CredX:
Acquisition and
Operation Risk Analytics

PGDDS: December-2018

By: Pravin Pawar and Aashish Sharma

Objective

Problem Statement

- CredX is a leading credit card provider that gets thousands of credit card applications 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’.

Objective

- The objective is to help CredX identify the right customers using predictive models.
- To build an application scorecard and identify the cut-off score below which one would not grant credit cards to applicants.
- We need to determine the factors affecting credit risk and create strategies to mitigate the acquisition risk and assess the financial benefit of the project.

Data Overview

Customer's Demographic Data

Dimensions:

12 Variables

71,295

Observations

- Information provided by the customer while applying for credit card

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")

Customer's Credit Bureau Data

Dimensions:

19 Variables

71,295

Observations

- This information is extracted by the institution while accessing the customers application

Variables	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 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 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 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 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")

Nature of Data and Data Quality Observations(1/2)

- Application ID is the common key between Demographic and Credit Bureau data
- Performance Tag is the target variable and its common in both data set. 1 signifies defaulted customer and 0 signifies non-defaulted customer
- Performance Tag has 1,425 row as blank, which means the performance is not mapped
- There are some duplicate entries of 3 Application Ids (765011468, 671989187 and 653287861)
  - The application ids can left as is since the rest of the variables have different values, also while creating a woe variables and model preparation application id would be dropped
- There are 65 observations in the Age variables which are less than 18 which seems to be incorrect. These values can be imputed with the appropriate mean of the corresponding Education and Profession
  - There are 64 non-default customer and 1 defaulted customer
  - As there are very less possibility that an under age person can be apply for a credit card

Nature of Data and Data Quality Observations (2/2)

- Income variable has 107 observation which are less than equal to 0. These values will be imputed with the mean value of the same gender, education and profession
- Education variables has 119 observations which are blanks. These values cannot be imputed as education cannot be derived from age, income, gender and profession
- Gender has 2 missing values, Marital Status has 6 missing values, No of dependents has 3 missing values and Type of residence has 8 missing values
  - Since these are very small missing values these will be dropped

Missing Value, Outlier Treatment and Derived Variable: Variable wise Treatment

Variable Name	Observation	Imputation or Treatment or Creation Method
Performance Tag	2% missing values in Defaulter which is out target / dependent variable.	As per problem statement, if applicant has gone 90 days past due (DPD) or worse in the past 12 months then that customer is marked as Defaulter. Basis this business rule the missing value are imputed
Age	1. 20 observation less than equal to 0 2. 45 observation grater than 0 but less than 18	1. For values less than equal to 0: Value imputed with the mean age for the most matching profile. Profile attributes are like, Education, Profession, Gender and Marital Status 2. For values greater than 0 and less than 18: Retained as is since there are applicant with age less than 18 but targeted as married which again raise question on data quality.
CC Utilization in 12M	Missing values for 1058 observation	Since all the missing values means that customer has not appeared in any of the variables hence imputing the missing values with 0 for each of the variable.
Trades Opened in 6M	Missing values for 1 observation	
Presence of open home loan	Missing values for 272 observation	
Outstanding Balance	Missing values for 272 observation	
Income per Dependent	There are multiple dependents for each of applicant, hence income per dependent could be a useful variable to know if the customer has sufficient income per dependent	Income of applicant / number of dependents
Have Secured Loans	It would be interesting to know if the customer have secured loans (car loan or home loans) or unsecured loans (personal loan or other unsecured loans which don’t have any collateral against the loan)	If the customer has any of home loan or car loan then marked as 1 else 0
Variable Name Corrections	The provided names were very lengthy and also there were some spelling mistakes in the variables	Shorter and relevant names were assigned and correction in the spelling mistakes were corrected as part of the data quality step

Weight of Evidence (WOE) and Information Value (IV) (1/3)

WOI and IV is a variables transformation and selection technique. Using this technique of the data set we can select useful variable and transform variables to get rid of outlier variables and missing values.

Stage 1

Creating groups for Continuous variable of data set

Stage 2

Calculating WOE Values:

ln(Event%/Non-Event%)

Stage 3

Calculating IV:

*∑(Event% - Non-Event%) * ln(Event%/Non-Event%)*

Stage 4

WOE and IV scores for each of the variable and corresponding groups

Bucketing continuous variables

Demographic Data:

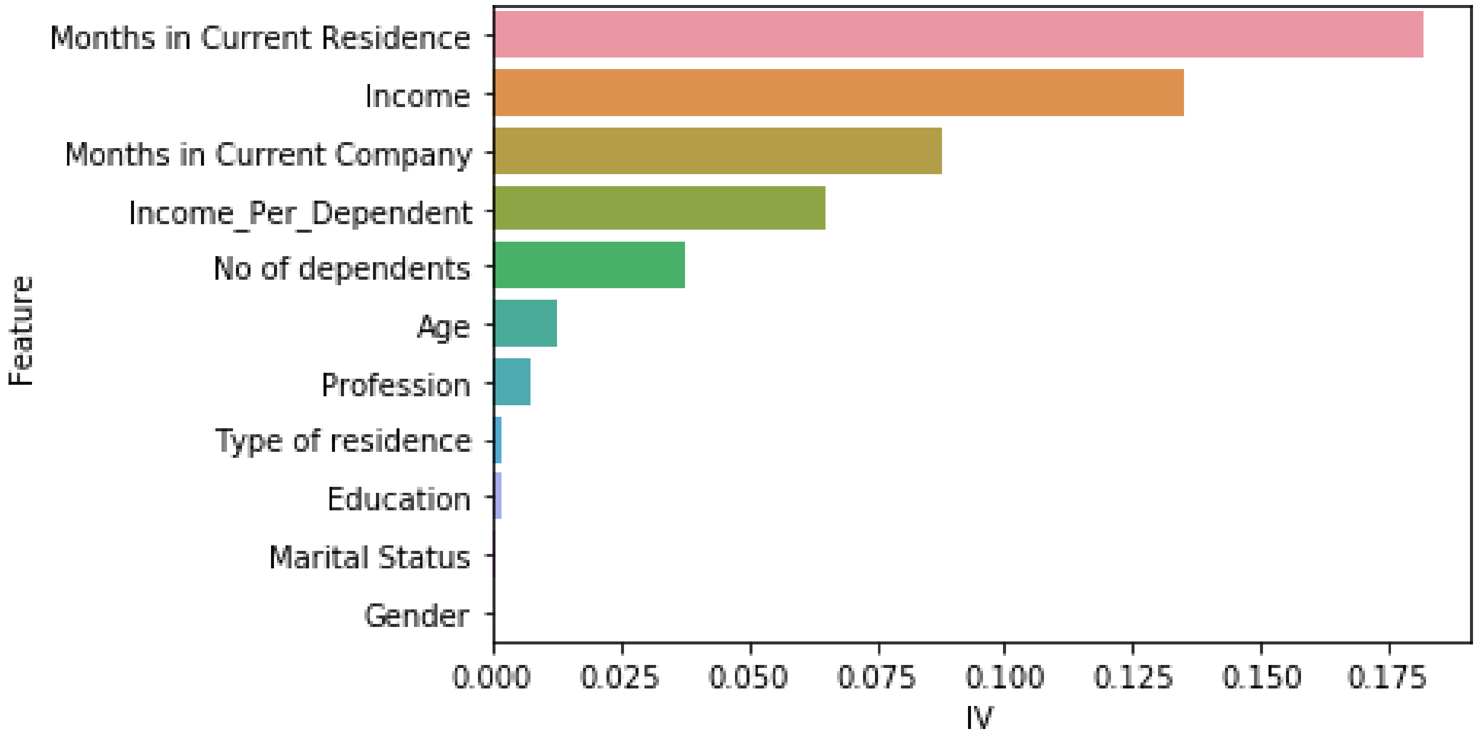
- Age: *0 to 18, 19 to 20, 21 to 25, 26 to 30, 31 to 40, 41 to 50, 51 to 60, 61 to 70 are the groups for Age variable*
- Income: *0 to 10, 11 to 20, 21 to 30, 31 to 40, 41 to 50, 51 to 61 are the groups for Income variable*
- Income_Per_Dependent: *0 to 10, 11 to 20, 21 to 30, 31 to 40, 41 to 50, 51 to 61 are the groups for Income per Dependent variable*
- Months in Current Company: *0 to 6, 7 to 12, 13 to 24, 25 to 36, 37 to 48, 49 to 60, 61 to 72, 73 to 133 are the groups for Months in Current Company*
- Months in Current Residence: *0 to 6, 7 to 12, 13 to 24, 25 to 36, 37 to 48, 49 to 60, 61 to 72, 73 to 84, 85 to 96, 97 to 108, 109 to 120, 121 to 132 are the groups for Months in Current Residence variable*

Credit Bureau Data:

- CC Utilization in 12M: *0 to 1, 2 to 10, 11 to 20, 21 to 30, 31 to 40, 41 to 50, 51 to 60, 61 to 70, 71 to 80, 81 to 90, 91 to 100, 101 to 120 are the groups for CC Utilization in 12M variable*
- Outstanding Balance: *0 to 10k, 10.1k to 50k, 50.1k to 100k, 100.1k to 200k to 200.1k to 300k, 300.1k to 400k, 400.1k to 500k, 500.1k to 1mn, 1.1mn to 2mn to 2.1mn to 3mn, 3.1mn to 4mn, 4.1mn to 5mn, 5.1mn to 6mn are the groups for Outstanding Balance*
- Total No of Trades: *0 to 5, 6 to 10, 11 to 15, 16 to 20, 21 to 25, 26 to 30, 31 to 35, 36 to 45 are the groups for Total No. of Trades*
- Trades Opened in 6M: *All the numbers till 9 are individual group and last group is from 10 to 12*
- Trades Opened in 12M: *Groups formed with interval of 2 starting from 0 and the last group is from 21 to 28*
- PL Trades Opened in 12M: *All the numbers till 9 are individual group and last group is from 10 to 12*
- Inquiries in 12M: *Groups formed with interval of 2 starting from 0 till 20.*

Weight of Evidence (WOE) and Information Value (IV) for Demographic Data

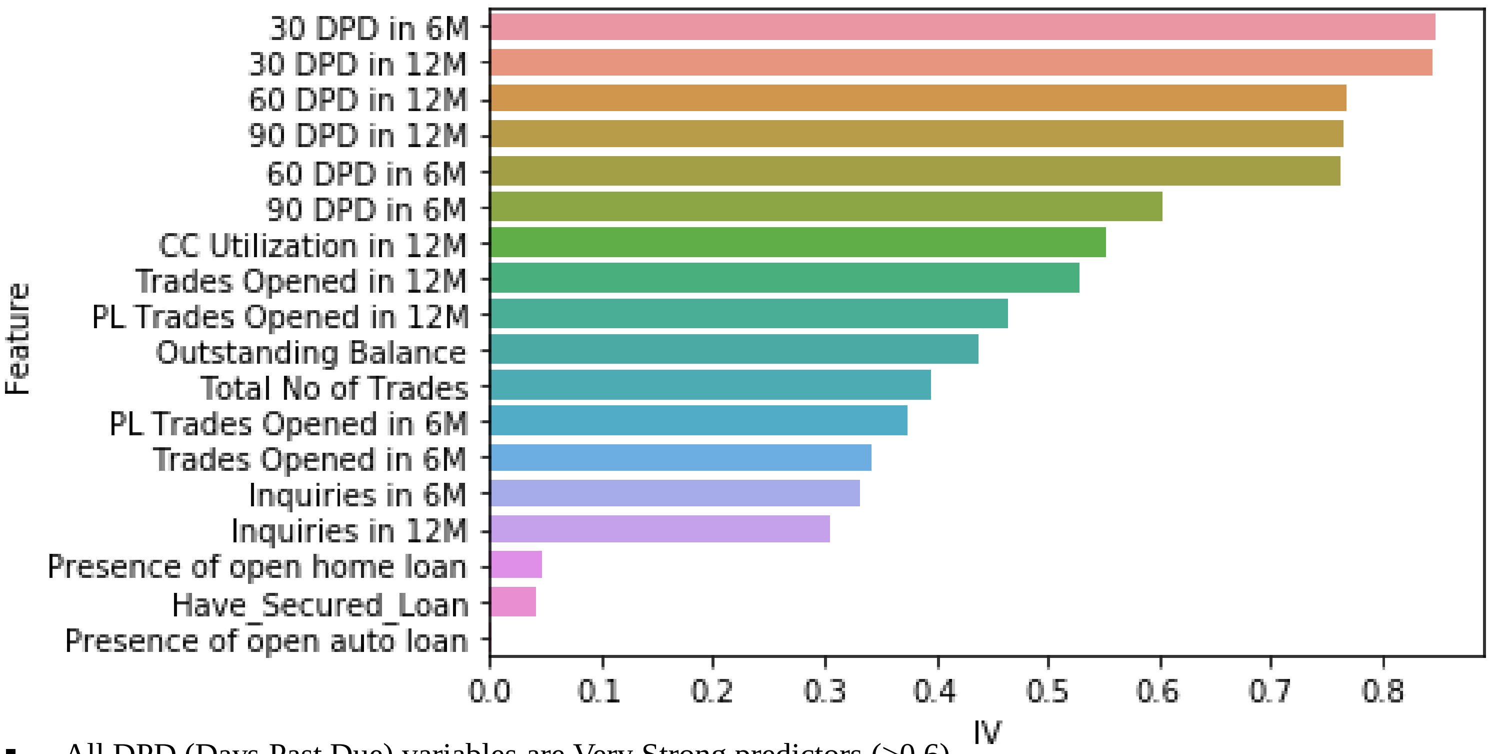
Feature	IV (Information Value)
Months in Current Residence	0.181553
Income	0.135013
Months in Current Company	0.087902
Income_Per_Dependent	0.065015
No of dependents	0.037736
Age	0.012391
Profession	0.007376
Type of residence	0.002033
Education	0.001575
Marital Status	0.000622
Gender	0.000092



- Months in Current Residence(0.18), Income(0.13) variable have Medium Predictor Power
- Where as Months in Current Company(0.09), Income_Per_Dependent(0.06) & No of dependents(0.04) variables are Weak Predictors.
- Remaining variables are not useful for Prediction.

Weight of Evidence (WOE) and Information Value (IV) for Credit Bureau

Feature	IV (Information Value)
30 DPD in 6M	0.848371
30 DPD in 12M	0.845984
60 DPD in 12M	0.768439
90 DPD in 12M	0.765552
60 DPD in 6M	0.761867
90 DPD in 6M	0.604573
CC Utilization in 12M	0.553363
Trades Opened in 12M	0.528978
PL Trades Opened in 12M	0.464845
Outstanding Balance	0.438889
Total No of Trades	0.395969
PL Trades Opened in 6M	0.374655
Trades Opened in 6M	0.341709
Inquiries in 6M	0.333308
Inquiries in 12M	0.304615
Presence of open home loan	0.046253
Have_Secured_Loan	0.040865
Presence of open auto loan	0.001917

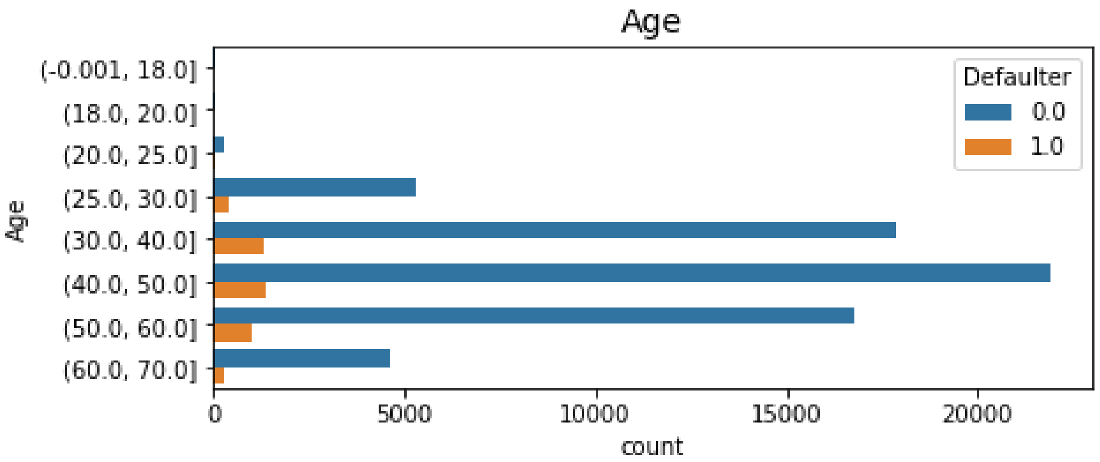


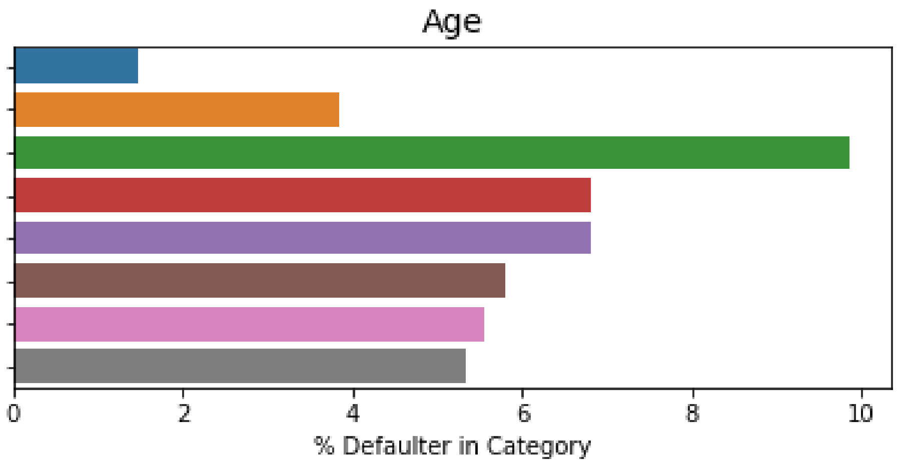
- All DPD (Days Past Due) variables are Very Strong predictors (>0.6)
- Apart from DPD variable Credit Card Utilization, Trade Opened in 12M, Personal Loan Trades Opened in 12M, Inquiries in 12M are strong predictors with information value greater than 0.5
- Months in Current Residence(0.18) & Income(0.13) variable have Medium Predictor Power
- Presence of open home loan, Presence of open auto loan variable scores very less which indicates they are not useful for prediction.

Univariate Analysis

For Customer Demographic and Credit Bureau Data

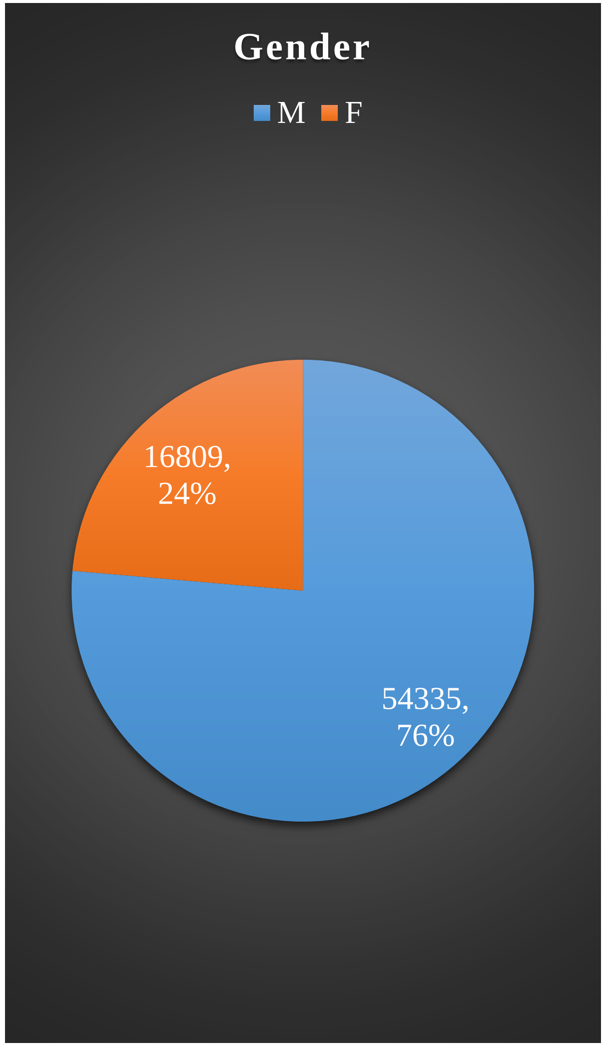
Applicant Age Distribution

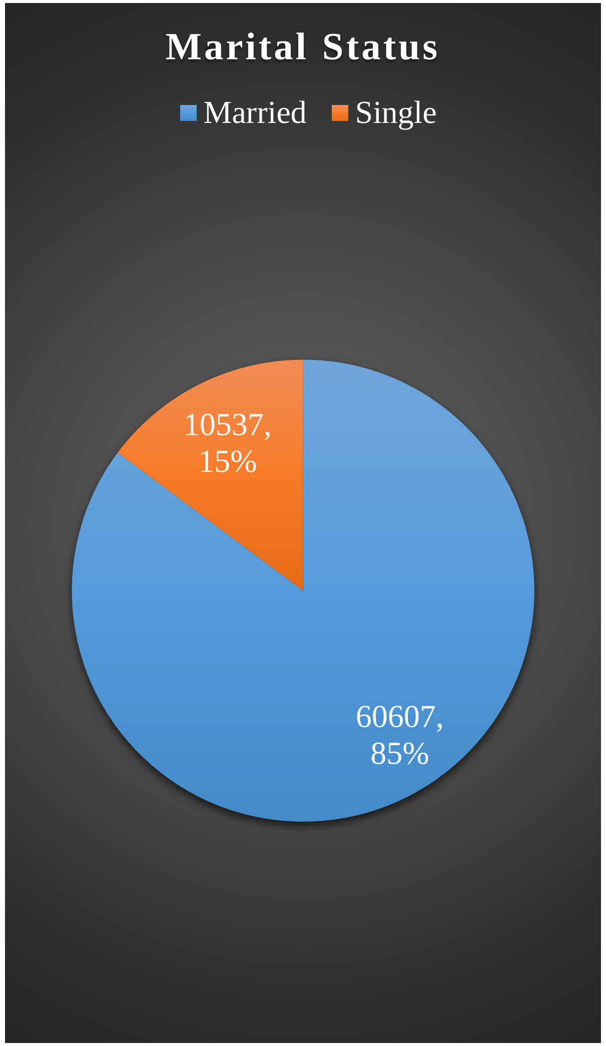


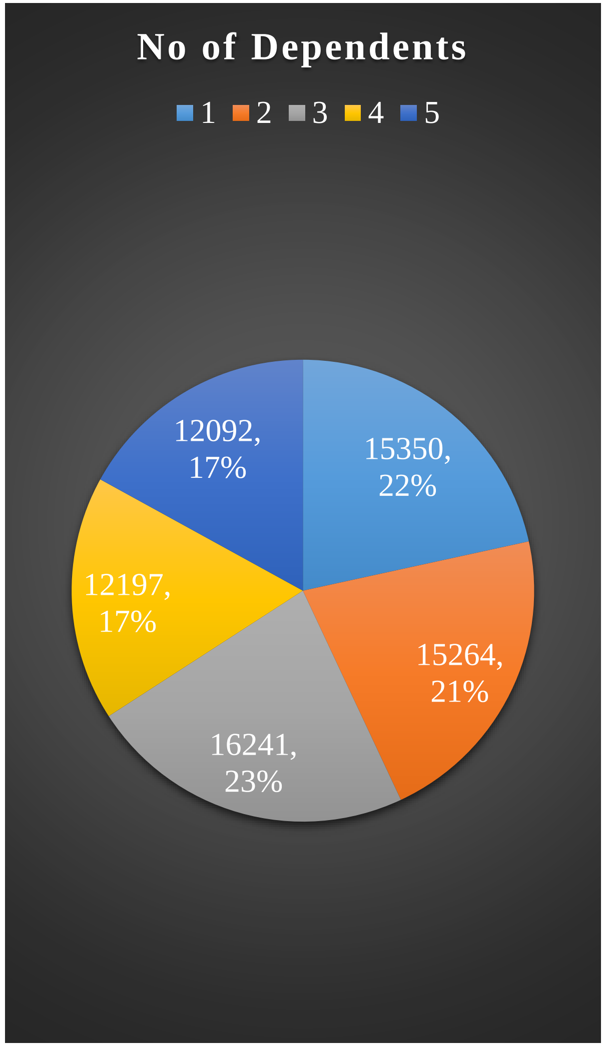


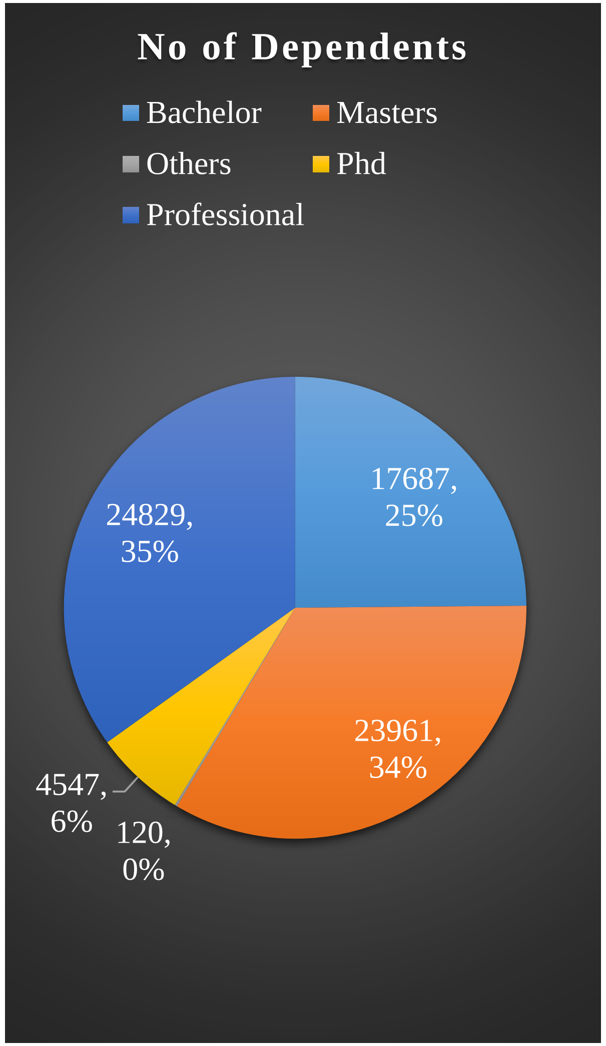
Age	No of Application	No of Defaulter	% Defaulter in Category	% Defaulter in All Defaultor
(-0.001, 18.0]	68	1	1.47	0.02
(18.0, 20.0]	52	2	3.85	0.05
(20.0, 25.0]	304	30	9.87	0.69
(25.0, 30.0]	5668	386	6.81	8.94
(30.0, 40.0]	19137	1303	6.81	30.17
(40.0, 50.0]	23237	1348	5.8	31.21
(50.0, 60.0]	17792	988	5.55	22.88
(60.0, 70.0]	4886	261	5.34	6.04

Gender, Marital Status, No. of Dependents and Education

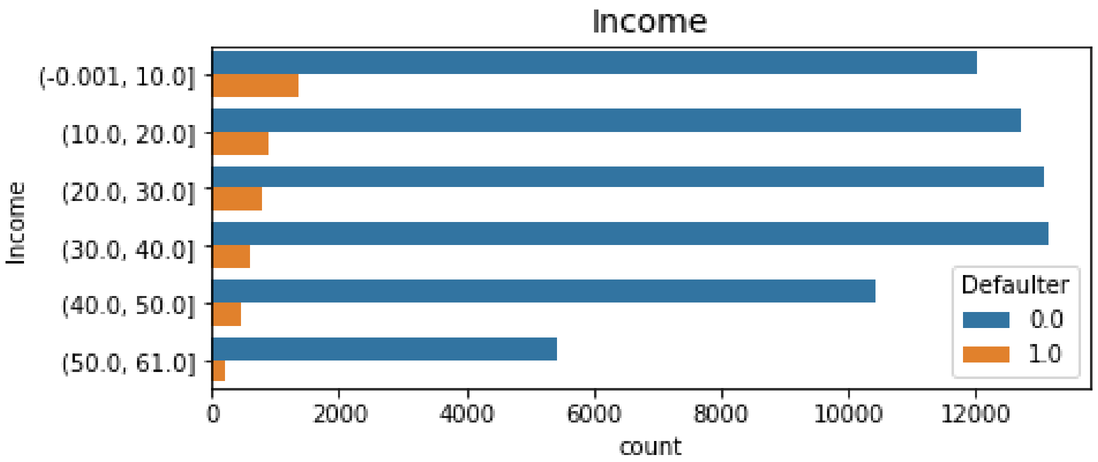


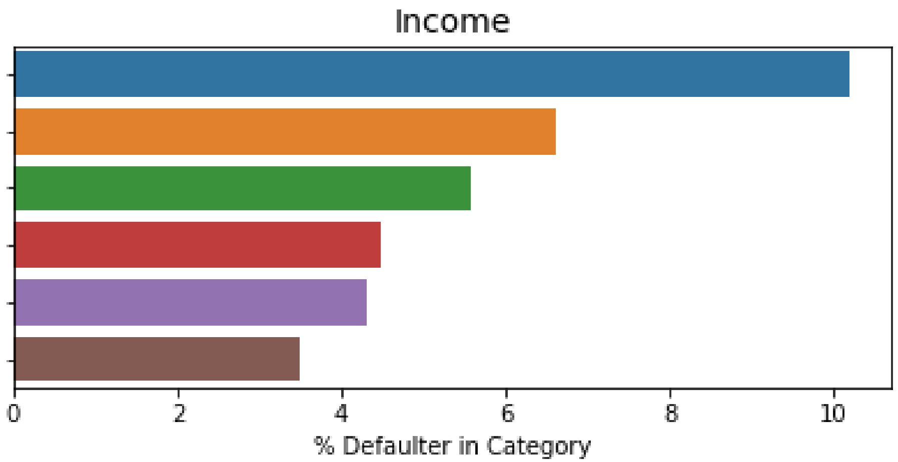






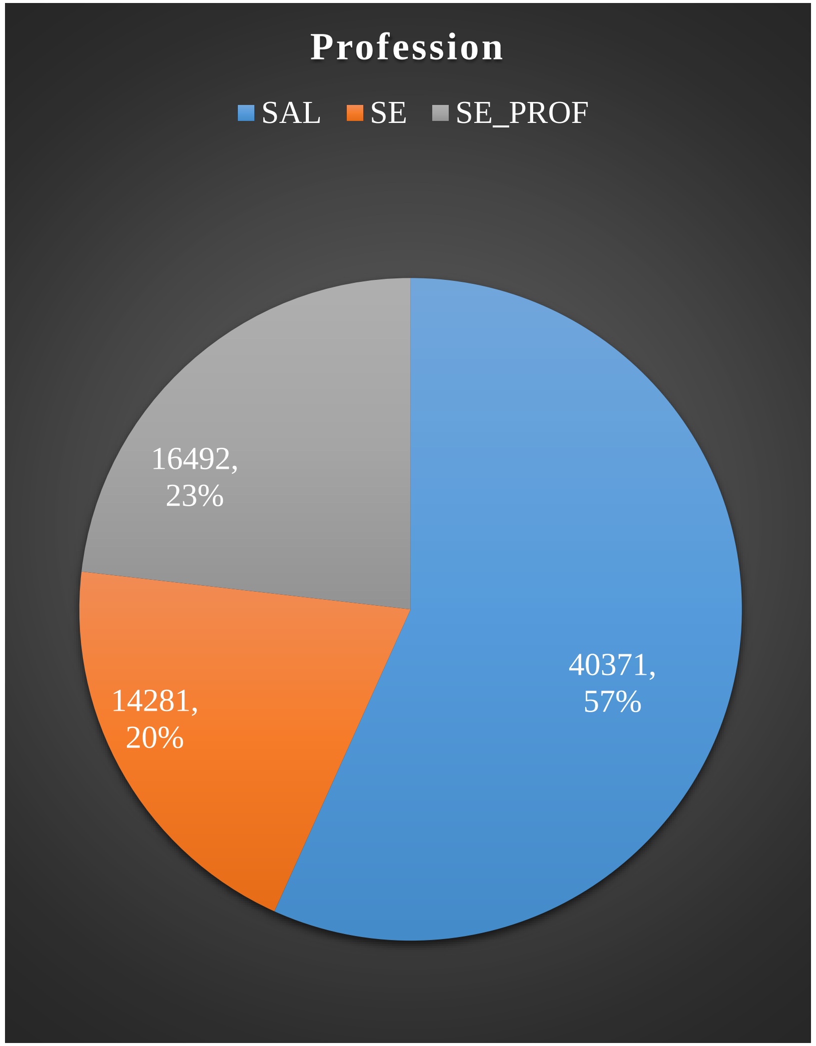
Income Distribution

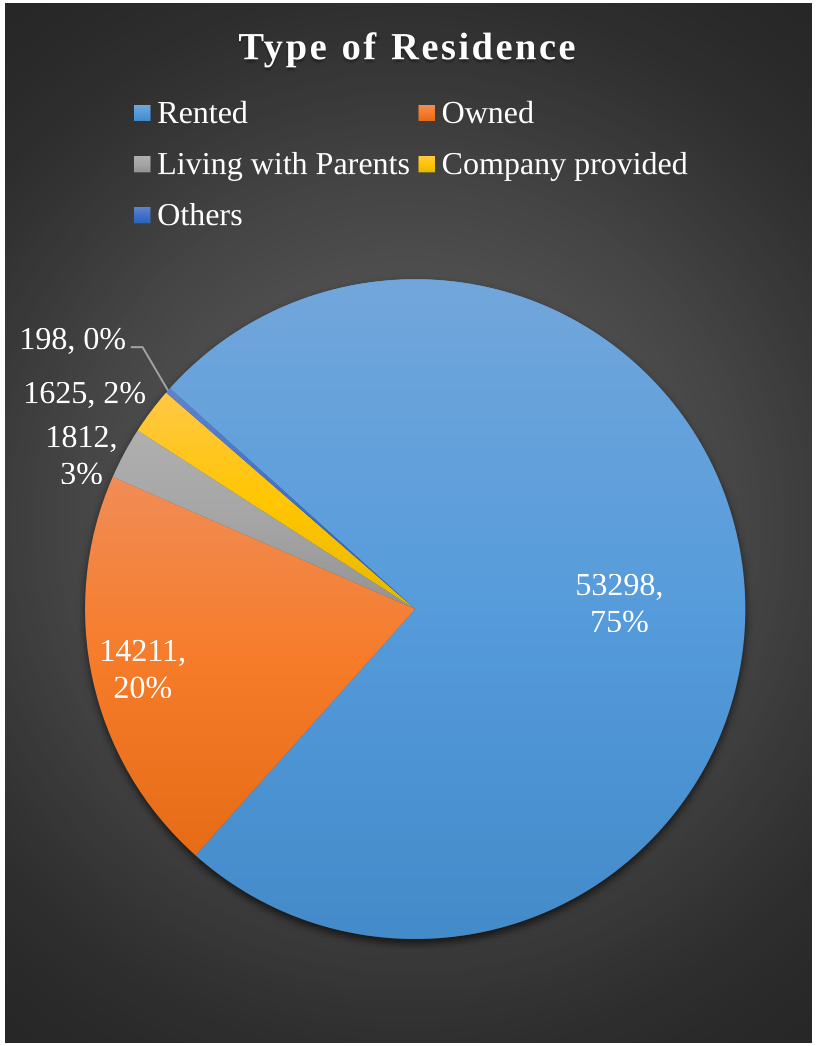


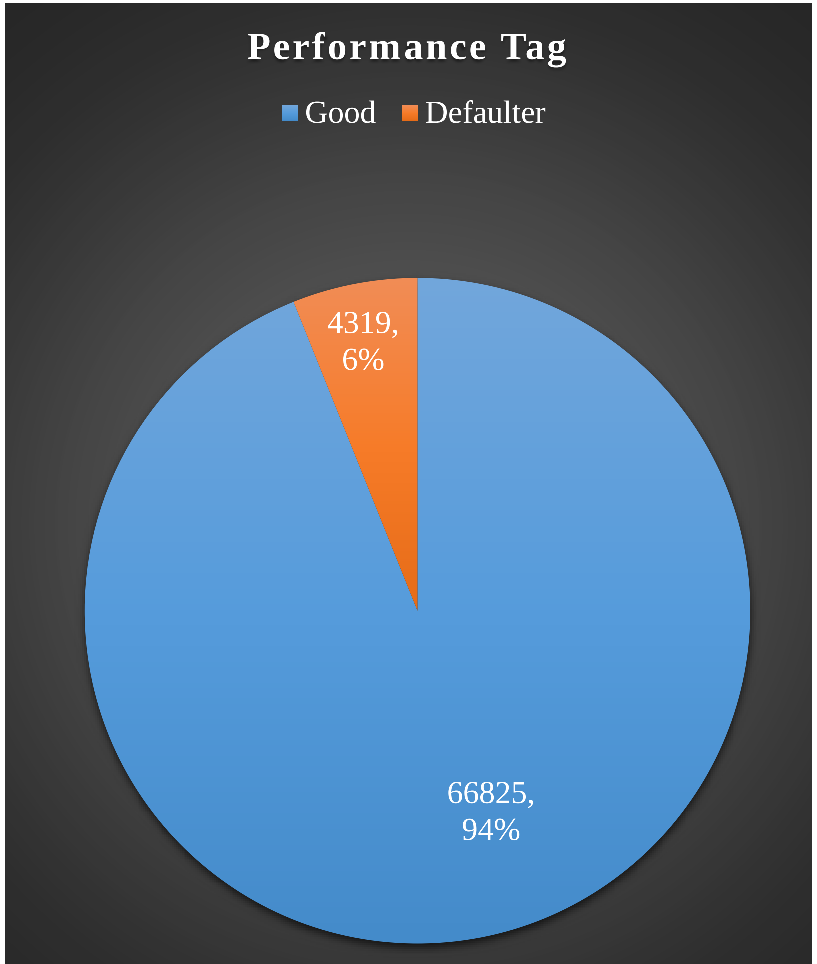


Income	No of Application	No of Defaulter	% Defaulter in Category	% Defaulter in All Defaultor
(-0.001, 10.0]	13375	1363	10.19	31.56
(10.0, 20.0]	13629	902	6.62	20.88
(20.0, 30.0]	13842	771	5.57	17.85
(30.0, 40.0]	13754	616	4.48	14.26
(40.0, 50.0]	10908	471	4.32	10.91
(50.0, 61.0]	5636	196	3.48	4.54

Profession, Type of Residence and Performance Tag

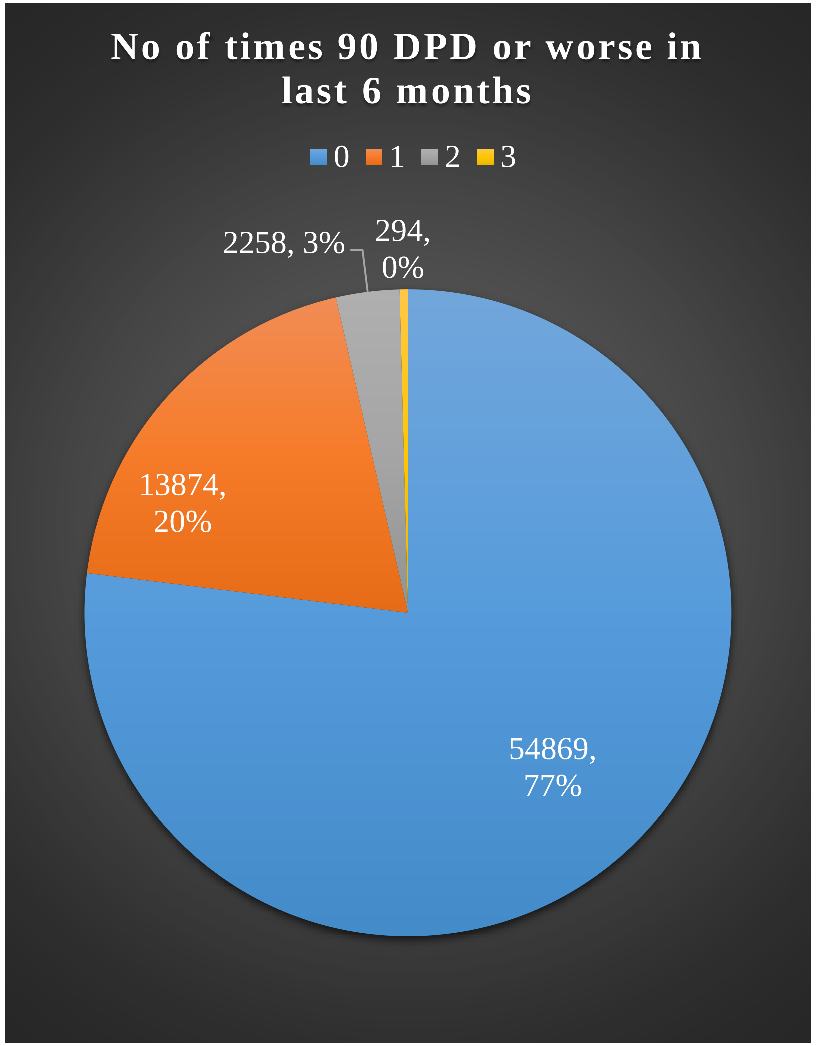


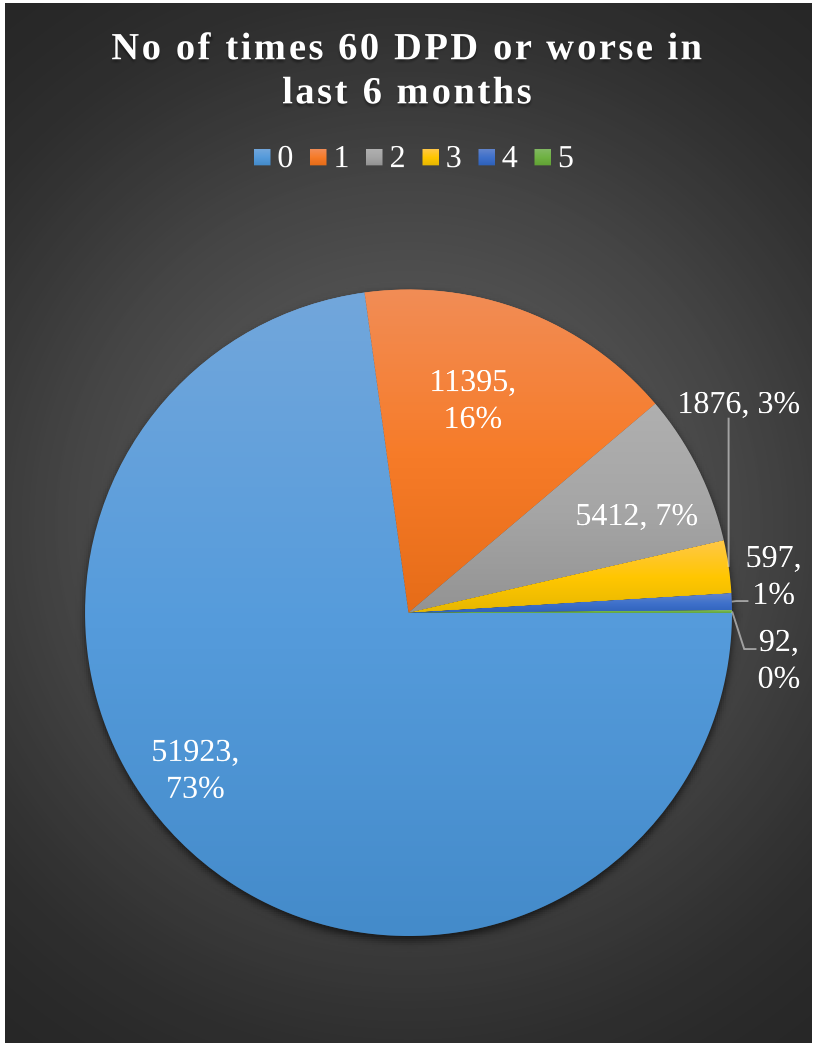


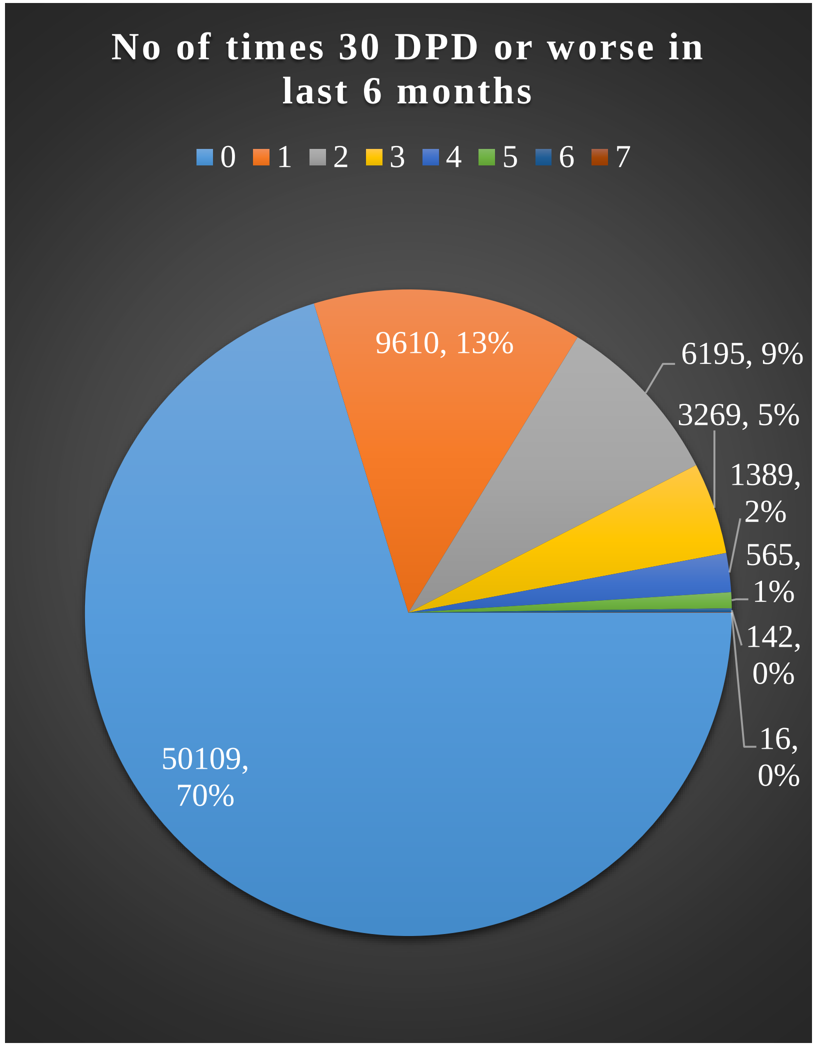


As it is evident from the distribution there is a clear class imbalance

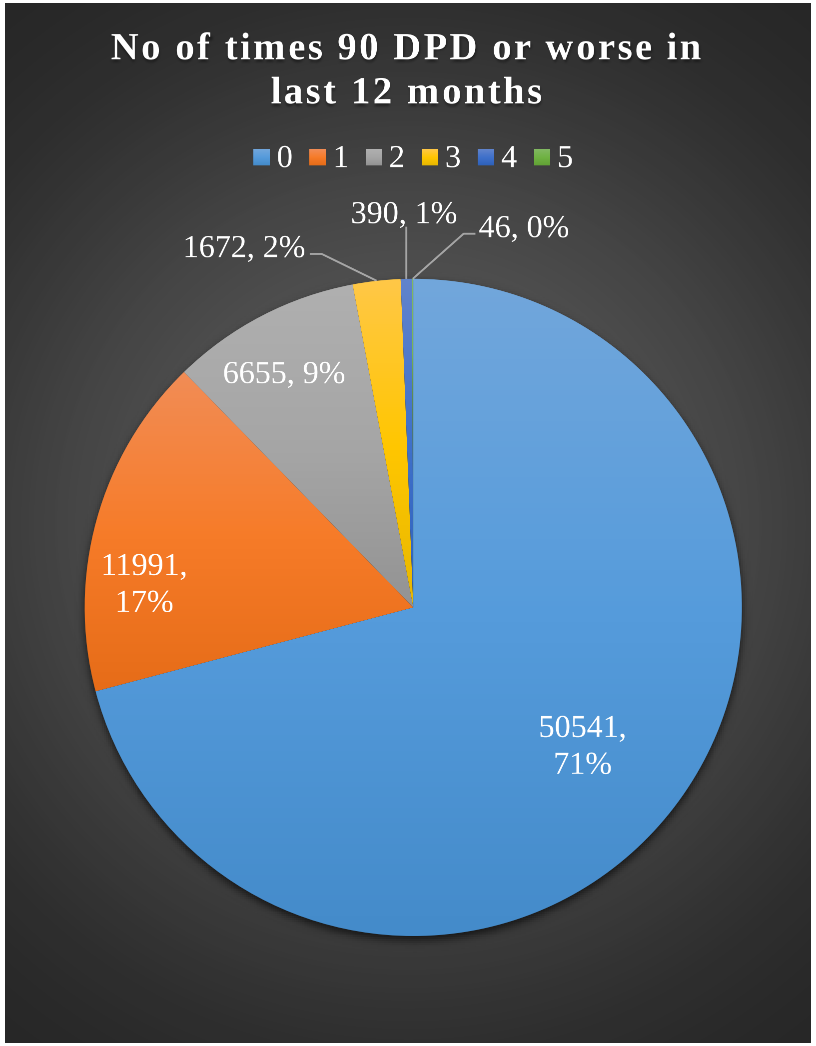
Credit Bureau Data: No of times 90, 60 and 30 DPD or worse in last 6 months

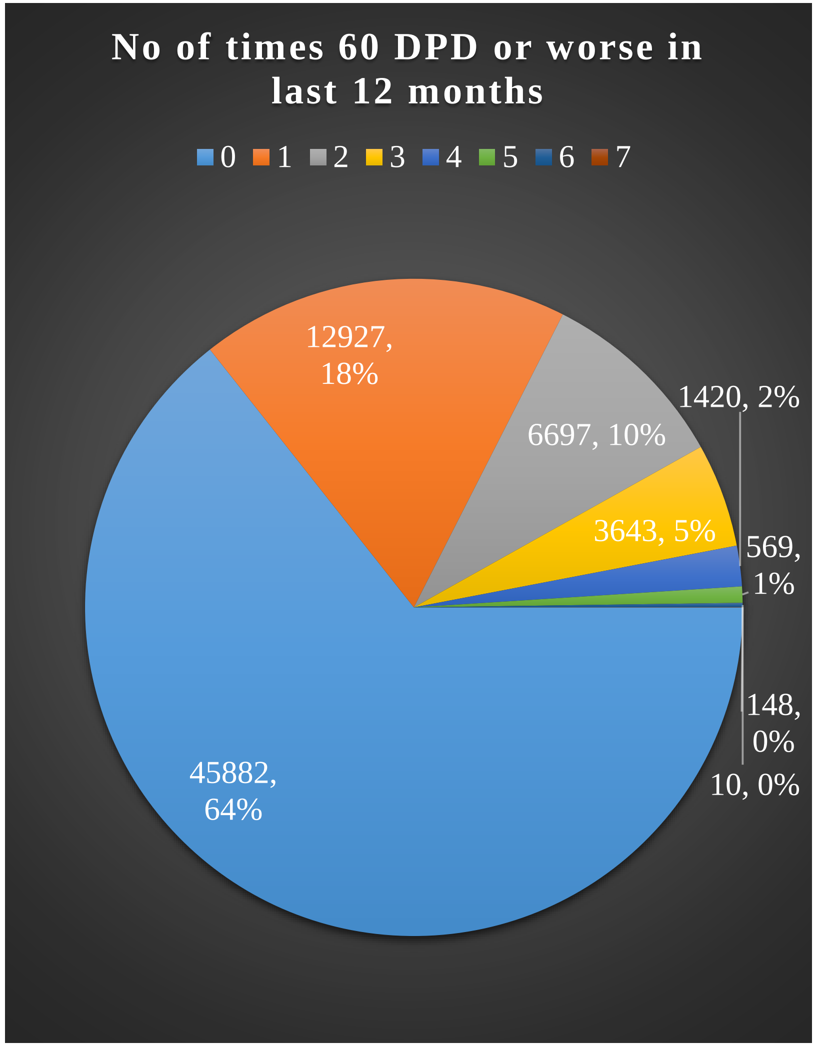


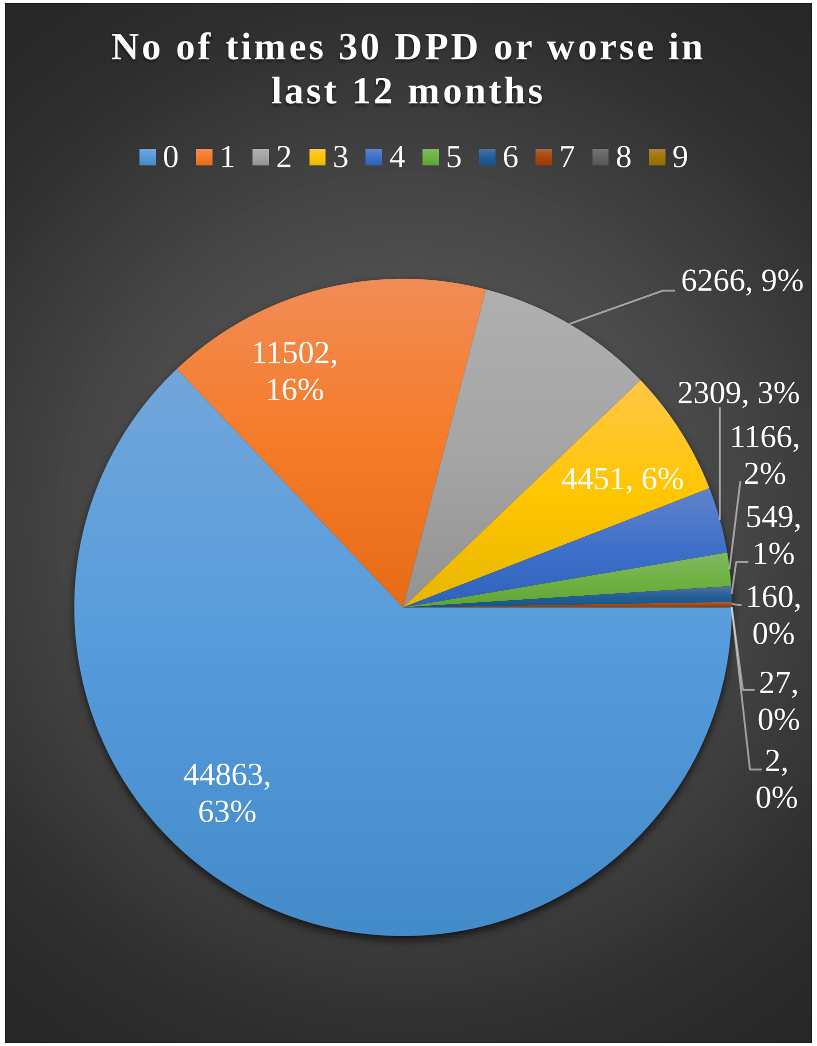




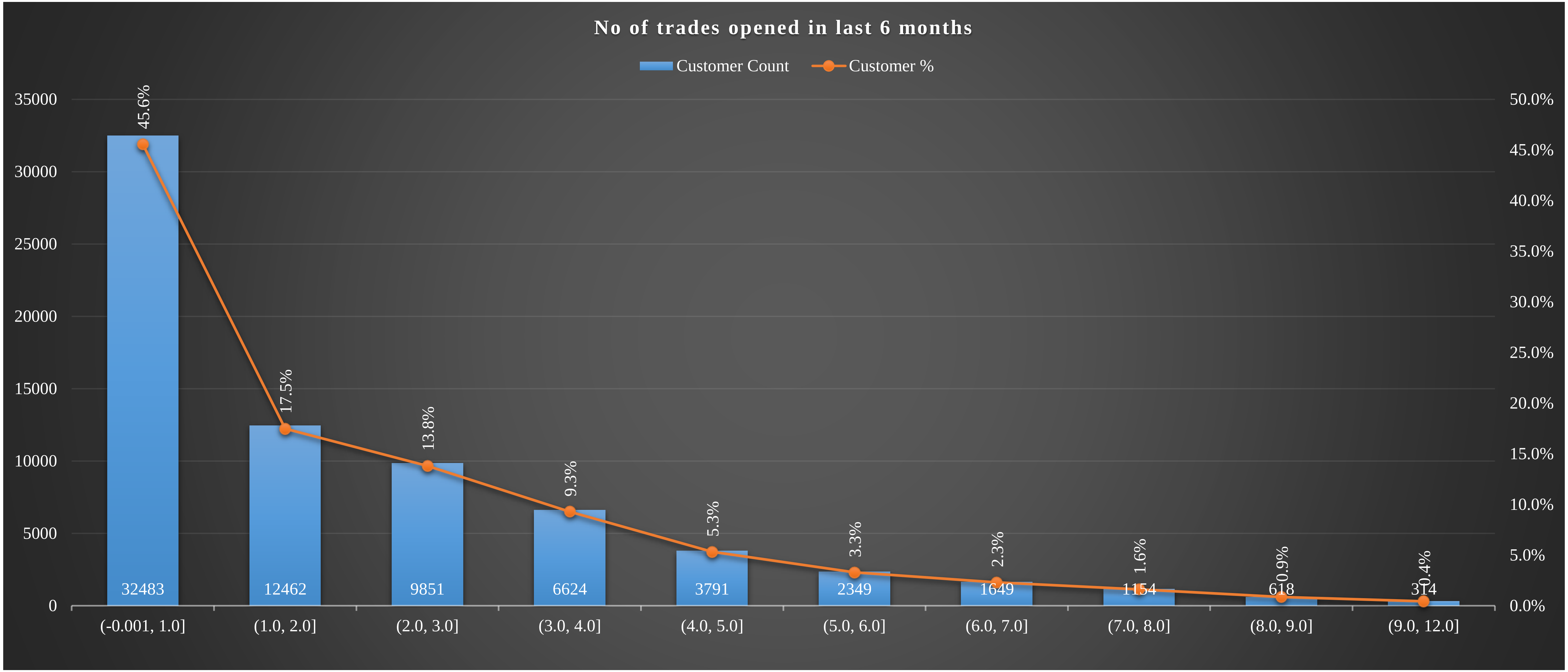
Credit Bureau Data: No of times 90, 60 and 30 DPD or worse in last 12 months



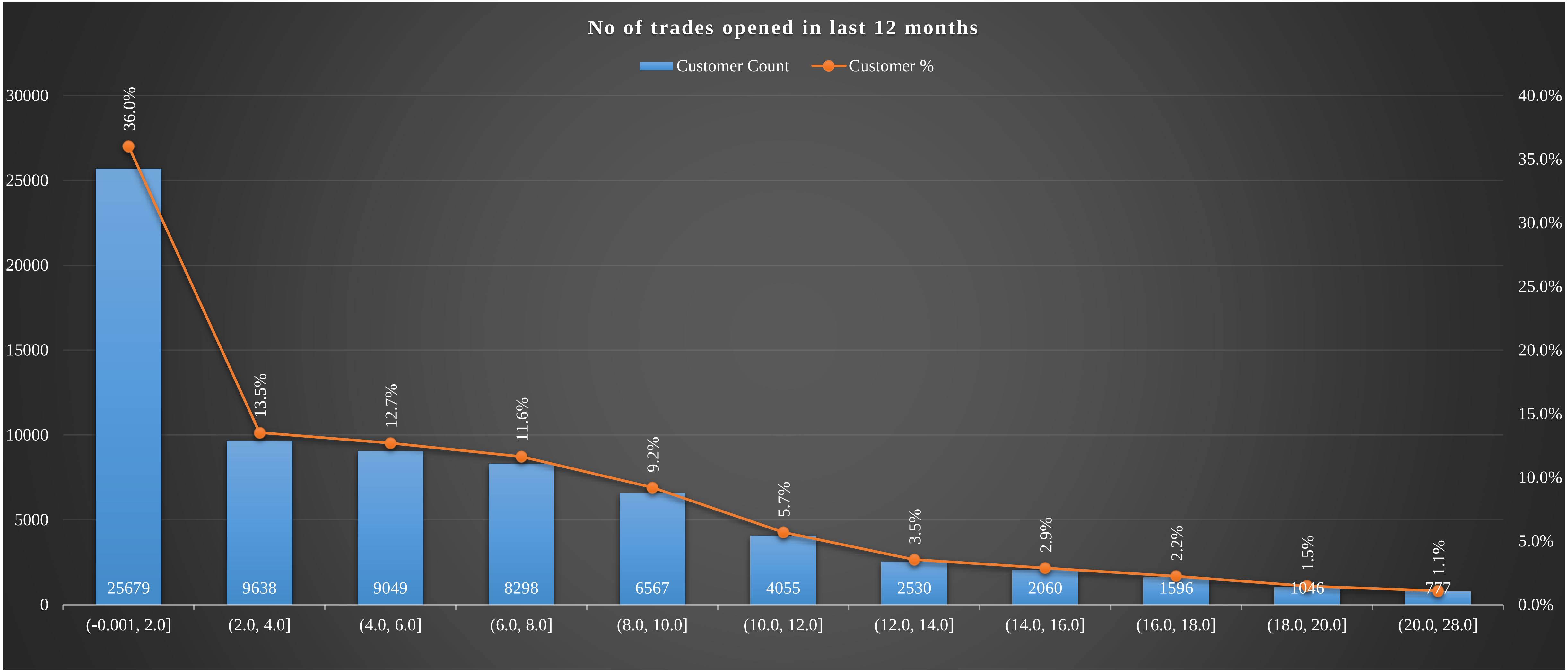




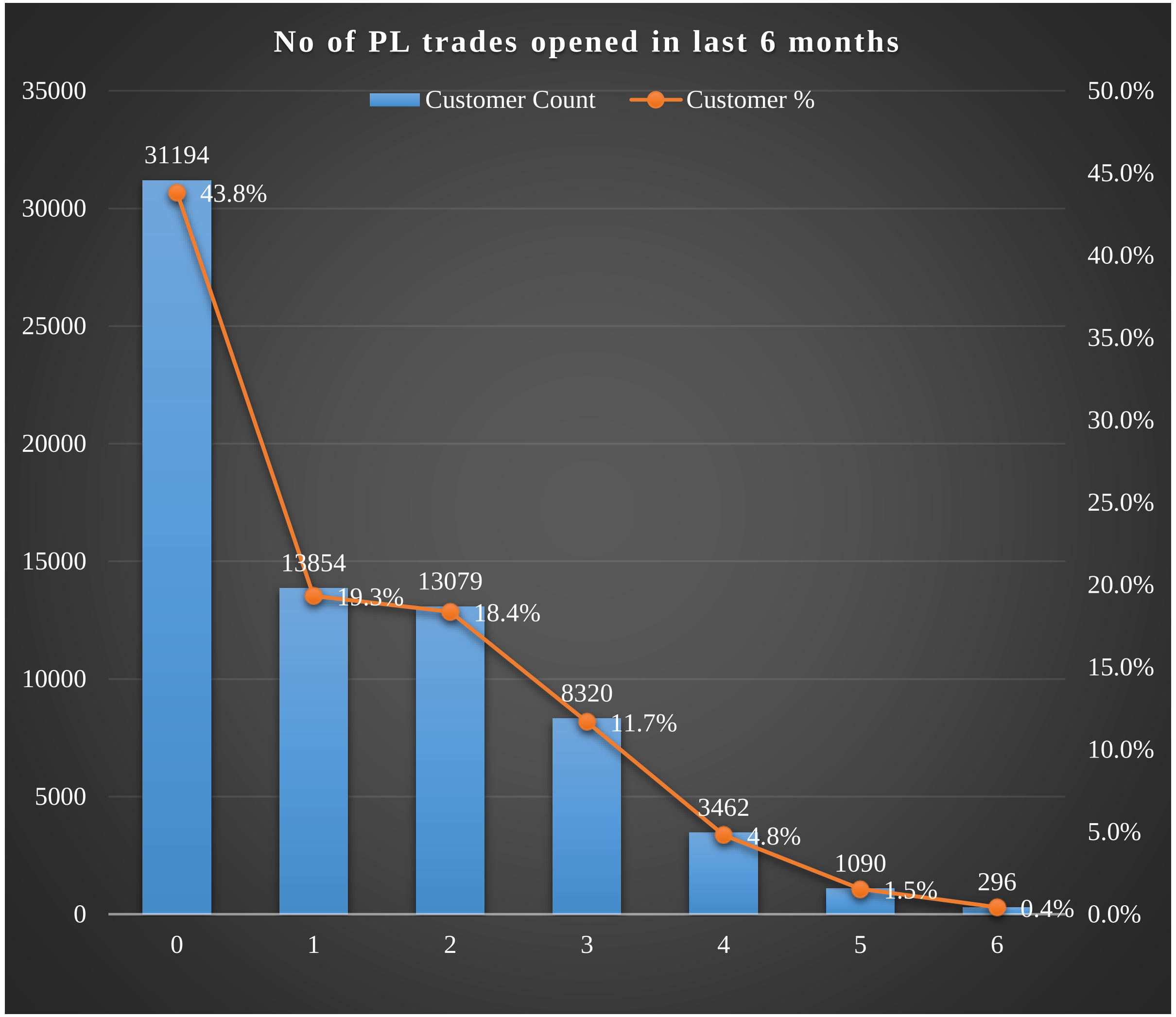
Credit Bureau Data: No of trades opened in last 6 months

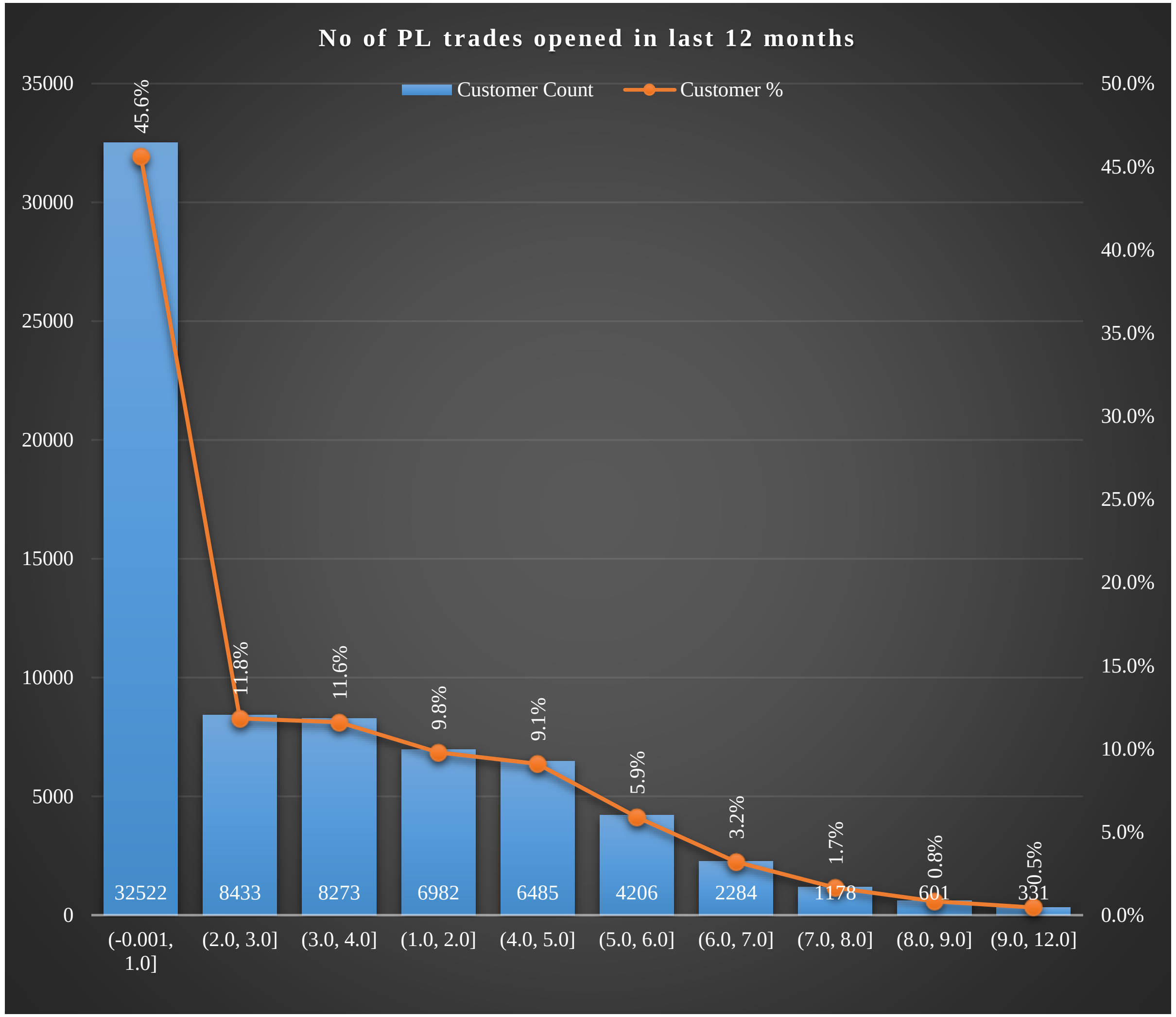


Credit Bureau Data: No of trades opened in last 12 months

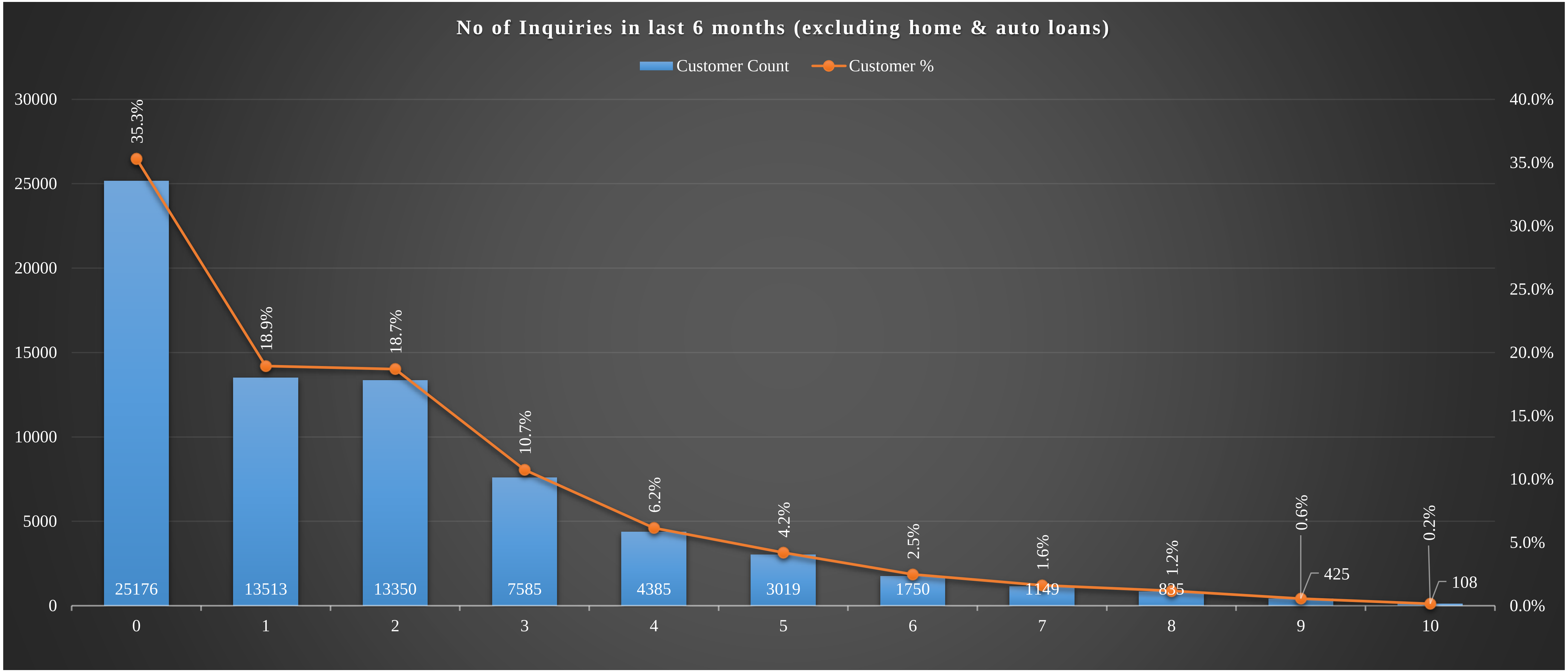


Credit Bureau Data: No of PL trades opened in last 6 and 12 months

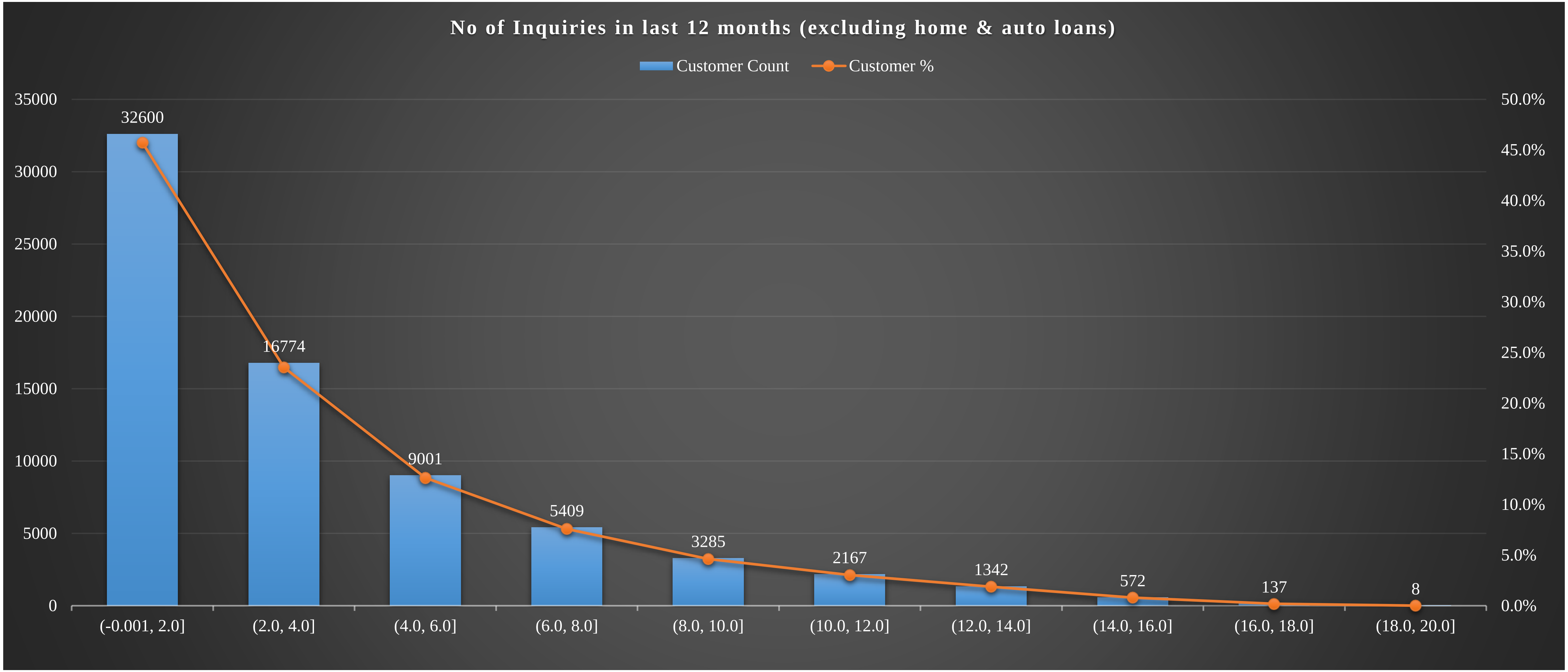




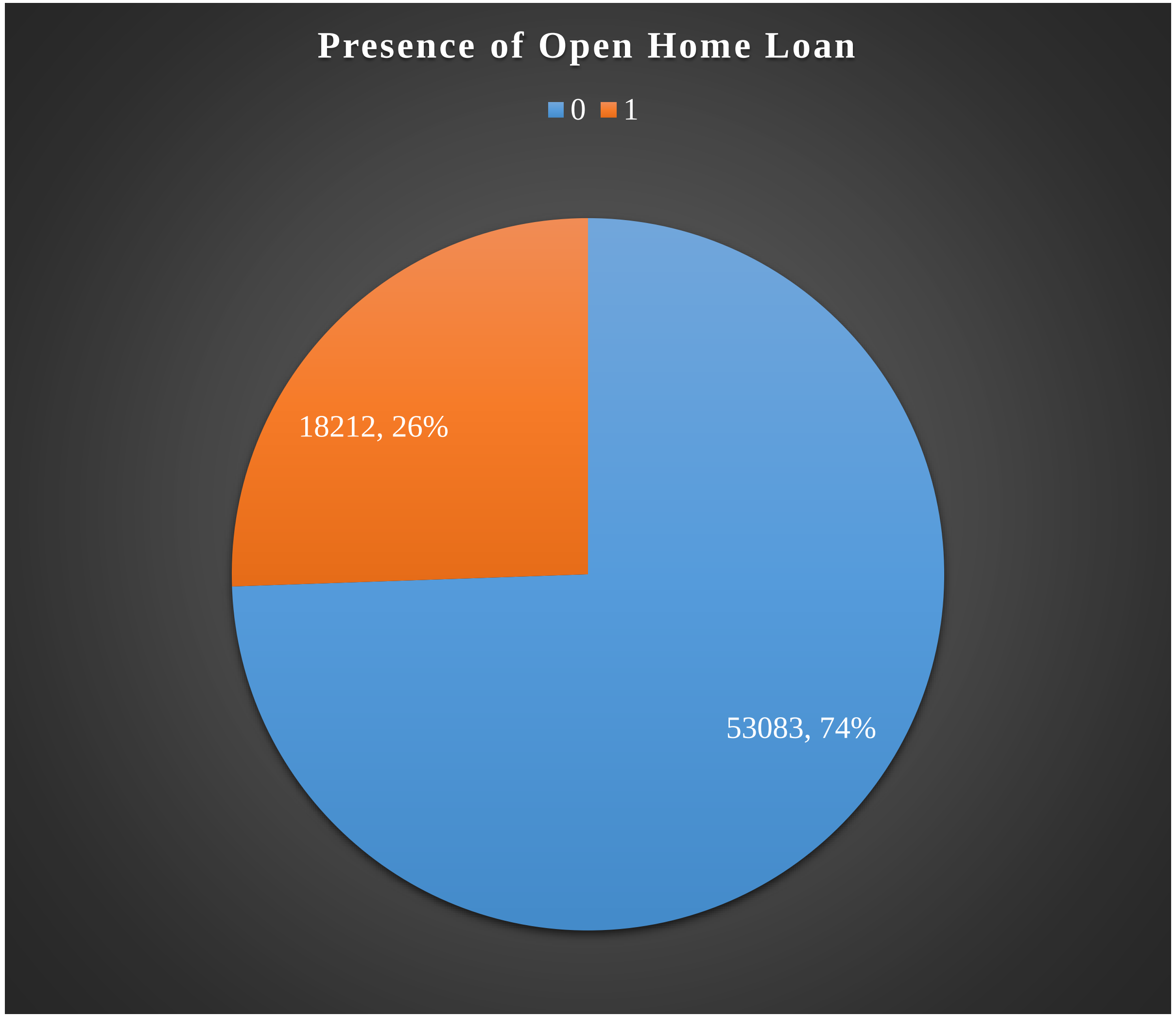
Credit Bureau Data: No of Inquiries in last 6 months (excluding home & auto loans)

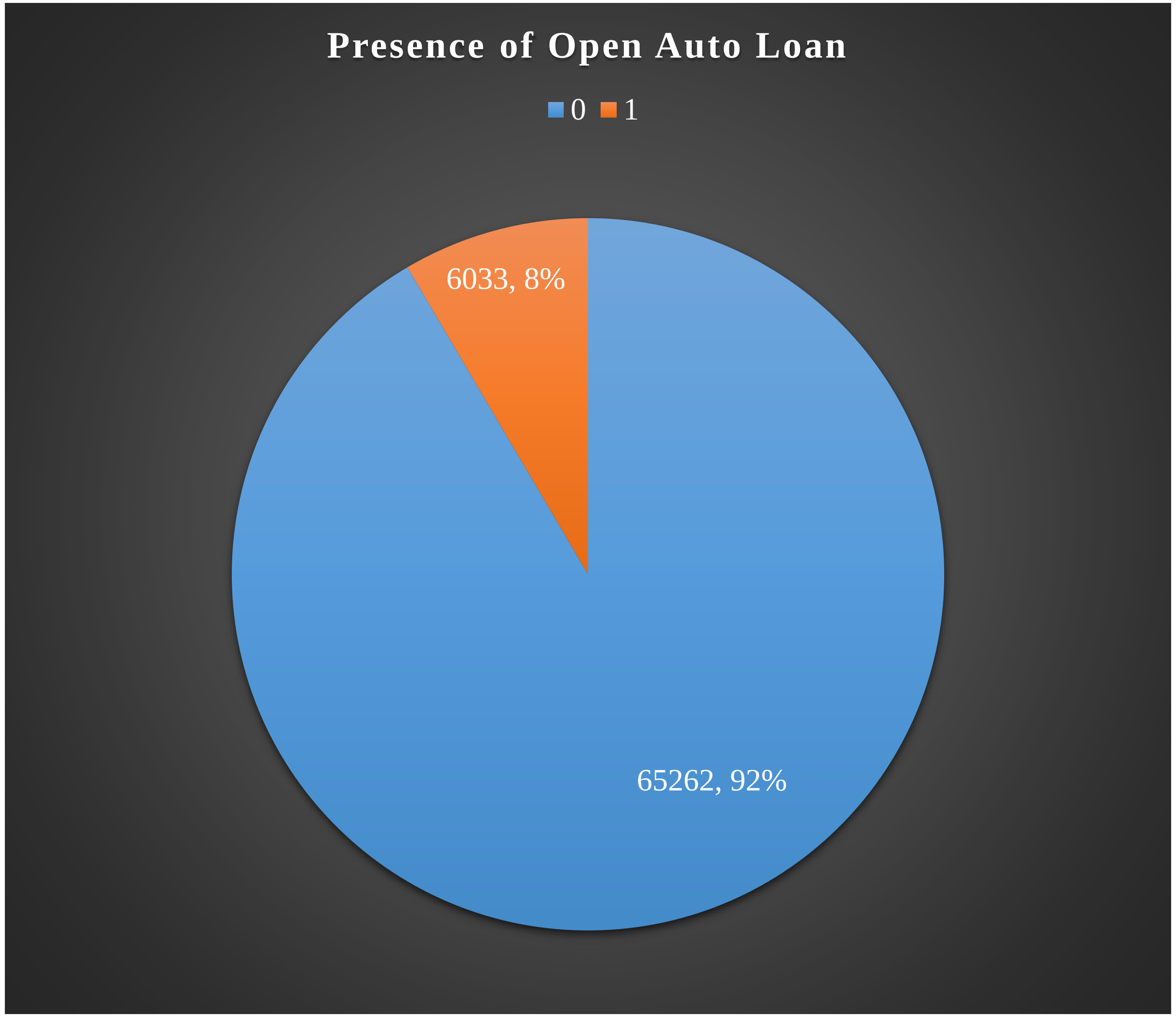


Credit Bureau Data: No of Inquiries in last 12 months (excluding home & auto loans)

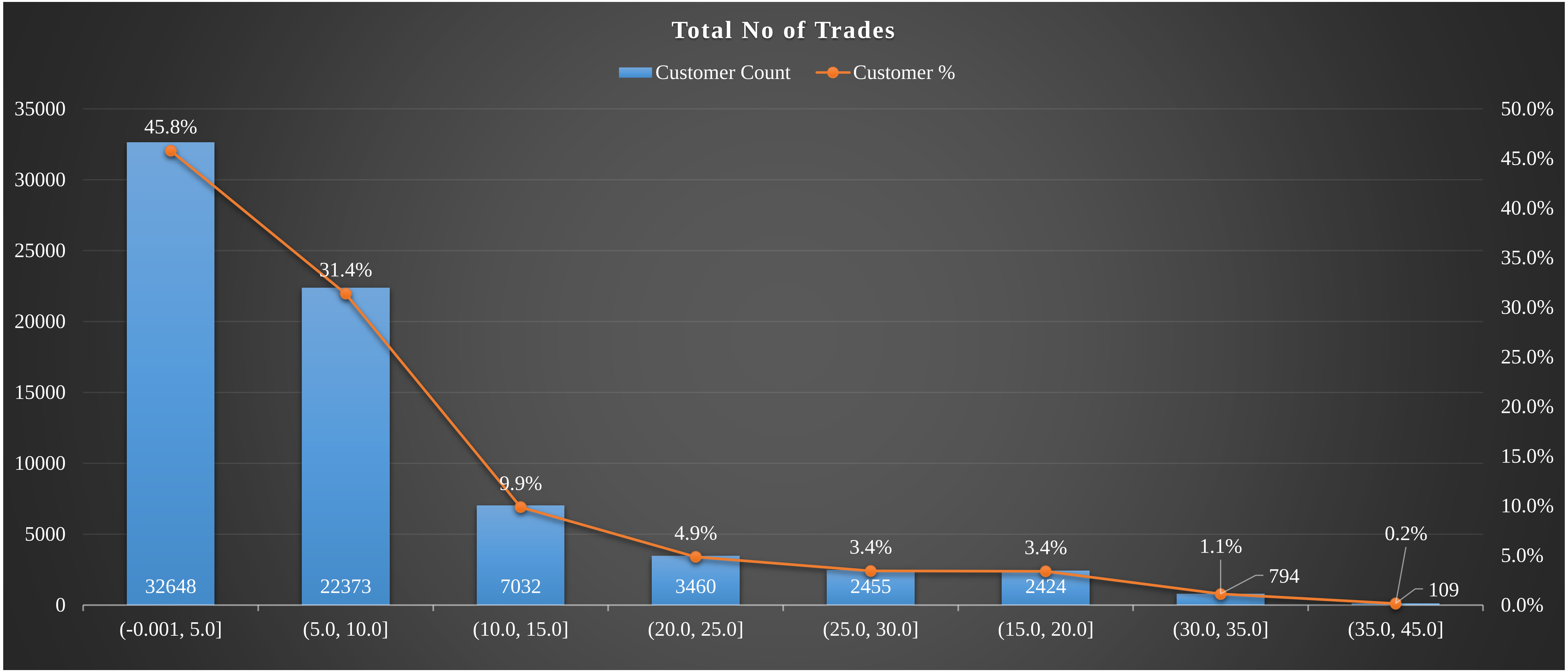


Credit Bureau Data: Presence of Open Home Loan, Auto Loan





Credit Bureau Data: Total No. Trades



Insights

Multivariate Analysis

Who is the Defaulter?

Top 6 Customer Profiles Contributing to 37% of Default Customers

7% defaults are caused by the applicants with:

- Gender: Male
- Age: 41 to 50
- Marital Status: Married
- Education: Professional

7% defaults are caused by the applicants with:

- Gender: Male
- Age: 41 to 50
- Marital Status: Married
- Education: Masters

6% defaults are caused by the applicants with:

- Gender: Male
- Age: 51 to 60
- Marital Status: Married
- Education: Masters

6% defaults are caused by the applicants with:

- Gender: Male
- Age: 51 to 60
- Marital Status: Married
- Education: Professional

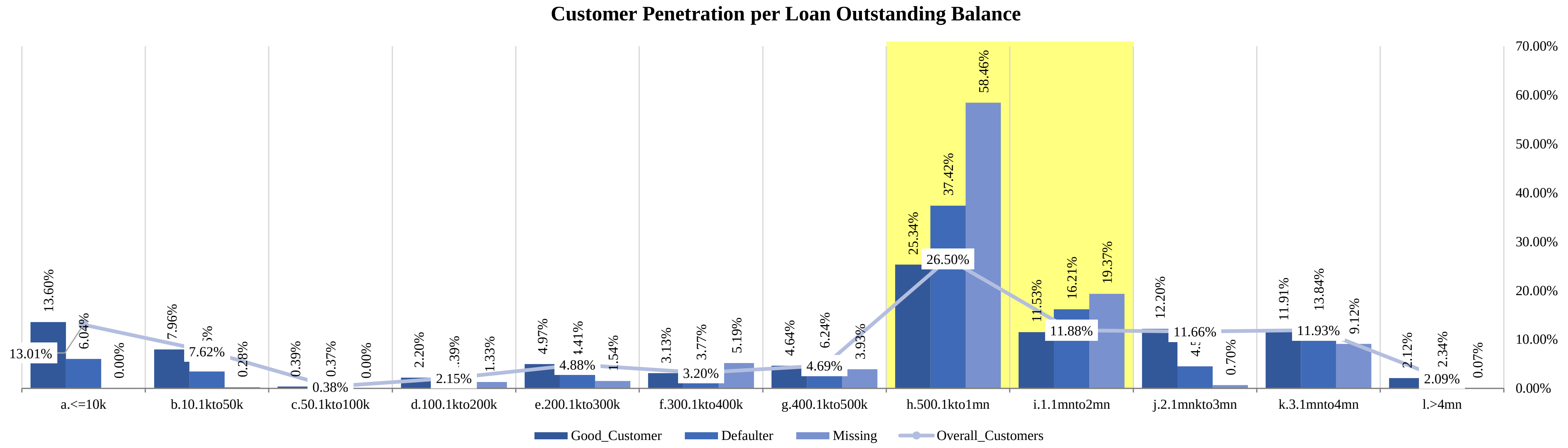
6% defaults are caused by the applicants with:

- Gender: Male
- Age: 31 to 40
- Marital Status: Married
- Education: Masters

6% defaults are caused by the applicants with:

- Gender: Male
- Age: 31 to 40
- Marital Status: Married
- Education: Professional

Outstanding Balance and Customer Penetration for Defaulters



- Every second defaulter (54% of the defaulters) have the outstanding balance between 500k to 2 Mn.
- 96% of the defaulter have neither car loan or home loan

Salary Group and Outstanding Balance for Defaulters

Outstanding_Balance_Grp	a.<=10		b.11to20		c.21to30		d.31to40		e.41to50		f.51to60	
a.<=10k	21	2.9%	18	3.0%	46	7.7%	40	8.2%	36	9.4%	17	10.6%
b.10.1kto50k	14	2.0%	19	3.1%	16	2.7%	22	4.5%	23	6.0%	8	5.0%
c.50.1kto100k	3	0.4%	2	0.3%	3	0.5%	2	0.4%			1	0.6%
d.100.1kto200k	5	0.7%	8	1.3%	9	1.5%	9	1.9%	5	1.3%	5	3.1%
e.200.1kto300k	28	3.9%	29	4.8%	24	4.0%	19	3.9%	22	5.7%	8	5.0%
f.300.1kto400k	31	4.3%	29	4.8%	23	3.8%	10	2.1%	15	3.9%	3	1.9%
g.400.1kto500k	32	4.5%	47	7.8%	39	6.5%	35	7.2%	21	5.5%	10	6.2%
h.500.1kto1mn	329	46.1%	230	38.1%	220	36.6%	169	34.8%	110	28.6%	45	28.0%
i.1.1mnto2mn	131	18.4%	104	17.2%	87	14.5%	74	15.3%	58	15.1%	24	14.9%
j.2.1mnkto3mn	9	1.3%	21	3.5%	25	4.2%	30	6.2%	35	9.1%	13	8.1%
k.3.1mnto4mn	97	13.6%	89	14.7%	87	14.5%	60	12.4%	50	13.0%	25	15.5%
l.>4mn	13	1.8%	8	1.3%	22	3.7%	15	3.1%	9	2.3%	2	1.2%
Grand Total	**713**	**100%**	**604**	**100%**	**601**	**100%**	**485**	**100%**	**384**	**100%**	**161**	**100%**

- Every 4th Defaulter has income less than 20k and outstanding balance greater than Rs. 500k and none of them has any of the secured loans
- Every 5th Defaulter come from a segment with outstanding balance less than Rs. 50k and salary between 31 to 60

Salary Group and Total Trades for Defaulters

Total_No_of_Trades_Grp	a.<=10		b.11to20		c.21to30		d.31to40		e.41to50		f.51to60	
a.<=5	140	19.6%	139	23.0%	170	28.3%	157	32.4%	139	36.2%	67	41.6%
b.6to10	354	49.6%	299	49.5%	265	44.1%	186	38.4%	147	38.3%	55	34.2%
c.11to15	162	22.7%	103	17.1%	87	14.5%	70	14.4%	33	8.6%	21	13.0%
d.16to20	9	1.3%	19	3.1%	15	2.5%	18	3.7%	17	4.4%	6	3.7%
e.21to25	23	3.2%	22	3.6%	33	5.5%	29	6.0%	21	5.5%	3	1.9%
f.26to30	19	2.7%	17	2.8%	23	3.8%	21	4.3%	18	4.7%	4	2.5%
g.31to35	5	0.7%	4	0.7%	8	1.3%	4	0.8%	9	2.3%	5	3.1%
i.>35	1	0.1%	1	0.2%								
Grand Total	**713**	**100%**	**604**	**100%**	**601**	**100%**	**485**	**100%**	**384**	**100%**	**161**	**100%**

- 12% of the defaulters have less than 5 trades and also income is greater than 30.
- Defaulters with trades between 6 to 15 trades contribute and salary less than 20 contribute to 31% of defaulters

Way Forward

- *Model Building*
- *Model Evaluation*
- *Application Scorecard*

Model Building

- Predicting Defaulter and Non Defaulter is **Classification Problem** under **Supervised Learning** category as we know the dependent/target variable.
- Will build 2 different classification model, one with Demographic Data and another with Demographic & Credit Bureau data
- Major problem in Dataset is **Class Imbalance (94:6)** which means out of 100 random sample there are 94 Non Defaulters and 6 Defaulters, so will use **Stratified Sampling** technique to split data in train & test set.
- We are not going to scale data as we are building model on WOE transformed value dataset.
- For Feature selection we are using **WOE & Information Value**, based on IV metrics of Week, Medium & Strong predictor will choose feature variable from both dataset.

Modelling Technique:

- Will use below mentioned 3 classification algorithms to fit data and based on different evaluation criteria will choose final model.
- **Logistic Regression** is easy to interpret and it will act as baseline model for classification problems
- **Random Forest** is one of the ensemble technique which build strong predictor using multiple week predictor/trees
- **Support Vector Machine** provides different kernel method which transform data from non liner space to linear space and fit model on that transformed dataset to segregate different classes.

Model Evaluation

- For module evaluation will use Train-Validate-Test approach using **GridSerachCV** and **Stratified K-Fold** Cross Validator
- Tune different **Hyper-Parameter** for different algorithms and verify the score
- There are multiple metrics to **evaluate classification model**
- Validate **Discriminatory Power** of model using Sensitivity & Specificity metrics as well as AUC curve.
- **Sensitivity** is true positive rate which measures proportion of actual (Defaulters) that are correctly identified.
- Where as **Specificity** deals with negative rate (True negative rate)
- **Area Under Curve (AUC)** measure how well target variable distinguish between two groups, higher the AUC i.e. 1, better the model.
- Check **Calibration Accuracy**, discriminatory power validate classification ability, where as accuracy used to validate how different actual and predicated defaults are
- **Stability** is nothing but consistency of model on unseen data, to evaluate stability will check performance of model of train dataset vs test dataset.

Application Scorecard

- Score is the Calibrated log of odds value predicated by model.
- The score reflects the increase or decrease in odds, with High Score Values reflect a low probability of default.
- To generate score card will build logistic regression model using customer single view data (complete data set Demographic & Credit Bureau) as we want regression coefficient to scale or calibrate the scorecard.
- We calibrate predicated log of odds such that at score of 400 good to bad odds is 10 to 1 and increase score by 20 points corresponds to doubling of good to bad odds.

$$\text{Score} = \sum_{i=1}^{n} \left(-\left(woei * \beta_i + \frac{a}{n}\right) * factor + offset/n\right)$$

$$\text{Factor} = \text{points to double}/Ln(2)$$

$$\text{Offset} = \text{score } - (factor * \ln(odds))$$

Odds	Score
10	400
20	420
40	440

Thank You