	0.164
No.of.times.60.DPD.or.worse.in.last.6.months	[0,0]	-0.343	0.075
	[1,5]	0.622	0.210
No.of.times.30.DPD.or.worse.in.last.6.months	[0,0]	-0.395	0.093
	[1,1]	0.467	0.131
	[2,7]	0.741	0.246
No.of.times.90.DPD.or.worse.in.last.12.months	[0,0]	-0.364	0.081
	[1,1]	0.508	0.136
	[2,5]	0.723	0.218
No.of.times.60.DPD.or.worse.in.last.12.months	[0,0]	-0.359	0.072
	[1,1]	0.215	0.081
	[2,7]	0.694	0.189
No.of.times.30.DPD.or.worse.in.last.12.months	[0,0]	-0.385	0.080
	[1,2]	0.284	0.103
	[3,9]	0.798	0.203
Avgas.CC.Utilization.in.last.12.months	[0,4]	-0.801	0.036
	[5,6]	-0.800	0.072
	[7,8]	-0.793	0.116
	[9,11]	-0.672	0.163
	[12,14]	-0.467	0.180
	[15,21]	-0.074	0.181
	[22,37]	0.476	0.210
	[38,51]	0.586	0.254
	[52,71]	0.565	0.297
	[72,113]	0.384	0.315

VARIABLE NAME	BINS	WOE	IV
No.of.trades.opened.in.last.6.months	[0,0]	-0.741	0.065
	[1,1]	-0.478	0.119
	[2,2]	0.236	0.130
	[3,3]	0.435	0.161
	[4,4]	0.522	0.193
	[5,12]	0.134	0.196
No.of.trades.opened.in.last.12.months	[0,0]	-0.896	0.031
	[1,1]	-1.018	0.142
	[2,2]	-0.815	0.205
	[3,4]	0.009	0.205
	[5,5]	0.205	0.208
	[6,7]	0.450	0.238
	[8,9]	0.572	0.282
	[10,12]	0.487	0.311
	[13,28]	0.005	0.311
No.of.PL.trades.opened.in.last.6.months	[0,0]	-0.678	0.149
	[1,1]	0.202	0.158
	[2,2]	0.438	0.201
	[3,6]	0.360	0.229
No.of.PL.trades.opened.in.last.12.months	[0,0]	-0.949	0.215
	[1,1]	-0.130	0.217
	[2,2]	0.250	0.224
	[3,3]	0.416	0.249
	[4,4]	0.501	0.285
	[5,5]	0.424	0.305
	[6,12]	0.238	0.312

Data Preparation Continued..

WOE Values

VARIABLE NAME	BINS	WOE	IV
No.of.Inquiries.in.last.6.months.exc.home.auto	[0,0]	-0.756	0.144
	[1,1]	0.177	0.150
	[2,2]	0.216	0.160
	[3,4]	0.508	0.214
	[5,10]	0.014	0.214
No.of.Inquiries.in.last.12.months.exc.home.auto	[0,0]	-1.143	0.233
	[1,1]	-0.025	0.233
	[2,2]	0.141	0.235
	[3,3]	0.167	0.239
	[4,4]	0.249	0.246
	[5,5]	0.580	0.278
	[6,8]	0.485	0.316
	[9,20]	0.015	0.316
Presence.of.open.home.loan	[0,0]	0.074	0.004
	[1,1]	-0.238	0.018
Outstanding.Balance	[0,7790]	-0.978	0.063
	[7791,56368]	-0.850	0.113
	[56516,392876]	-0.079	0.113
	[392909,590337]	0.272	0.122
	[590343,777912]	0.466	0.149
	[777936,976123]	0.415	0.169
	[976147,1362729]	0.390	0.188
	[1362827,2962087]	-0.391	0.201
	[2962089,3289690]	-0.803	0.246
	[3289827,5218801]	0.295	0.256

VARIABLE NAME	BINS	WOE	IV
Total.No.of.Trades	[1,1]	-1.058	0.030
	[2,2]	-1.016	0.096
	[3,3]	-0.701	0.141
	[4,4]	-0.447	0.159
	[5,5]	-0.047	0.159
	[6,7]	0.218	0.166
	[8,8]	0.461	0.183
	[9,10]	0.542	0.223
	[11,19]	0.425	0.250
	[20,44]	-0.066	0.250
Presence.of.open.auto.loan	[0,0]	0.012	0.000
	[1,1]	-0.137	0.002
Age	[-3,30]	-0.043	0.000
	[31,35]	0.044	0.000
	[36,38]	0.068	0.001
	[39,41]	0.077	0.001
	[42,44]	-0.056	0.002
	[45,47]	-0.006	0.002
	[48,50]	-0.009	0.002
	[51,53]	-0.139	0.004
	[54,57]	0.046	0.004
	[58,65]	-0.013	0.004
Gender	F	0.029	0.000
	M	-0.009	0.000

WOE Values

VARIABLE NAME	BINS	WOE	IV
Marital.Status	Married	-0.0036	0.0000
	Single	0.0212	0.0001
No.of.dependents	[1,1]	0.0469	0.0005
	[2,2]	-0.0911	0.0022
	[3,3]	0.0527	0.0029
	[4,4]	-0.0287	0.0030
	[5,5]	0.0075	0.0030
Income	[-0.5,5]	0.3024	0.0095
	[6,10]	0.2682	0.0171
	[11,16]	0.0755	0.0178
	[17,21]	0.0875	0.0186
	[22,26]	0.0092	0.0186
	[27,31]	0.0702	0.0191
	[32,36]	-0.1358	0.0208
	[37,41]	-0.2656	0.0268
	[42,48]	-0.1772	0.0300
	[49,60]	-0.3648	0.0418
Education	NA	0.0148	0.0000
	Bachelor	0.0202	0.0001
	Masters	-0.0008	0.0001
	Others	0.5304	0.0007
	Phd	-0.0086	0.0007
	Professional	-0.0155	0.0008

VARIABLE NAME	BINS	WOE	IV
Profession	NA	0.000	0.000
	SAL	-0.027	0.000
	SE	0.089	0.002
	SE_PROF	-0.013	0.002
Type.of.residence	NA	0.000	0.000
	Company provided	0.079	0.000
	Living with Parents	0.071	0.000
	Others	-0.519	0.001
	Owned	-0.003	0.001
	Rented	-0.002	0.001
No.of.months.in.current.residence	[6,9]	-0.275	0.033
	[10,28]	0.504	0.065
	[29,49]	0.303	0.076
	[50,72]	0.139	0.078
	[73,97]	0.133	0.080
	[98,126]	-0.078	0.081
No.of.months.in.current.company	[3,5]	0.103	0.001
	[6,12]	0.178	0.004
	[13,19]	0.203	0.009
	[20,26]	0.037	0.009
	[27,33]	-0.079	0.010
	[34,40]	0.024	0.010
	[41,47]	-0.161	0.012
	[48,53]	-0.222	0.016
	[54,61]	-0.229	0.021
	[62,133]	0.063	0.022

Information Value

VARIABLE NAME	IV
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	0.31600
Avgas.CC.Utilization.in.last.12.months	0.31466
No.of.PL.trades.opened.in.last.12.months	0.31231
No.of.trades.opened.in.last.12.months	0.31129
Outstanding.Balance	0.25590
Total.No.of.Trades	0.25015
No.of.times.30.DPD.or.worse.in.last.6.months	0.24642
No.of.PL.trades.opened.in.last.6.months	0.22899
No.of.times.90.DPD.or.worse.in.last.12.months	0.21836
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	0.21448
No.of.times.60.DPD.or.worse.in.last.6.months	0.20984
No.of.times.30.DPD.or.worse.in.last.12.months	0.20262
No.of.trades.opened.in.last.6.months	0.19562
No.of.times.60.DPD.or.worse.in.last.12.months	0.18926
No.of.times.90.DPD.or.worse.in.last.6.months	0.16381
No.of.months.in.current.residence	0.08057
Income	0.04181
No.of.months.in.current.company	0.02183
Presence.of.open.home.loan	0.01753
Age	0.00380
No.of.dependents	0.00300
Profession	0.00212
Presence.of.open.auto.loan	0.00166
Application.ID	0.00131
Type.of.residence	0.00089
Education	0.00079
Gender	0.00026
Marital.Status..at.the.time.of.application.	0.00008

Information value is a useful technique to select important variables in a predictive model. It helps to rank variables on the basis of their importance.

The table shows the information value attached to each variable in the dataset.

Dummy variable creation

Dummy variables were created for the below mentioned variables.

Variables for which dummies were created
Age
Income
No of months in current residence
No of months in current company
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
No of trades opened in last 6 months
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)
Outstanding Balance
Total No of Trades
No of dependents
Education
Profession
Type of residence

For all the variables that were binned looking at the WOE values, we also created dummy variables for them. We also created dummies for all the categorical variable in the combined dataset.

Model creation activity

The dataset will be divided into 70:30, 70% of the data points will be used for training and 30% will be used for testing. We will use three machine learning algorithm to build models, these are logistic regression, decision tree and the famous ensemble algorithm called random forest. We will not use any black box techniques for model building.

Logistic Regression

For logistic regression we will use STEP AIC to reduce the number of variables first and then will go on and manually reduce the variables based on High VIF and insignificance value of ($p > 0.05$). Once we have all the significant variables we will go ahead and finalize the model. This model will be used for testing purpose.

Decision Tree

For decision tree we will first look at if it's a binary decision tree or a multiway decision tree. We will use recursive partitioning to build the tree and check what are the class that gets attached to a leaf. Our main aim would be to increase the homogeneity of the dataset.

Random Forest

For random forest we will use the celebrated technique of bagging, a bootstrapped aggregation technique which chooses random samples of observations from a dataset. This will ensure that sampling is uniform and it will ensure that the sample is diverse. We will not dedicate any data point for testing as random forest works on the concept of out of bag error.

Data Modelling Continued..

In total there were six models created for finding the right variables along with the right algorithm. Four models were created on logistic regression algorithm by doing a variety of slicing and dicing of variables. One model was created by using decision tree algorithm and one model was created by using random forest algorithm.

Model-1 (Logistic Regression):

1.We used only the demographic variables to create this model. There were 16 sub-models that were created to reach at the final variables.

2. **The most important variables were**

Mth_in_res_binned6_9, Mth_in_res_binned98_126, Mth_in_comp_binned27_33, Mth_in_comp_binned41_47, Mth_in_comp_binned48_53, Mth_in_comp_binned54_61, Income_binned32_36, Income_binned37_41, Income_binned42_48, Income_binned49_60

3. Sensitivity was at 0.51814, Specificity was at 0.63539 and Accuracy was at 0.6304 for this model.

Model-2 (Logistic Regression):

1.We used only the credit bureau variables to create this model. There were 20 sub-models that were created to reach at the final variables.

2. **The most important variables were**

No.of.times.30.DPD.or.worse.in.last.6.months1, No.of.times.30.DPD.or.worse.in.last.6.months2_7, No.of.times.30.DPD.or.worse.in.last.12.months3_9, Avg_Utility_12Mths22_37, Avg_Utility_12Mths38_51, Avg_Utility_12Mths52_71, Avg_Utility_12Mths72_113, Noofenq_12Mths_exc_hmat1, Noofenq_12Mths_exc_hmat2, Noofenq_12Mths_exc_hmat3, Noofenq_12Mths_exc_hmat4, Noofenq_12Mths_exc_hmat5, Noofenq_12Mths_exc_hmat6_8, Noofenq_12Mths_exc_hmat9_20, Outstanding.Balance_binned2962089_3289690.

3. Sensitivity was at 0.65574, Specificity was at 0.58890 and Accuracy was at 0.5926 for this model.

Model-3 (Logistic Regression):

1.We used WOE values and took all the variables to create this model. There were 6 sub-models that were created to reach at the final variables.

2. **We see that "Number of enquiries 12 months excluding home and auto loan" along with "number of PL trades opened in 12 last months" and "average utility 12 months" seems to be the most significant variables.**

3. This model was not evaluated. This model was created to check the important variable that comes out, when we use WOE values for model building exercise.

Model-4 (Logistic Regression):

1.We used all the variables to create this model. There were 33 sub-models that were created to reach at the final variables.

2. **The most important variables were**

Mth_in_res_binned6_9, Mth_in_res_binned73_97, No.of.times.30.DPD.or.worse.in.last.6.months1, No.of.times.30.DPD.or.worse.in.last.6.months2_7,Avg_Utility_12Mths22_37,Avg_Utility_12Mths38_51, Avg_Utility_12Mths52_71, Avg_Utility_12Mths72_113, Noofenq_12Mths_exc_hmat1,Noofenq_12Mths_exc_hmat2, Noofenq_12Mths_exc_hmat3, Noofenq_12Mths_exc_hmat4, Noofenq_12Mths_exc_hmat5, Noofenq_12Mths_exc_hmat6_8, Noofenq_12Mths_exc_hmat9_20, Outstanding.Balance_binned2962089_3289690

3. Sensitivity was at 0.65873, Specificity was at 0.60061 and Accuracy was at 0.6031 for this model.

Model-5 (Decision Tree):

1. .We used all the variables to create this model. We used the 70/30 principal to divide the data into training and test. There were 3 sub models created taking different criteria into consideration.
2. We used recursive partitioning to correctly classify the data points. The accuracy level remained same for all the three sub models.
3. The first sub model used all the default hyper-parameters which included Gini index. The second sub model was built taking information gain as a hyper-parameter. The last sub model had additional hyper-parameters like minsplit, minbucket and complexity.
4. The accuracy of all the models stood at 95.78%.

Model-6 (Random Forest):

1.We used all the variables to create this model. We used the 80/20 principal to divide the data into training and test. Though we actually do not need a test data as such for random forest algorithm.
2. **Hyper-parameters considered are**

a) na.action, this is used to omit NA's from the dataset while calculating
b) do.trace, this is used to see while the algorithm is running and creating trees
c) ntree, this is for the number of trees to be produced
d) mtry, the number of attributes it will try on before making a split
e) proximity, the closeness between data-points is not considered, hence it is false

3. Out of bag error for each tree stood at 4.23%.

Model validation activity

Below are the validation technique that will be used for each machine learning algorithm

ML_Algorithm	Evaluation_Techniques
Logistic Regression	Sensitivity, Specificity & Accuracy %
	Gain & Lift Chart
	KS Statistics
Decision Tree	Truncation & Pruning
	Gini Index
	Information Gain
Random Forest	Out of bag error
	K-fold Validation

1. For logistic regression the accuracy stood at 60.31%
2. For Decision Tree the accuracy stood at 95.78%. We got the same percentage of accuracy while using gini index and information gain
3. For Random Forest the accuracy stood at 95.78%. The out of bag error on an average for each tree stood at 4.23%. The random forest created 1000 trees for this purpose.

Application Scorecard

Application Scorecard will be built taking three things into consideration, which are how do you tag a customer as good or bad, riskiness of the portfolio looking at the payment dates and What period of time do you consider to come to a conclusion about the riskiness.

These queries will be answered by the concept of roll rate matrix. The effect of roll backs and roll over on a portfolio. We will use the concept of vintage curve to come to a conclusion on, if a customer should be tagged as a defaulter or vice versa.

The below table holds the variables that were used to create the scorecard.

Sr No	Variable Name
1	**Mth_in_res_binned6_9**
2	**No.of.times.30.DPD.or.worse.in.last.6.months1**
3	**No.of.times.30.DPD.or.worse.in.last.6.months2_7**
4	**Avg_Utility_12Mths22_37**
5	**Avg_Utility_12Mths38_51**
6	**Avg_Utility_12Mths52_71**
7	**Avg_Utility_12Mths72_113**
8	**Noofenq_12Mths_exc_hmat5**
9	**Noofenq_12Mths_exc_hmat6_8**
10	**Outstanding.Balance_binned2962089_3289690**

Parameters used to build the scorecard are:-

1.Base Score of 400 (Parameter – point0)
2.Score doubling with every 20 points (Parameter – odds0)
3.Good to Bad odds stands at 10 to 1 (Parameter - pdo)

Scorecard was built on the model output derived through logistic regression.