CAPSTONE PROJECT – BFSI FINAL SUBMISSION

Pradnya Paithankar

Dr. Rajendra Warke

Vaibhav Palkar

BUSINESS UNDERSTANDING

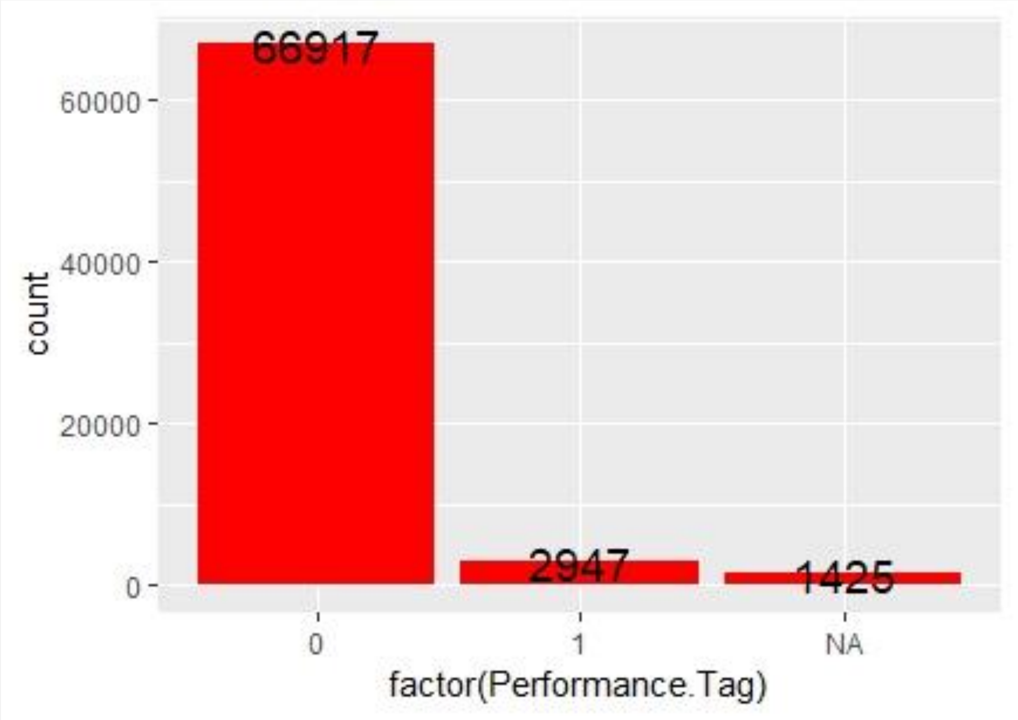
- CredX a leading credit card provider has experienced an increase in credit loss.
- The CEO believes that the best strategy to mitigate credit risk is to 'acquire the right customers'.
- Objectives : To help Credx to
    - Identify the right customers during acquisition using predictive models
    - Using past data of the bank's applicants, determine the factors affecting credit risk
    - Create strategies to mitigate the acquisition risk
    - Assess the financial benefit of prediction analysis project

DATA UNDERSTANDING

- Demographic Data – credit card application data, provided by customer
  - Age, Gender, Marital Status, No of dependents, Income, Education, Profession, Type of residence, No of months in current residence, No of months in current company
- Credit Bureau data
  - # 90 DPD or worse in last 6 months, # 60 DPD or worse in last 6 months, #30 DPD or worse in last 6 months, # 90 DPD or worse in last 12 months, # 60 DPD or worse in last 12 months, #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), Presence of open home loan, Outstanding Balance, Total No of Trades
- Common fields
  - Application ID, Performance Tag (If customer is 90 days past due or worse in the past 12-months (i.e. defaulted) after getting a credit card)
- Total number of input records after merging both the data sets - 71289

DATA UNDERSTANDING

- Dependent variable : Performance tag
  - 0 : Not defaulted / Good customers
  - 1 : Defaulted / Bad customers
  - **Blank** : Customers whose applications were rejected
    - Removed as these applications were rejected
    - Model would be run on these records to validate the rejection
- Ratio of defaulted customers
  - (# records with performance tag = 1) / (# records with performance tag = 0 and 1)
  - Default rate of input data - 4.22%



DATA UNDERSTANDING

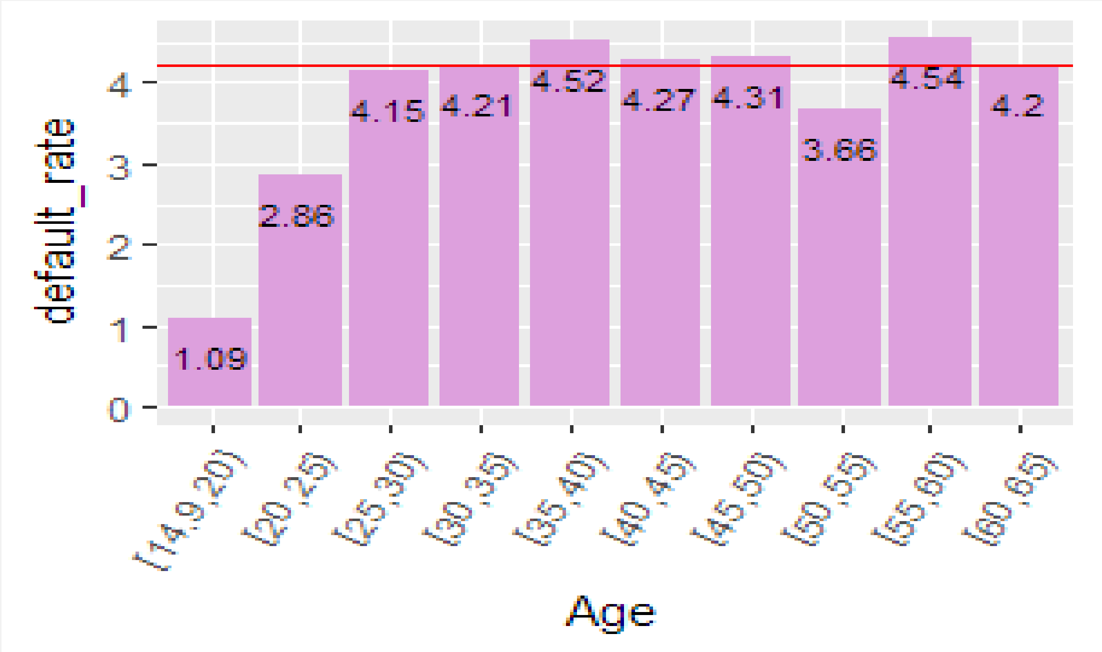
- 272 applicants do not have any record of loans with credit bureau
  - First timer loan applicants
- 751 applicants (excluding 272 above) do not have credit card utilization record
  - First timer in credit card application
- ** Both these missing records to be imputed with woe values during model building

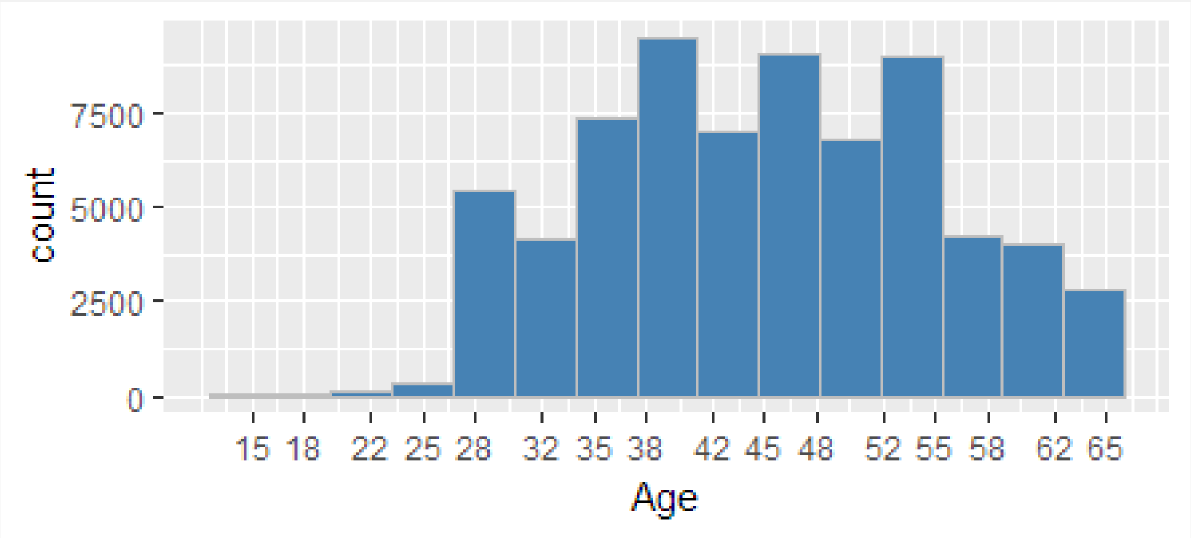
BASIC DATA CLEANING

Parameter	Data issue	Action taken
Application ID	3 duplicate records	Deleted as Insignificant count
Age	20 records have age as 0 and -3	Replaced with modal category
Gender	1 blank record	Deleted as just 1 record
Education	118 blank records	Replaced with modal category
Marital status	6 blank records	Replaced with modal category
Type of residence	8 blank records	Replaced with modal category
# dependents	3 blank records	Deleted as Insignificant count
Profession	13 blank records	Replaced with modal category
Income	81 records having income as -0.5	Replaced with mean income
number of months in current company	10 Outlier records at exact 100 percentile	Capped with 99^{th} percentile value

69858 Records after basic data cleaning

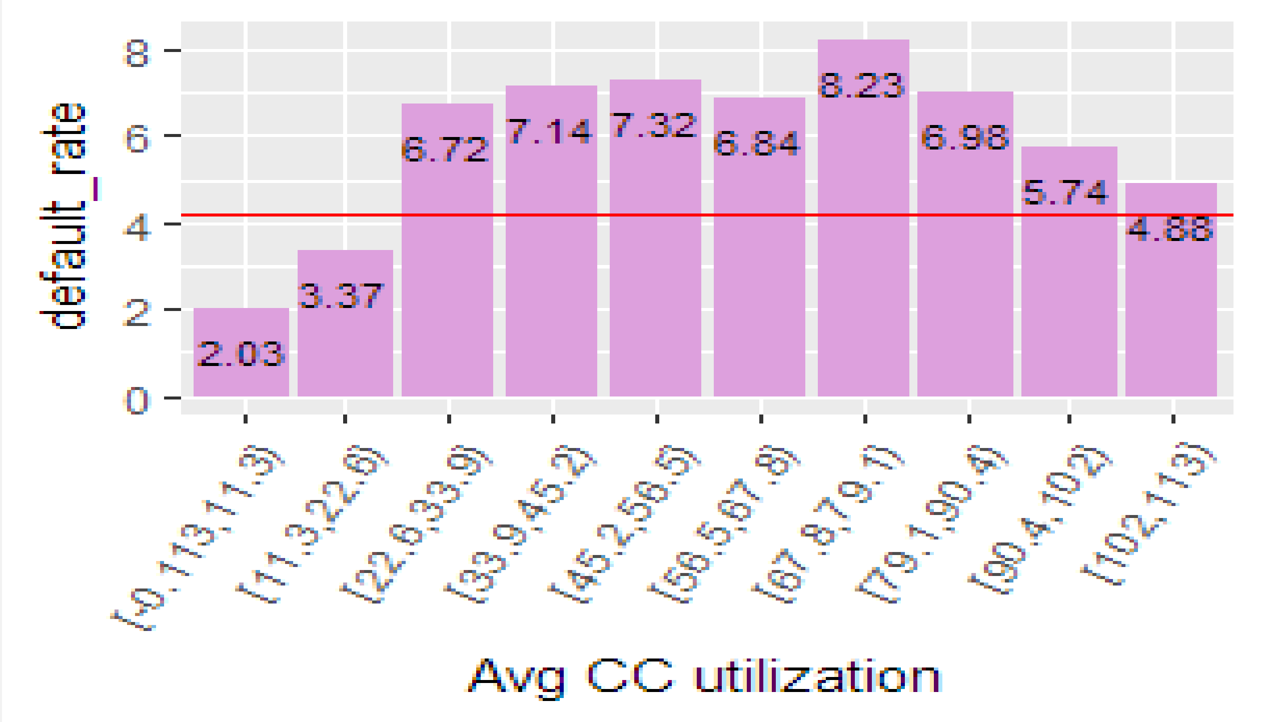
UNIVARIATE ANALYSIS

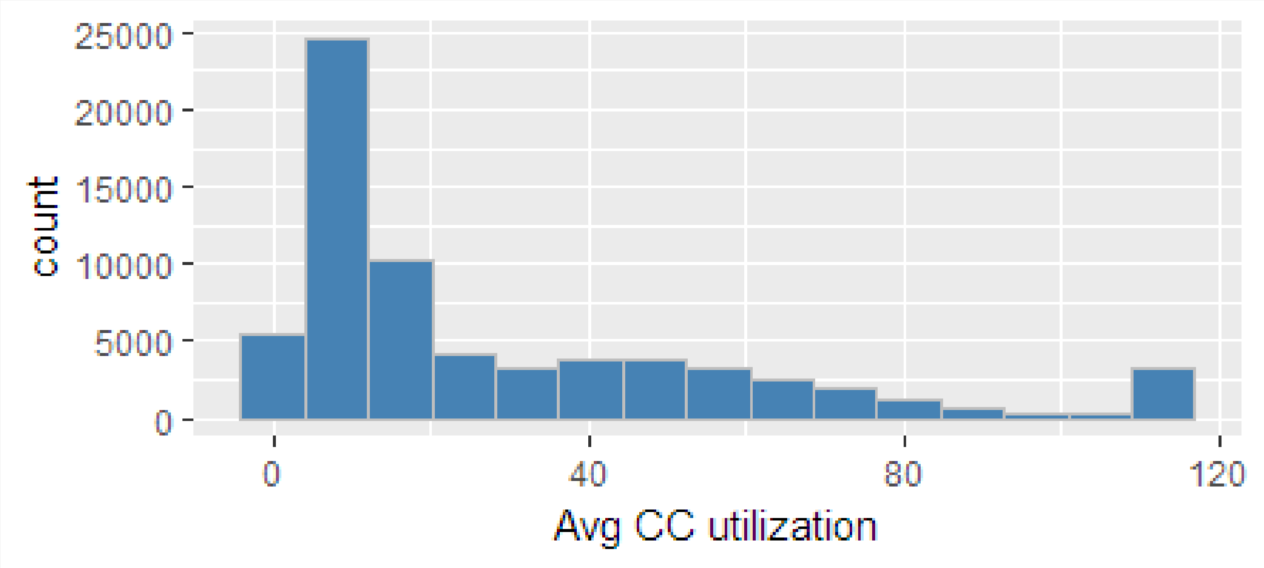




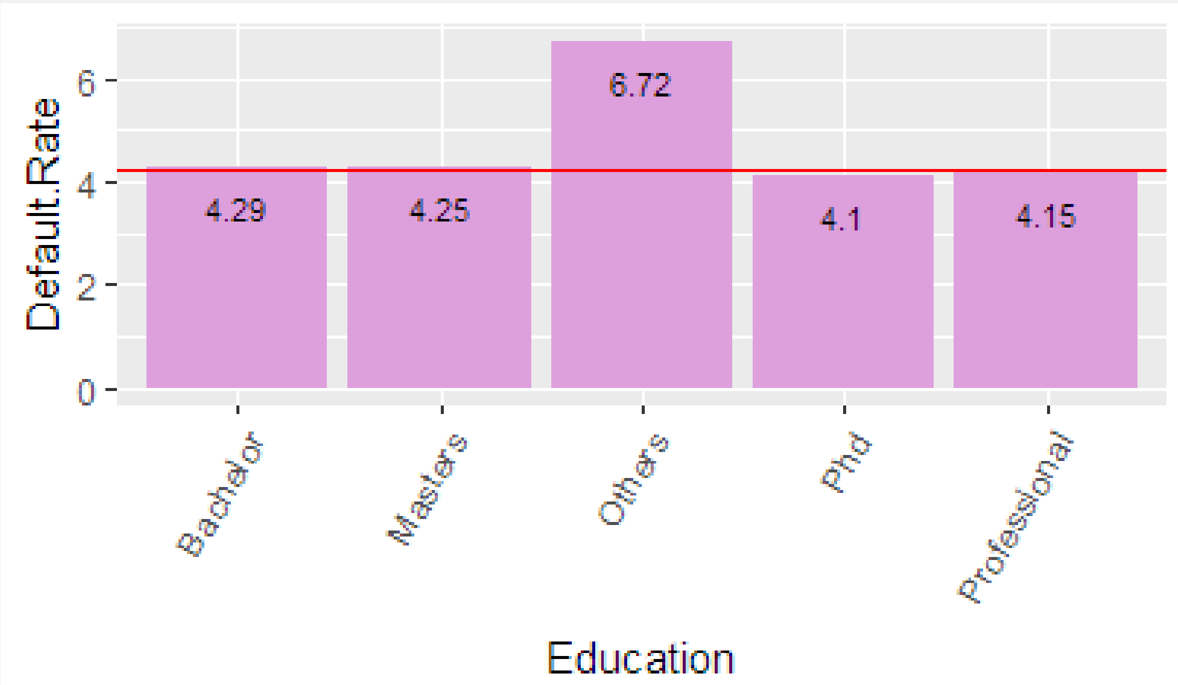
People with age 35-40, 55-60 have default rate slightly higher than total default rate

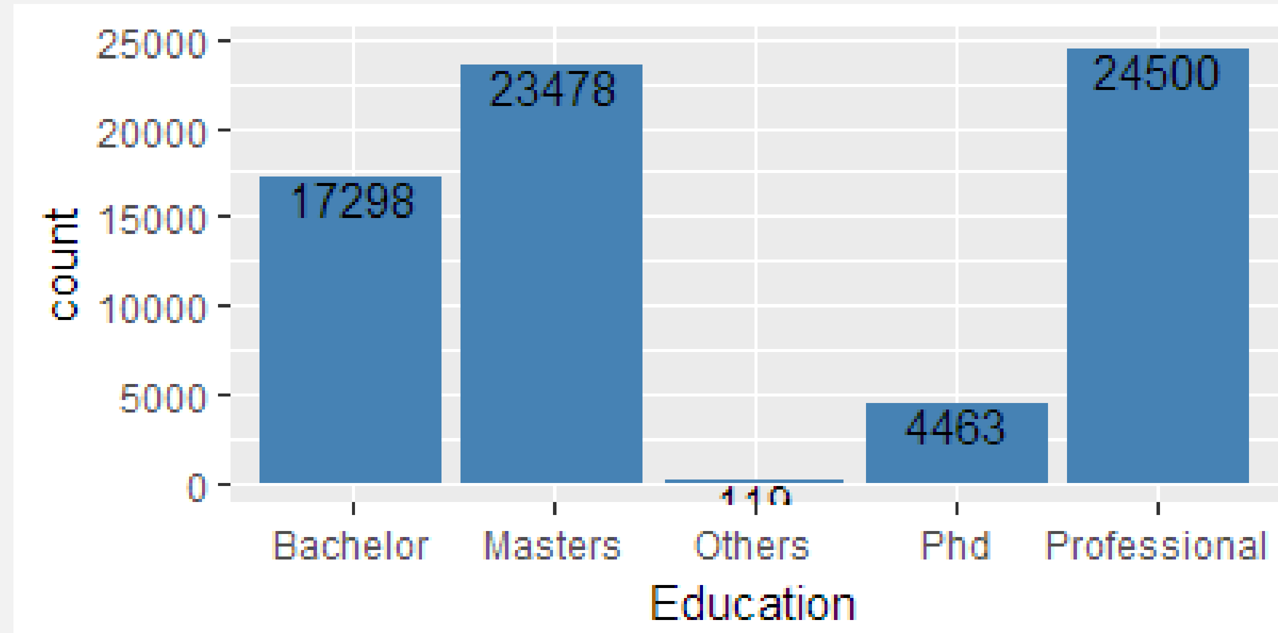
People with avg CC utilization have default rate higher than total default rate





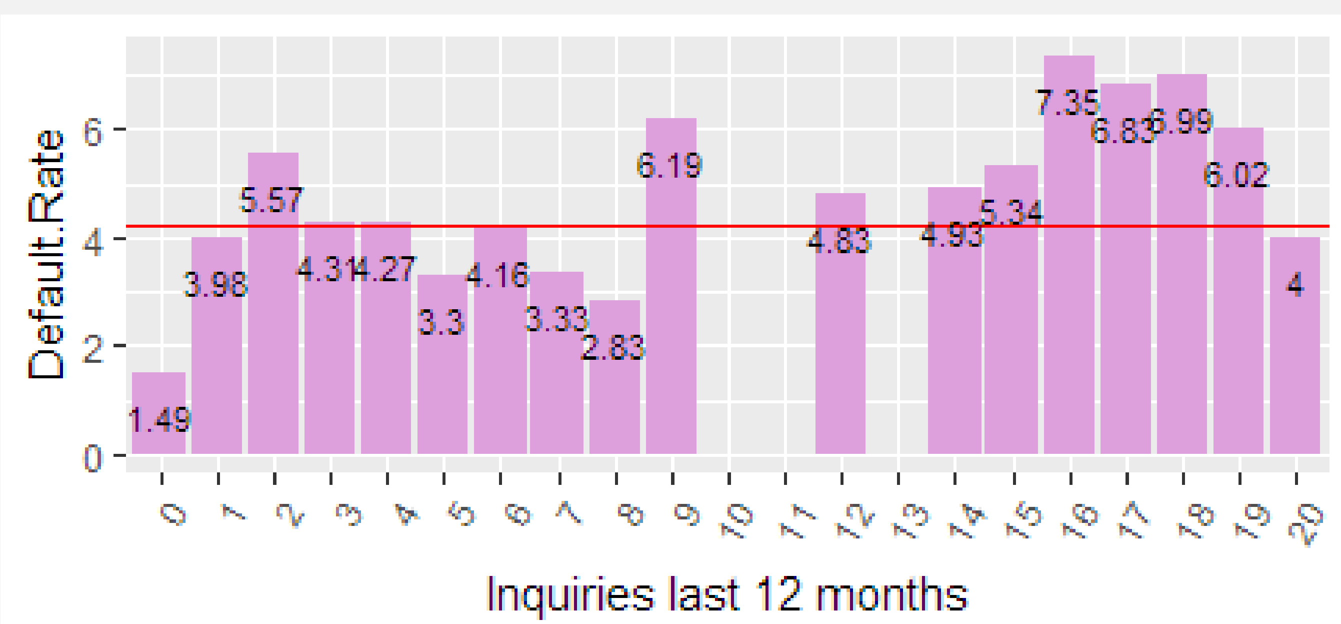
UNIVARIATE ANALYSIS

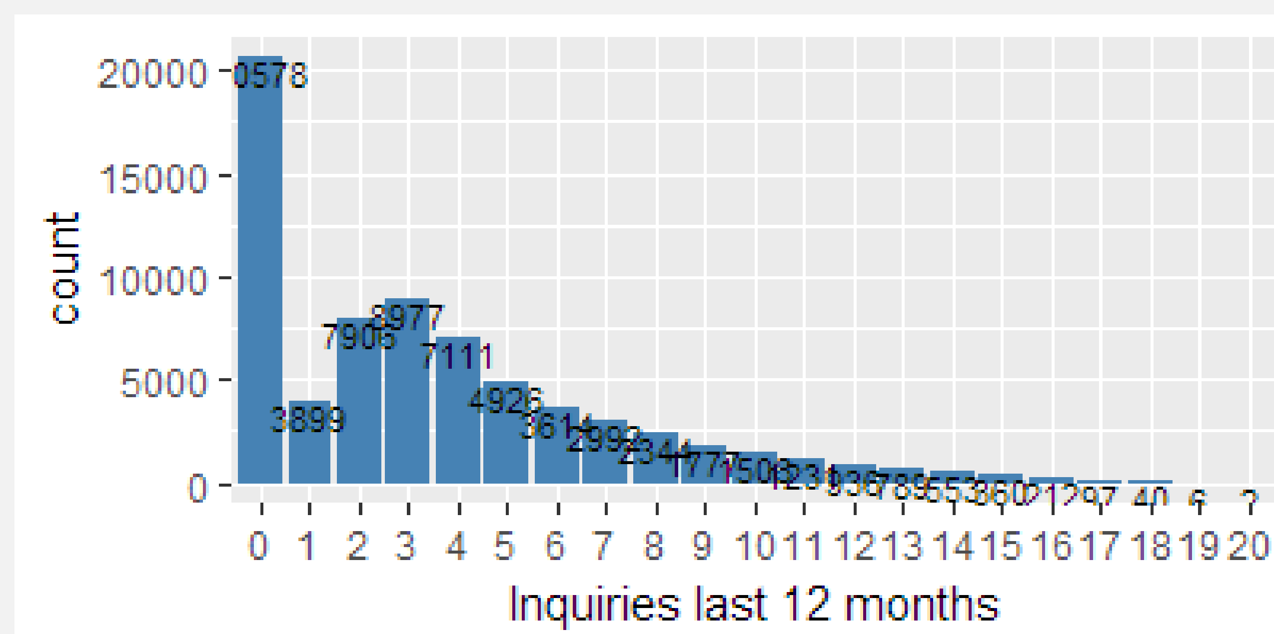




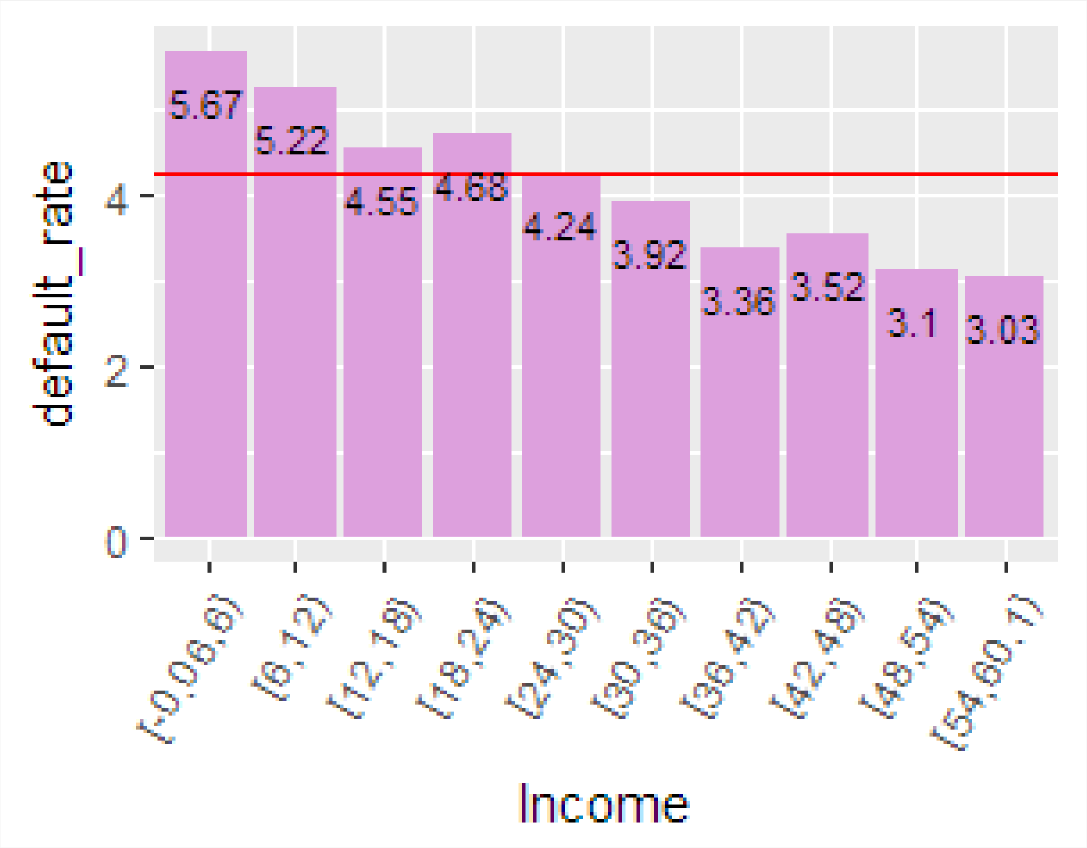
People with EDU as other have default rate higher than total default rate

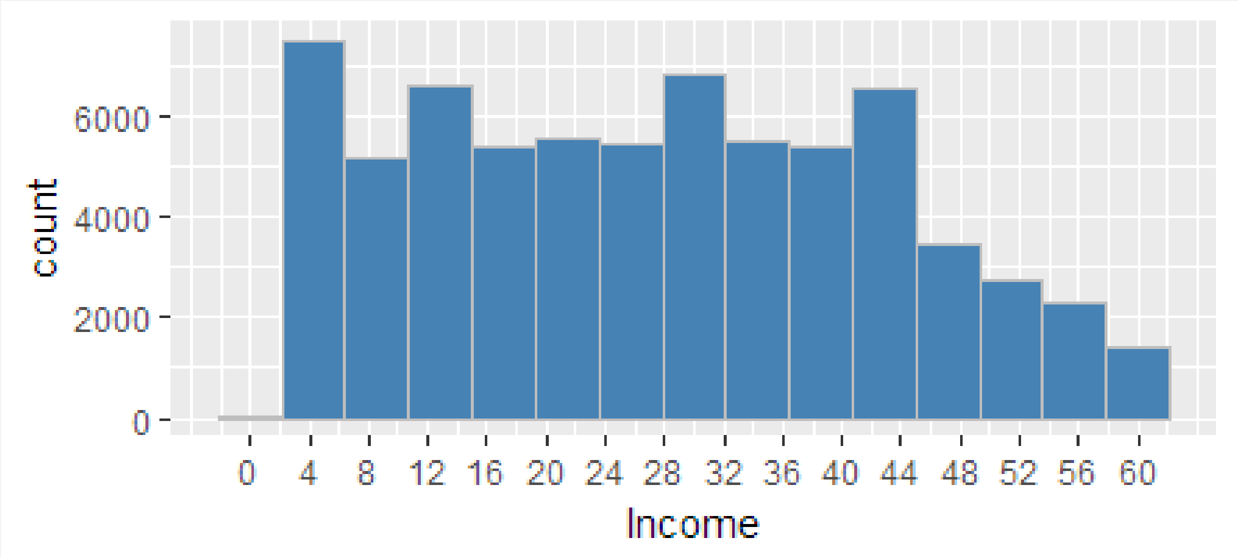
People with Inquiries 2 and 9 to 19 have default rate higher than total default rate



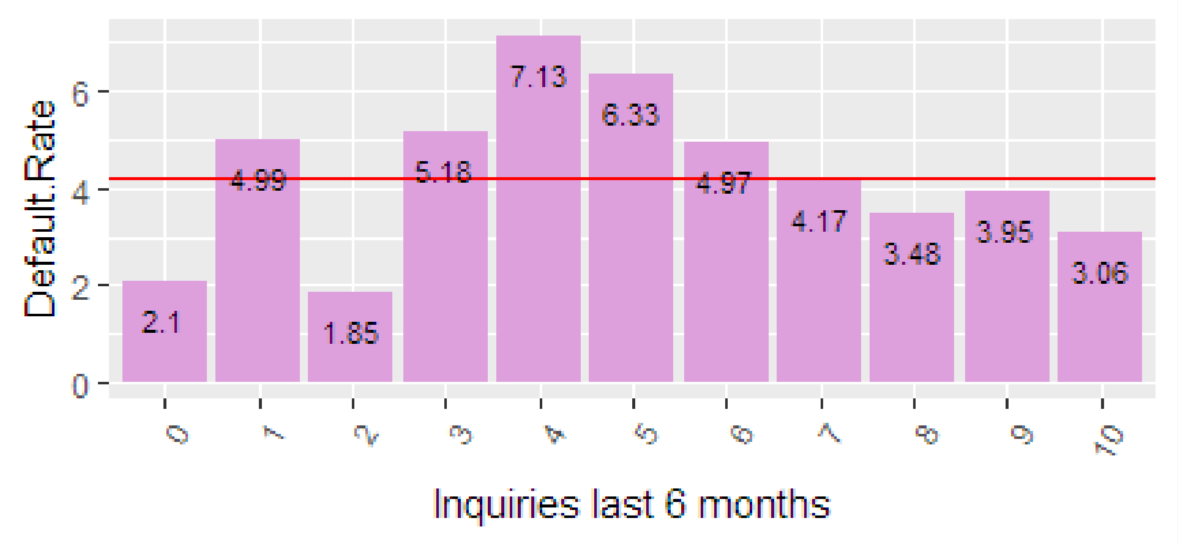


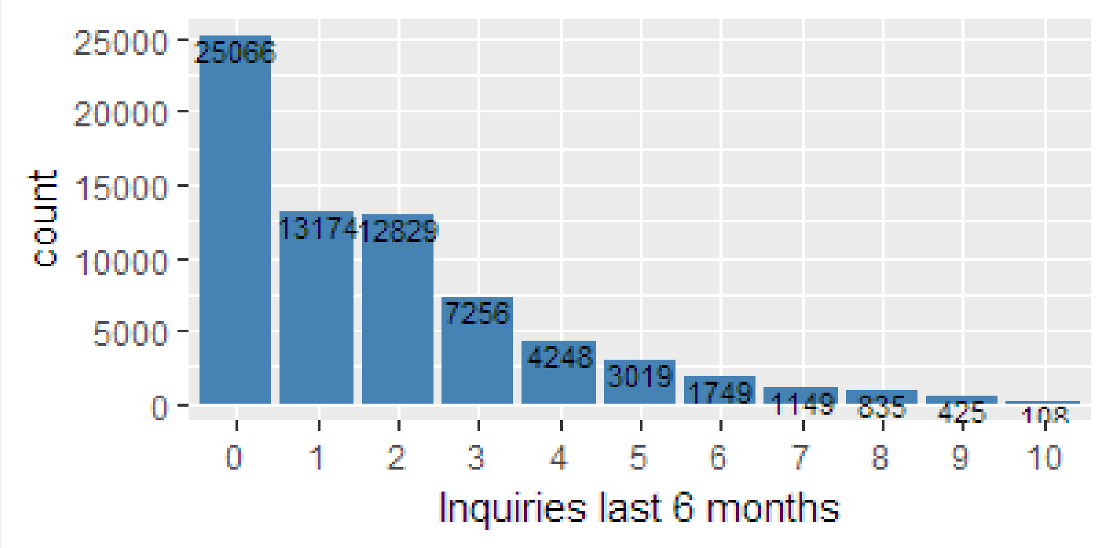
UNIVARIATE ANALYSIS





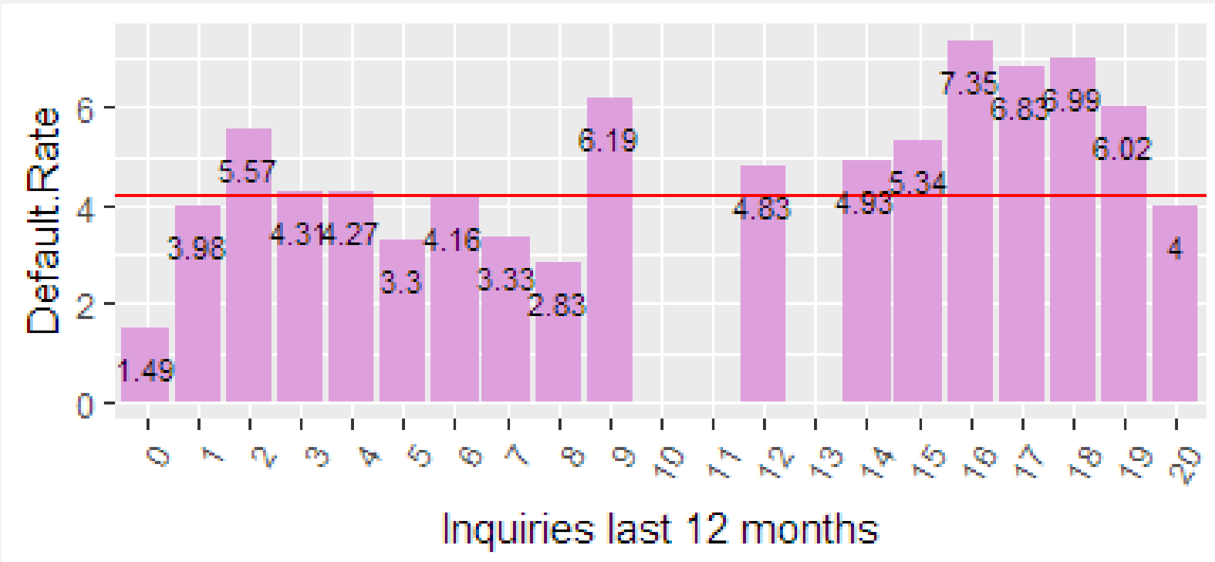
People with less income upto 24 have default rate higher than total default rate

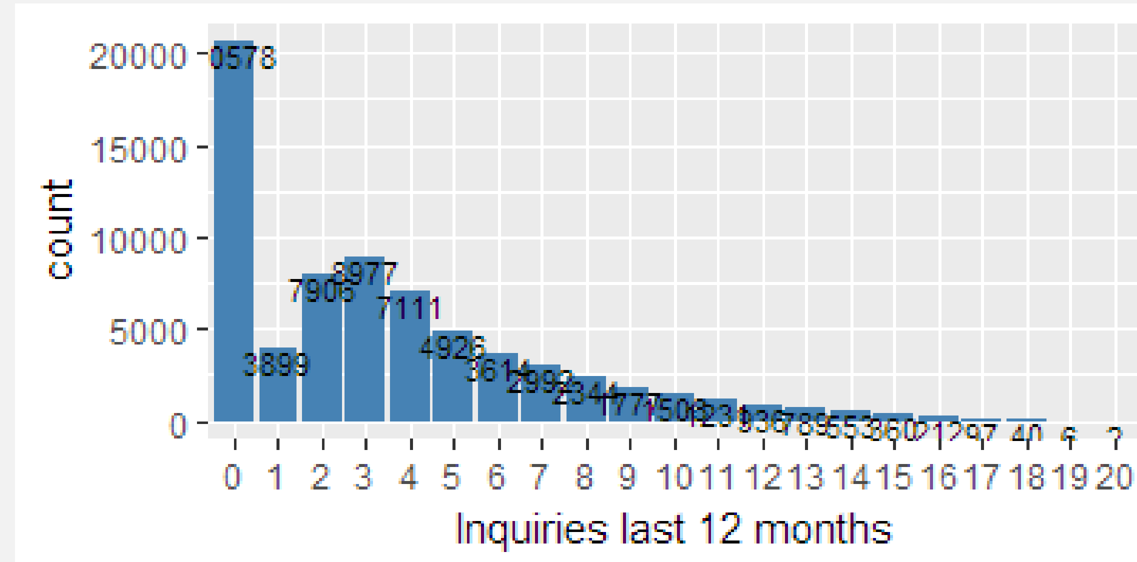




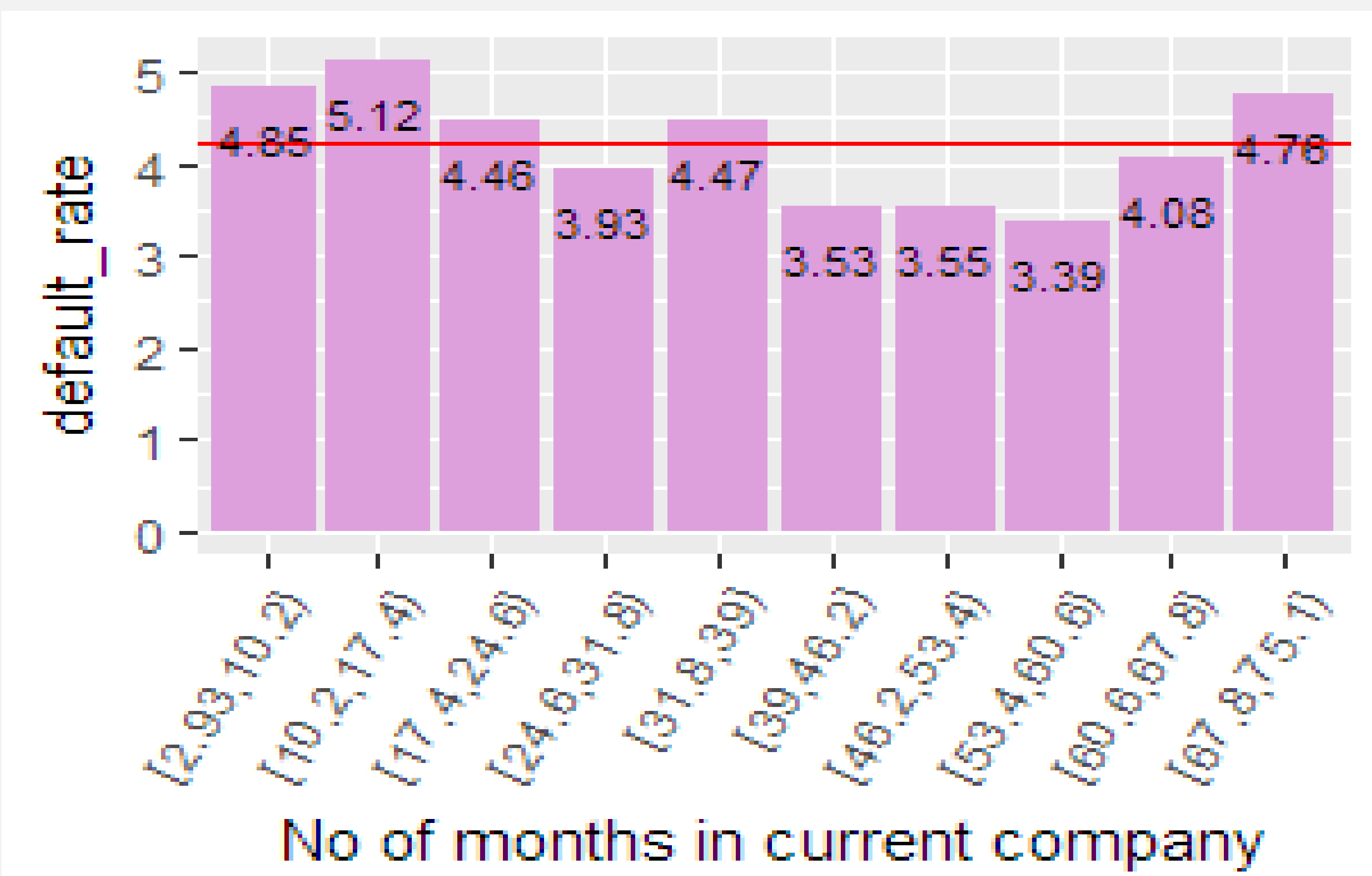
People with Inquiries Inquiries 1 and 3 to 6 have default rate higher than total default rate

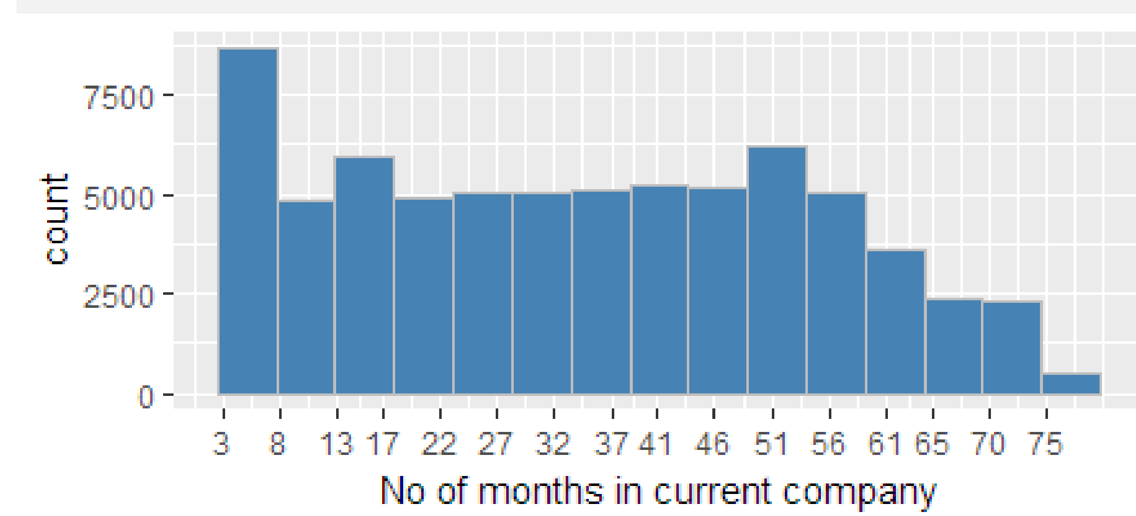
UNIVARIATE ANALYSIS





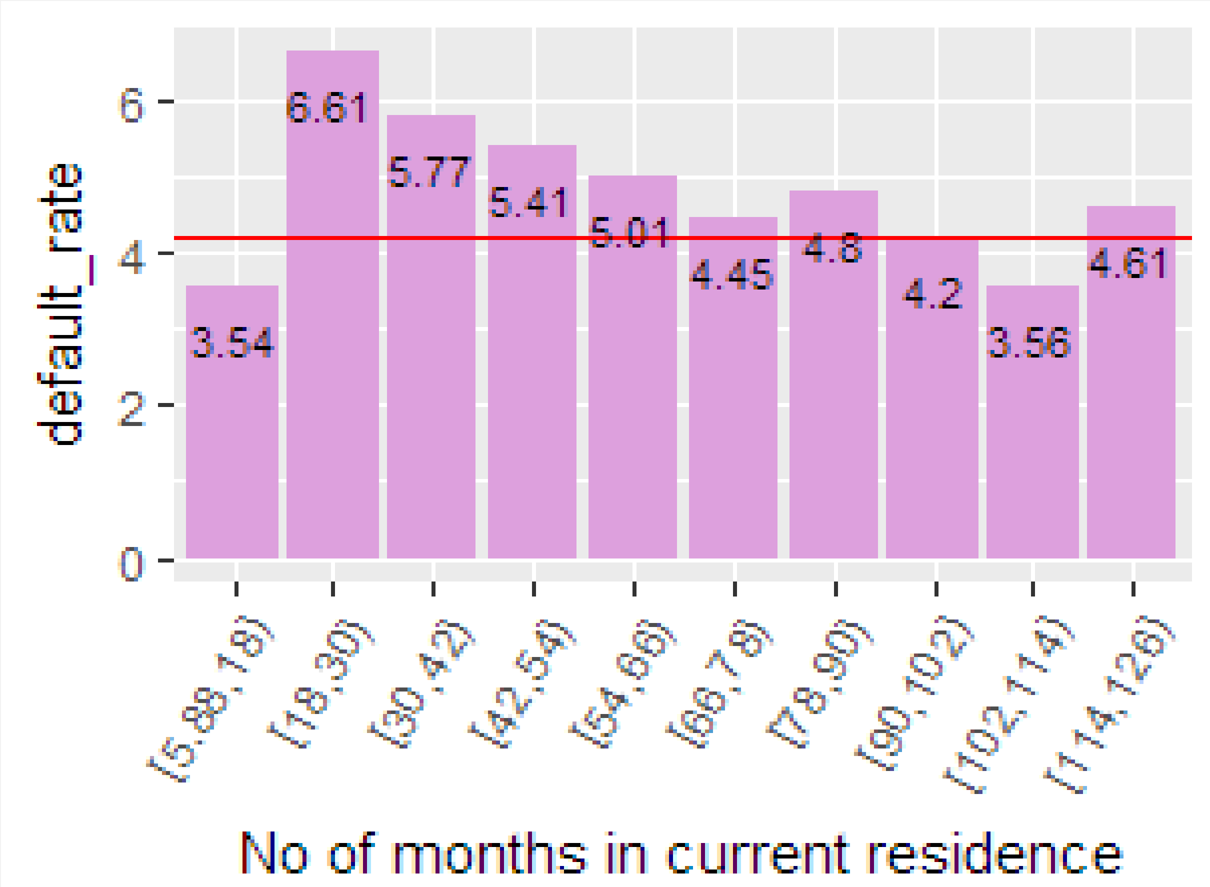
People with 2,9,12 and 14 to 19 Inquiries have default rate higher than total default rate

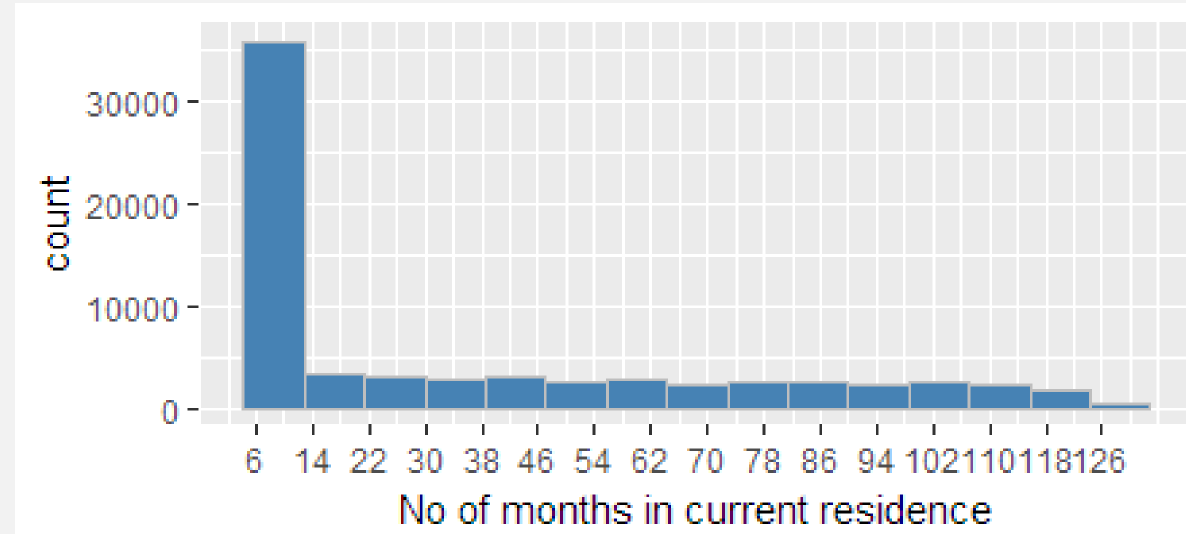




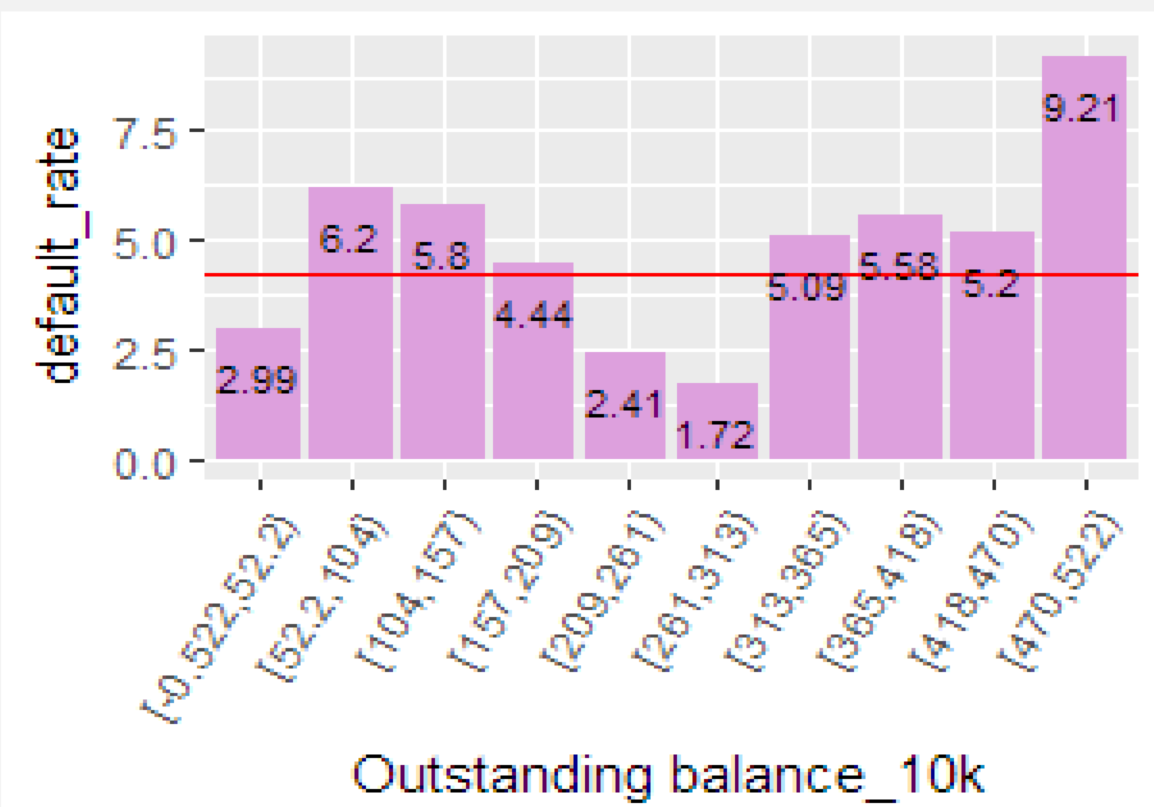
People with current company from 3-24 and 67- 75 have default rate higher than total default rate

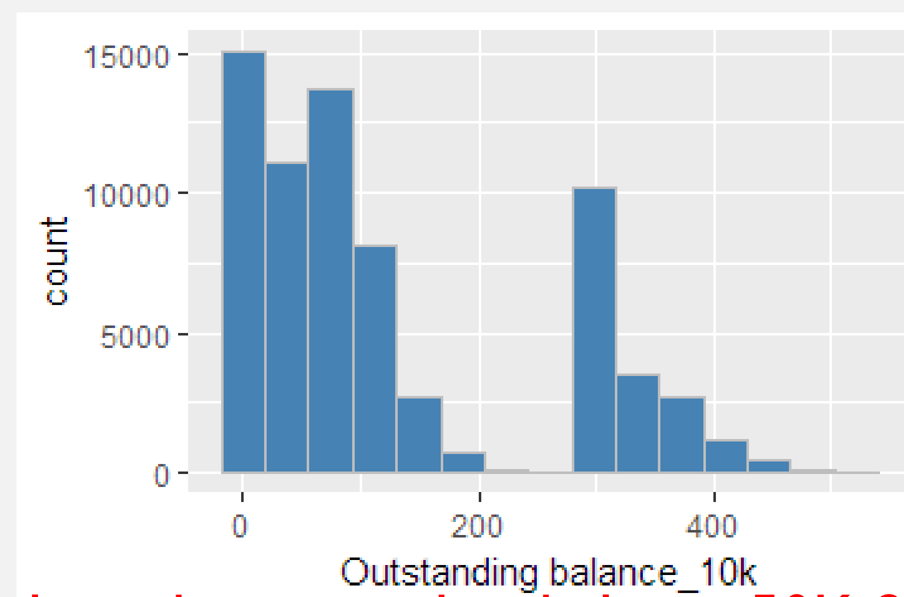
UNIVARIATE ANALYSIS





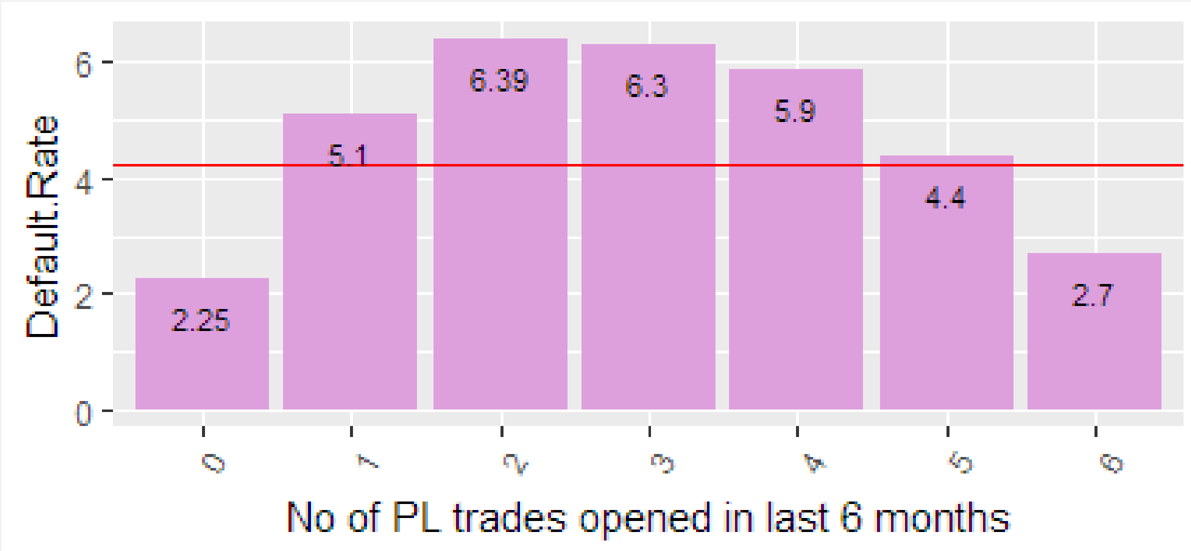
People with current residence 18-90 and 114-126 have default rate higher than total default rate

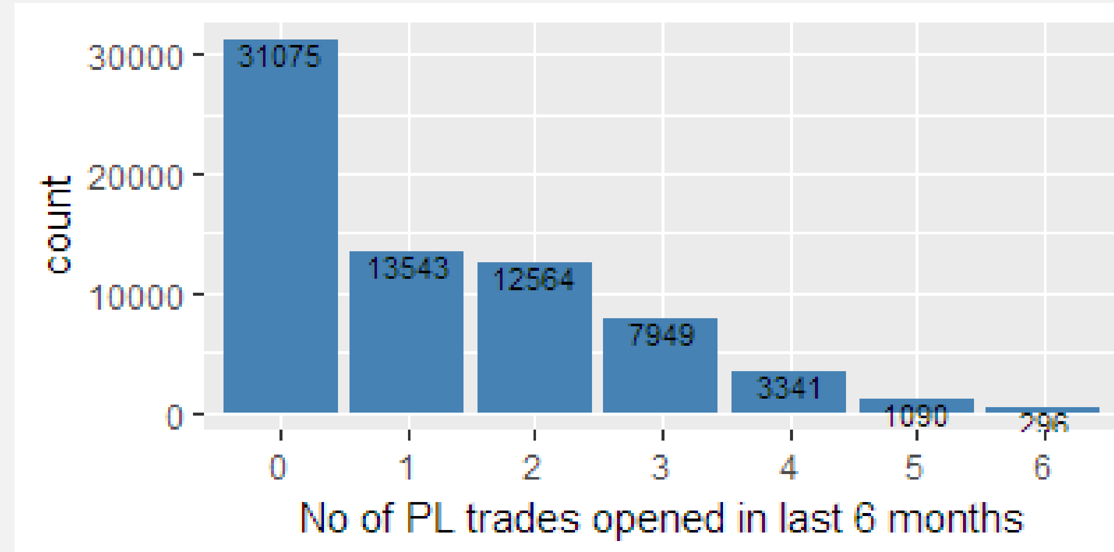




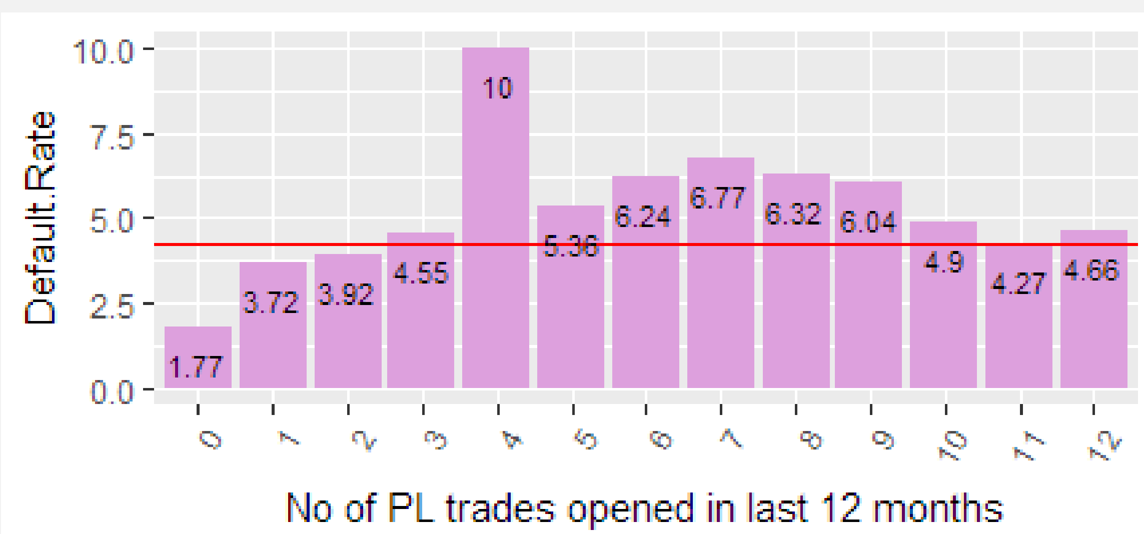
People with outstanding balance 50K-209 and 313k-522 have default rate higher than total default rate

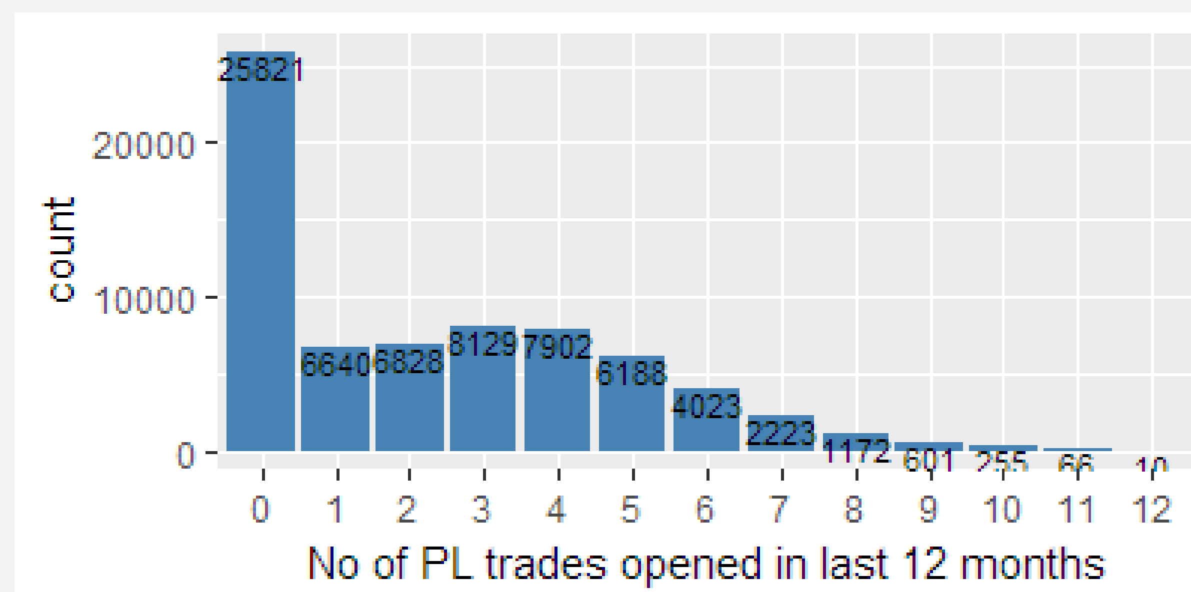
UNIVARIATE ANALYSIS





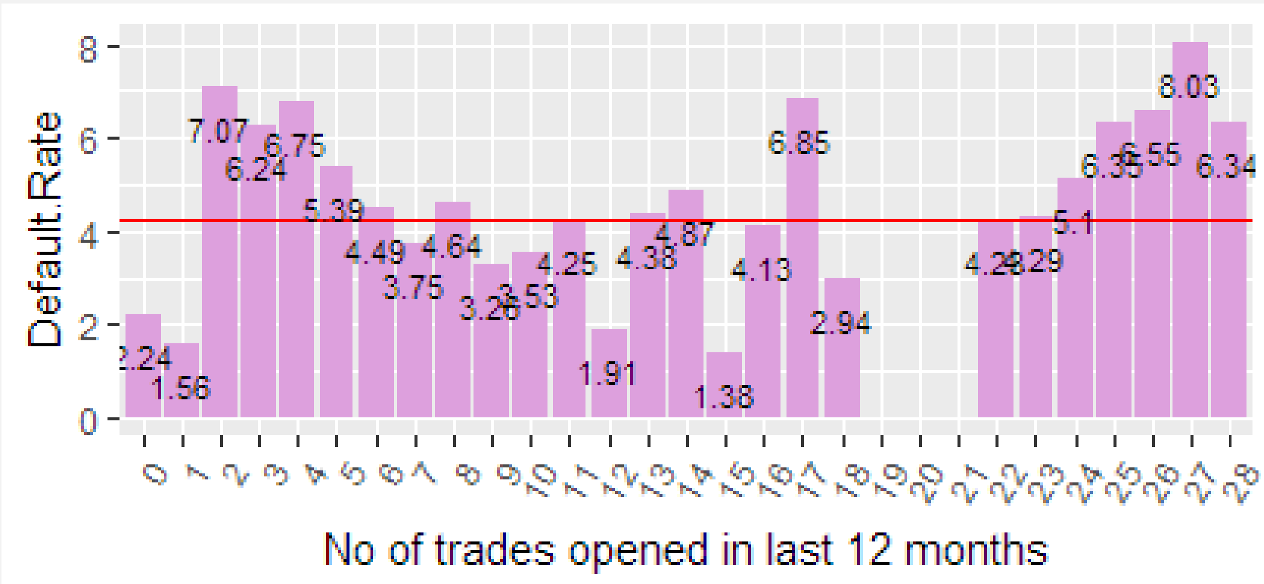
People with no of PL trades open 6m between 1-4 have default rate higher than total default rate

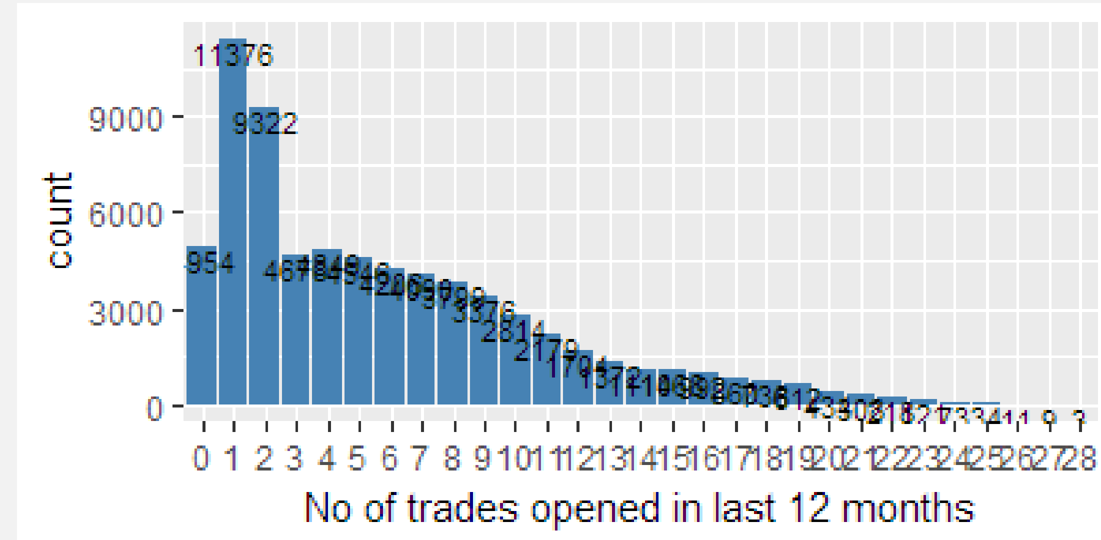




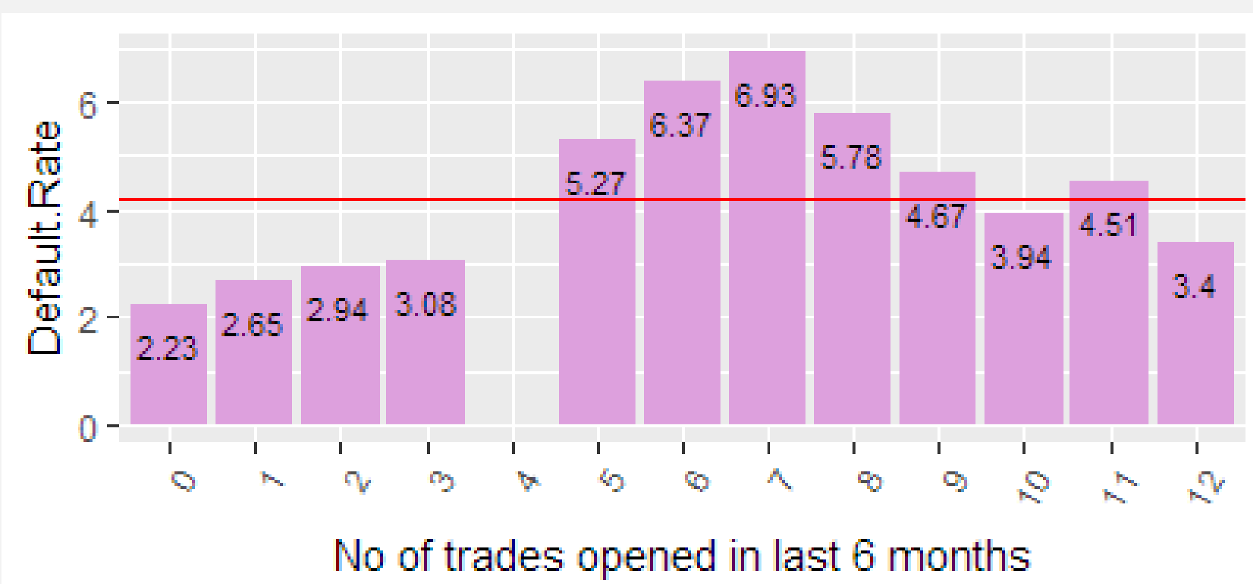
People with no of PL trades open 12m between 4-10 have default rate higher than total default rate

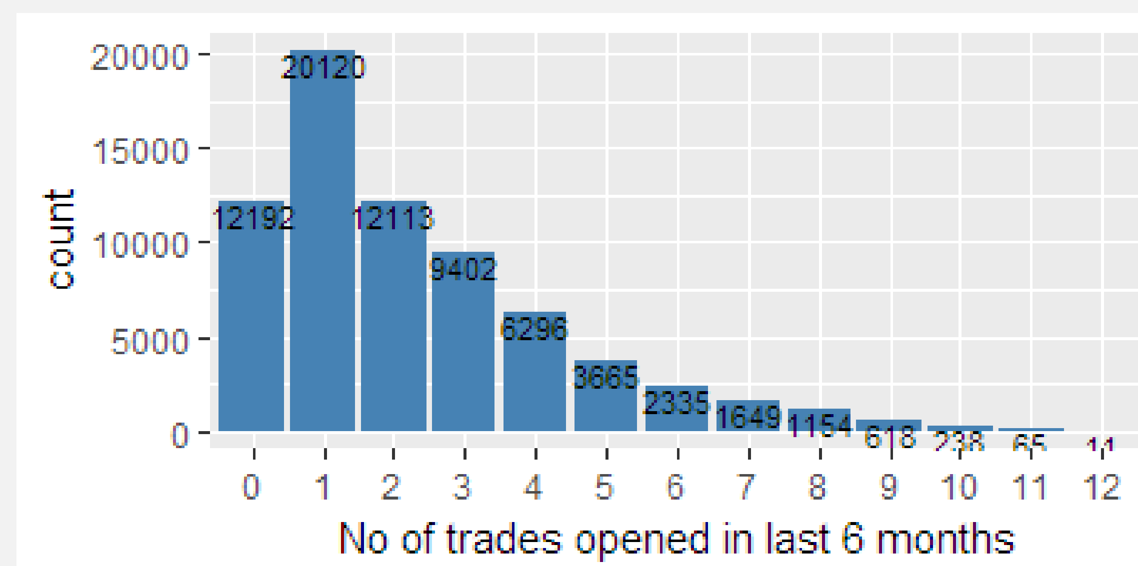
UNIVARIATE ANALYSIS





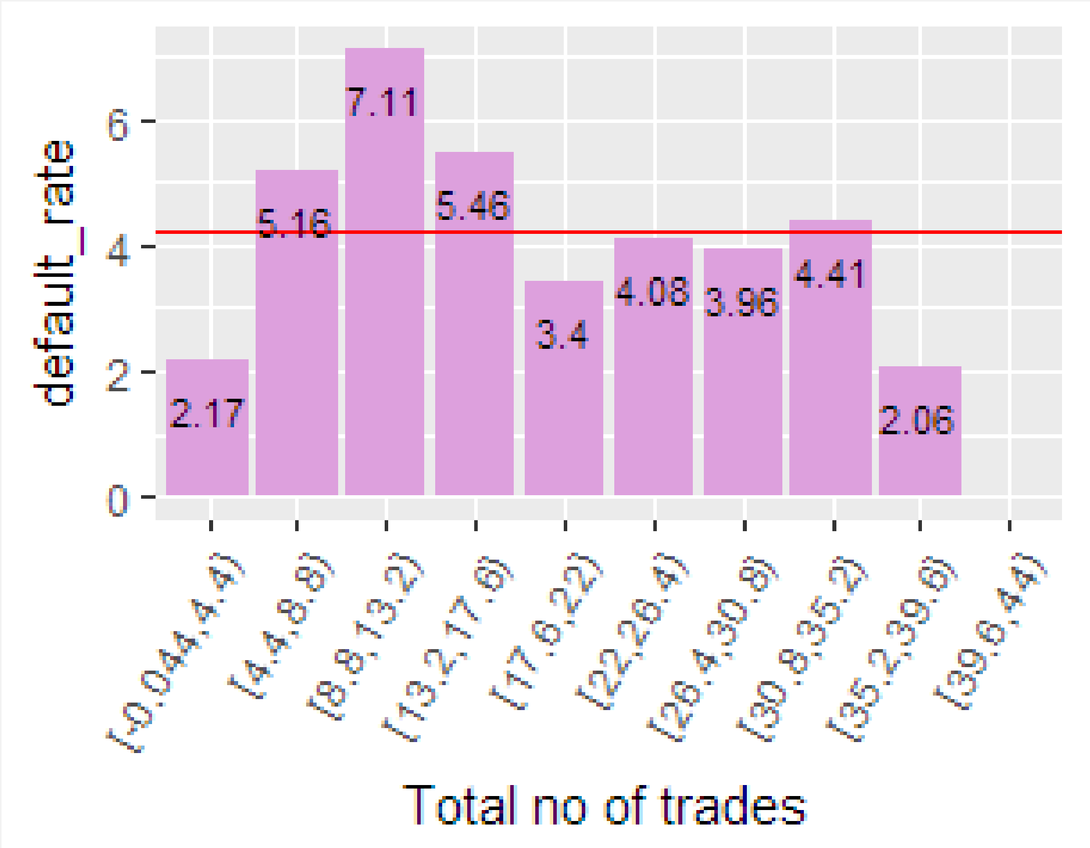
People with no of trades open 12m 2-6 , 8, 11, 14, 17, and 24-28 have default rate higher than total default rate

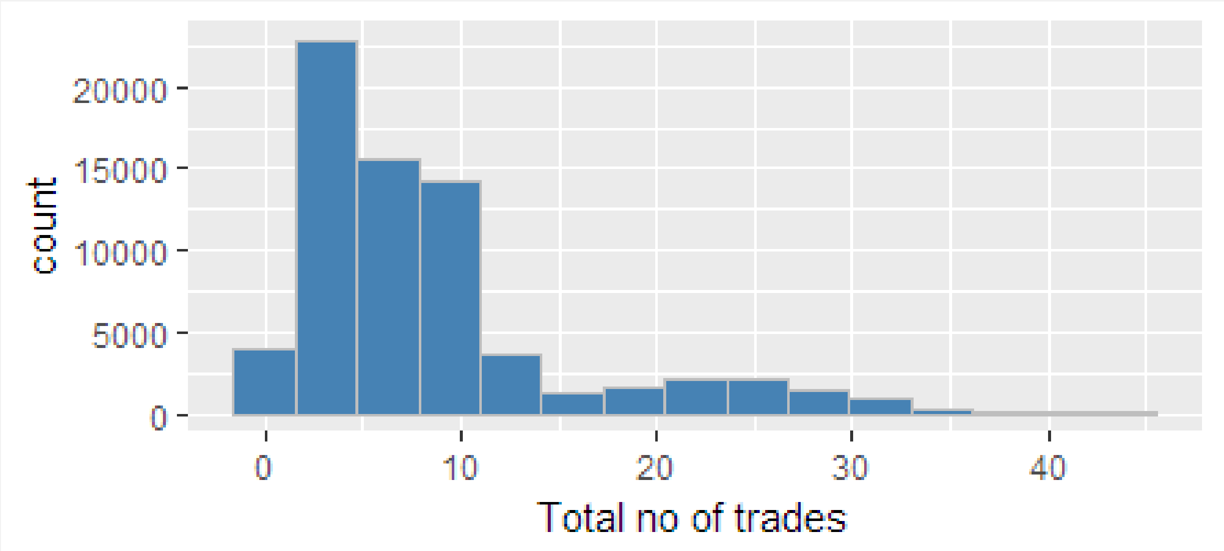




People with no of trades open 6m 5-9 & 11 have default rate higher than total default rate

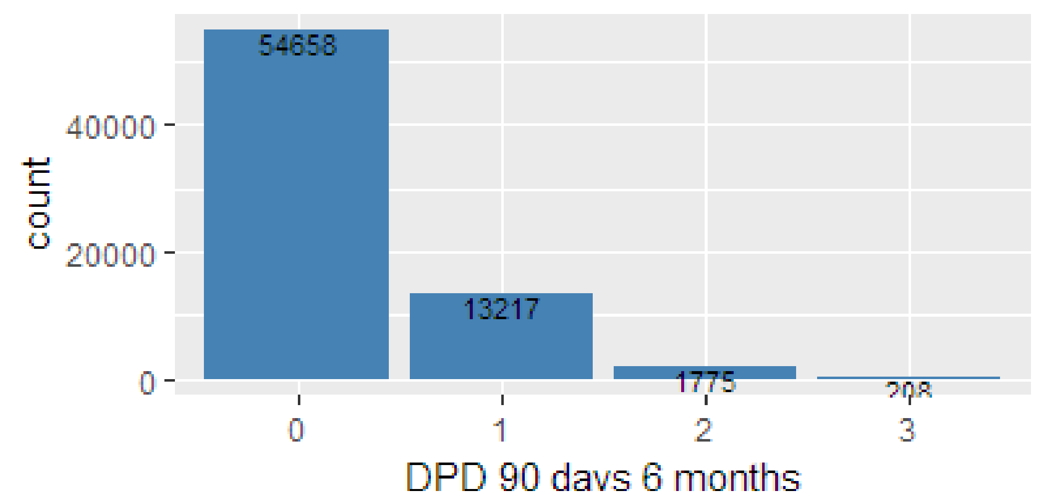
UNIVARIATE ANALYSIS

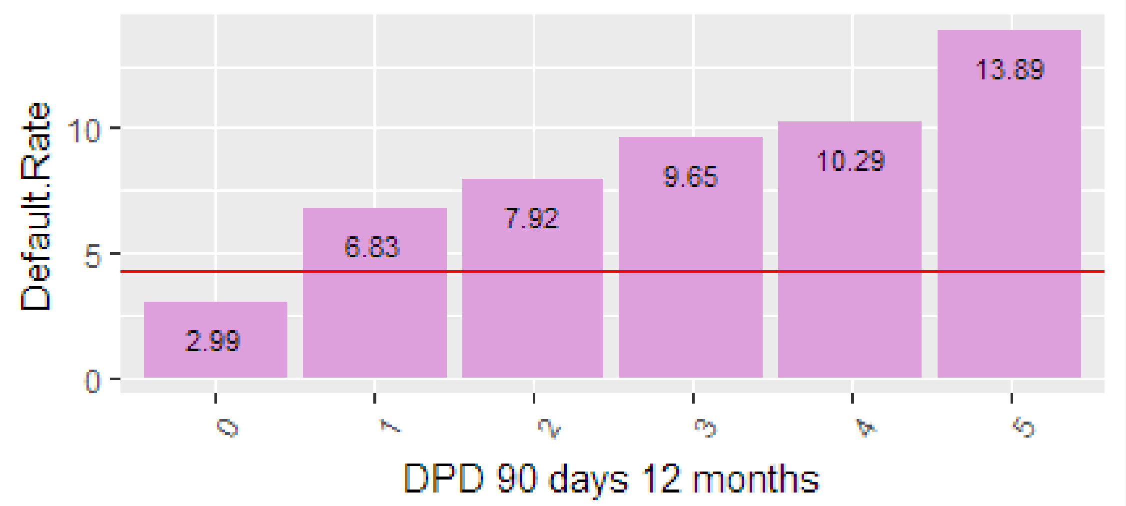


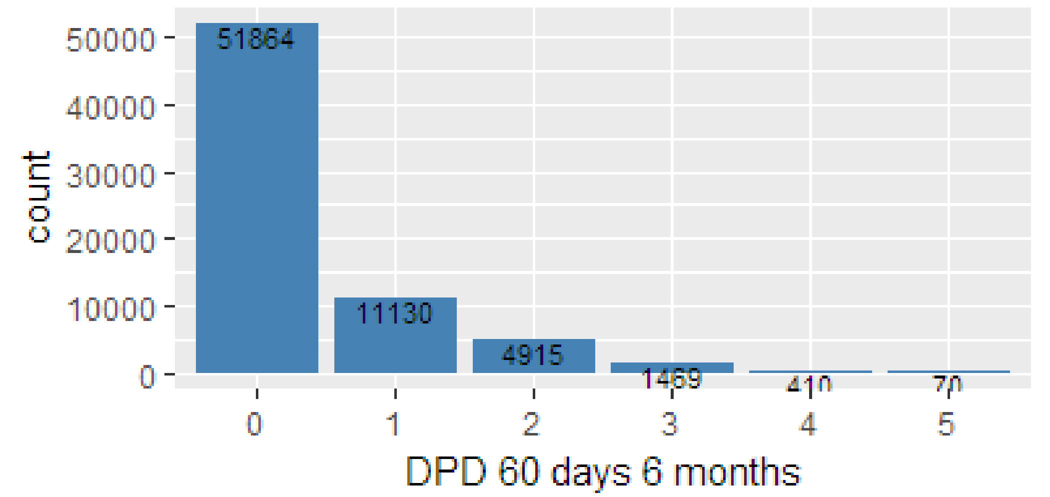


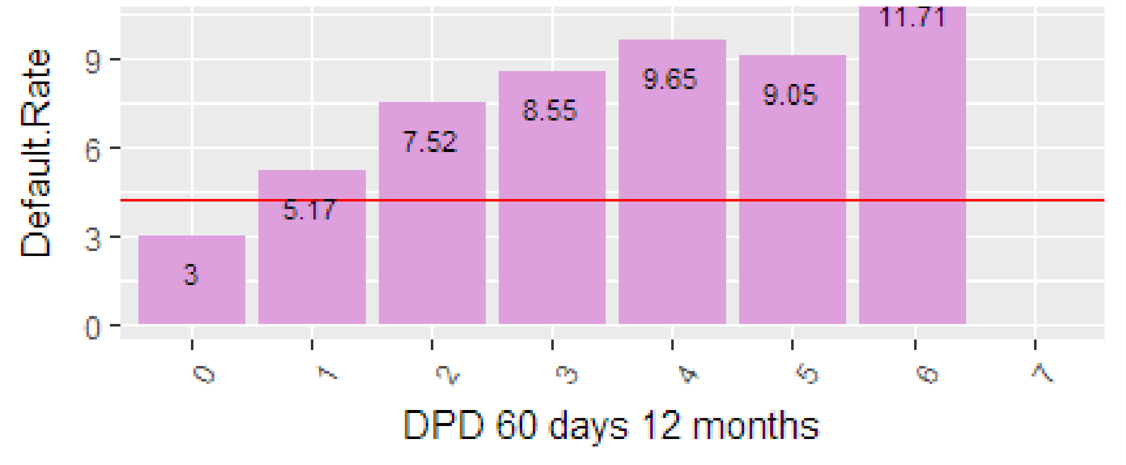
People with total trades 4-18 & 22-26 & 30-35 have default rate higher than total default rate

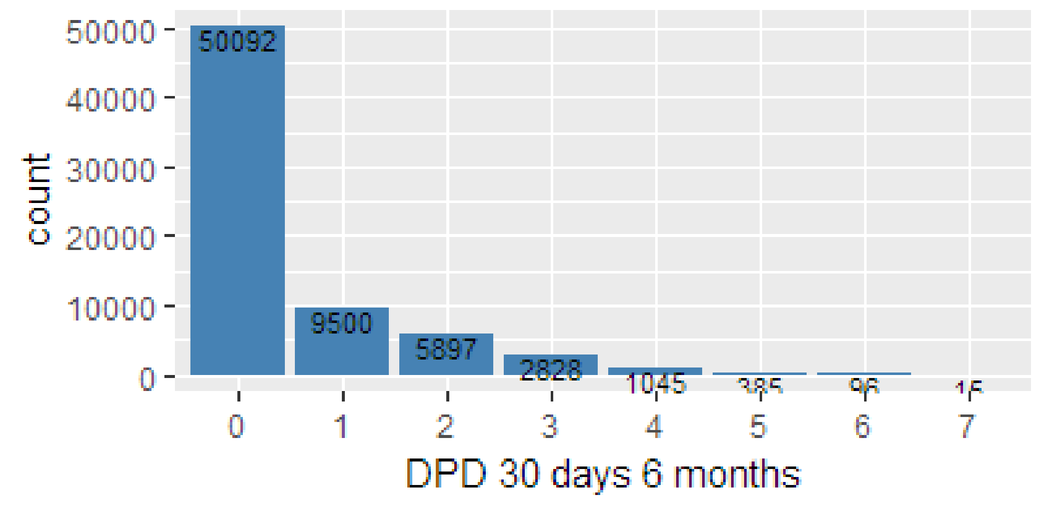
UNIVARIATE ANALYSIS

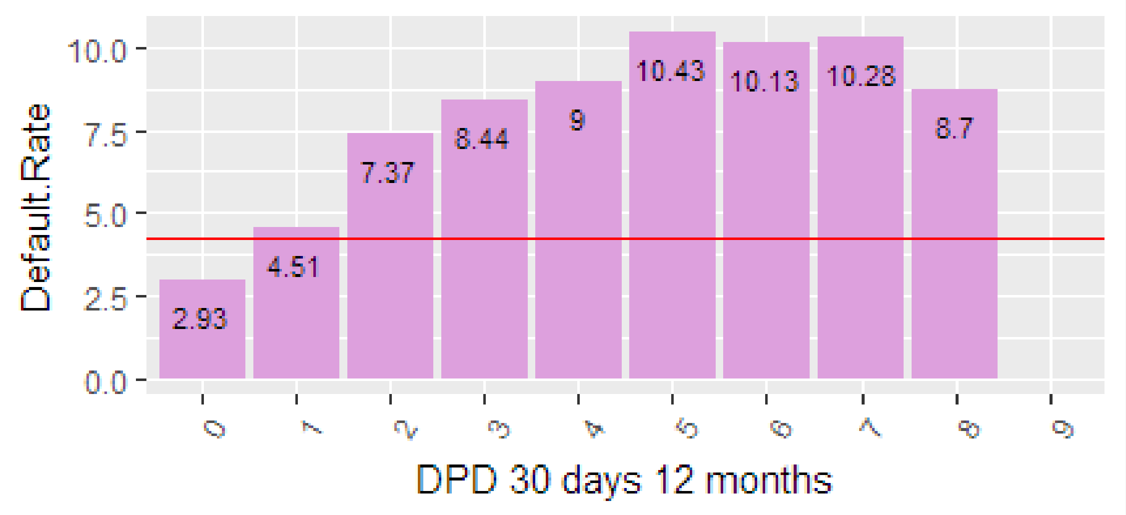






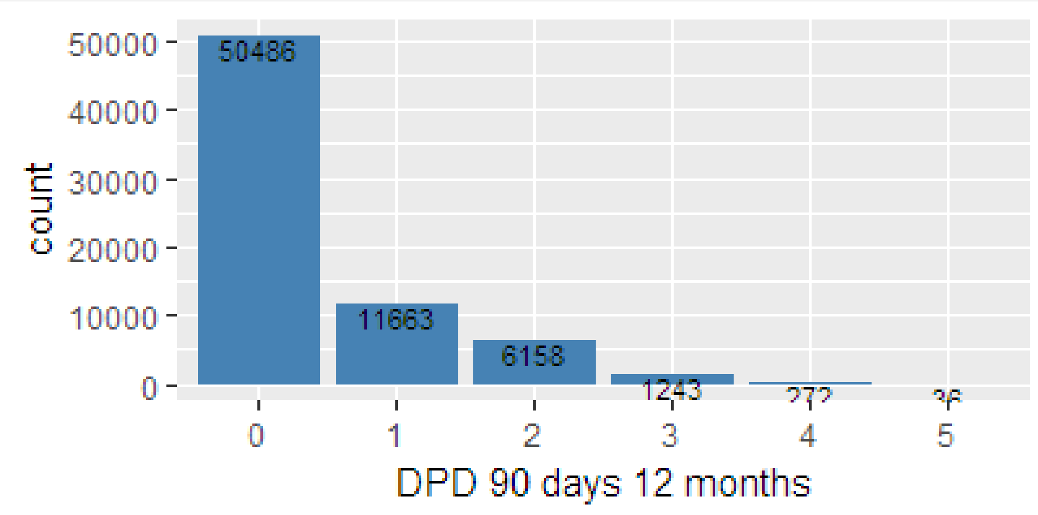


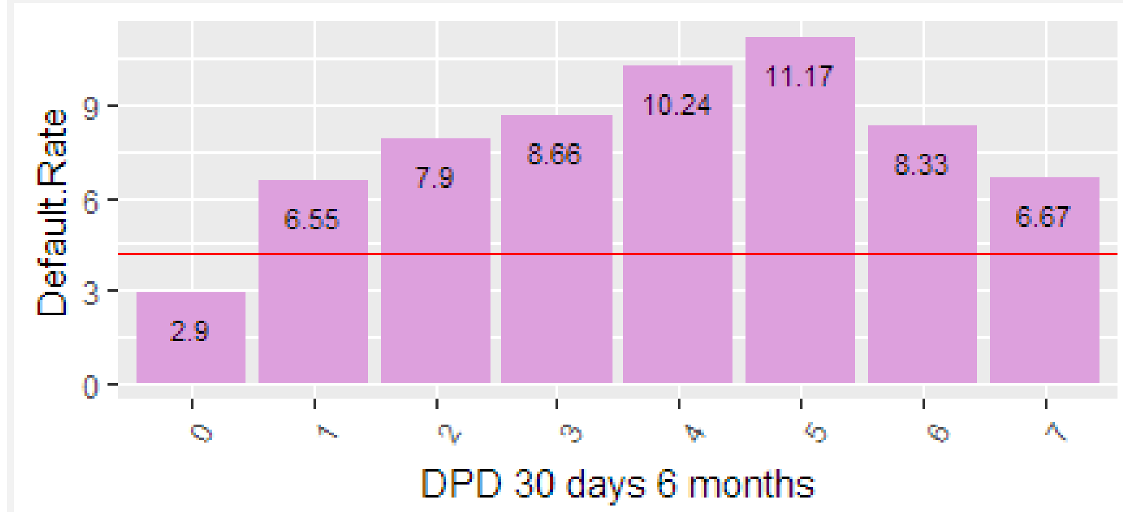


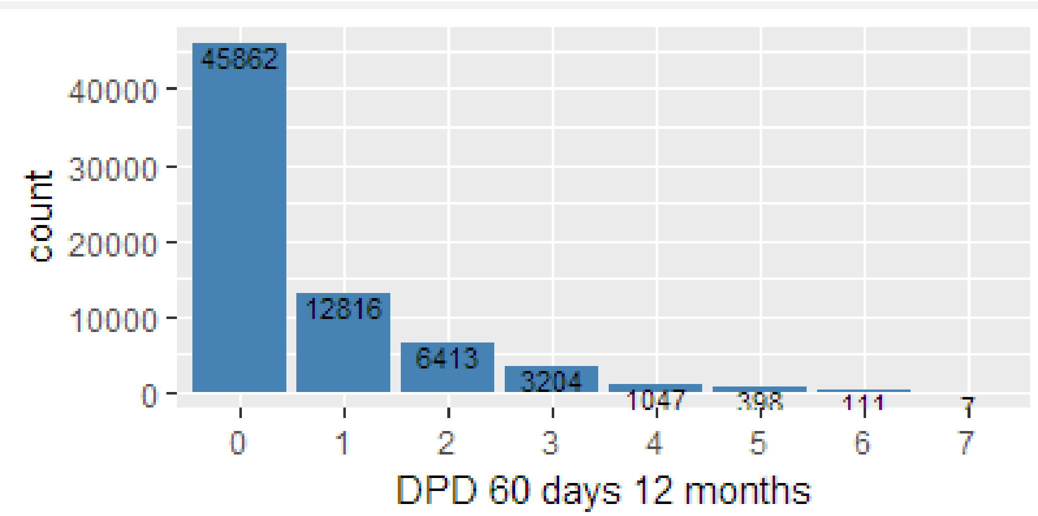


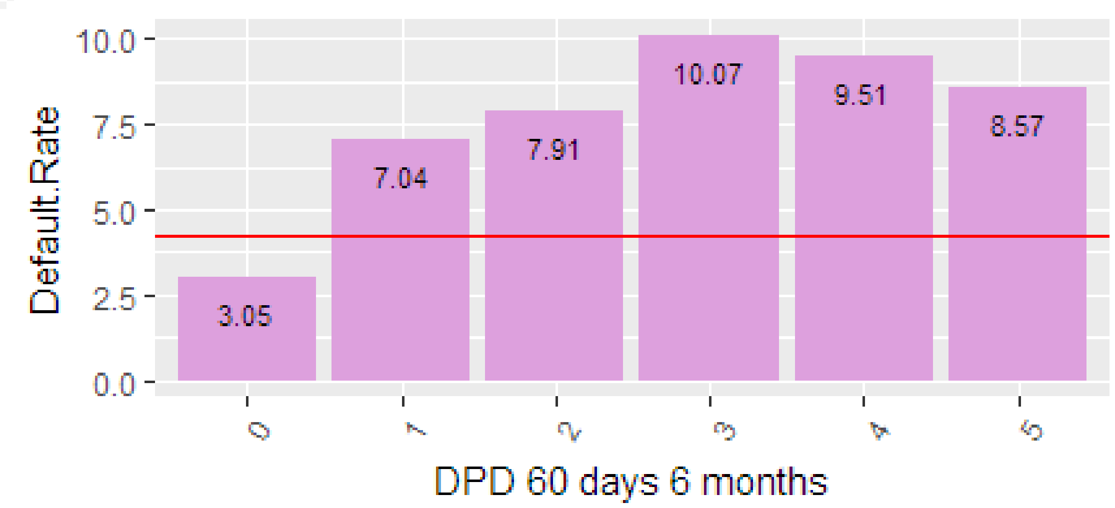
People 30/60/90 dpd in 6 months for more than once have default rate higher than total default rate

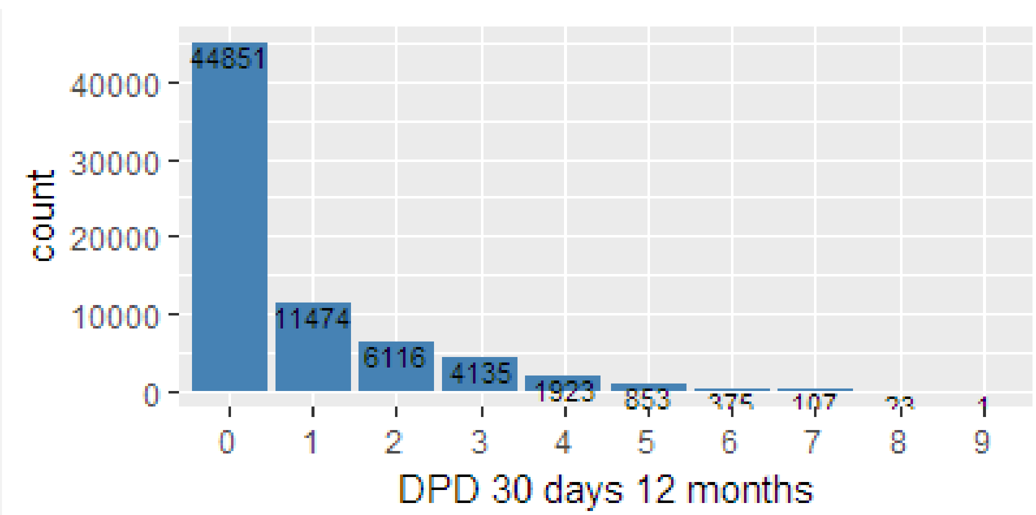
UNIVARIATE ANALYSIS

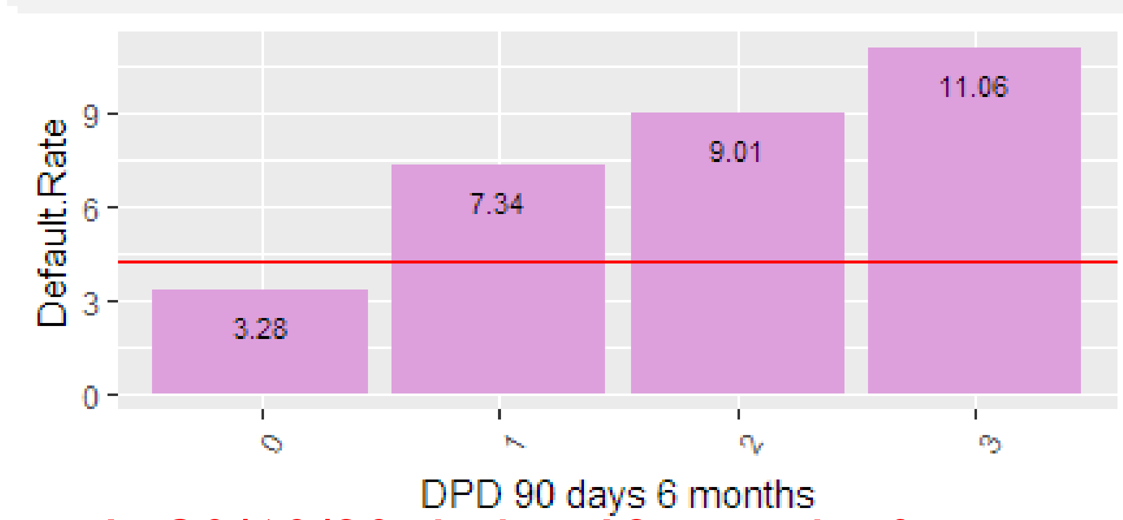










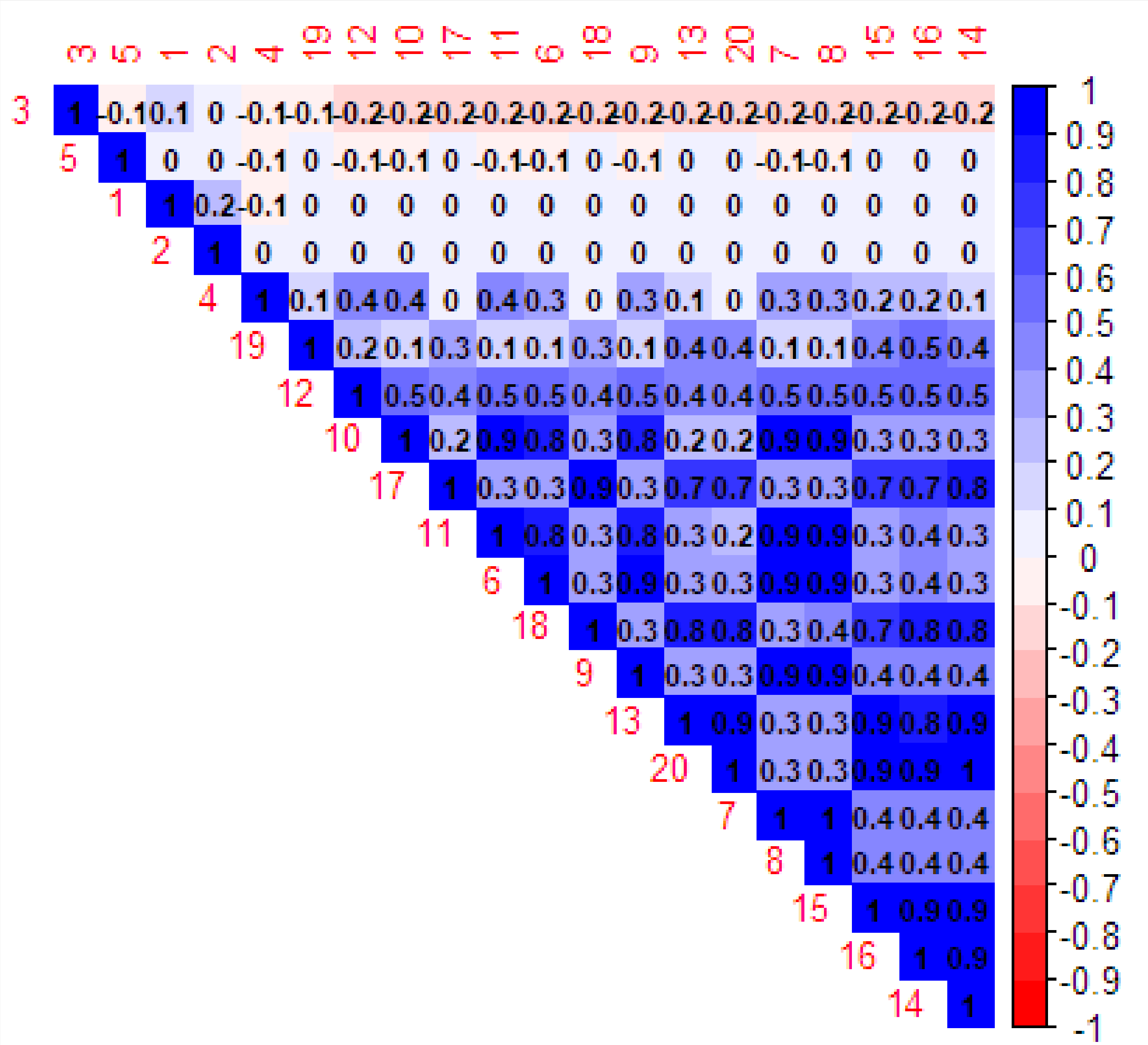


People 30/60/90 dpd in 12 months for more than once have default rate higher than total default rate

UNIVARIATE DATA ANALYSIS - OBSERVATIONS

- Possible Influential variables from basic understanding of domain and EDA
  - Average credit card utilization
  - Inquiries in last 12 months
  - Outstanding balance
  - No of PL trades open in 6 and 12 months
  - No of trades open in 6 and 12 months
  - 30/60/90 DPDs in 6 / 12 months
  - Total number of trades

BIVARIATE ANALYSIS – CORRELATION PLOT



1	Age
2	No.of.dependents
3	Income
4	No.of.months.in.current.residence
5	No.of.months.in.current.company
6	No.of.times.90.DPD.or.worse.in.last.6.months
7	No.of.times.60.DPD.or.worse.in.last.6.months
8	No.of.times.30.DPD.or.worse.in.last.6.months
9	No.of.times.90.DPD.or.worse.in.last.12.months
10	No.of.times.60.DPD.or.worse.in.last.12.months
11	No.of.times.30.DPD.or.worse.in.last.12.months
12	Avgas.CC.Utilization.in.last.12.months
13	No.of.trades.opened.in.last.6.months
14	No.of.trades.opened.in.last.12.months
15	No.of.PL.trades.opened.in.last.6.months
16	No.of.PL.trades.opened.in.last.12.months
17	No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.
18	No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.
19	Outstanding.Balance
20	Total.No.of.Trades

BIVARIATE ANALYSIS - OBSERVATIONS

- Full correlation (correlation value = 1)
  - No.of.times.60.DPD.or.worse.in.last.6.months, No.of.times.30.DPD.or.worse.in.last.6.months
  - No.of.trades.opened.in.last.12.months, Total.No.of.Trades
- Highly corelated with each other (correlation value = 0.7 to 0.9)
  - Set 1
    - 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
  - Set 2
    - 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.,
    - Total.No.of.Trades

INFORMATION VALUE ANALYSIS

- WOE and IV calculation outcome (20 bins)

Variable	IV	Significance
Avgas.CC.Utilization.in.last.12.months	0.3239047451	Strong predictive Power
No.of.trades.opened.in.last.12.months	0.3024276978	Strong predictive Power
No.of.PL.trades.opened.in.last.12.months	0.2984879891	Strong predictive Power
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	0.2964871353	Strong predictive Power
Outstanding.Balance	0.2862708723	Strong predictive Power

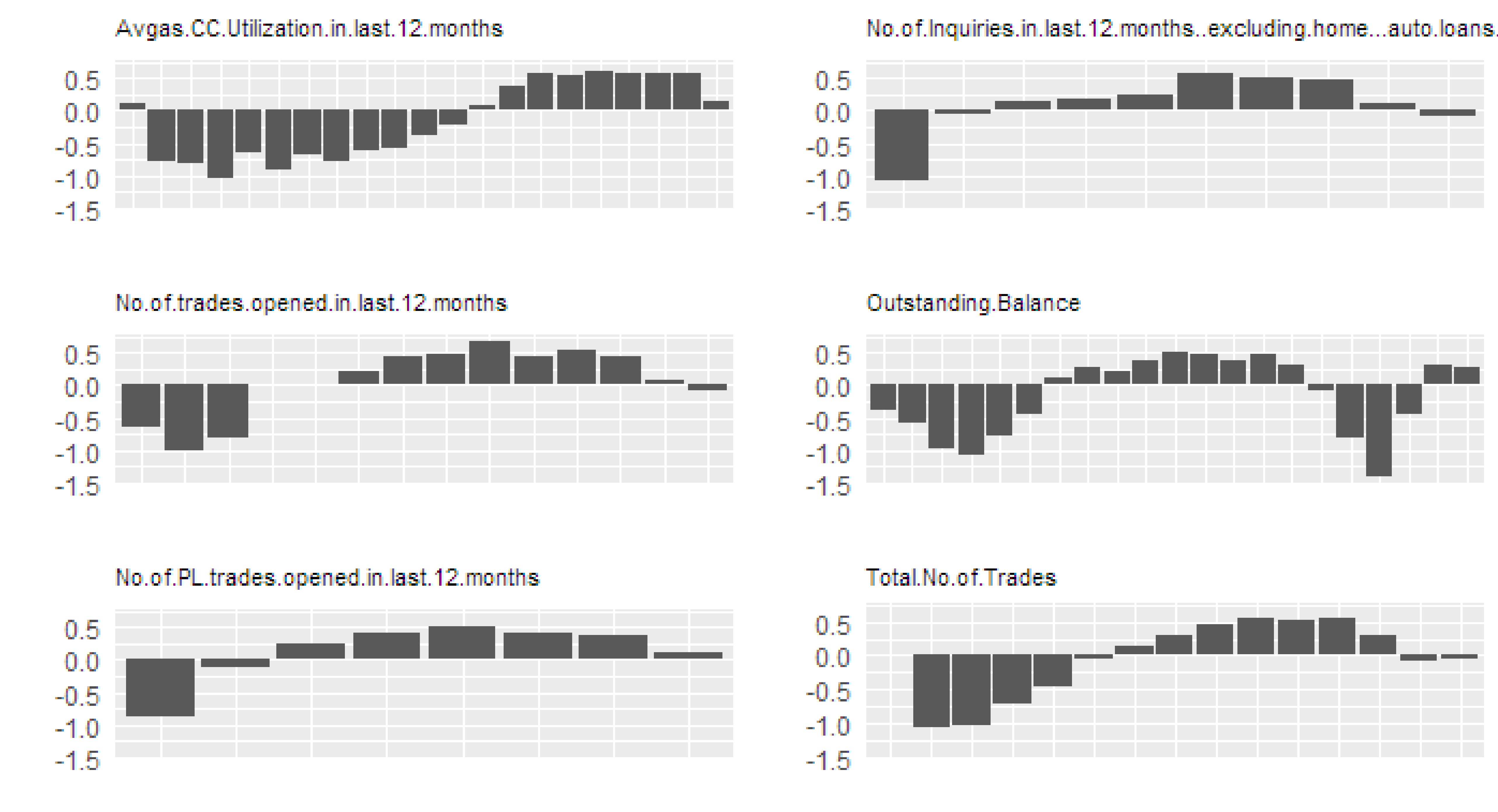
INFORMATION VALUE ANALYSIS

Variable	IV	Significance
Total.No.of.Trades	0.2511089002	Medium predictive Power
No.of.times.30.DPD.or.worse.in.last.6.months	0.2431741060	Medium predictive Power
No.of.PL.trades.opened.in.last.6.months	0.2211861971	Medium predictive Power
No.of.times.30.DPD.or.worse.in.last.12.months	0.2170640155	Medium predictive Power
No.of.times.90.DPD.or.worse.in.last.12.months	0.2138979793	Medium predictive Power
No.of.times.60.DPD.or.worse.in.last.6.months	0.2093563473	Medium predictive Power
No.of.Inquiries.in.last.6.months..excluding.home...auto. loans.	0.2081105253	Medium predictive Power
No.of.trades.opened.in.last.6.months	0.1895297957	Medium predictive Power
No.of.times.60.DPD.or.worse.in.last.12.months	0.1874348436	Medium predictive Power
No.of.times.90.DPD.or.worse.in.last.6.months	0.1601387901	Medium predictive Power

INFORMATION VALUE ANALYSIS

Variable	IV	Significance
No.of.months.in.current.residence	0.0817927246	Weak predictive Power
Income	0.0472964466	Weak predictive Power
No.of.months.in.current.company	0.0280788083	Weak predictive Power
Presence.of.open.home.loan	0.0176318803	Not useful for prediction
Age	0.0088733497	Not useful for prediction
Application.ID	0.0043142655	Not useful for prediction
No.of.dependents	0.0026513486	Not useful for prediction
Profession	0.0022283304	Not useful for prediction
Presence.of.open.auto.loan	0.0016570499	Not useful for prediction
Type.of.residence	0.0009247539	Not useful for prediction
Education	0.0007834615	Not useful for prediction
Gender	0.0003263165	Not useful for prediction
Marital.Status..at.the.time.of.application.	0.0000960168	Not useful for prediction

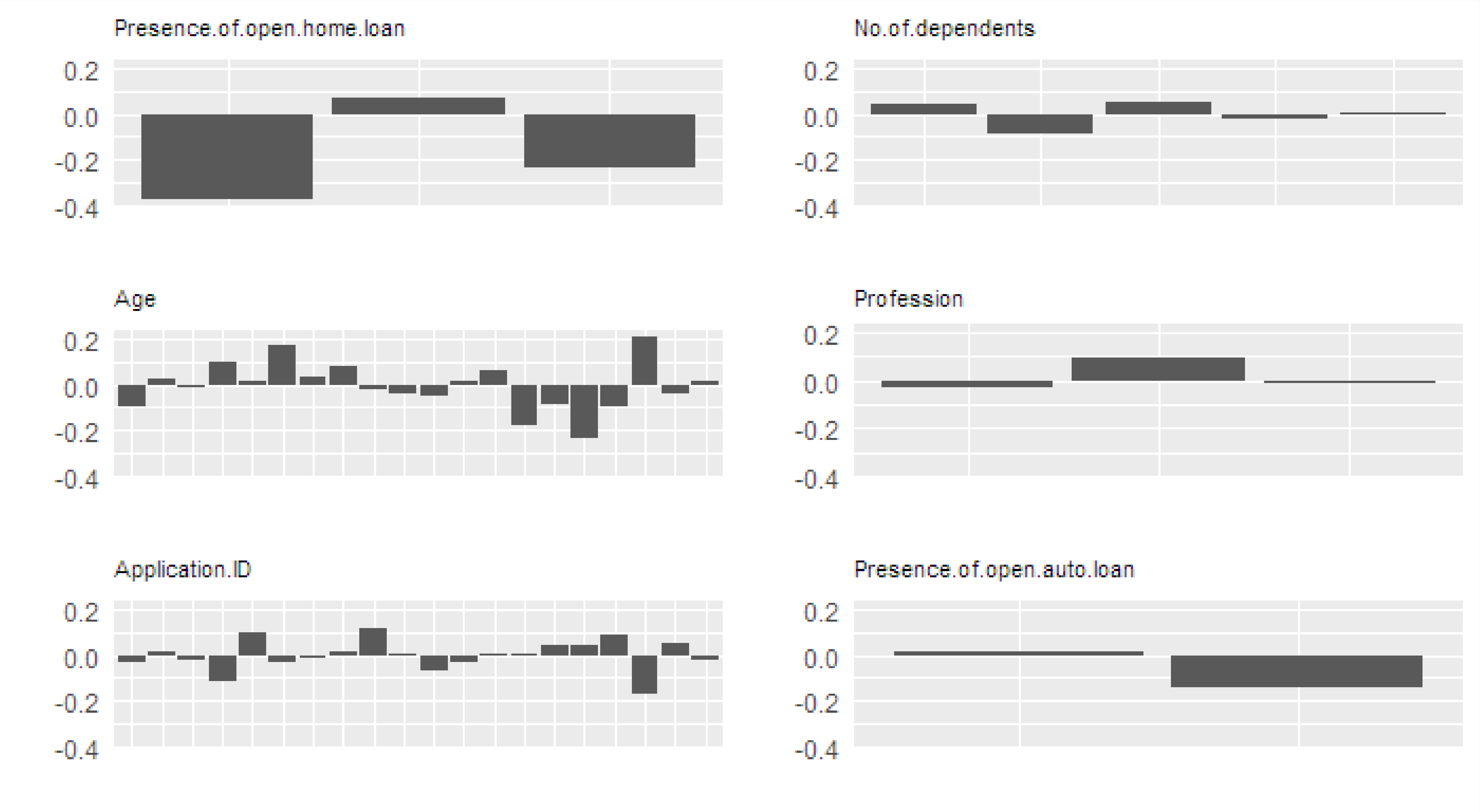
WOE BINS GRAPHS



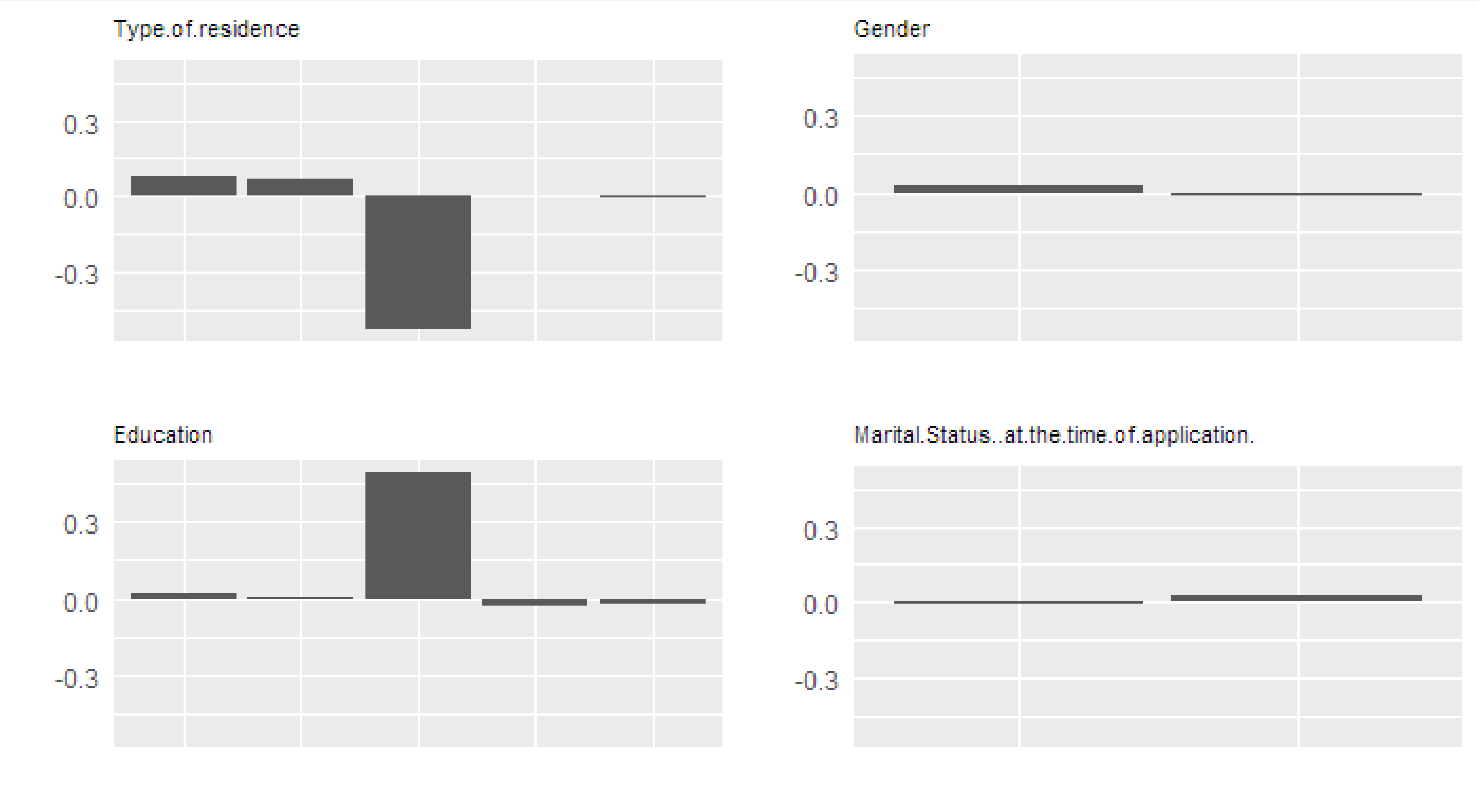
WOE BINS GRAPHS

WOE BINS GRAPHS

WOE BINS GRAPHS



WOE BINS GRAPHS



ROADMAP FOR MODEL BUILDING

- Step 1 – Additional data preparation (other than the cleaning already done)
  - Removal of rejected applications from model data (Performance Tag = NA)
  - Feature Engineering for logistic regression model
    - WOE binning and substitution for credit bureau numeric parameters having missing values (272 records of no earlier loan and 751 records with no earlier credit card)
    - Creating dummy variable for factor parameters
    - Scaling for other numeric parameters (which do not have any missing values)
  - Missing value imputation with WOE values for logistic regression model
    - Credit bureau parameters with missing values (272 records of no earlier loan and 751 records with no earlier credit card)

ROADMAP FOR MODEL BUILDING

- Step 2 – Create baseline model with logistic regression
  - Divide data in train and test (70:30)
  - Use stepAIC and VIF and P value to find significant predictive variables
  - Predict default probability on test data
  - Define optimal cut off of probability
  - Note the model performance parameters as baseline
    - Accuracy, Sensitivity / recall, Specificity, Precision , KS statistics, Gain and lift
- Step 3 – Run Logistic regression model only on demographic data
  - This is useful for identify predictive parameters for the applicants who do not have credit history record with credit bureau

ROADMAP FOR MODEL BUILDING

- Step 4 - Generate decision Tree model
  - Run model on original data (no scaling, no WOE transformation)
  - Usage of homogeneity measures to select attributes
    - Entropy, Gini Index, Information gain
  - Hyperparameters tuning to improve model performance
  - Truncation and Pruning to reduce overfitting and model complexity
- Step 5 – Generate Random forest model to improve overall performance
  - Run model on original data (no scaling, no WOE transformation)
  - Hyperparameters tuning to improve model performance
- Step 6 – Generate SVM model (Optional)
  - Generate SVM model if PC configuration supports
  - Usage of linear, polynomial and radial kernels to achieve tradeoff between bias and variance

ROADMAP FOR MODEL EVALUATION

- Key parameters for model evaluation (Default rate is considered as "yes")

Business objective	Model characteristic	Evaluation parameter
Reduce credit loss – Reject bad applications	Predict maximum defaulters	Sensitivity
Increase revenue - Reduce rejection of good applications	Predict Non-Defaulters correctly	Specificity
Reduction of acquisition cost – Automated rejection process for risky applicants	Identify minimum target number of applications to locate maximum defaulters	Gain and lift chart
A tradeoff between business /revenue growth and possible credit loss	Discriminative power of model	KS statistics

- Accuracy would not be used as a evaluation parameter for following reasons
  - Default rate of the base data is 4.22%
  - Even with no model used and by accepting all applications, accuracy of approx 95% would be achieved

MODEL BUILDING DATA MANIPULATION

- Used Logistic Regression and Random Forest Models
- Performed both Models on
  - Raw Data,
  - Balance Data,
  - Raw WOE Data
  - Balance WOE Data
- Scaled and created dummy variables on Test and Train Data – For logistic regression
- Split the data into Train and Test Data
- Balanced Data using ROSE package
  - By balancing data adjusted default rate to 15 to 20 %

MODEL PREPARATION – LOGISTIC REGRESSION DEMOGRAPHIC DATA

- Ran demographic model using all variables on training dataset(70%)
- Removed insignificant variable using stepAIC method
- Improved model iteratively
  - Removing multi-collinearity using VIF values
  - Choosing most significant variable using p-value
- Selected final model with all significant variables
  - Income,
  - No.of.months.in.current.residence,
  - No.of.months.in.current.company

MODEL EVALUATION – LOGISTIC REGRESSION DEMOGRAPHIC DATA

- Created prediction based on final model on test data set
- Checked Accuracy, Sensitivity and Specificity using different values of cut off
- Selected final cut off and got Accuracy, Sensitivity and Specificity for each models as per shown below,

Logistic Regression Demographic Data	
Accuracy	5605764
Sensitivity	0.5542986
Specificity	0.5608529

MODEL PREPARATION – RANDOM FOREST

- Ran random forest model with ntree-100, mtry-5 on train data set
- Plotted Imp variables plot to get idea about important variable
- Validated Random forest model on test dataset
- Predicted default customer on Test Dataset
- Define optimal cutoff to maximize sensitivity and specificity
- Model evaluation done using metrics Accuracy, Sensitivity, Specificity, KS-statistics

MODEL PREPARATION LOGISTIC REGRESSION

- Ran model using all variables on training dataset(80%)
- Removed insignificant variable using stepAIC method
- Improved model iteratively
  - Removing multi-collinearity using VIF values
  - Choosing most significant variable using p-value
- Selected final model with all significant variable

MODEL PREPARATION LOGISTIC REGRESSION

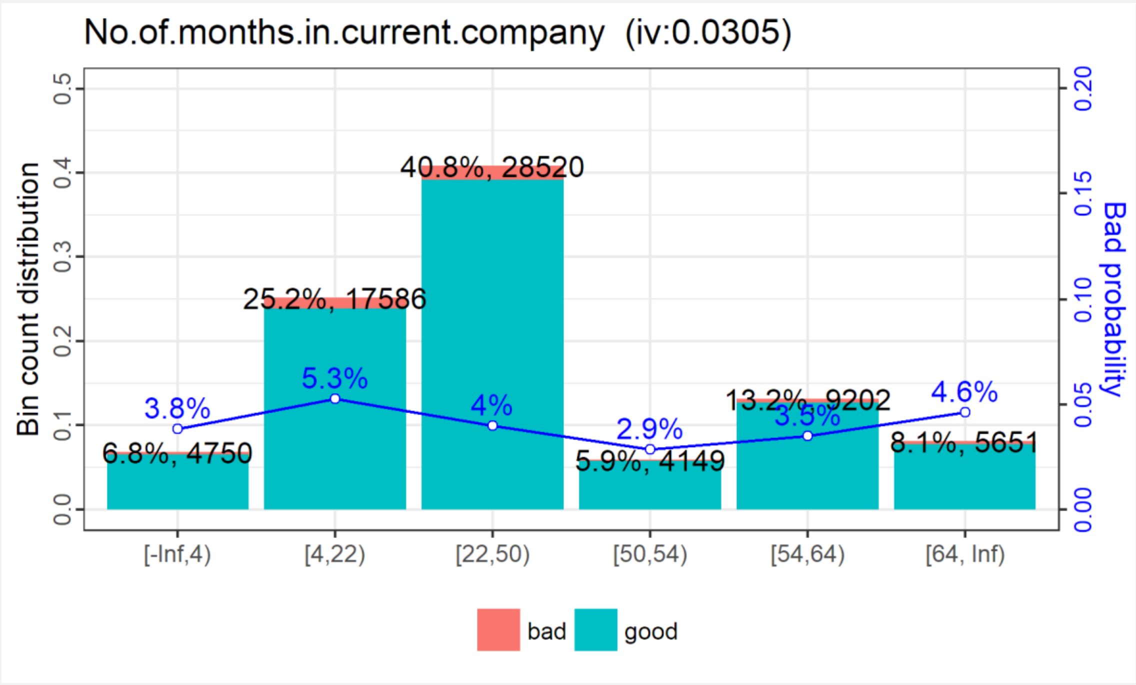
- Logistic Regression on Raw Data significant variables
  - Income, No.of.times.90.DPD.or.worse.in.last.12.months, No.of.times.30.DPD.or.worse.in.last.12.months, Avgas.CC.Utilization.in.last.12.months, No.of.PL.trades.opened.in.last.6.months, No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.
- Logistic Regression on Raw Balance Data significant variables
  - Income, No.of.months.in.current.residence, No.of.months.in.current.company, No.of.times.30.DPD.or.worse.in.last.12.months, Avgas.CC.Utilization.in.last.12.months, No.of.PL.trades.opened.in.last.12.months, No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.

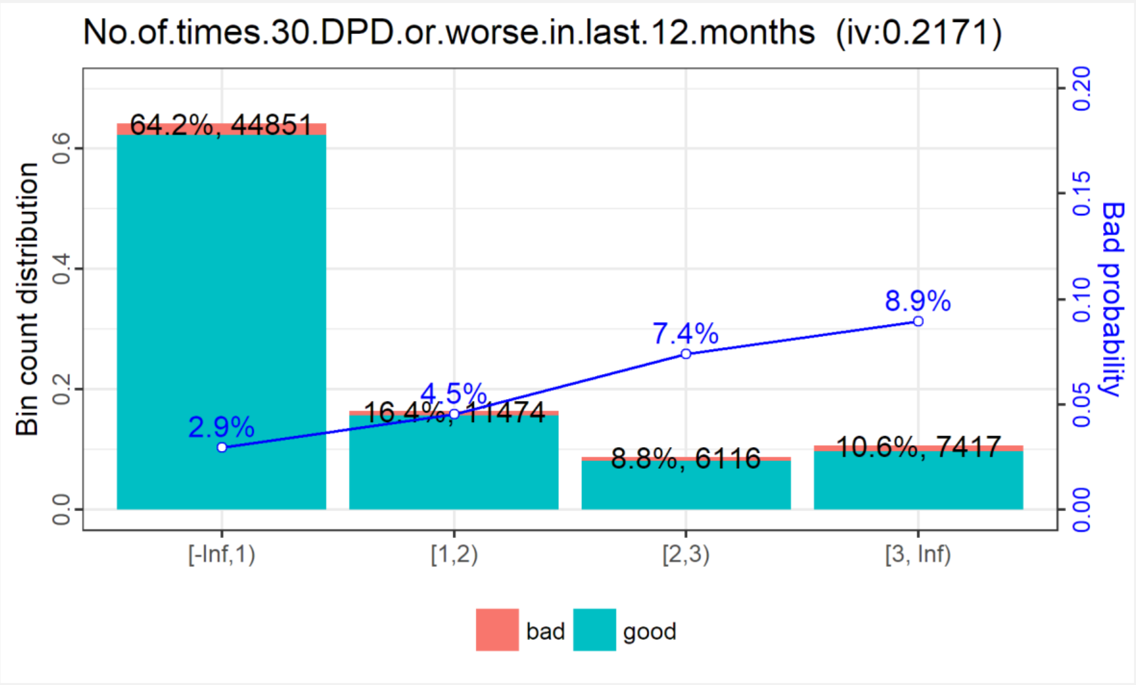
MODEL PREPARATION LOGISTIC REGRESSION

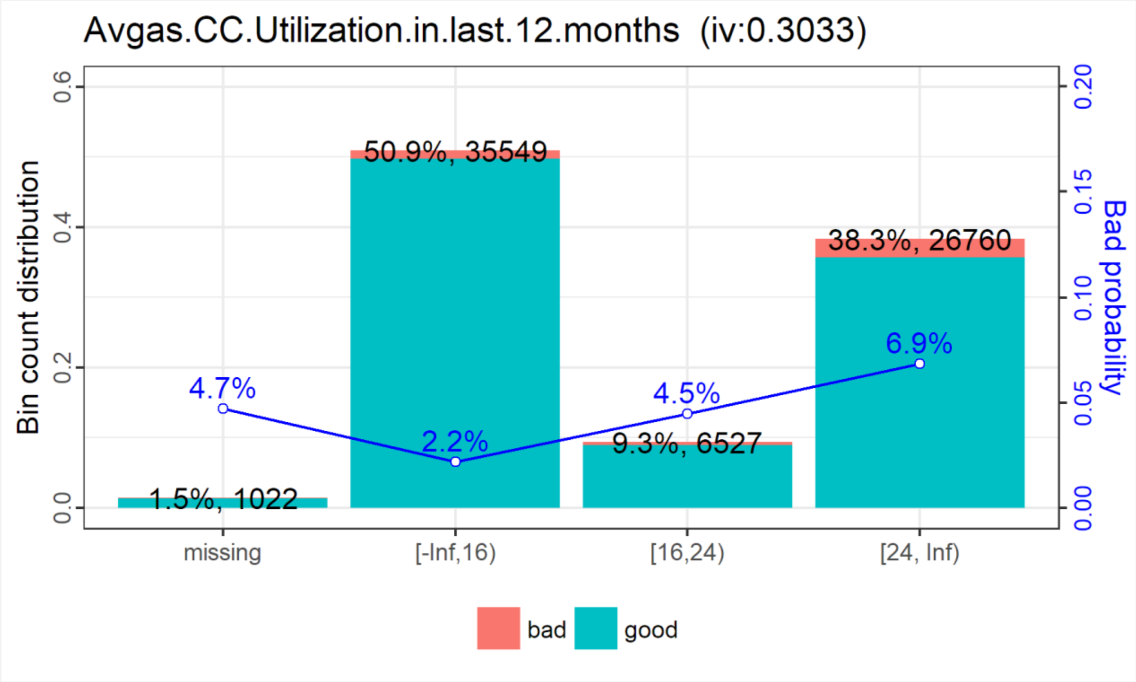
- Logistic Regression on Raw WOE Data significant variables
  - No.of.months.in.current.company_woe, No.of.times.30.DPD.or.worse.in.last.12.months_woe, Avgas.CC.Utilization.in.last.12.months_woe, No.of.Inquiries.in.last.12.months..excluding.home...auto.loans._woe
- Logistic Regression on WOE Balance Data significant variables
  - Age_woe, No.of.dependents_woe, Profession_woe, No.of.months.in.current.company_woe, No.of.times.30.DPD.or.worse.in.last.12.months_woe, Avgas.CC.Utilization.in.last.12.months_woe, No.of.Inquiries.in.last.12.months..excluding.home...auto.loans._woe, Outstanding.Balance_woe

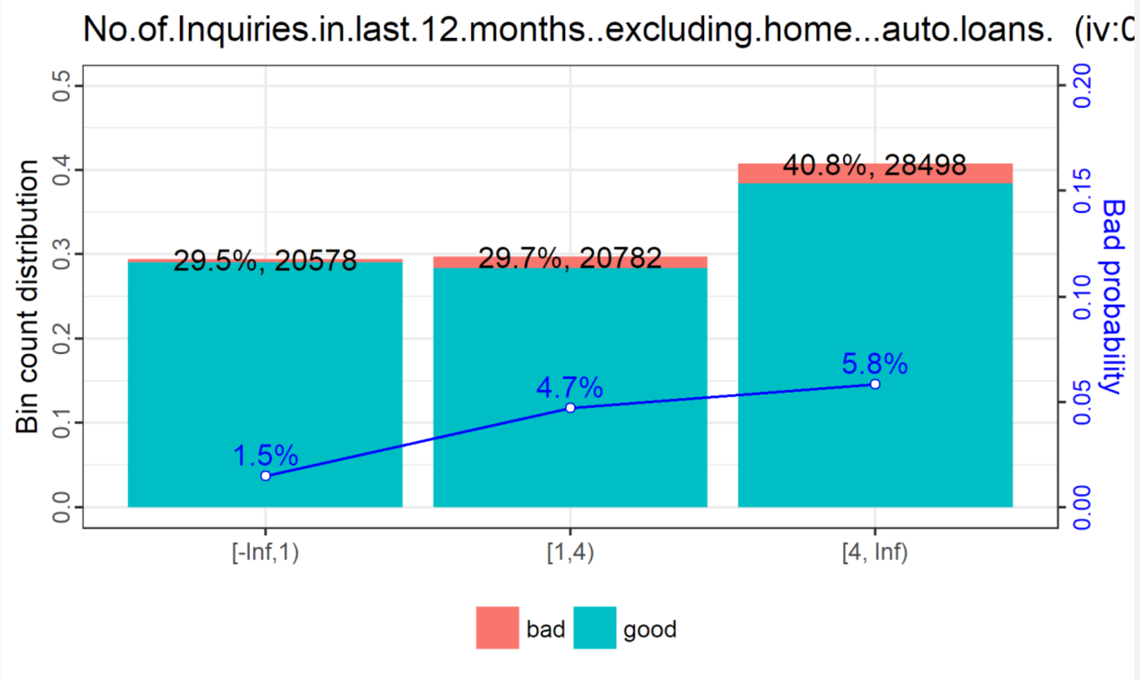
MODEL PREPARATION LOGISTIC REGRESSION

- WOE Raw Data Graphs for GOOD and BAD probability distribution of significant variables



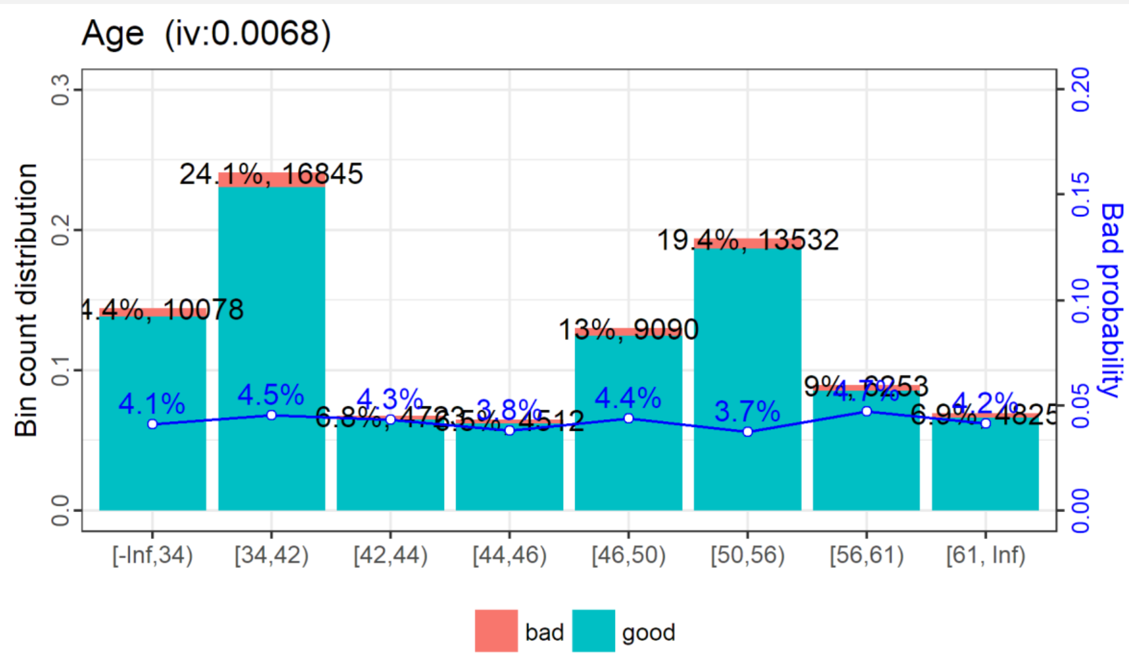


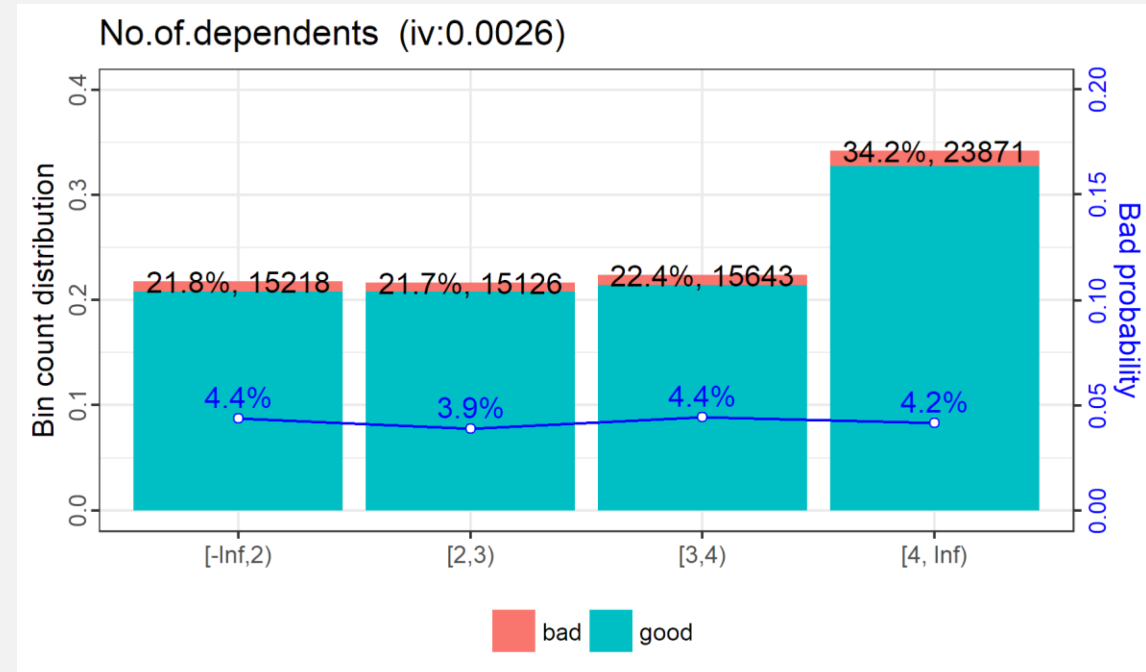


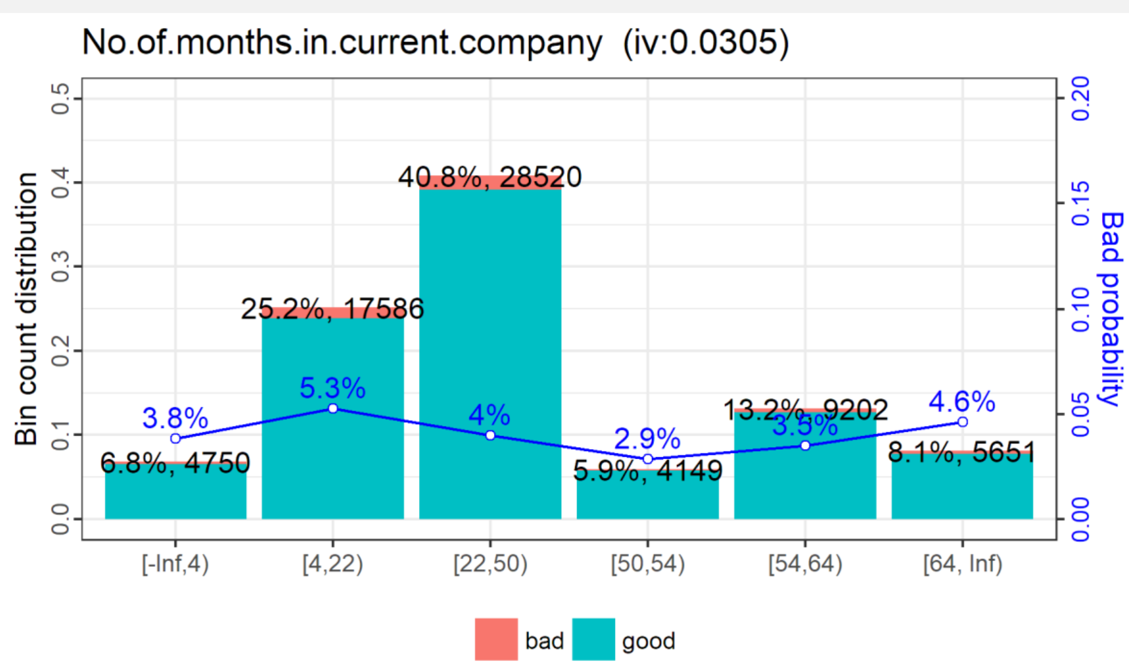


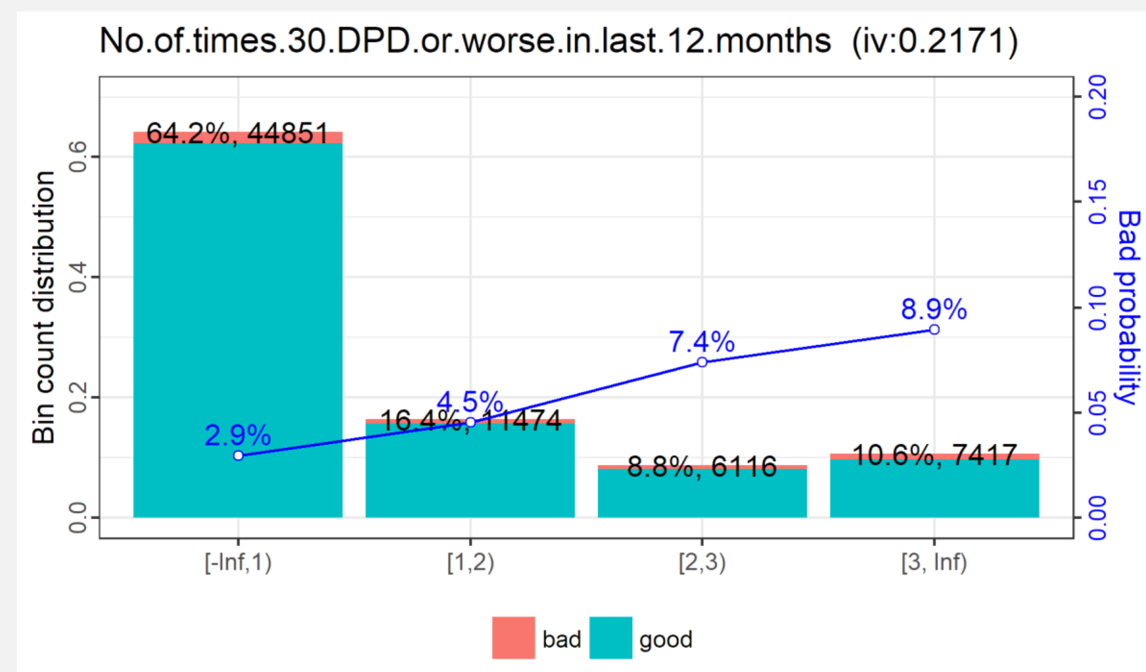
MODEL PREPARATION LOGISTIC REGRESSION

- WOE Balance Data Graphs for GOOD and BAD probability distribution of significant variables

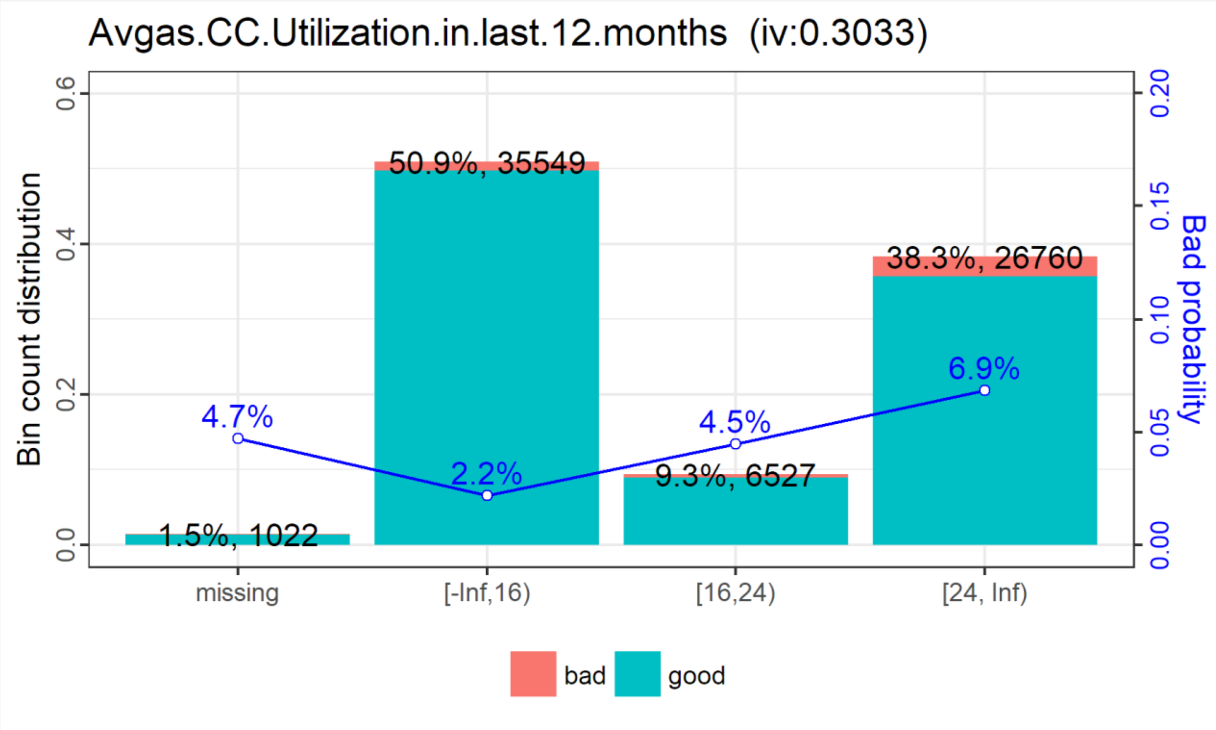


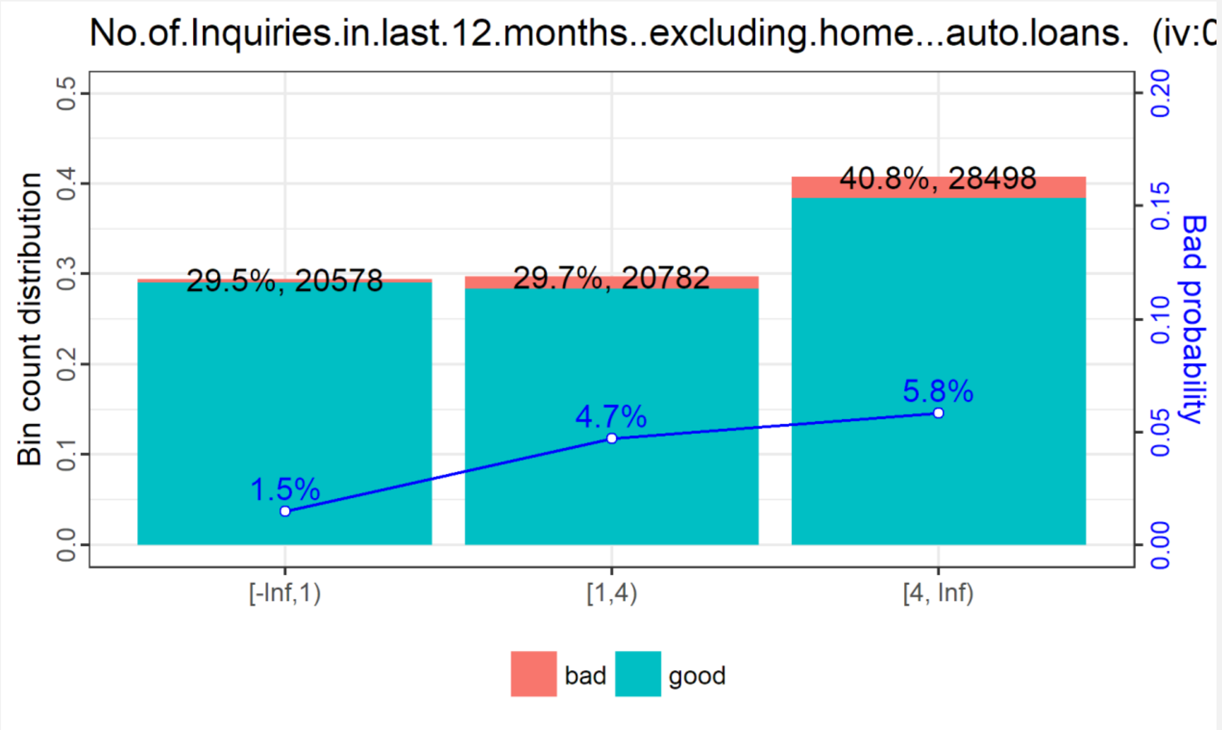


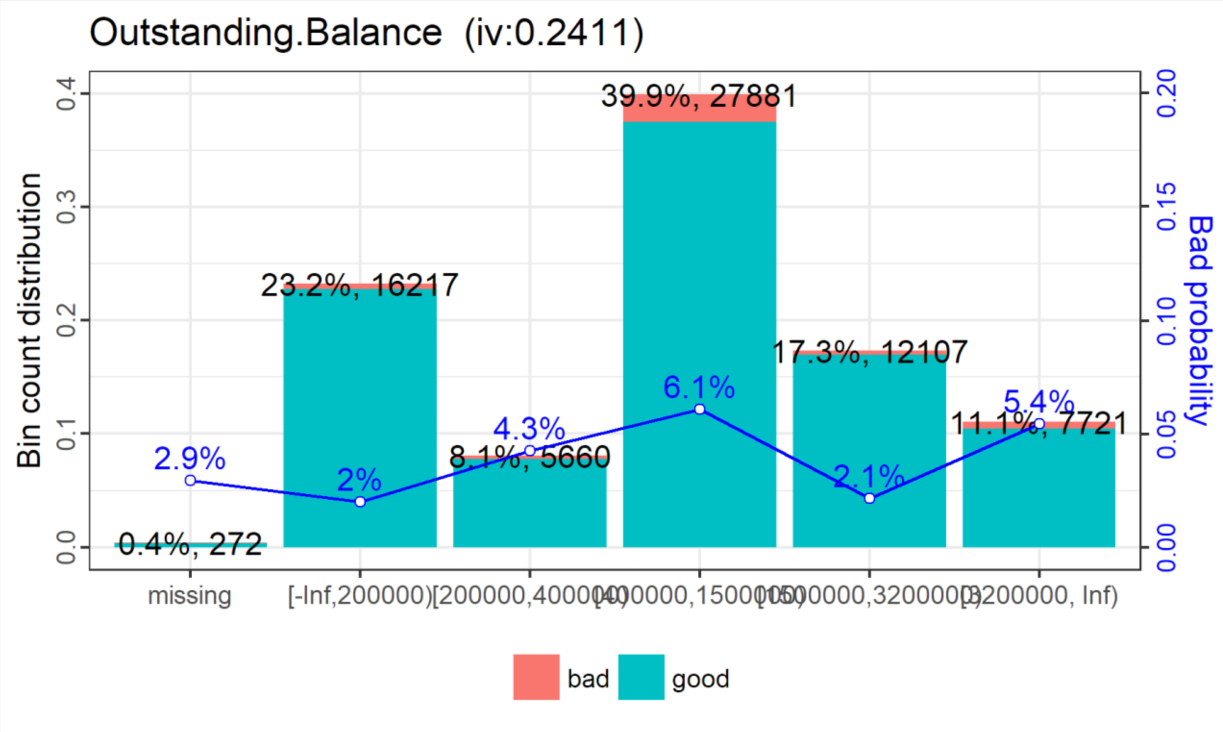




MODEL PREPARATION LOGISTIC REGRESSION







MODEL EVALUATION

- Created prediction base on final model on test data set
- Checked Accuracy, Sensitivity and Specificity using different values of cut off
- Selected final cut off and got Accuracy, Sensitivity and Specificity for each models as per shown below,

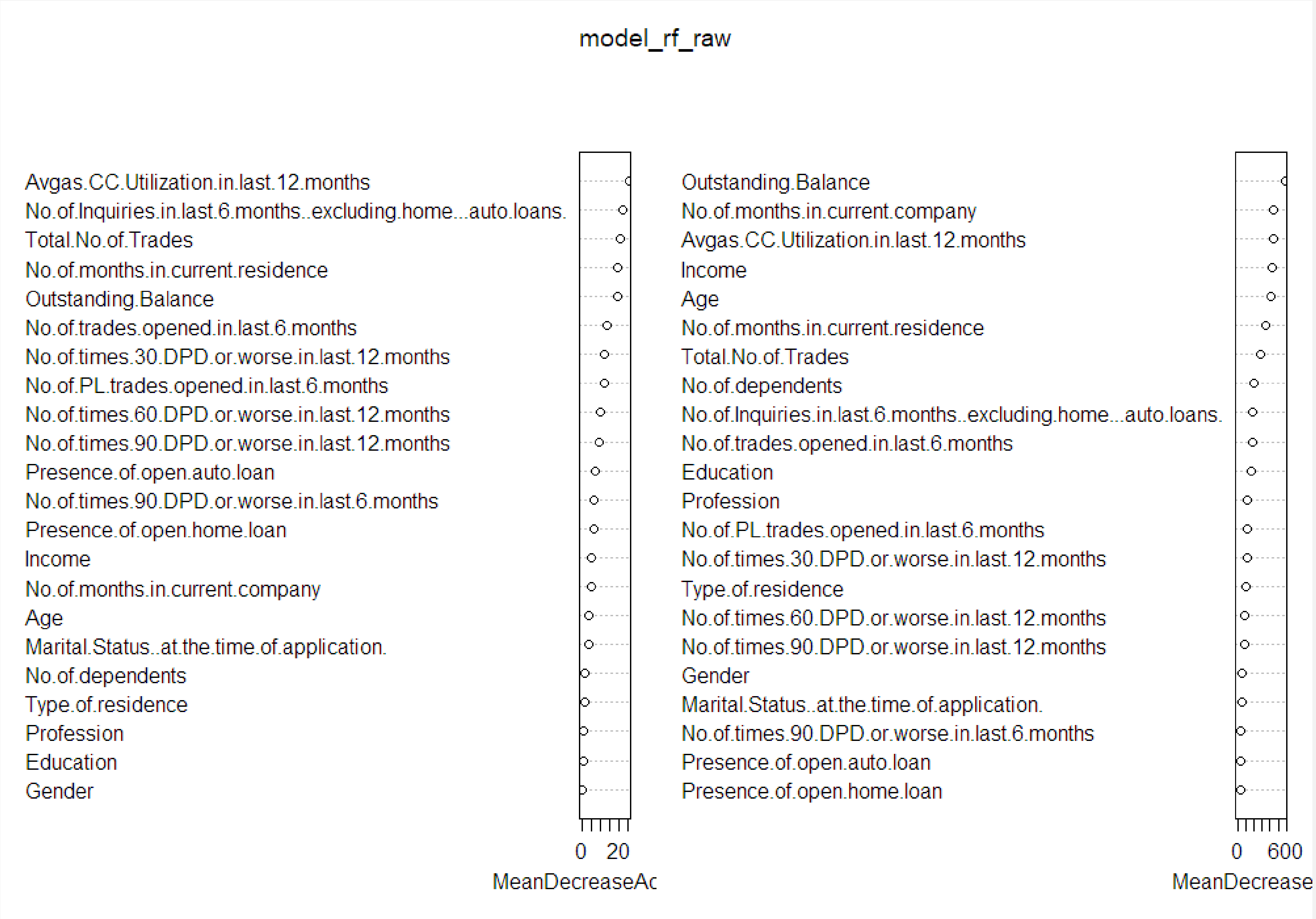
Logistic Regression Raw Data		Logistic Regression Raw balance		Logistic Regression Raw woe		Logistic Regression woe balance	
Accuracy	0.6055184	Accuracy	0.6124165	Accuracy	0.6452652	Accuracy	0.6391096
Sensitivity	0.6428571	Sensitivity	0.6207483	Sensitivity	0.6281834	Sensitivity	0.6417657
Specificity	0.6038713	Specificity	0.6120489	Specificity	0.646017	Specificity	0.6389927

- Model Discriminative power using KS-statistics and Gain

Logistic Regression Raw Data		Logistic Regression woe balance		Logistic Regression Raw balance		Logistic Regression Raw woe	
KS-statistics	Gain top 4 decile	KS-statistics	Gain top 4 decile	KS-statistics	Gain top 4 decile	KS-statistics	Gain top 4 decile
24.67	40.8	23.27	40.8	27.42	40.9	28.07	40.8

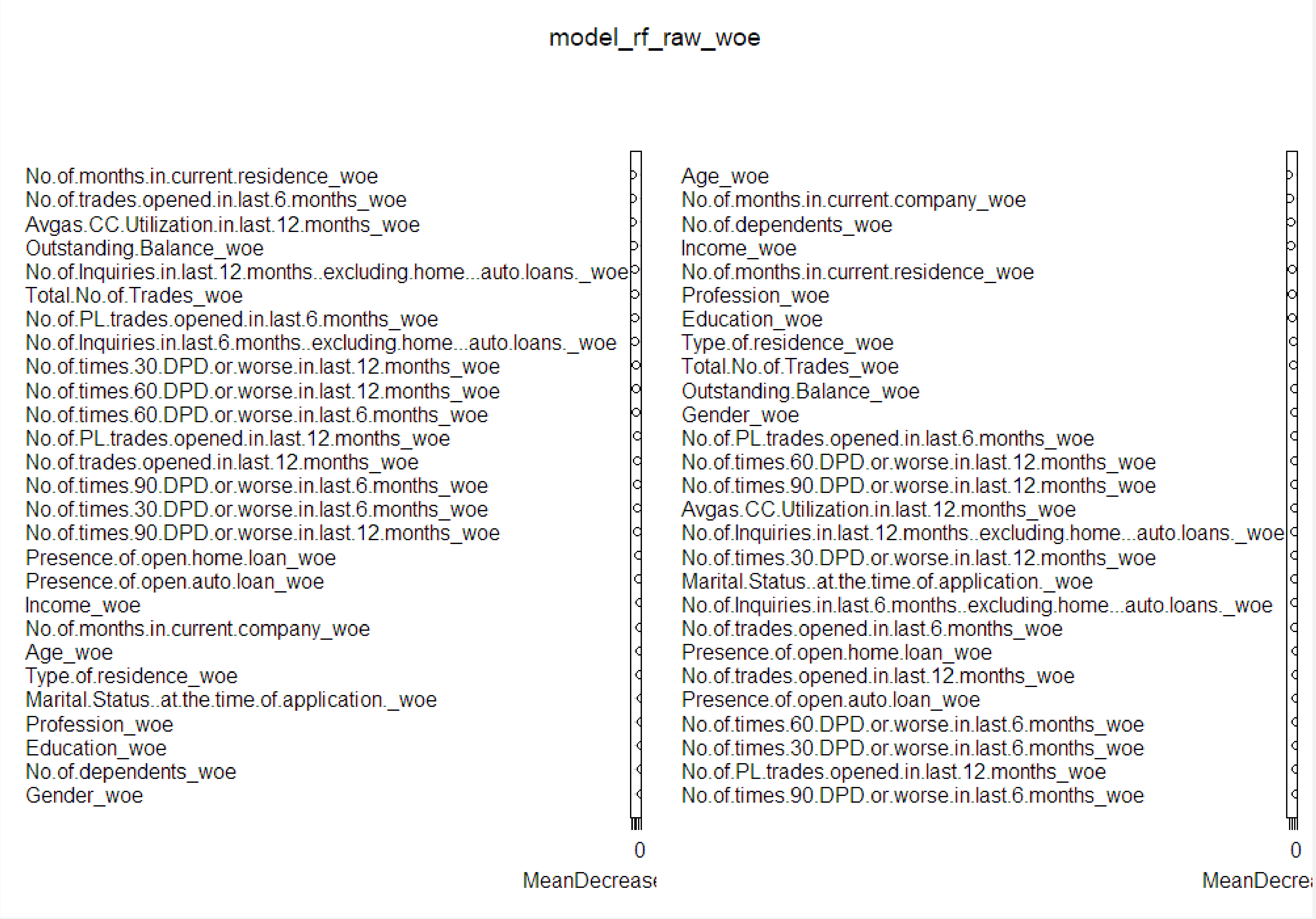
MODEL PREPARATION – RANDOM FOREST

- Important variables obtain from random Forest Models for RAW data set



MODEL PREPARATION – RANDOM FOREST

- Importance variables obtain from random Forest Models for WOE data set



MODEL – RANDOM FOREST

- Created prediction base on final model on test data set
- Checked Accuracy, Sensitivity and Specificity using different values of cut off
- Selected final cut off and got Accuracy, Sensitivity and Specificity for each models as per shown below,

Random Forest raw		Random Forest Raw balance		Random Forest Raw_woe		Random Forest woe balance	
Accuracy	0.5669325	Accuracy	0.6115635	Accuracy	0.5856417	Accuracy	0.6267268
Sensitivity	0.6207483	Sensitivity	0.6088435	Sensitivity	0.6264856	Sensitivity	0.5670628
Specificity	0.5645585	Specificity	0.6116835	Specificity	0.583844	Specificity	0.583844

- Model Discriminative power using KS-statistics and Gain

Random Forest raw		Random Forest Raw balance		Random Forest Raw_woe		Random Forest woe balance	
KS-statistics	Gain top 4 decile	KS-statistics	Gain top 4 decile	KS-statistics	Gain top 4 decile	KS-statistics	Gain top 4 decile
18.53	41.5	22.05	41.3	21.03	41.1	19.64	40.8

APPLICATION SCORECARD

- **Basis of scorecard** :
- good to bad odds of 10 to 1 at a score of 400 doubling every 20 points.

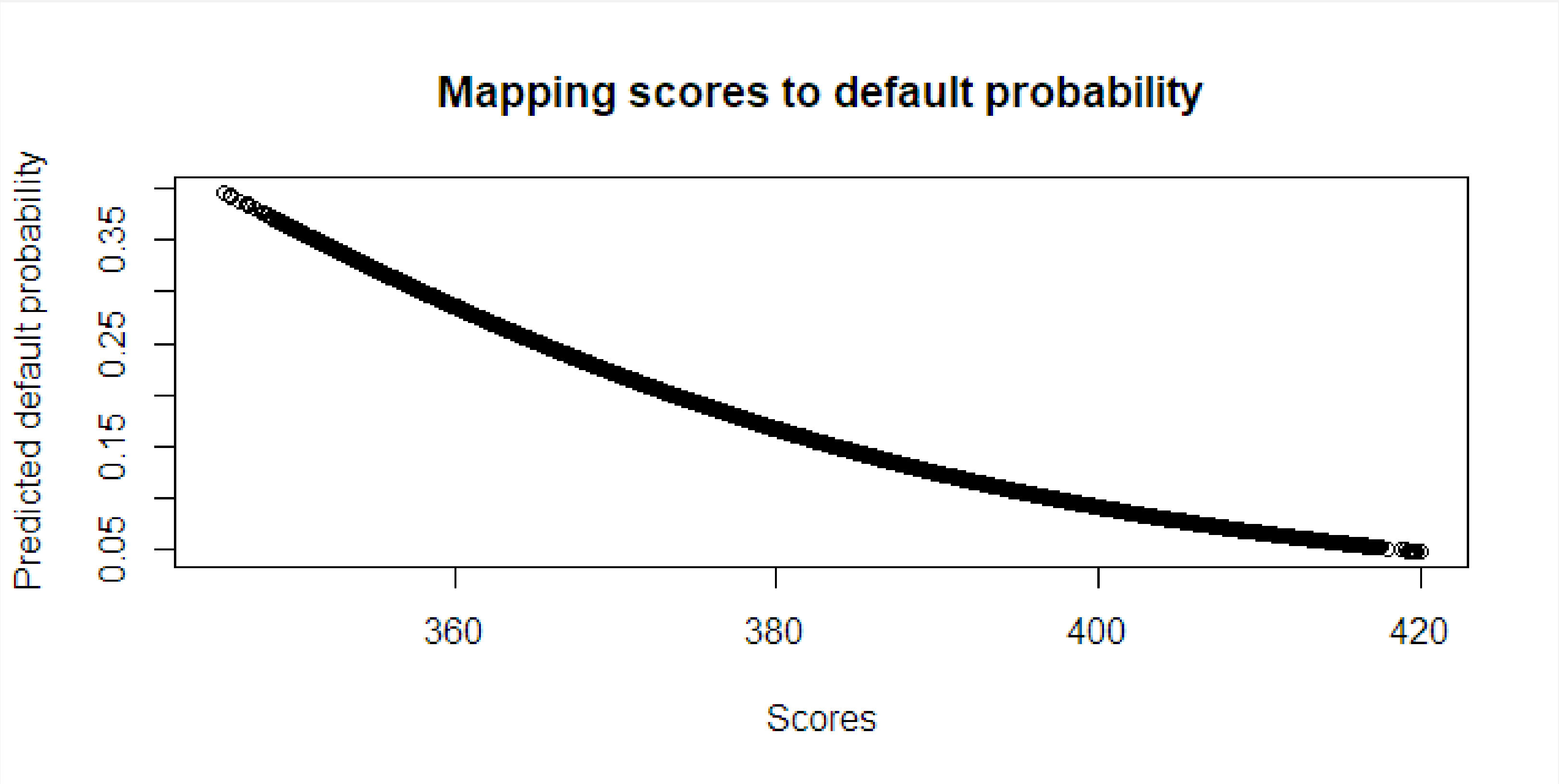
Goods	Bads	Total	P Good	P Bad	Odds	ln Odds	Score
10	1	11	0.909091	0.090909	10	2.302585	400
20	1	21	0.952381	0.047619	20	2.995732	420

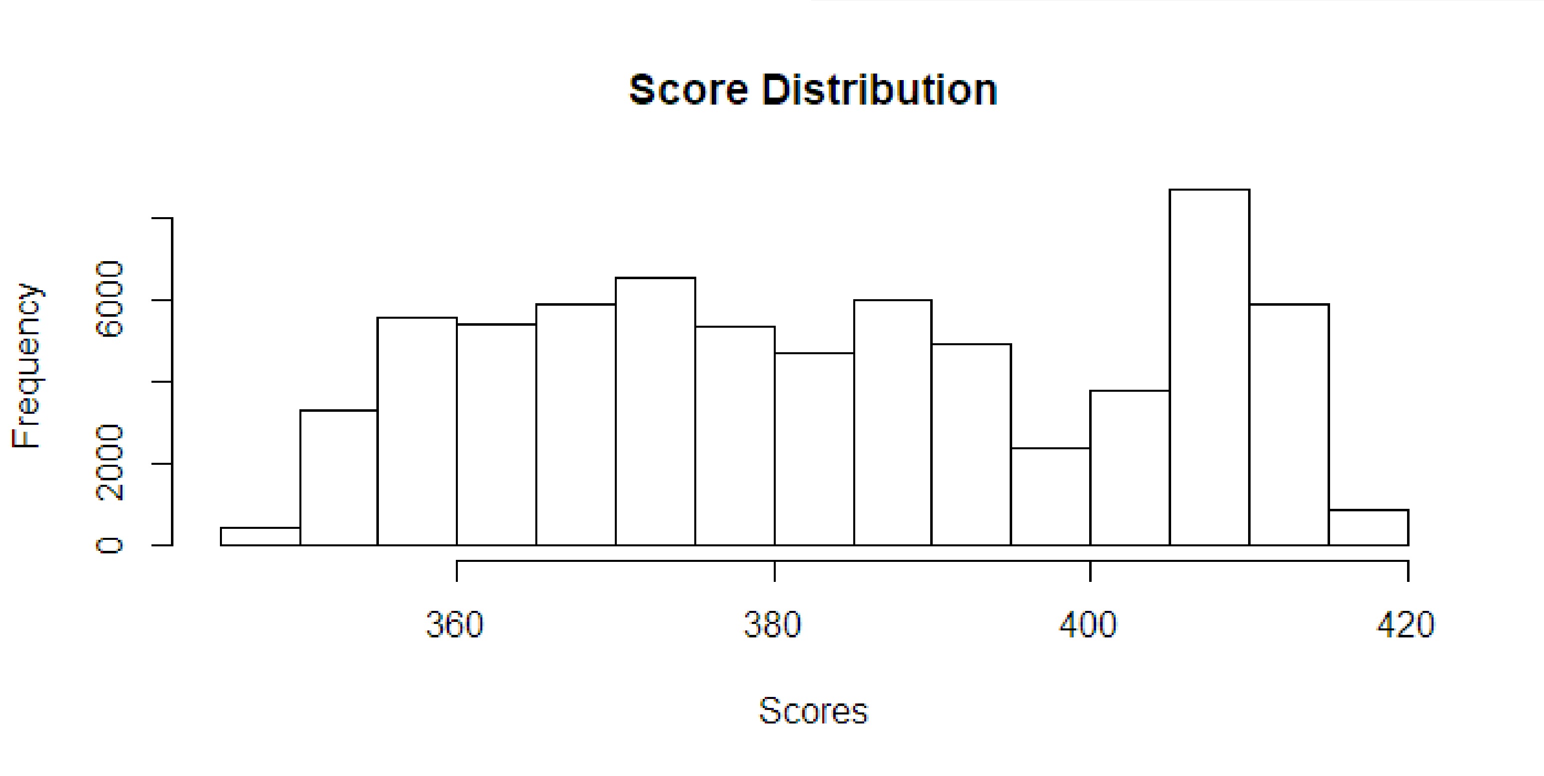
- Solving equations
- 400 = m*2.302585093 + c
- 420 = m*2.995732274 + c
- m = multiplication factor = 28.8539
- C = offset / intercept = 333.56144
- Equation of scorecard
- Score = 28.8539* log odds + 333.56144

APPLICATION SCORECARD

- Based on evaluation parameters, best model selected for application scorecard is Logisitic Regression model for WOE balanced data
- Score range based on the model predictions
- min score : 345.6618
- max score : 419.9896
- Score cutoff corresponding to bad probability cut off used in model (0.19570994) :374.314

APPLICATION SCORE CARD



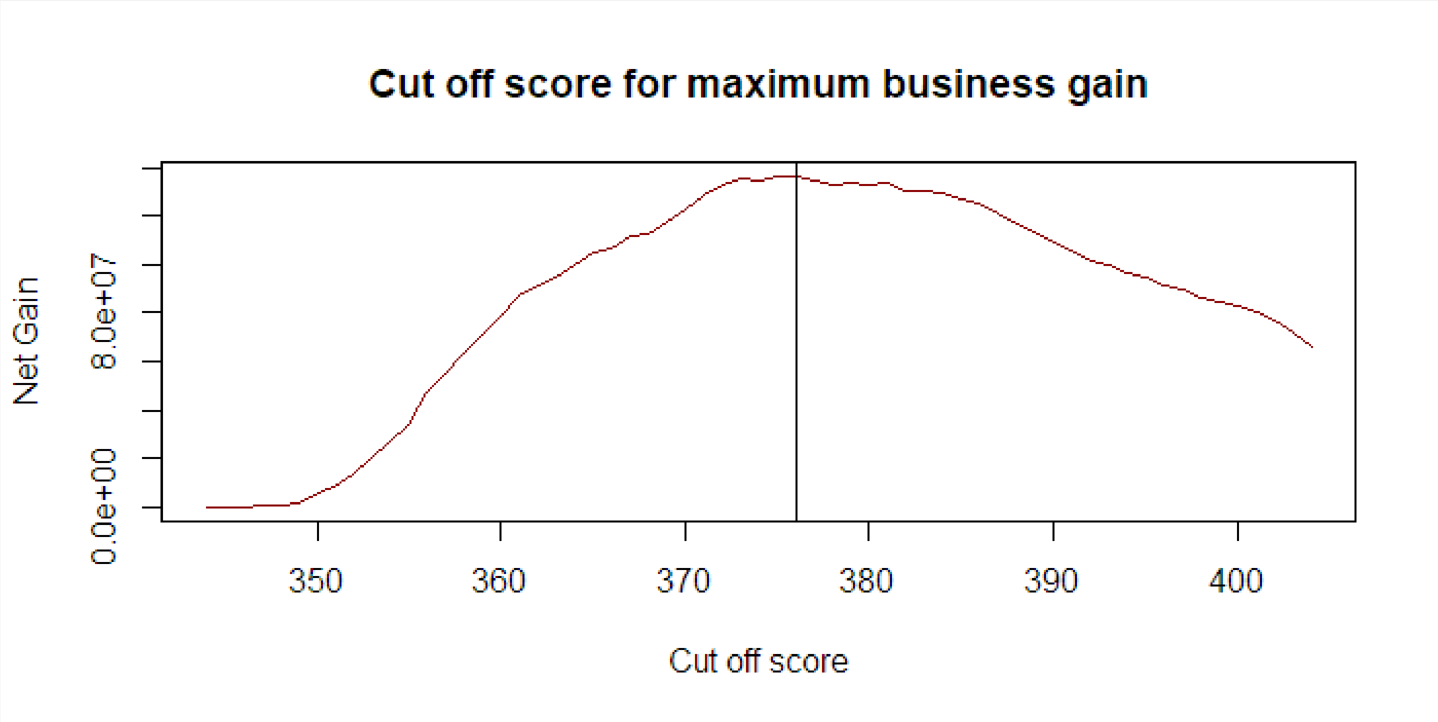


TUNING CUTOFF SCORE BASED ON FINANCIAL GAINS

- Finding the optimal cutoff where net gain in business would be maximum
- ASSUMPTIONS
  - Average credit limit per customer = 100000
  - Avg credit utilization per customer = 50%
  - Bank Earning per transaction = 1.5%
  - Business gain for bank = Credit limit* avg credit util*earning per transaction*good customers
  - Credit loss for bank = bad customers * 4 months(120 days) credit util
  - NET GAIN = Business gain from good customers – credit loss due to bad customers

TUNING CUTOFF SCORE BASED ON FINANCIAL GAINS

- Final cutoff based on tuning = 376 (Customers with score more than or equal to 376 would be provided credit card, all below 376 would not be granted)



- Accepted customers with new cutoff : 41402
- Rejected customers with new cutoff :28456
- With scorecard, acceptance ratio = 59.26% (All would be possibly good customers)
- Original acceptance ratio = 98%
- Acceptance process with score card would be automatic and more stringent leading to maximum business gain

FINANCIAL BENEFIT WITH MODEL - REJECTED CUSTOMERS

- Created prediction based on final model for rejected customers
- Possible accepts (score> 374)
  - 966 customers from customer rejected are good customers
- possible rejects (score<= 374))
  - 459 customers from customer rejected are bad customers
- Approval rate : (966/1425)*100 = 67.78%
- Possible business gain by implementation of model = Business gain from the possibly accepted customers
- = 966*100000*0.5*1.5*12/100 = Rs 8694000

ASSESSMENT OF THE MODEL

- Automatic approval process to total acceptance = 41402+ 966 = 42368 out of 71283
  - Total Acceptance ratio – 59.4%
- If 100 applicants are there,
  - Accepted = 59
  - Rejected = 41
- Business gain due to accepted customers based on earlier assumptions = 59*100000*0.5*1.5*12/100 = 531000
- Prevented credit loss for rejected customers = 41*100000*0.5*4 = 8200000
- Business loss due to rejected applicants = 41*100000*0.5*1.5*12/100 = 369000