CredX-Acquisition Analytics
BFSI – CAP STONE PROJECT

Submitted by:

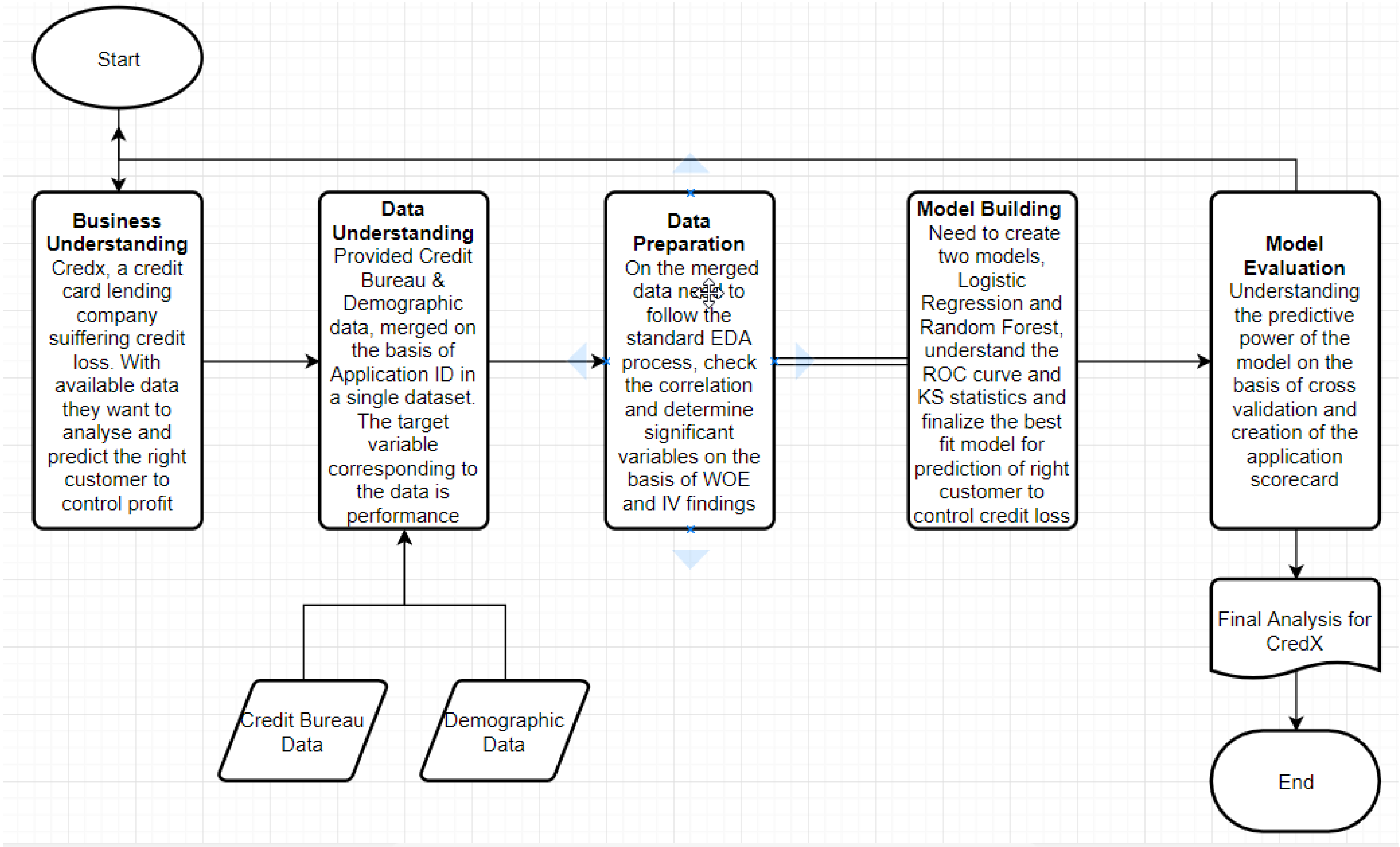
Saptarshi Banerjee

Vibhor Srivastava

Jaideep Pal

Vivek Rastogi

Approach - Flow Diagram : CRISP DM Framework



Business Understanding

Business Understanding

Problem Statement: CredX, one of the renowned credit card company facing credit loss since they are not able to correctly identify appropriate credit risk for applicants during past few years. It's CEO feels the most important step is to "acquire the right customers".

Aim

1) The aim is to automate the process of predicting the right customers using past data of the bank's applicants.
2) To understand different factors affecting the credit risk so that the right customer is chosen.
3) Create the appropriate strategies to mitigate acquisition risk and assess the financial benefit of the project.

Data sets

We have two sets of data

1) **Demographic / Application Data** – Acquired from the customer at the time of filling the application while applying for the loan. Giving us demographic details, giving us the ability to understand correlations between default with demographic attributes.
2) **Credit Bureau** – Obtained from Credit Bureau, at a customer level, with details of types and number of delinquencies. Giving us an idea of the financial health of the customer. So that we can map "which attributes of financial health can predict default"

Methodology - Data Understanding & Data Preparation

Collect Relevant Data & Integrate data Files

Demographic Data - 71295 obs. of 12 variables & Credit Bureau Data - 71295 obs. of 19 variables:

- **Primary Key**
  - Application id is the primary key common across both data sets and is "identical" in both
  - Duplicated Application id's -765011468, 653287861 & 671989187, deleted 6 records from both data sets
  - Merged data files on Application id as primary key
- **Target Variable**
  - Primary tag is target variable, is same in both data sets - Removing one.
  - 1425 rows have NA's in performance tag, saving as rejected population to be used for Model Validation
  - Overall Default rate is 2947/(2947+66917) = 0.04218195

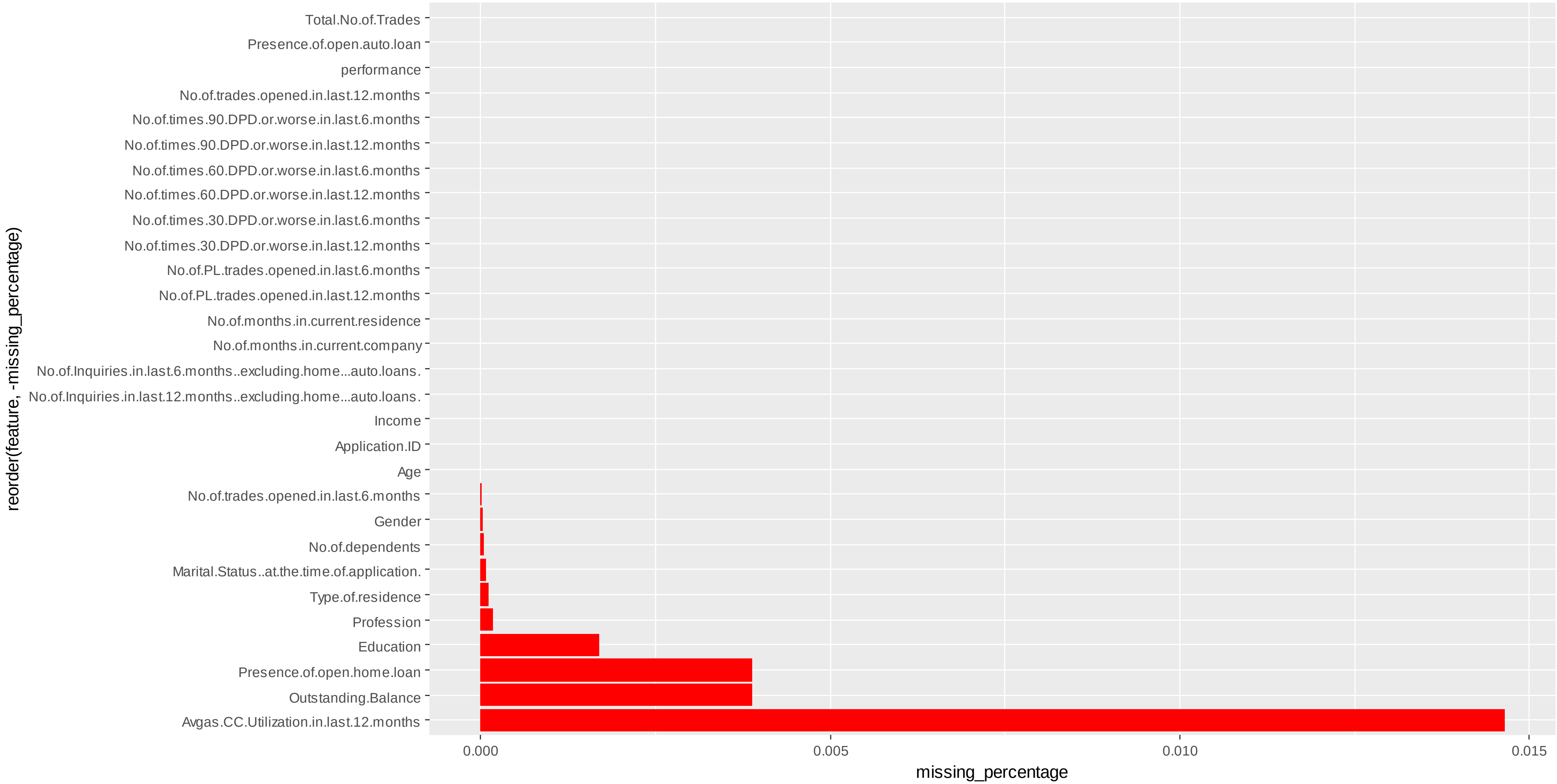
Verify Data Quality

Data has NA's, Blanks, Negative values, Zero's in certain variables

- **Missing Value**
  - We have two types of missing values – Blanks & NA's - Replacing all blank fields with NA's in both data sets

  Application data – Missing Values
  - Missing values are less than 2% in "Application data" variables - Age, Gender, No.of.Dependents,, Marital status, Residence, & Profession and Education. Approaches used are as follows
    a) Imputing rows with "NA" values for a few variables – One by one, after analyzing
    b) Replacing NA's with "others" where such category exits
    c) Creating separate data frame for NA values for further analysis
    d) Studying the NA values, w.r.t other variables to understand how they can be meaningfully replaced e.g. looking at age as a factor of "marital status" and "median of Qualification" etc.

Missing Values



- Application Data - Missing Values Account for less than 2%, ok to impute
- Credit Bureau Data – CC Utilization MV’s depict CC not used

Methodology - Data Understanding & Data Preparation

Explore Data [1] - Summarize, Create graphs : Construct and Format the data [2]

- **Numeric variables** (Age, Income, Outstanding Balance, No. months - Residence & Co., No Trades, Avg. CC Utilization)
  - Plotted Histograms and Box Plots [1]– To study spread for binning and outlier treatment
  - Outlier Treatment [2]– Age Capped - 97% : No. months Residence – 91% : Outstanding Balance - 97%
  - Correlation plots [1]
  - Binning [2] - Ordinal Variables (e.g. Age & Income) – Binning explored to ensure even spread
  - Scaling the variables before modeling [2]
- **Categorical variables**
  - Bar charts [1] depicting a) No. of prospects and b) Percentage defaults for each category – To Study impact on defaults
  - Creating Dummy variables to covert the data into numeric [2]

WOE & IV (using scorecard package and GLM model)

- Created a separate database with WOE and IV values for all variables
- Replaced actual values with WOE values
- Missing values and outliers automatically get replaced by WOE
- Binning in a way that extracts best IV of each bin
- Plotted the variables with WOE values
- The IV values give the importance of each variable and for each category with in the variable

Correlation: Continuous Variables

Very low correlation

a) Amongst the continuous variables

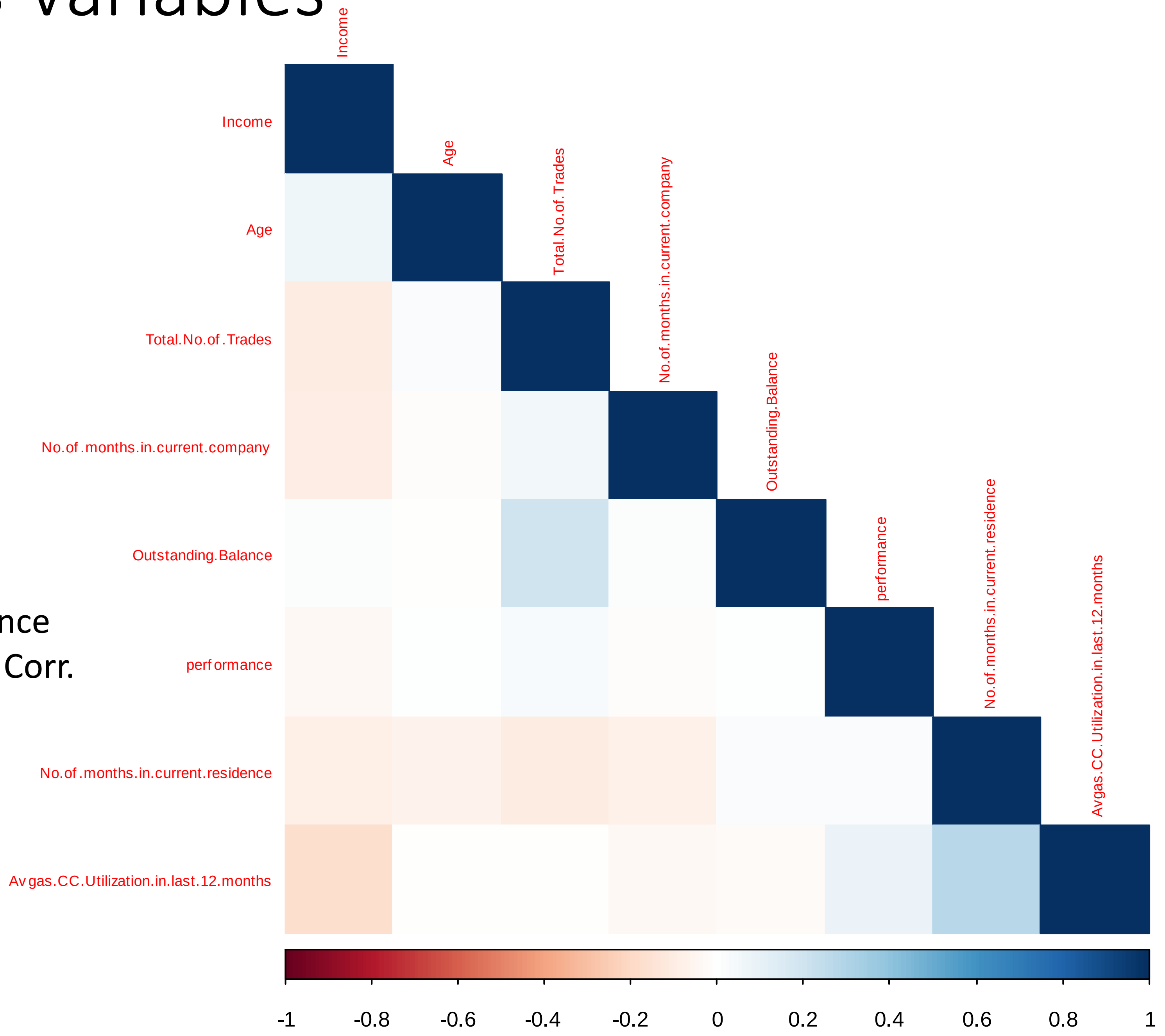
b) Target variable Vs Continuous Variables

Variables that are slightly correlated

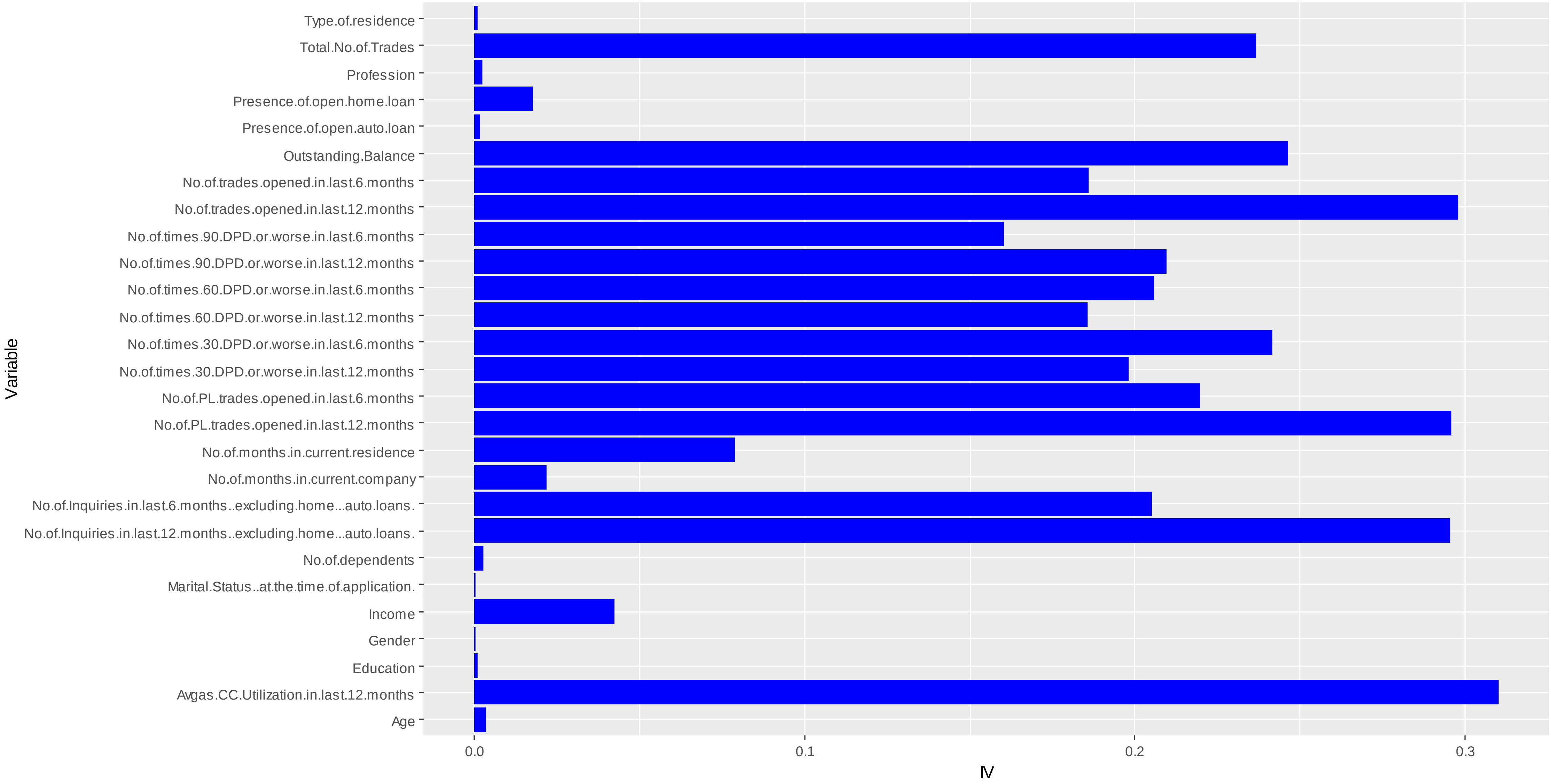
a) Outstanding Balance and Total # of trades

b) CC utilization in 12m & No. months in residence

c) Income and CC utilization in 12m – Negative Corr.



Imp. Predictor Variables @ Information Value



Important Variables : Avg. CC utilization in last 12m, No. Trades – 12 m, No. PL trades -12m, No. inquiries -12 months, Outstanding Balance, No 30 DPD in 6m, Total No. Trades, No PL trades in 6m, No. 90 DPD in 6m

Methodology – Model Building, Evaluation & Application Score Card

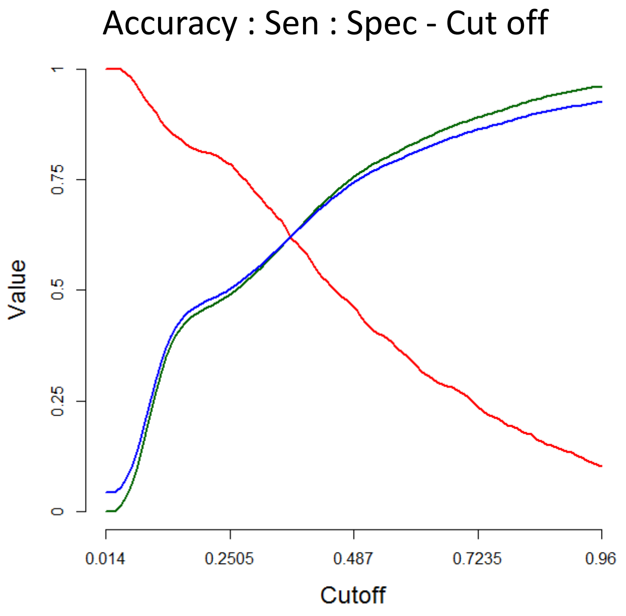
Model Building

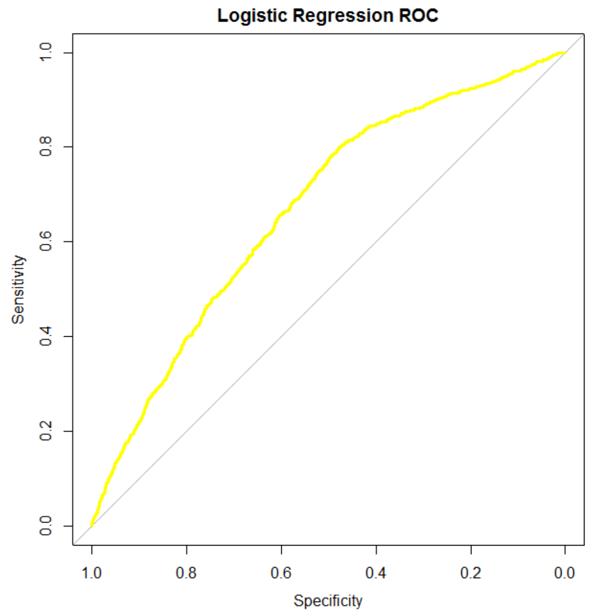
	Slide
• Model 1 – Logistic Regression on Overall data	10
• Tried different iterations with binning variables etc. : better scores with out binning	
• Model 2 – Random Forest on Overall data	11 & 12
• Model 3 - SMOTE to enhance the no. of defaults to 50% & Random Forest : No Improvement	13
• Model 4 - Logistic Regression on Overall data (WOE scores) : **Best Model**	14
• Final Application score card on the Logistic Regression (using WOE scores)	15
Testing the Application Scores on **Rejected Vs Approved** population	
• Financial Benefit and Takeaways	16

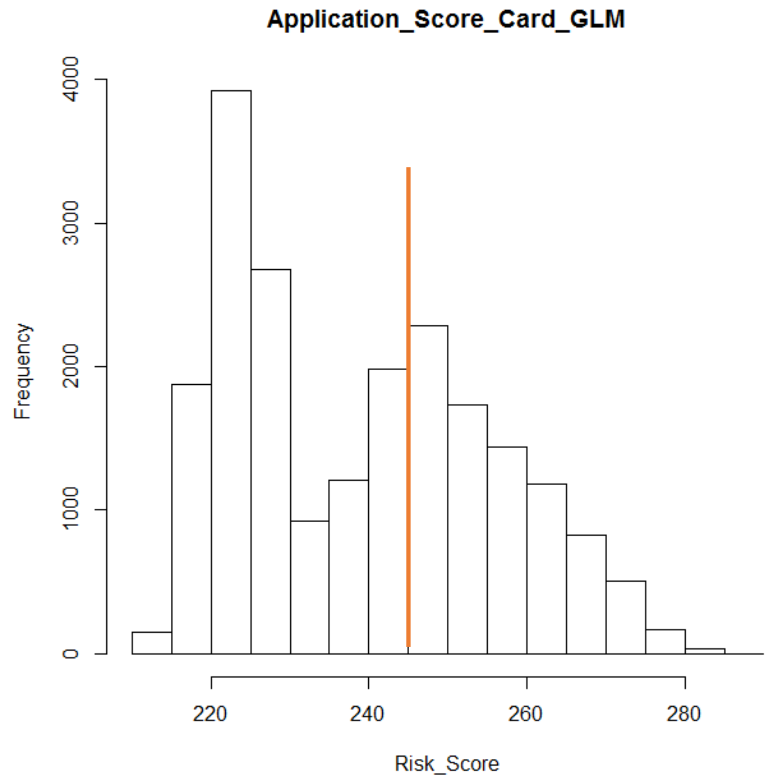
Assumptions

- Class Label 1 is considered as Default
- ~1400 customers, with missing values in performance tag have been considered as rejected populations
- Since missing values were a very small percentage of populations less than (3%), we have imputated then (except of Avg. CC utilizations, which were replaced by 0)

Model 1 - GLM





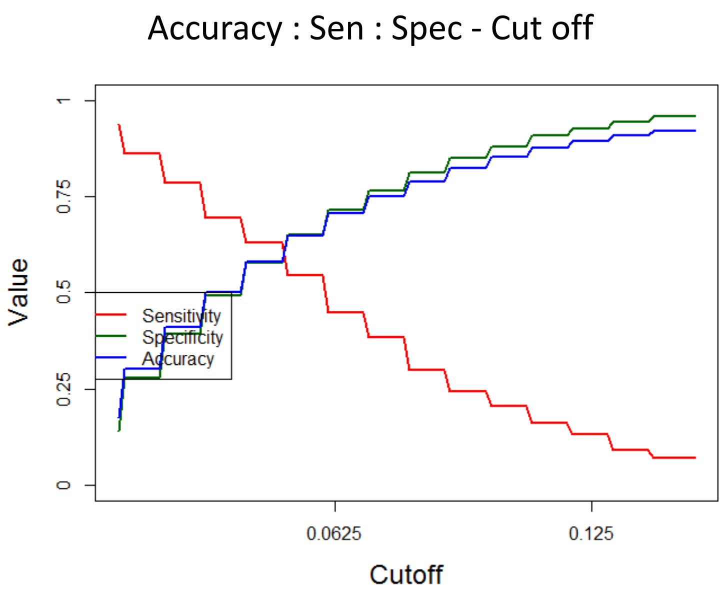


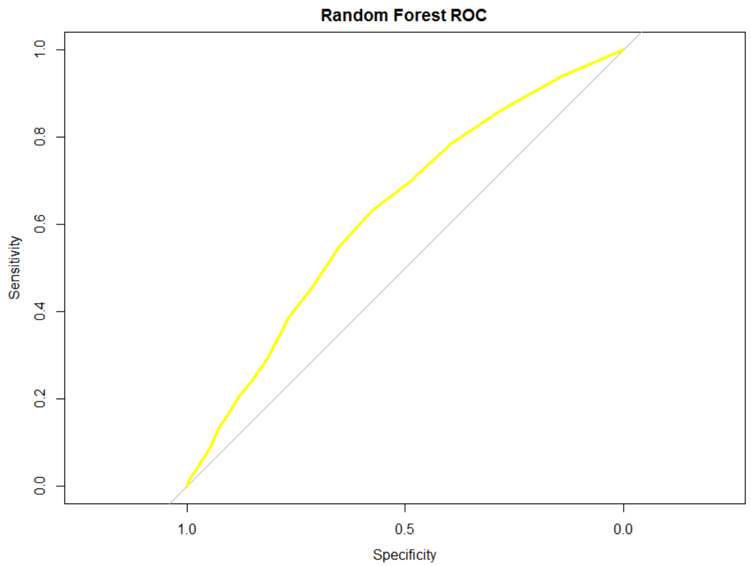
Takeaways : -

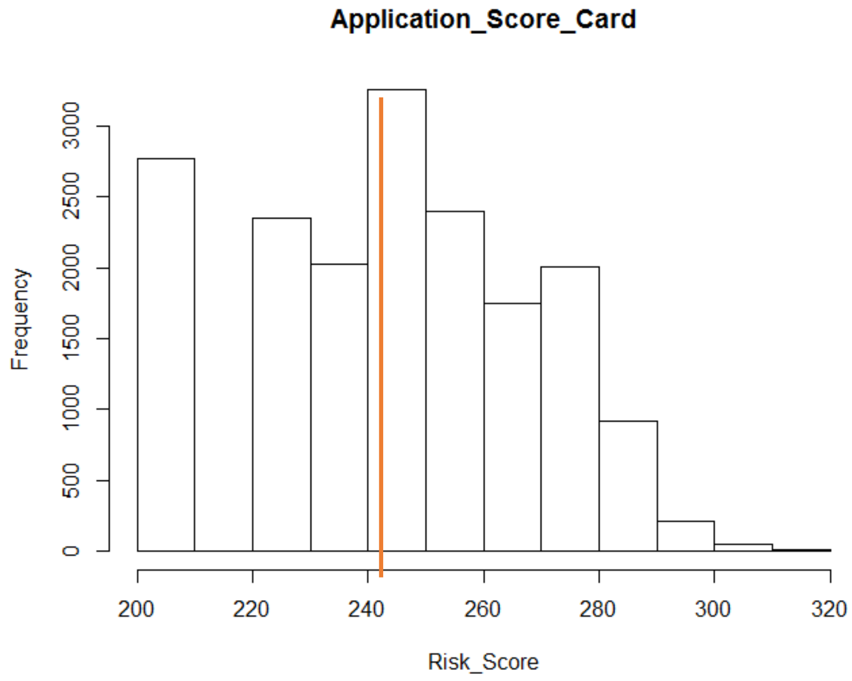
- GLM is the Best model
- At a cut off of 0.04464646 gives the
  - acc 0.6231413
  - sens 0.617214
  - spec 0.6234026
- ROC - Area under the curve : 0.6704

- Application score card
  - Range – 211.7 to 286.1
  - Cut off - 245

Model 2 – Random Forest







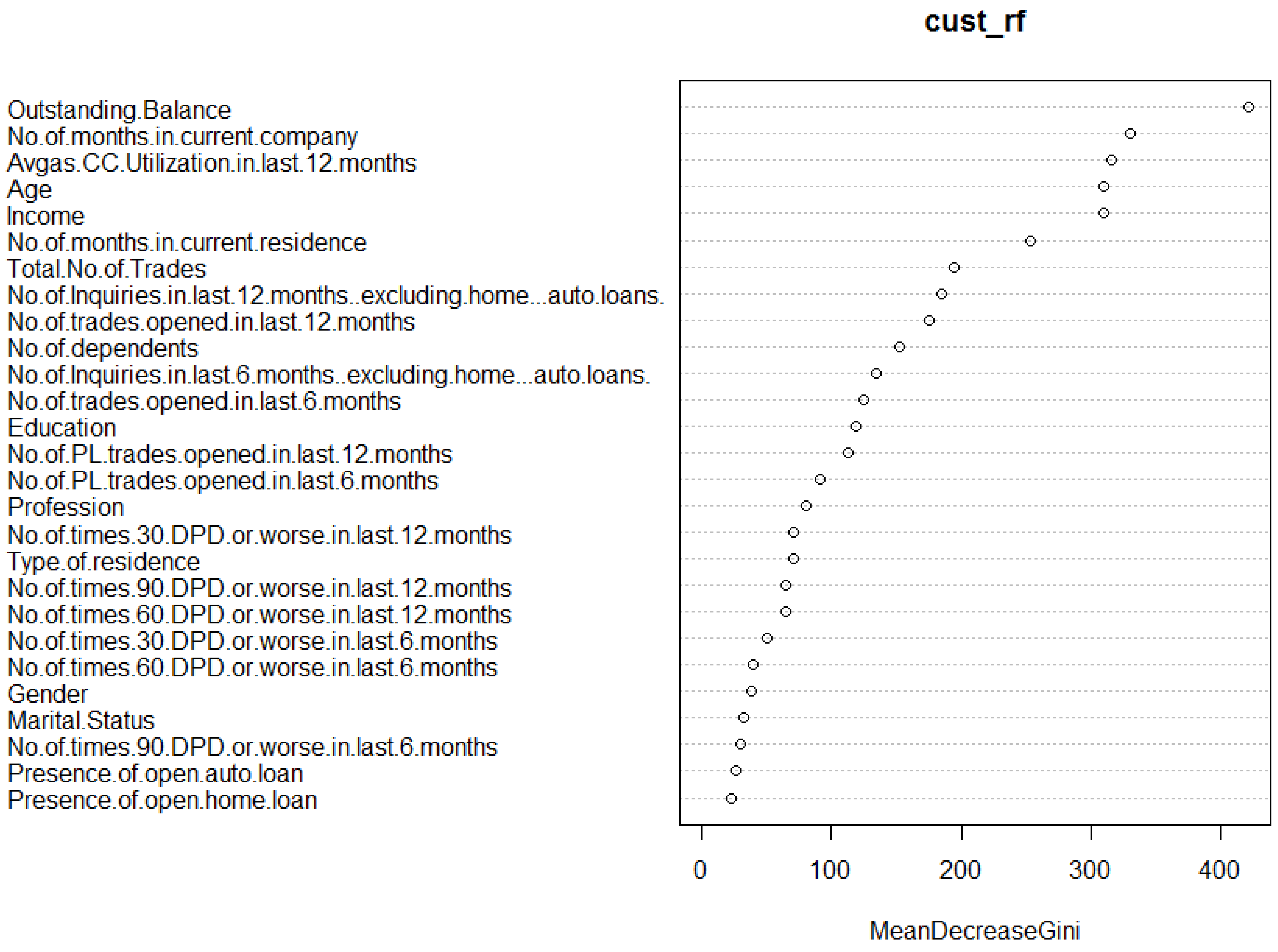
Takeaways : -

- GLM performed marginally better
- At a cut off of 0.04111111 gives the
  - acc 0.6290323
  - sens 0.5771417
  - spec 0.5793311

- ROC - Area under the curve : 0.6262

- Application score card
  - Range – 201.0 to 314.4
  - Cut off - 242

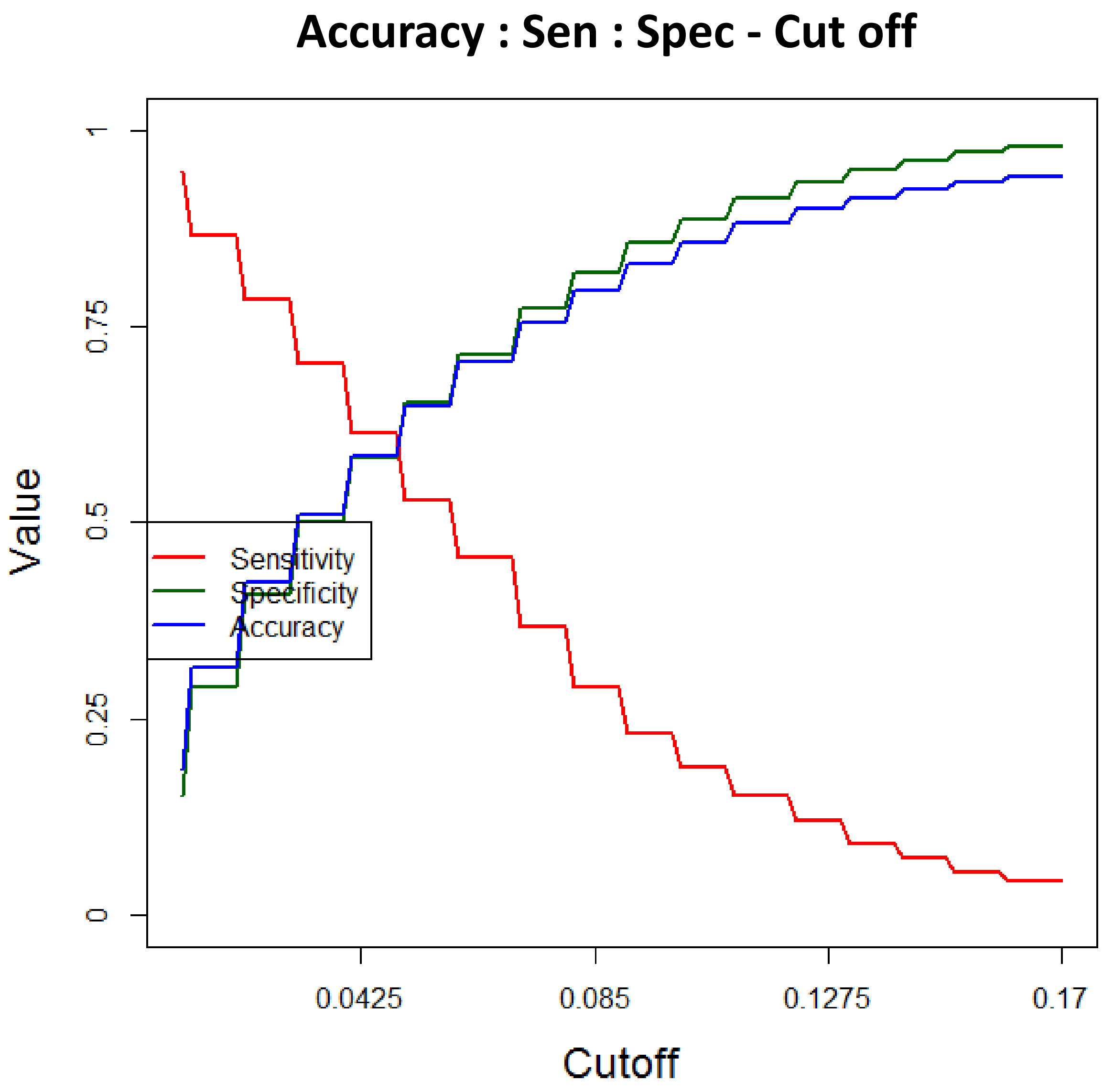
Model 2 – Random Forest – Variable Importance (Mean Decrease Accuracy)



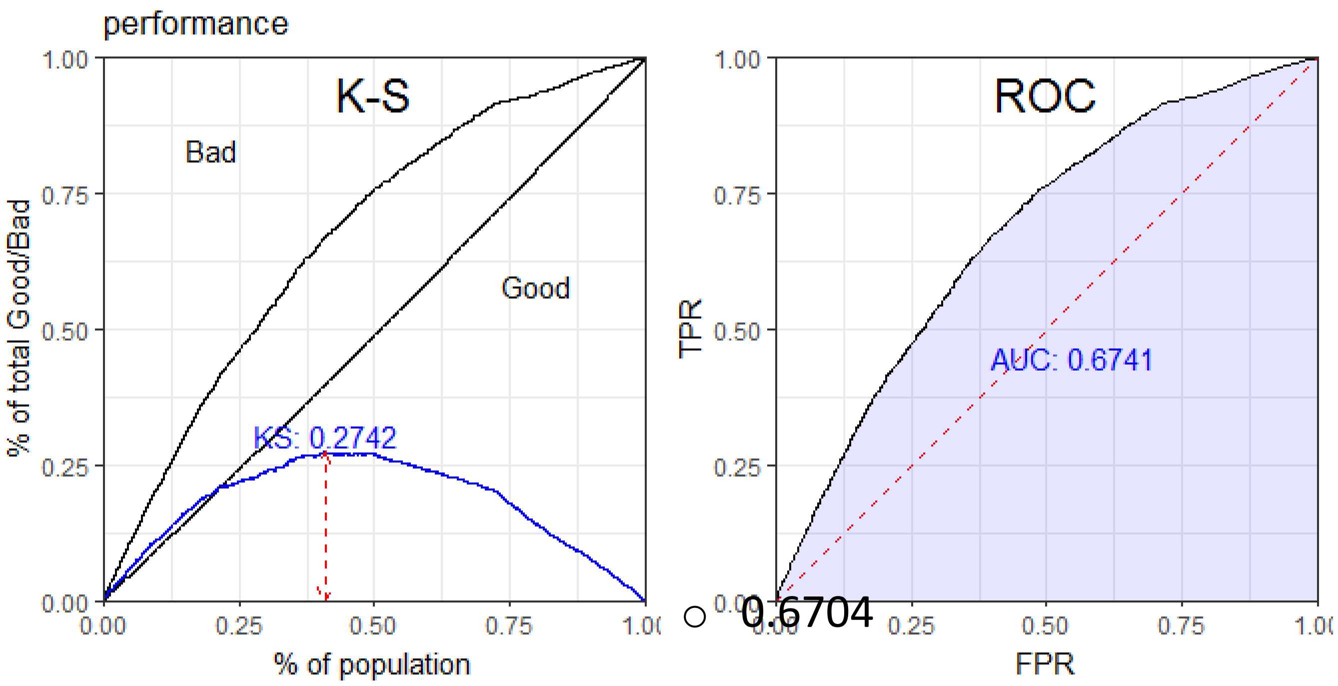
Model 3 – SMOTE + Random Forest

Takeaways : -

- SMOTE did not improve model performance
- Results very similar to Random forest with out SMOTE
- At a cut off of 0.04070707 gives the
  - acc 0.5720158
  - sens 0.6198157
  - spec 0.5720158



Model 4 – Logistic Regression (WOE data) & Application Score Card



○ 0.6704

Take aways :-

This is by far the best model

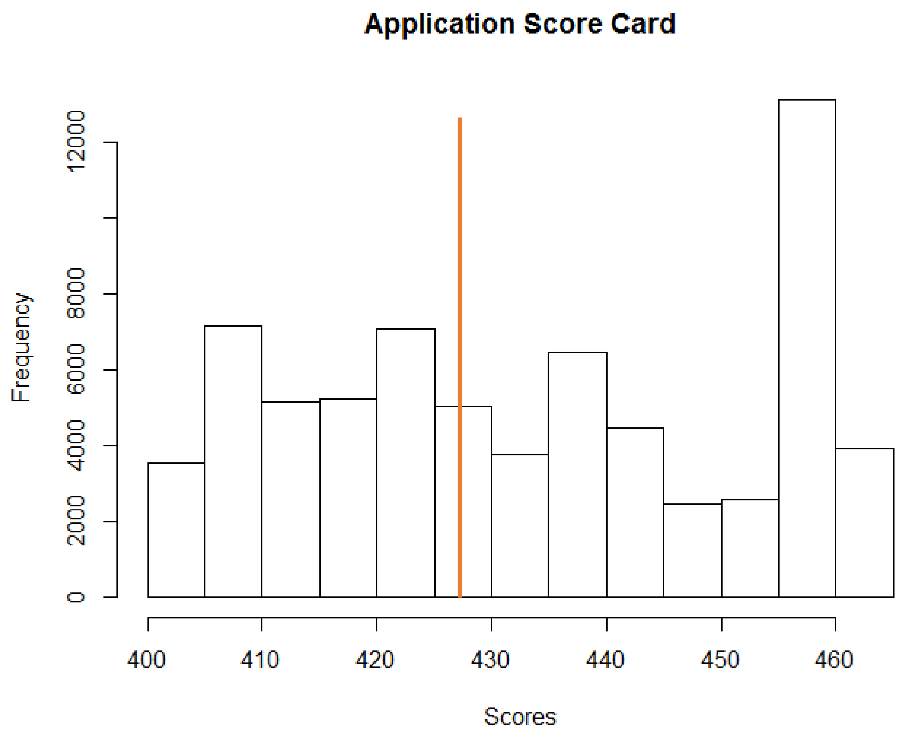
KS 0.2742 :

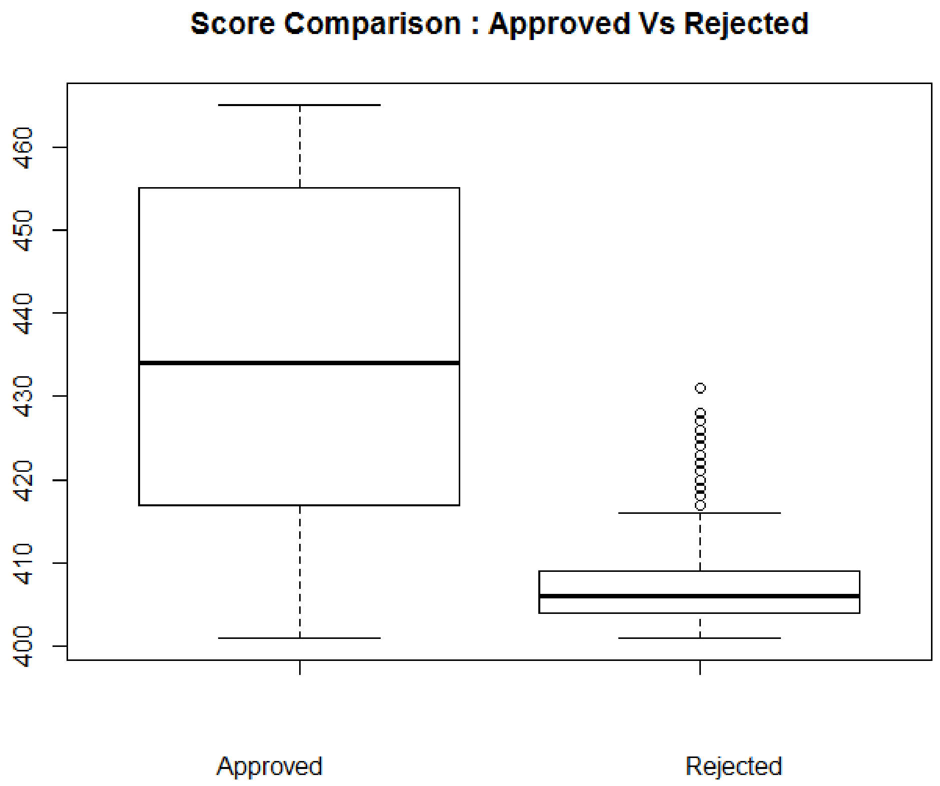
Gini 0.3384 :

AUC 0.6741 (vs a AUC of 0.6704 of GLM on normal data)

Application Score Card – Next Slide

Final – Application Score Card (Logistic Regression on WOE & IV)





Takeaways : -

- The score card is well distributed across population
- Recommendation – Cut off of 425 : All values between 420 and 425 to be reviewed by Underwriter / Decision maker

- The score card clearly distinguishes between the Approved and Rejected Customers
- The model and score card are performing well in terms of distinguishing the Approved and rejected population

Financial Benefit : Model and Score Card

Takeaways : -

- Application Risk Score has bins spread evenly across – Recommended Cut off of 425 : All values between 420 and 425 to be reviewed by Underwriter / Decision maker and identify important variables
- The important predictors are as follows
  1. No.of.times.90.DPD.or.worse.in.last.12.months
  2. No.of.times.30.DPD.or.worse.in.last.12.months
  3. Avgas.CC.Utilization.in.last.12.m
  4. No.of.PL.trades.opened.in.last.12.months
  5. No.of.Inquiries.in.last.12.months..excluding.home...auto loans.
  6. Outstanding Balance
  7. Income
  8. No of months in current company

Financial Benefit :-

1. If model is in place vs current reality (4% default rate) : we will be able to reduce the default rate by 62%
2. We can grant credit to 15% customers whom we are currently rejecting
3. The model is balanced and the scores are easy to comprehend, making it easier to understand and implement across the Decision Makers
4. The Model will automatically approve 62% application and reject 37%. Thereby significantly improving the speed and accuracy of the transactions
5. We recommend a manual over ride for 5% application below the cut off

Assumptions :-

1. Class Label 1 is considered as Default
2. ~1400 customers, with missing values in performance tag have been considered as rejected populations
3. Since missing values were a very small percentage of populations less than (3%), we have imputed then (except of Avg. CC utilizations, which were replaced by 0)

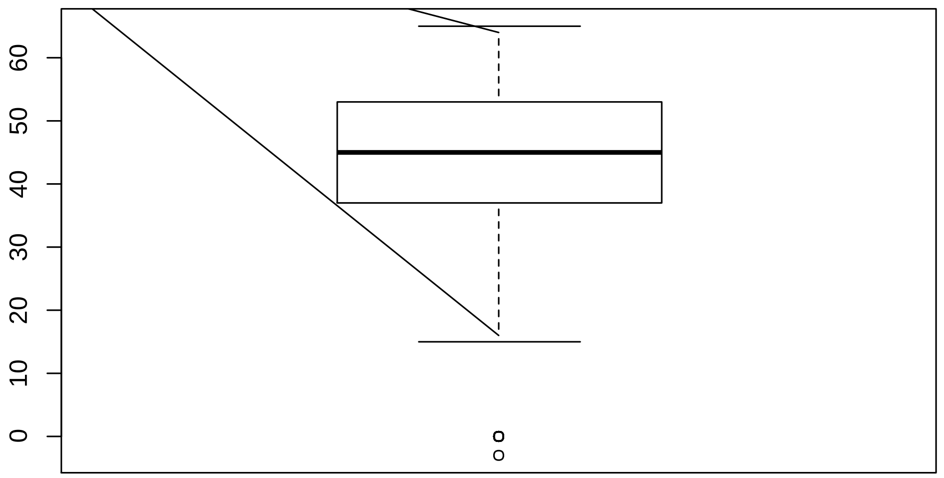
Annexure

Data Dictionary

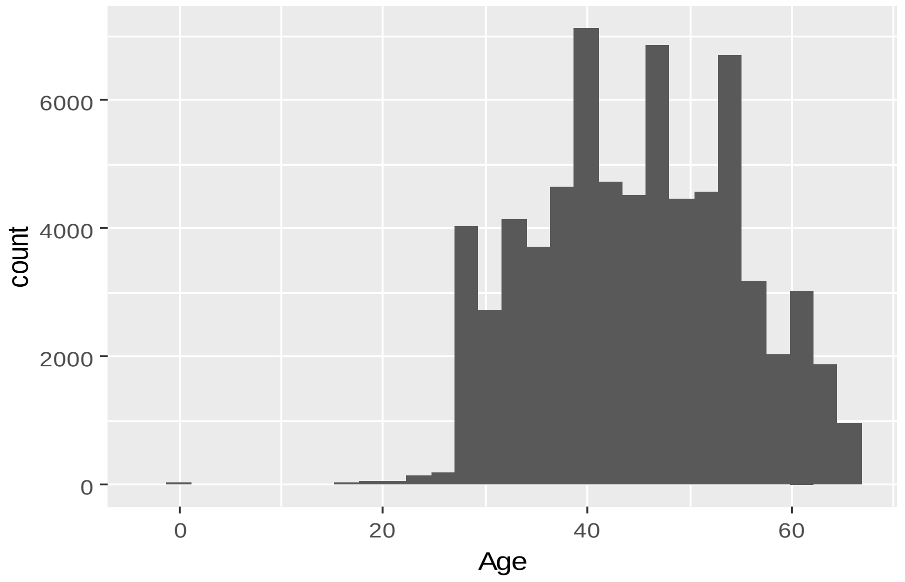
Variable	Meaning
Application.ID	Unique ID of the customers after merging both the datasets
No.of.times.90.DPD.or.worse.in.last.6.months	Number of times customer has not payed dues since 90days in last 6 months
No.of.times.60.DPD.or.worse.in.last.6.months	Number of times customer has not payed dues since 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")
Age	Age of customer
Gender	Gender of customer
Marital.Status	Marital status of customer (at the time of application)
No.of.dependents	No. of direct dependents 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	Status of customer performance (" 1 represents "Default") after merging both the datasets
binning.age	Binned the Age of customer in multiple ranges
binning.income	Binned the Age of customer in multiple bands

Continuous - Age

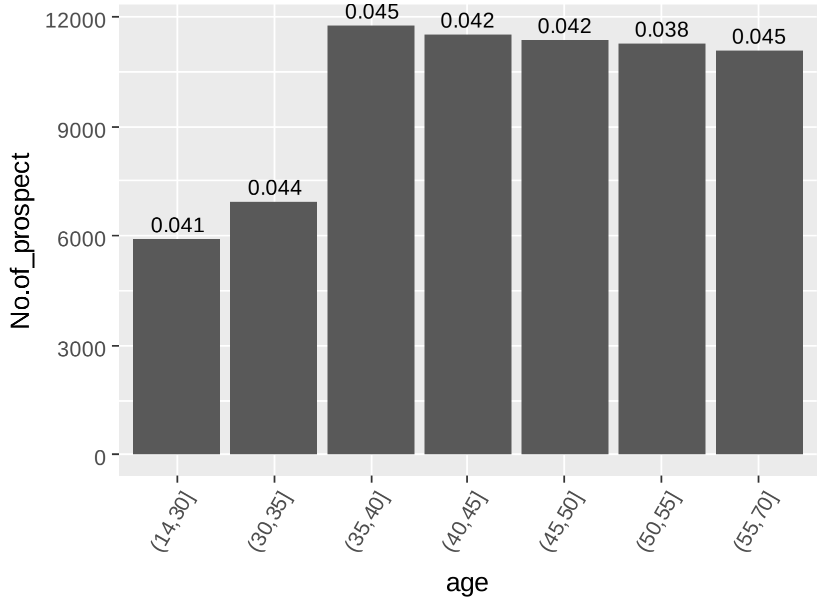
Box Plot



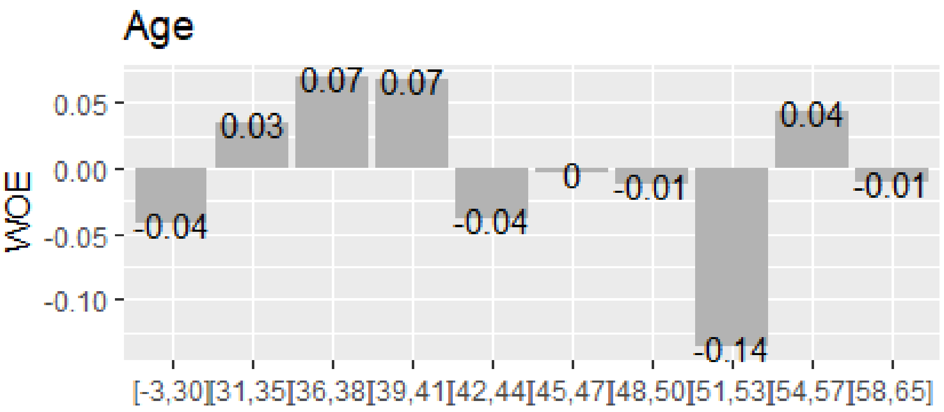
Histogram



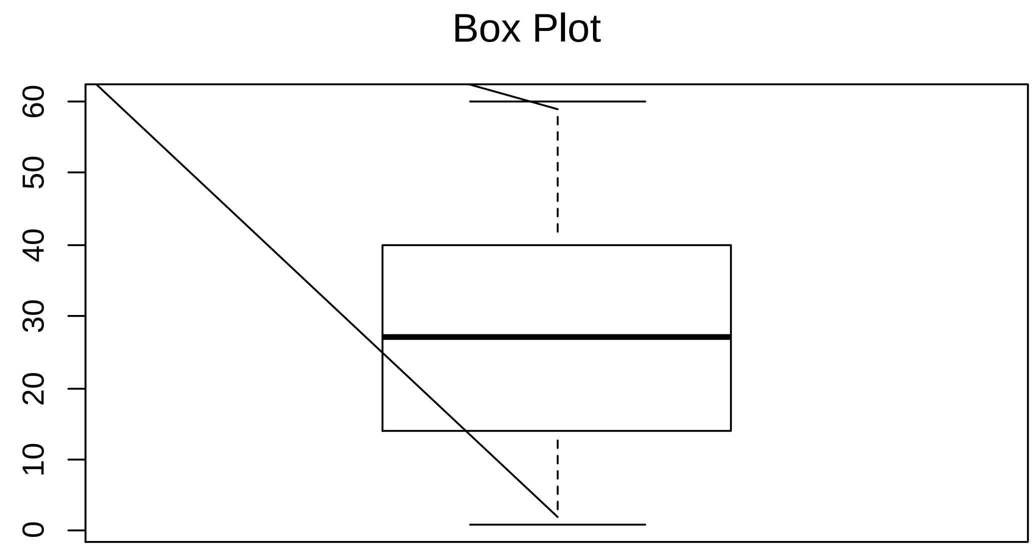
Bar Chart

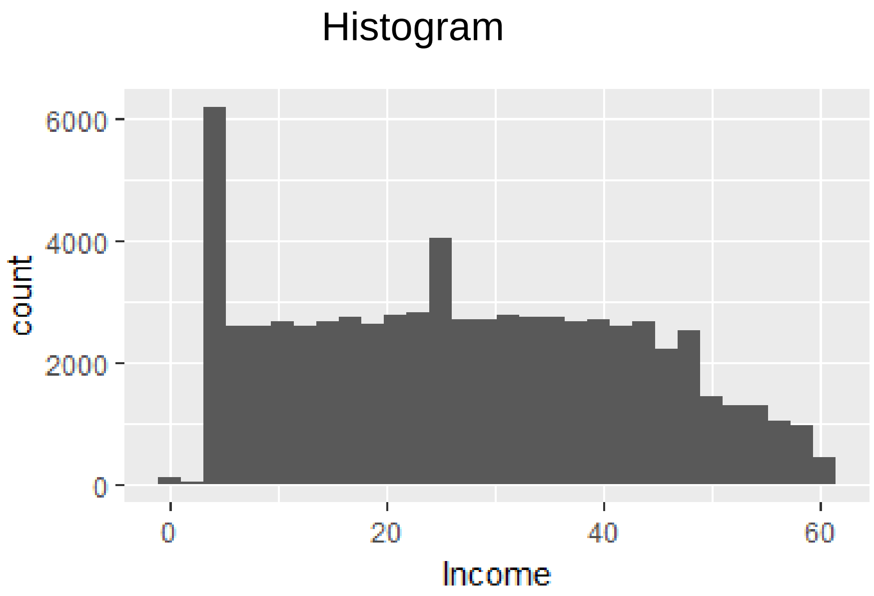


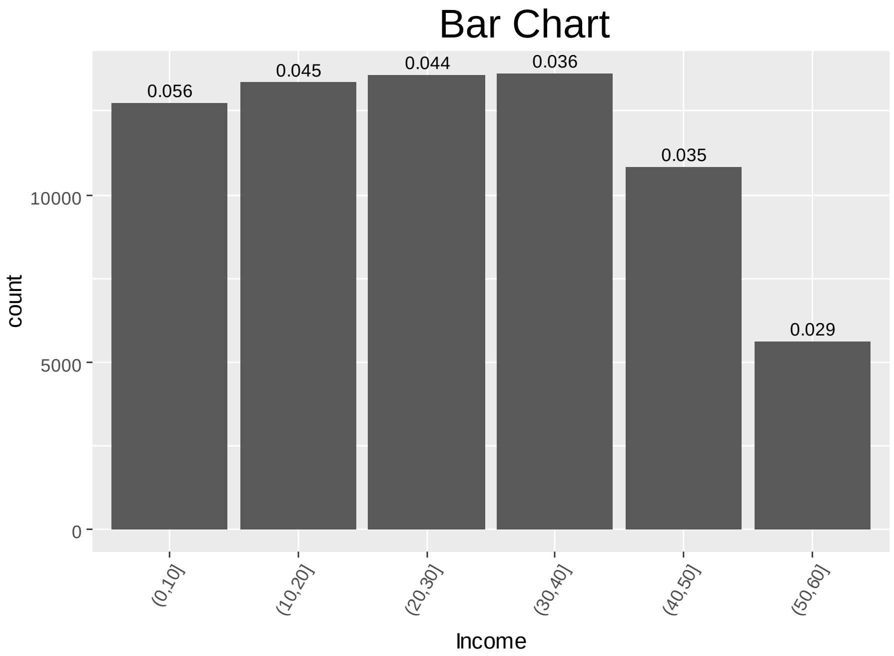
WOE

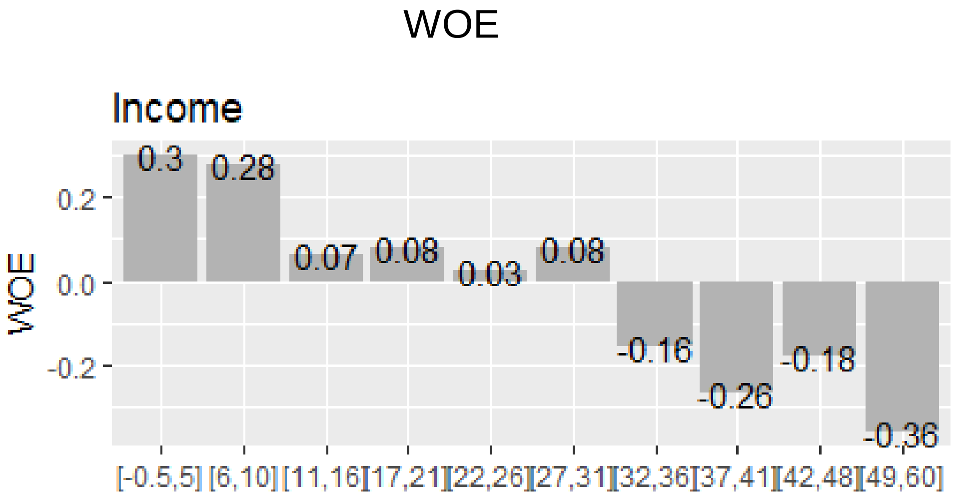


Continuous - Income



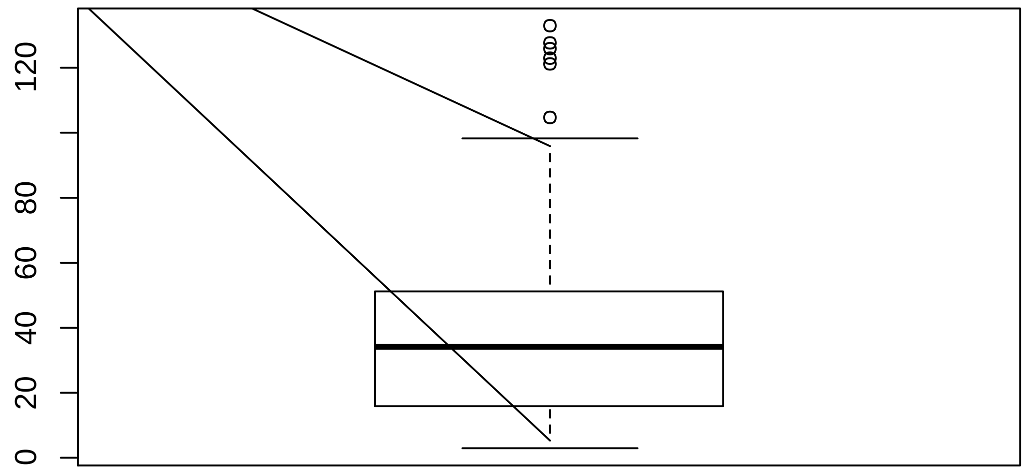




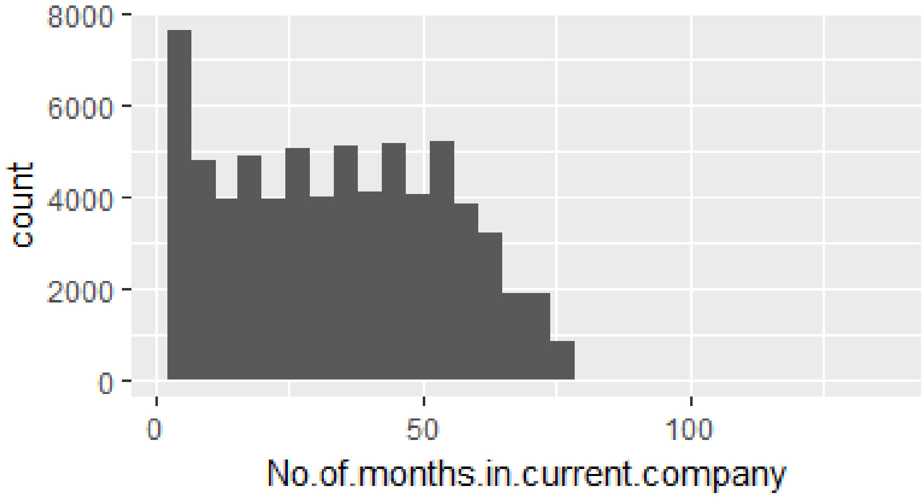


Continuous – No of months in current company

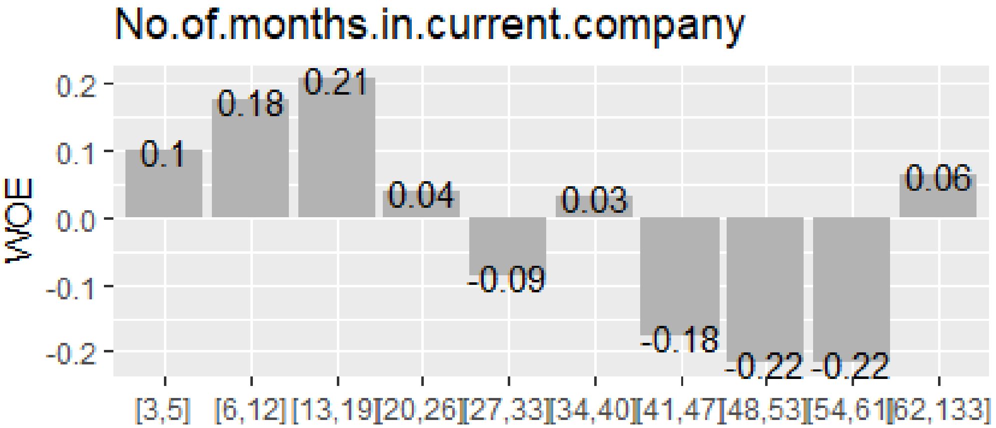
Box Plot



Histogram

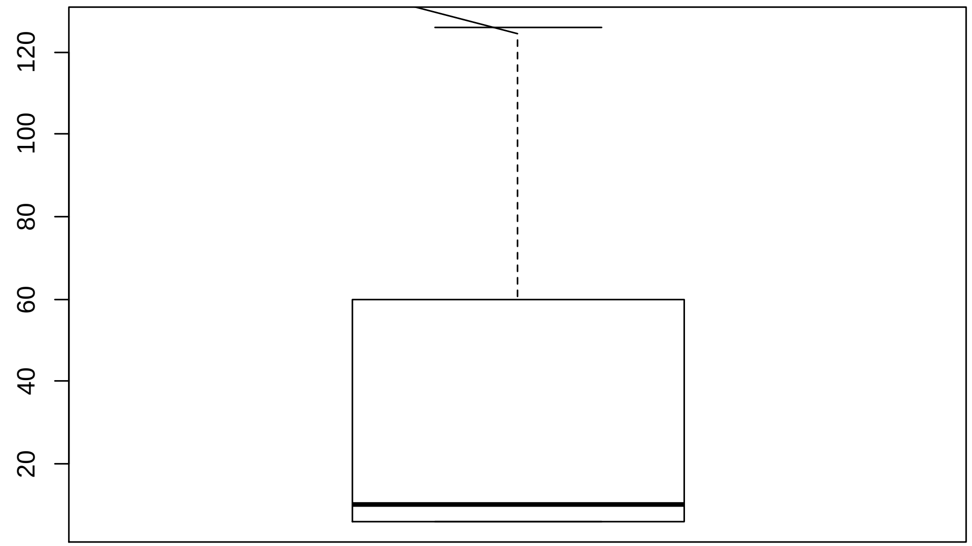


WOE

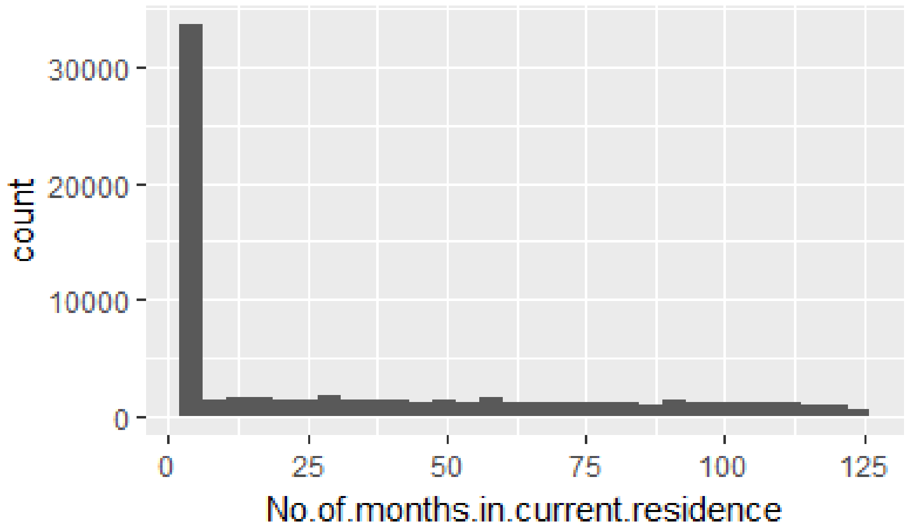


Continuous – No of months in current residence

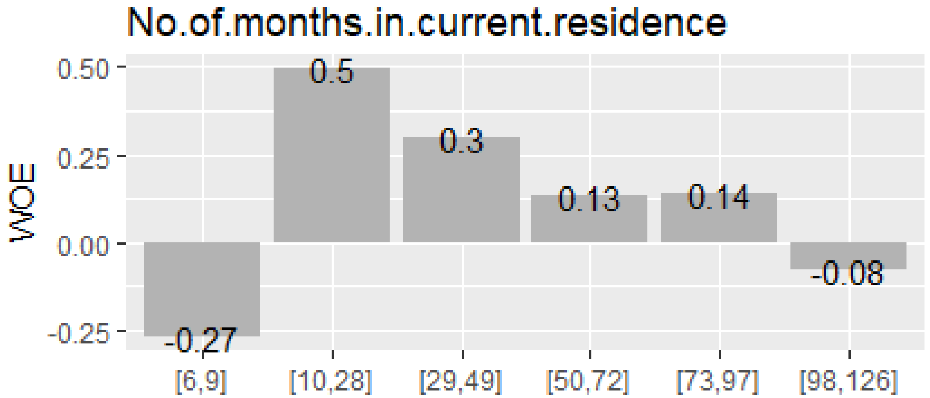
Box Plot



Histogram

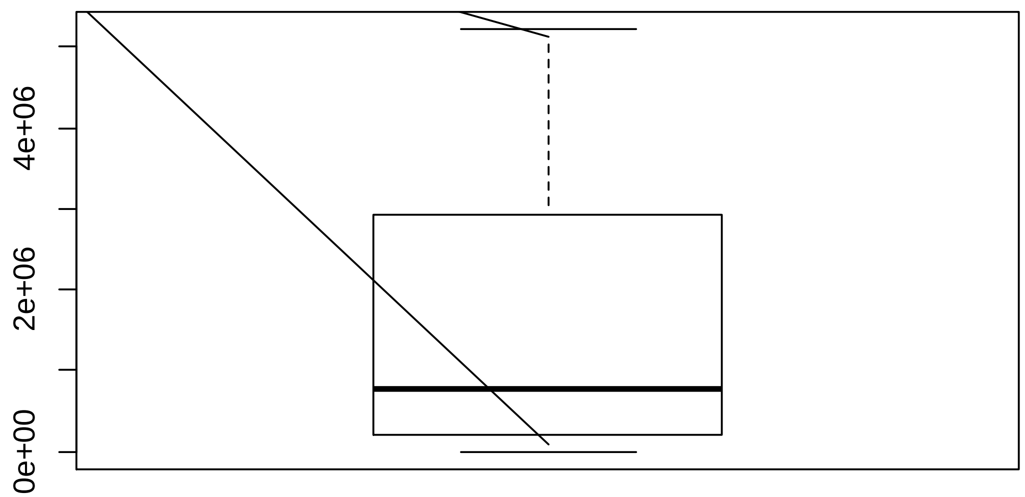


WOE

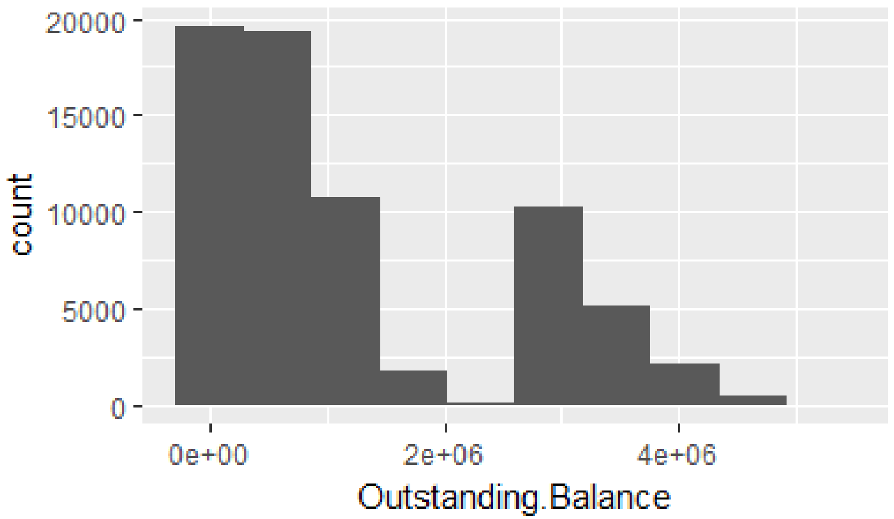


Continuous – Outstanding Balance

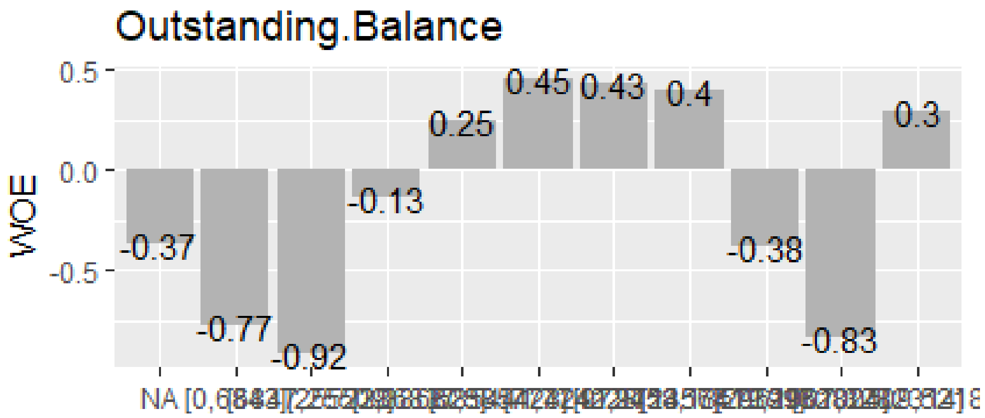
Box Plot



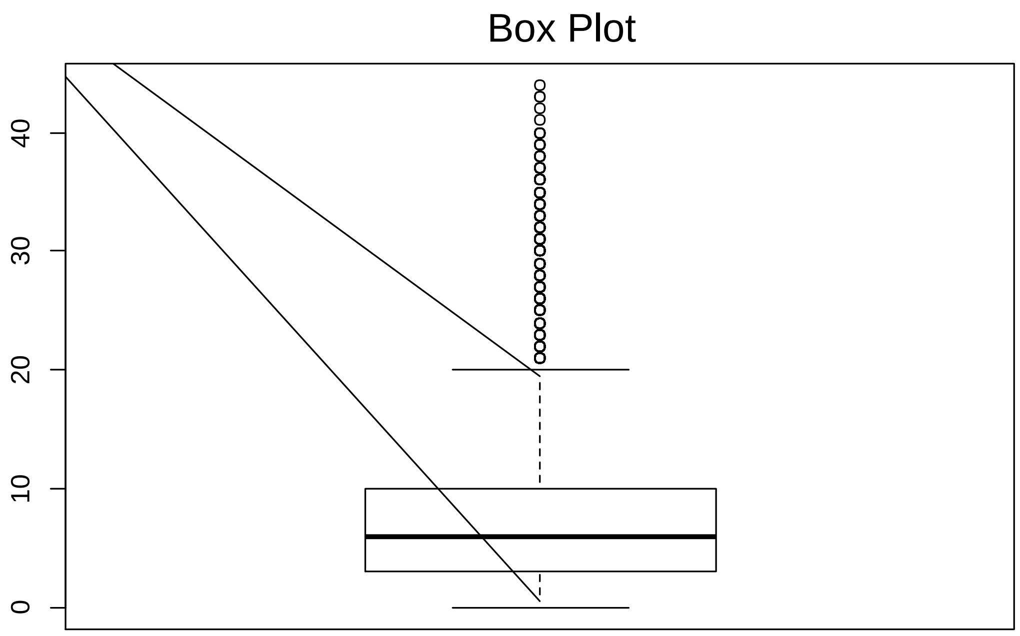
Histogram

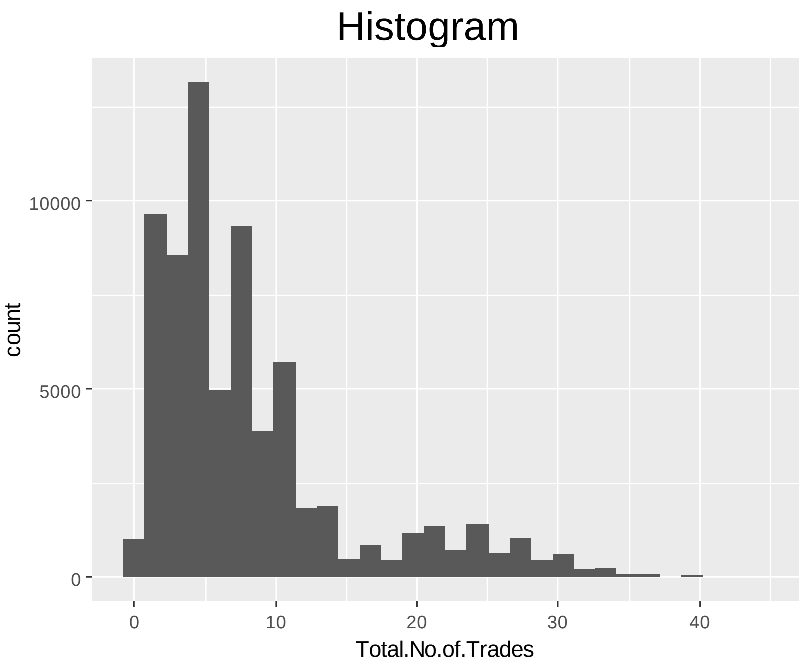


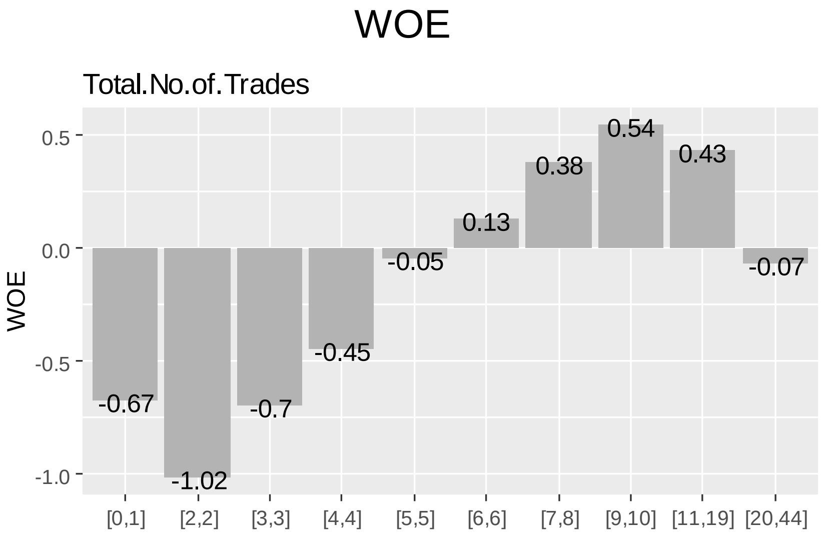
WOE



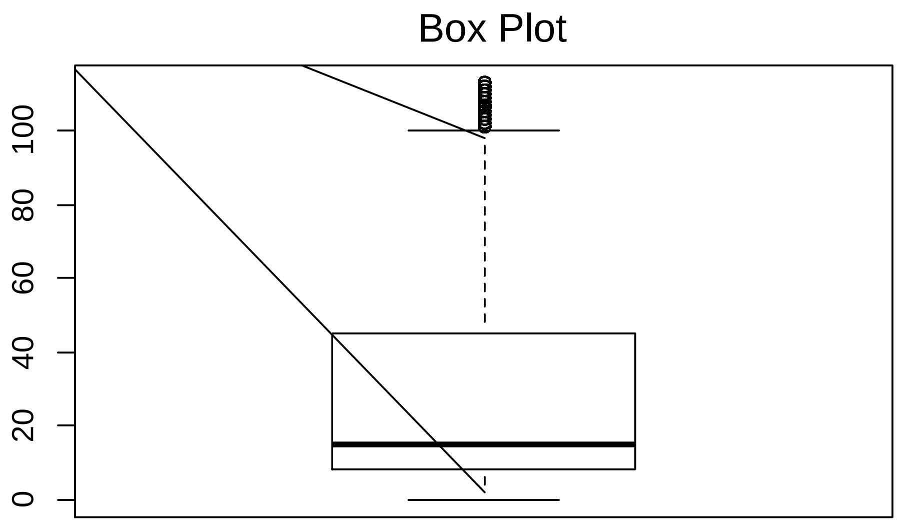
Continuous – Total No of Trades

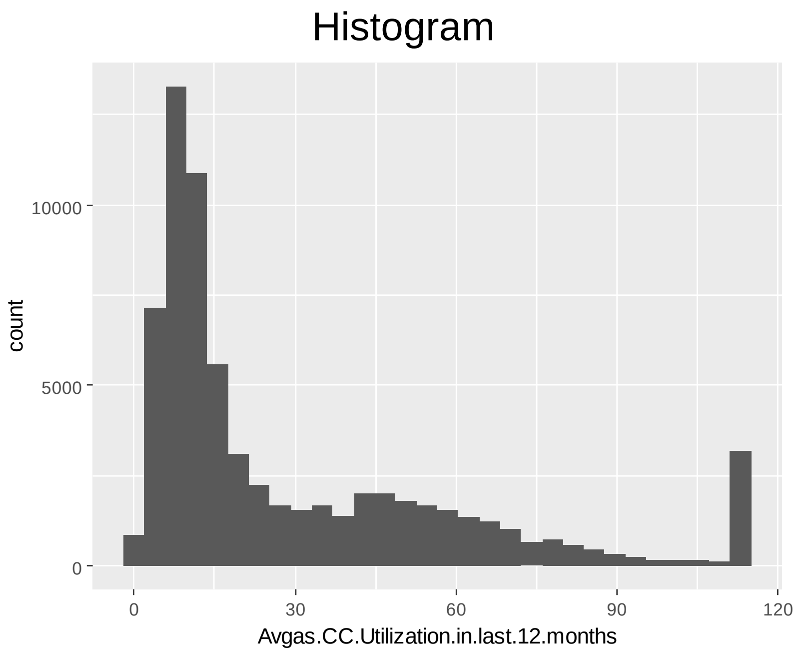


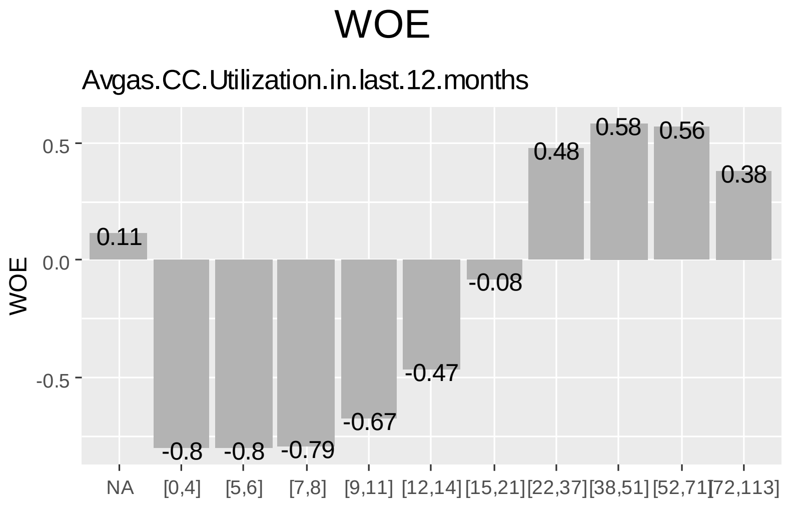




Continuous – Avg. CC Utilisation

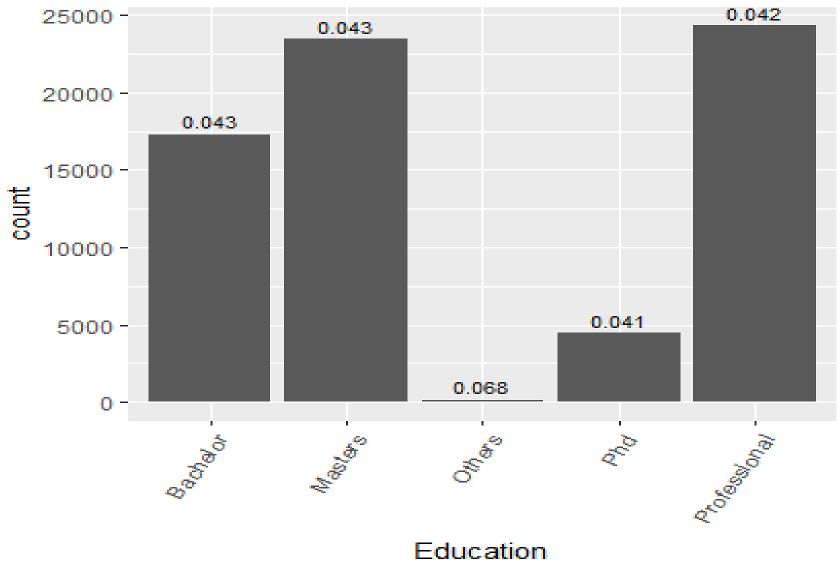




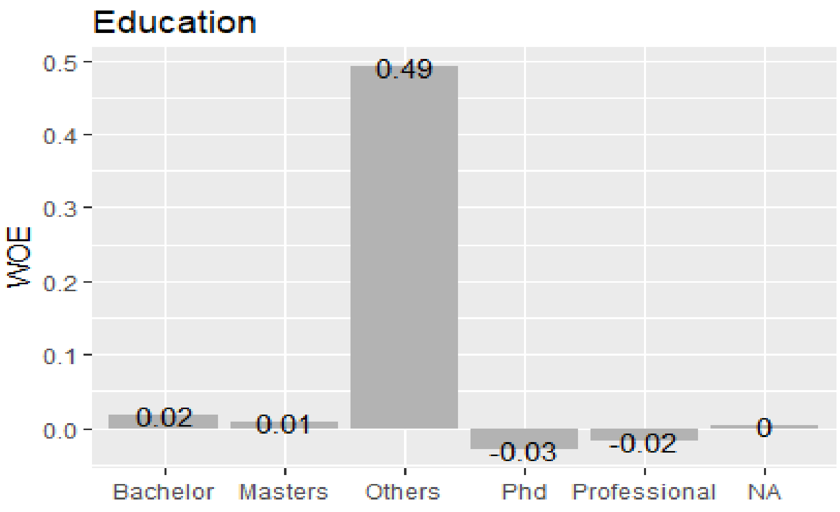


Categorical – Education & Gender

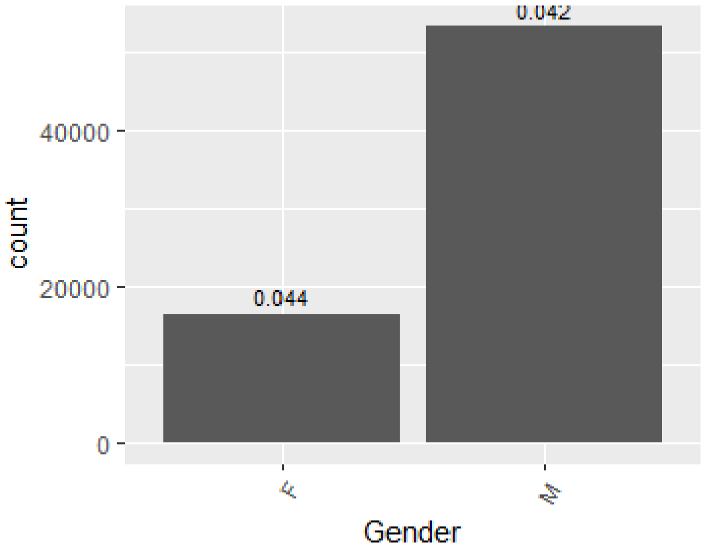
Bar Chart - Education



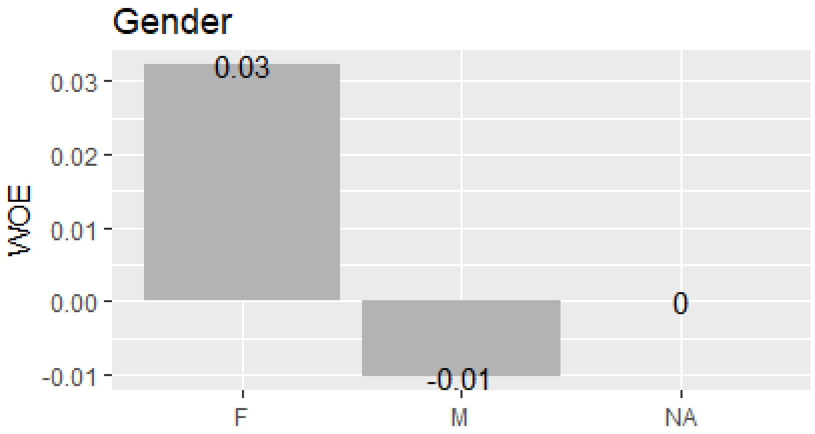
WOE- Education



Bar Chart - Gender

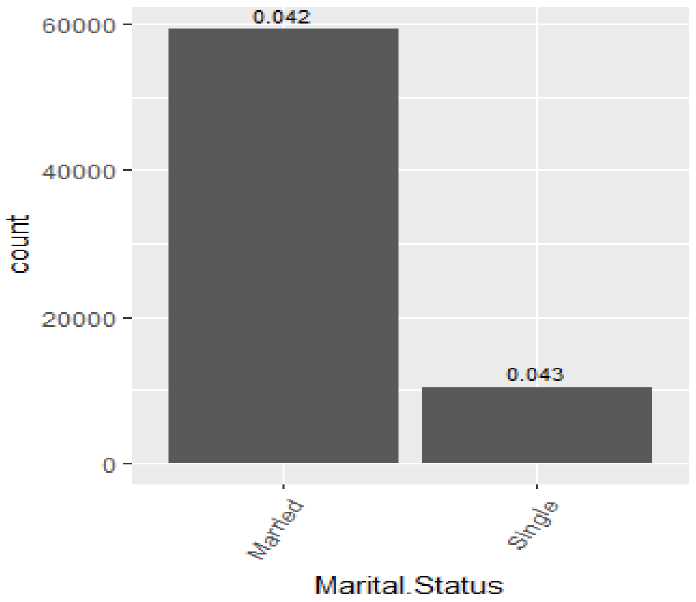


WOE- Gender

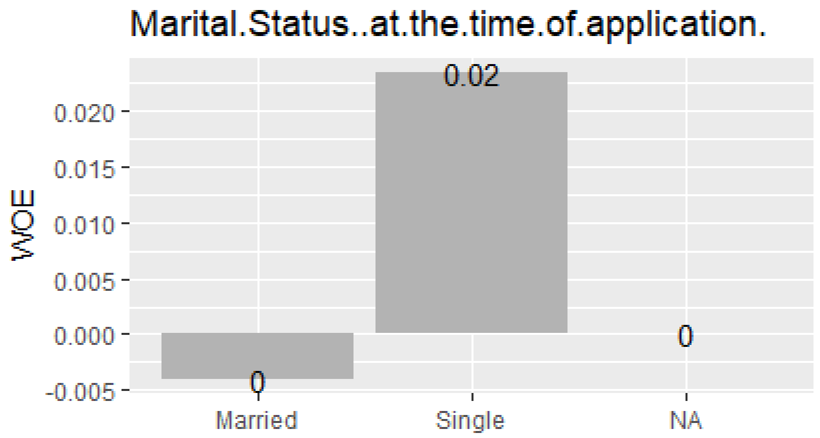


Categorical – Marital & No of dependents

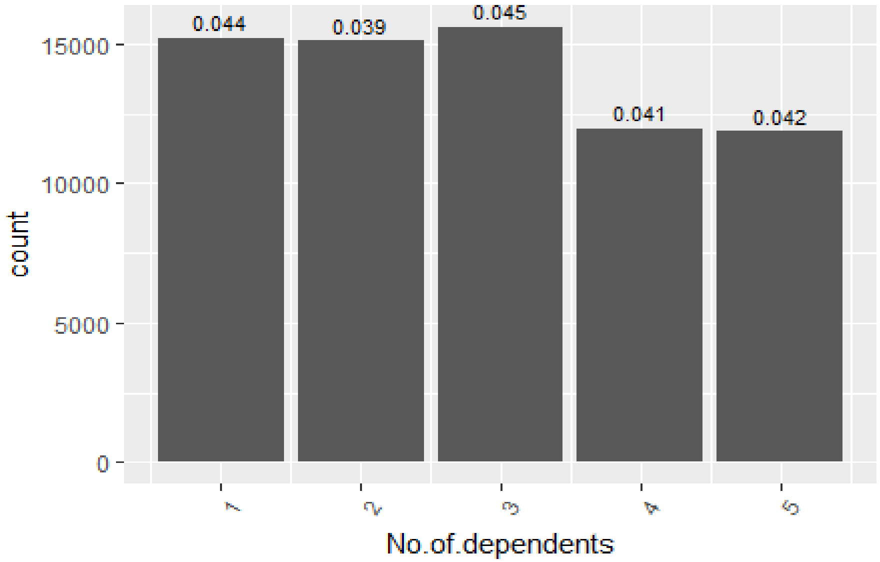
Bar Chart – Marital Status



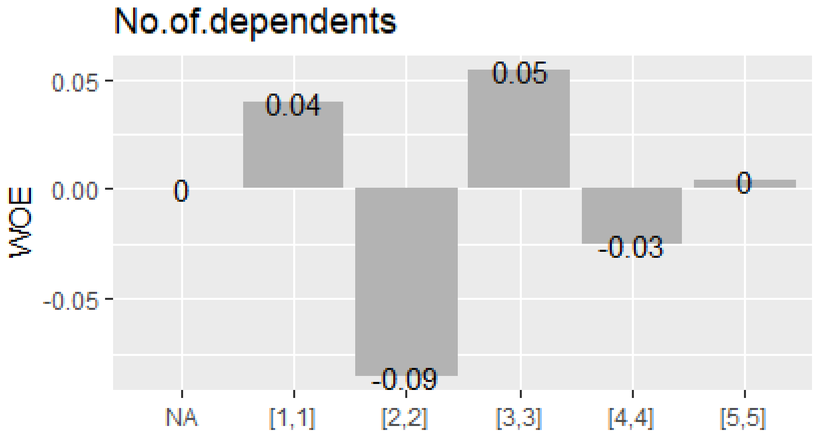
WOE- Marital Status



Bar Chart – No of dependents

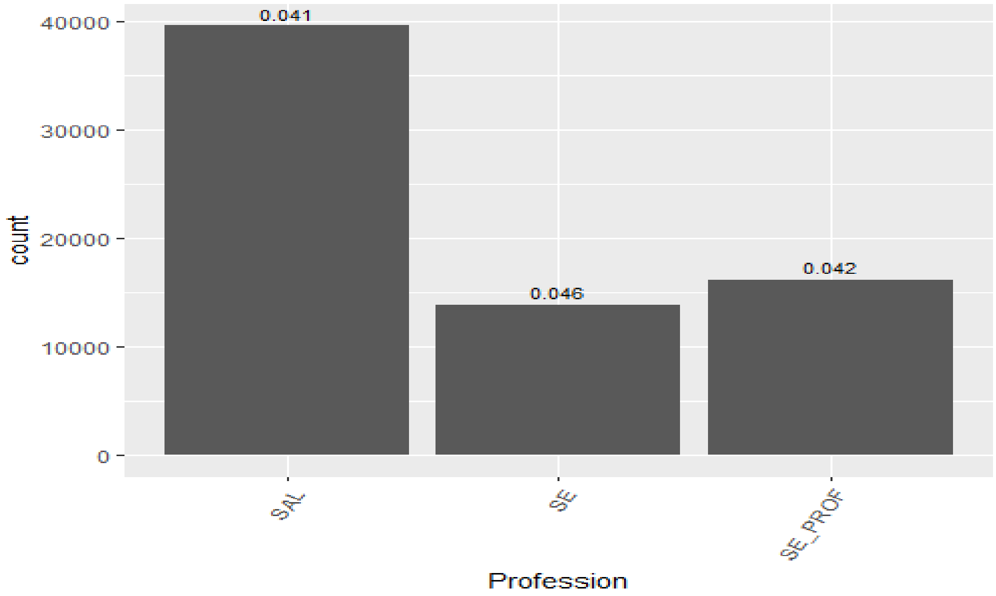


WOE- No of dependents

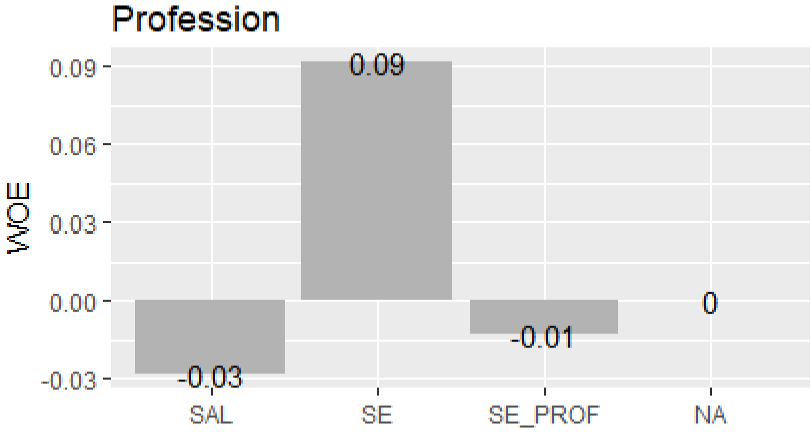


Categorical – Profession & Type of Residence

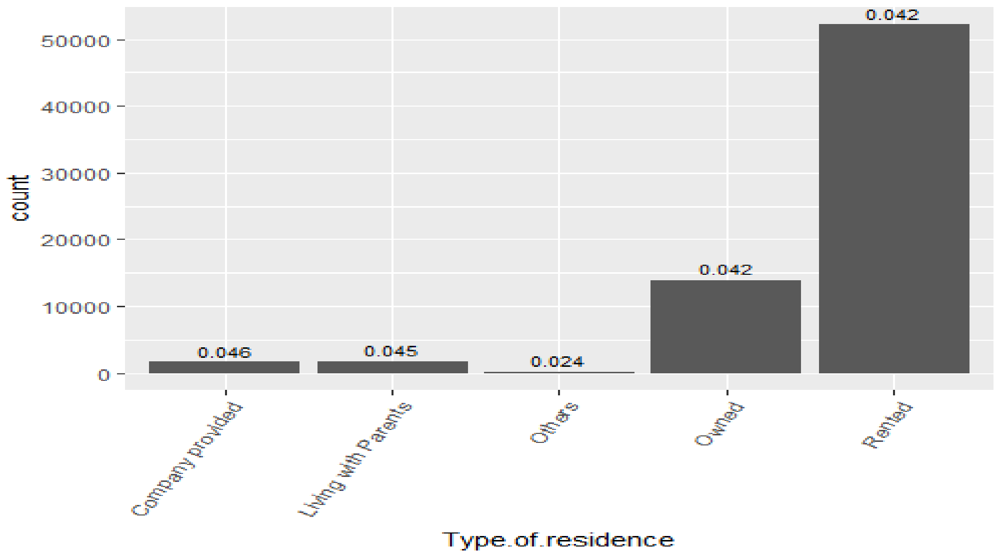
Bar Chart – Profession



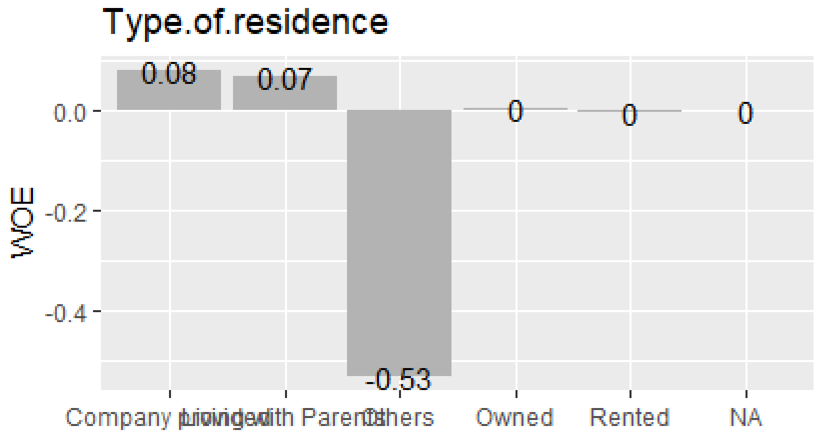
WOE- Profession



Bar Chart – Type of residence

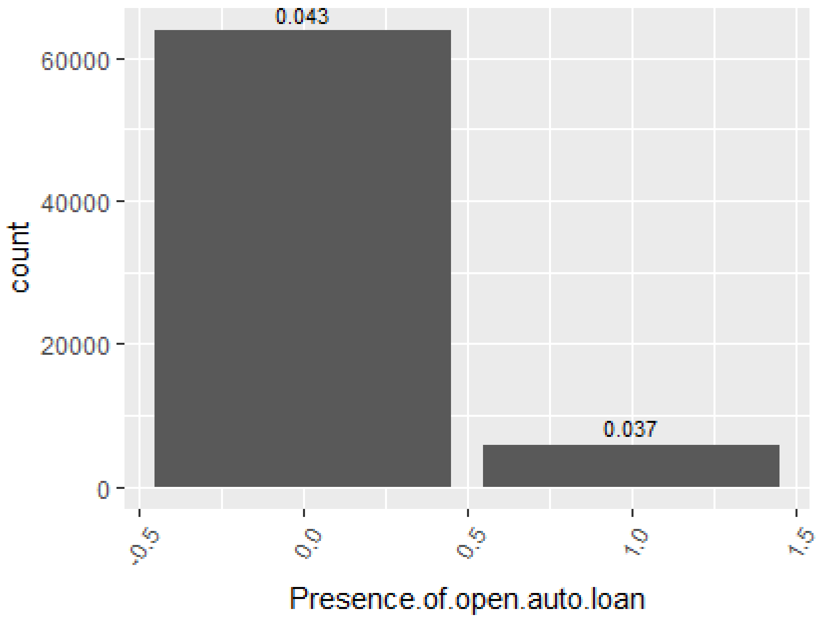


WOE- Type of residence

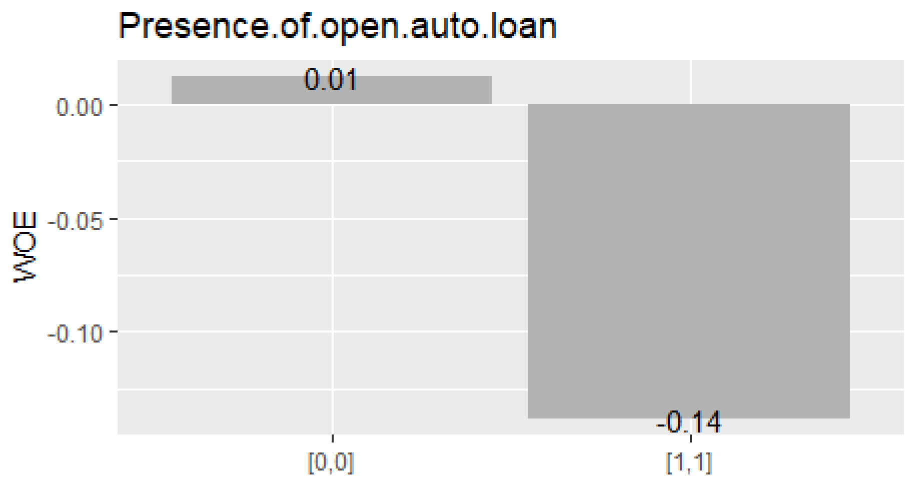


Categorical – Presence of Open Auto Loan & No of times 30 DPD in 12m

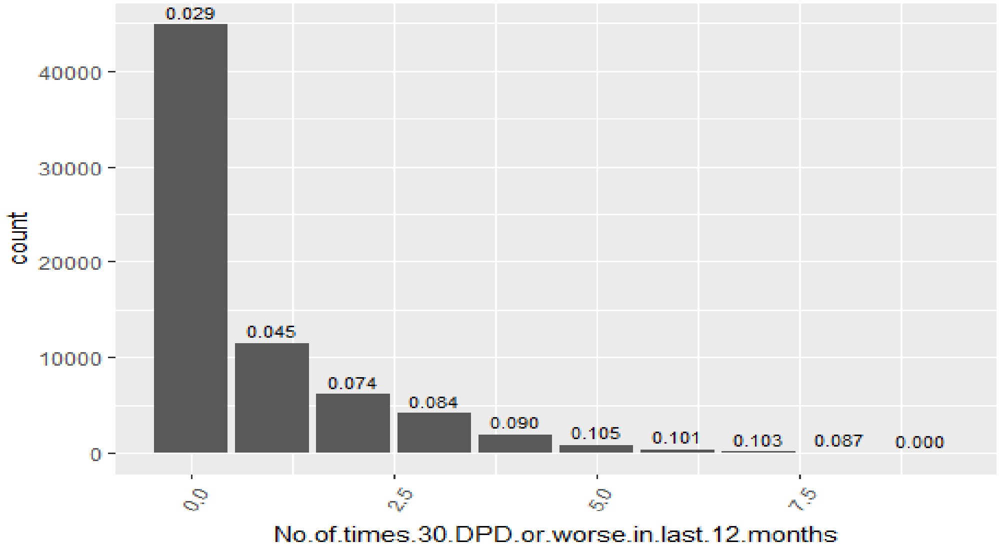
Bar Chart – Presence.of.open.auto.loan



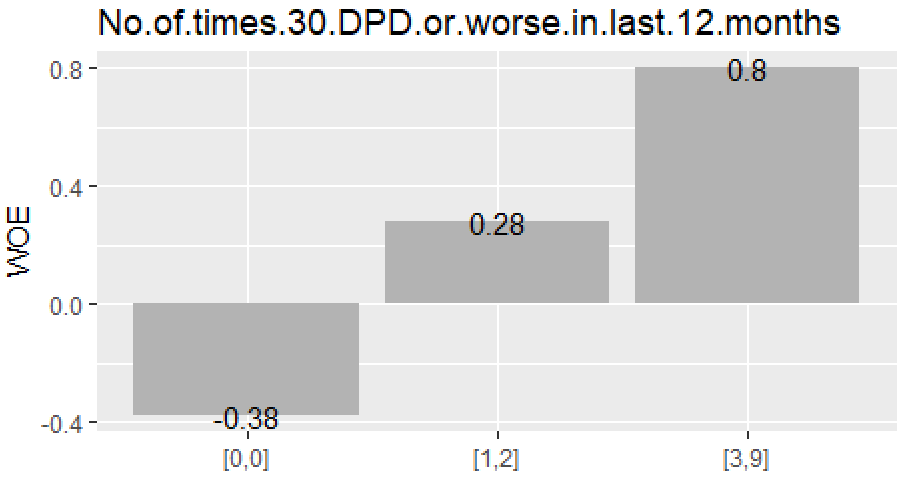
WOE- Presence.of.open.auto.loan



Bar Chart – No.of.times.30.DPD.or.worse.in.last.12.months

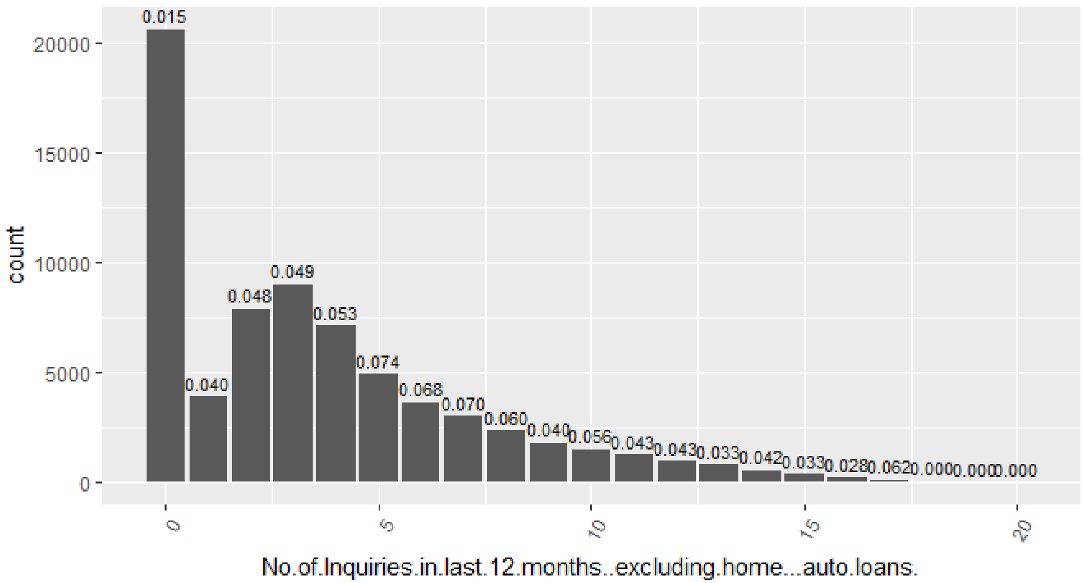


WOE- No.of.times.30.DPD.or.worse.in.last.12.months

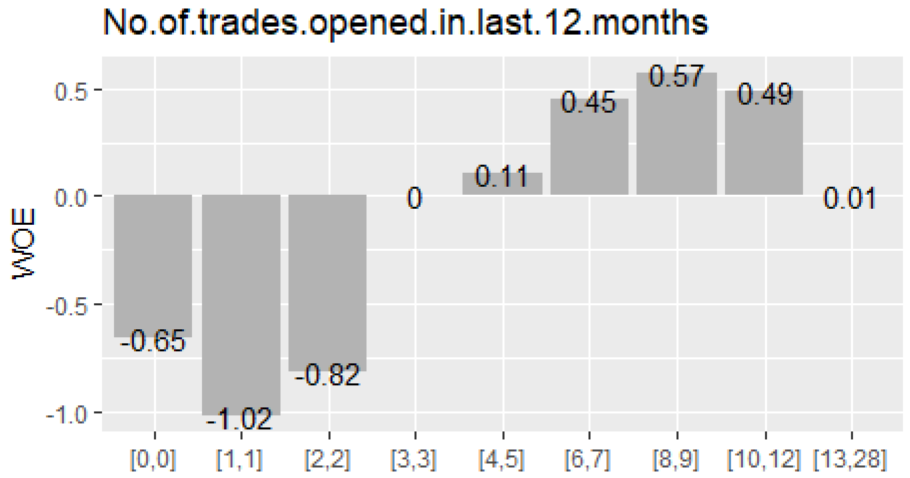


Categorical – No of Inquiries in 12m & No of Trades in 6m

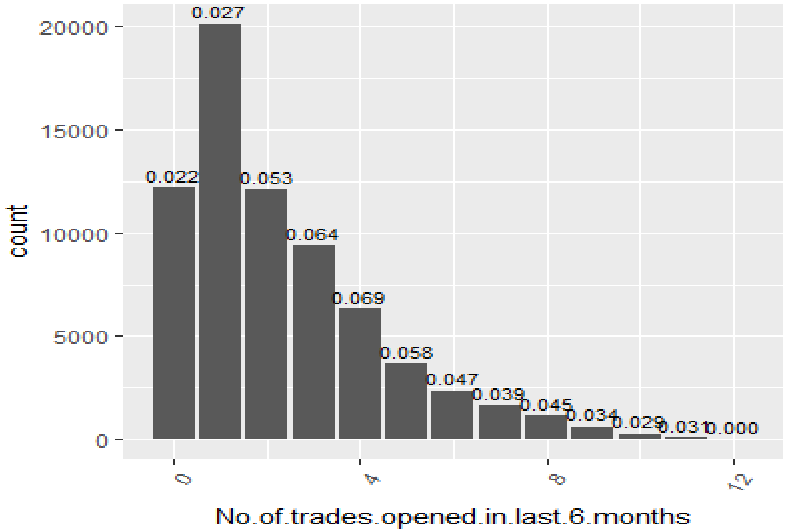
Bar Chart - No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.



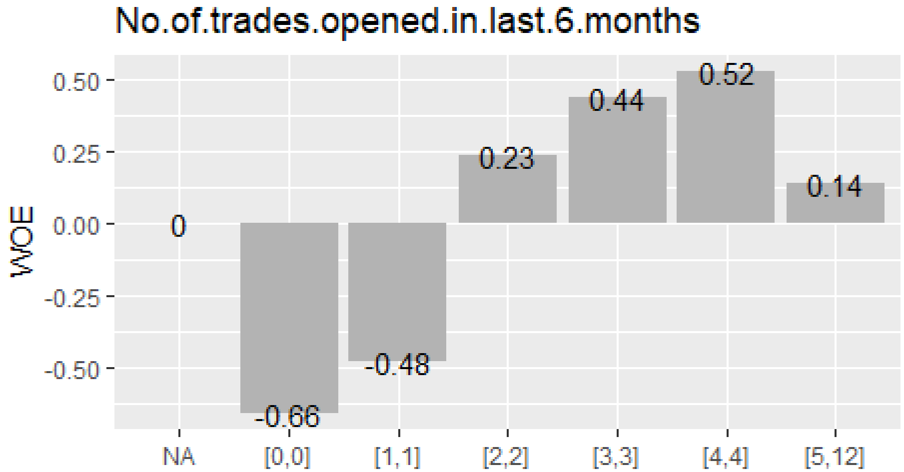
WOE- No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.



Bar Chart – No.of.trades.opened.in.last.6.months

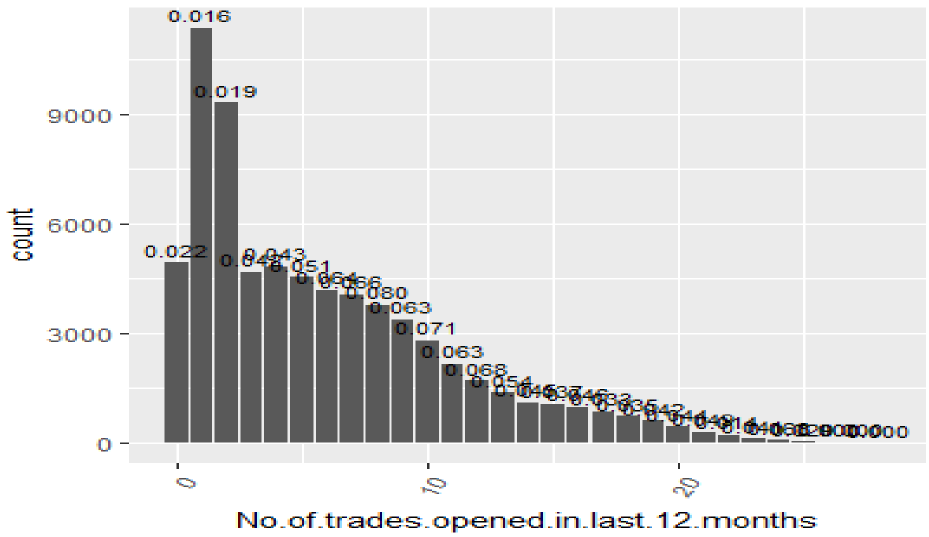


WOE- No.of.trades.opened.in.last.6.months

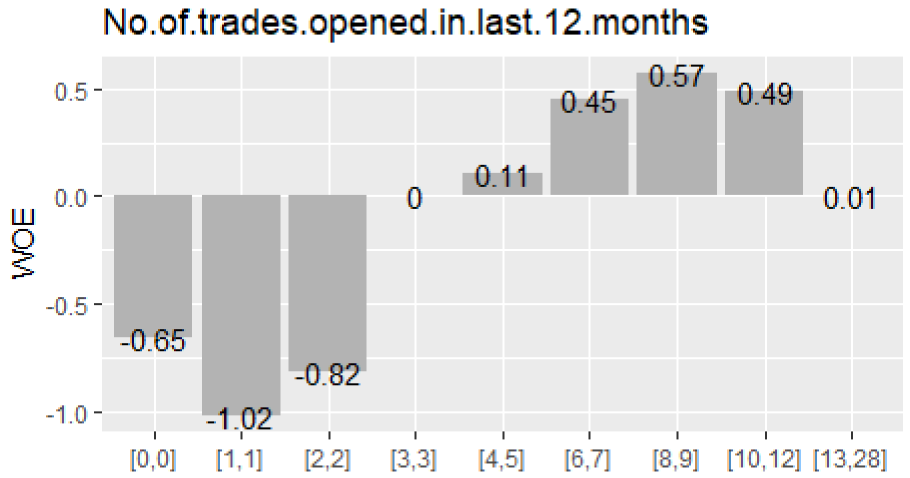


Categorical – No of Trades in 12m & No of Inquiries in 6m

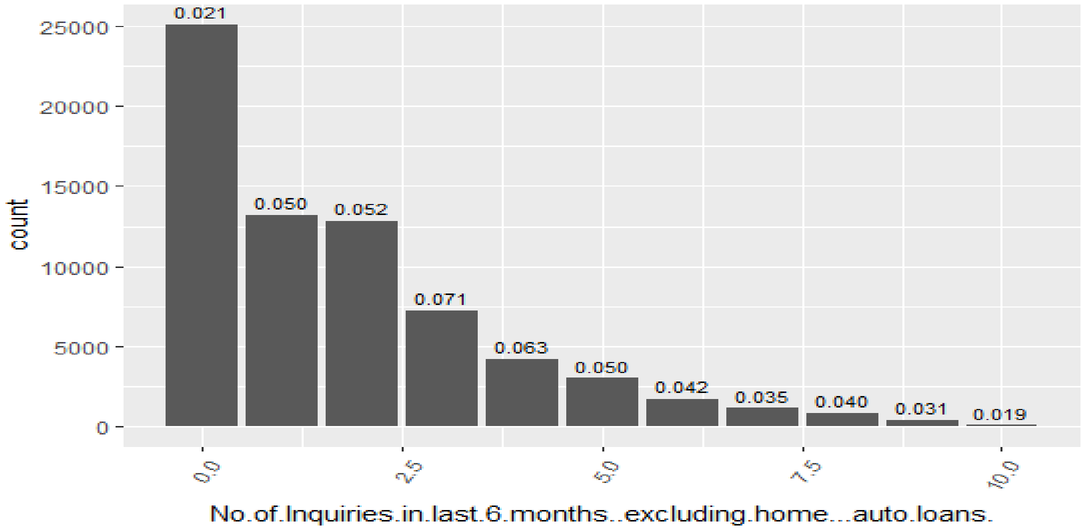
Bar Chart – No.of.trades.opened.in.last.12.months



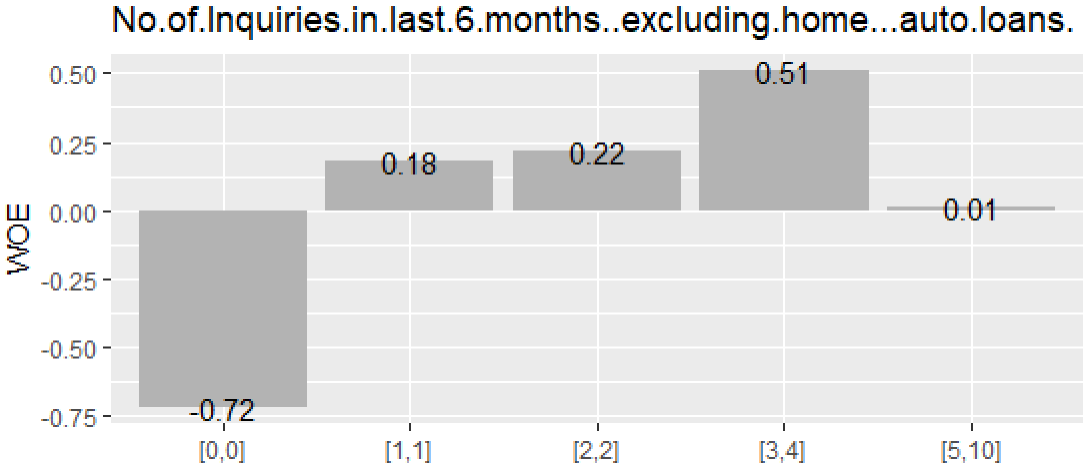
WOE- No.of.trades.opened.in.last.12.months



Bar Chart – No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.

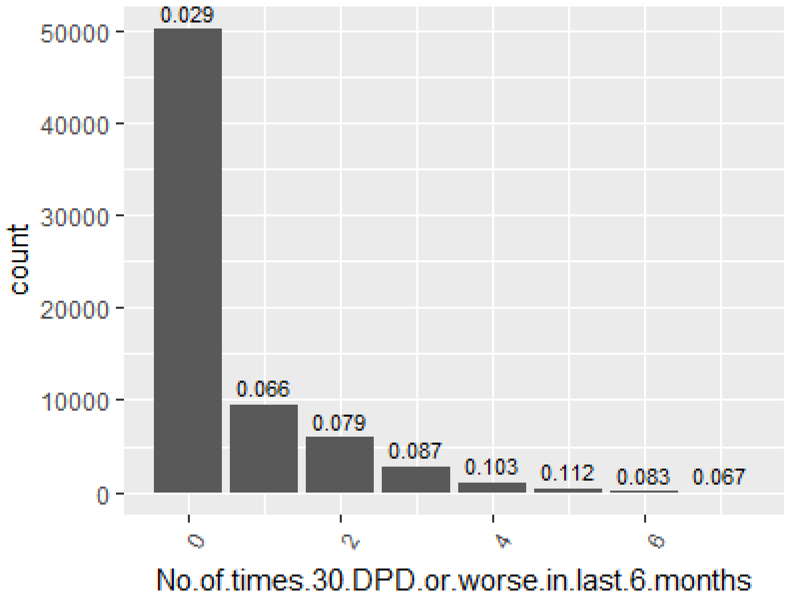


WOE- No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.

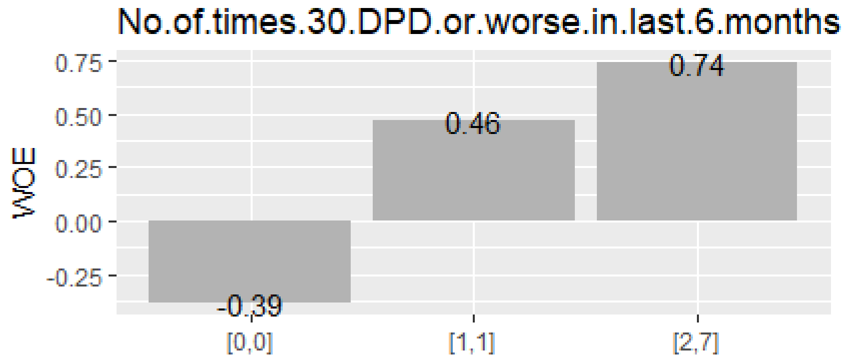


Categorical – No of Times 30 DPD in 6m & No of times DPD in 12m

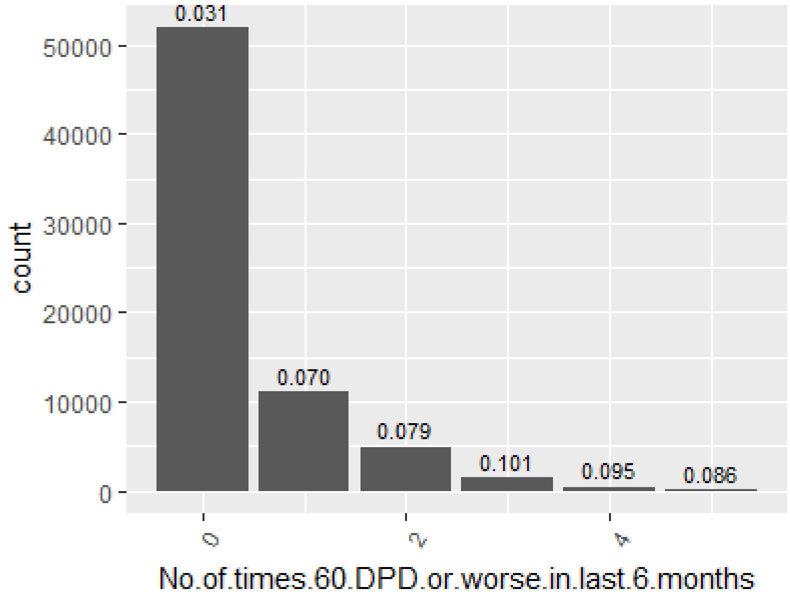
Bar Chart - No.of.times.30.DPD.or.worse.in.last.6.months



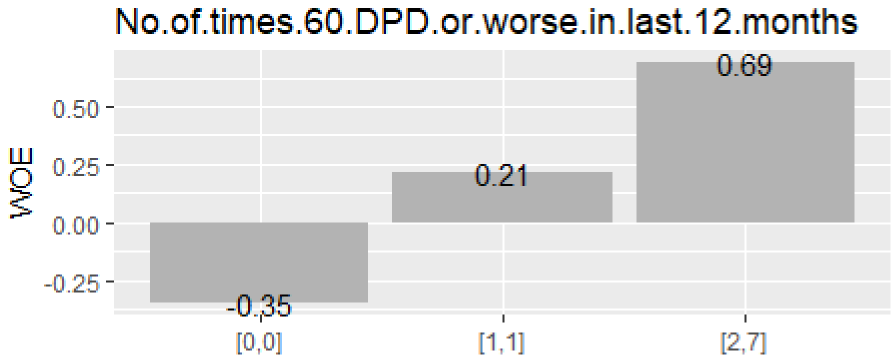
WOE- No.of.times.30.DPD.or.worse.in.last.6.months



Bar Chart - No.of.times.60.DPD.or.worse.in.last.12.months

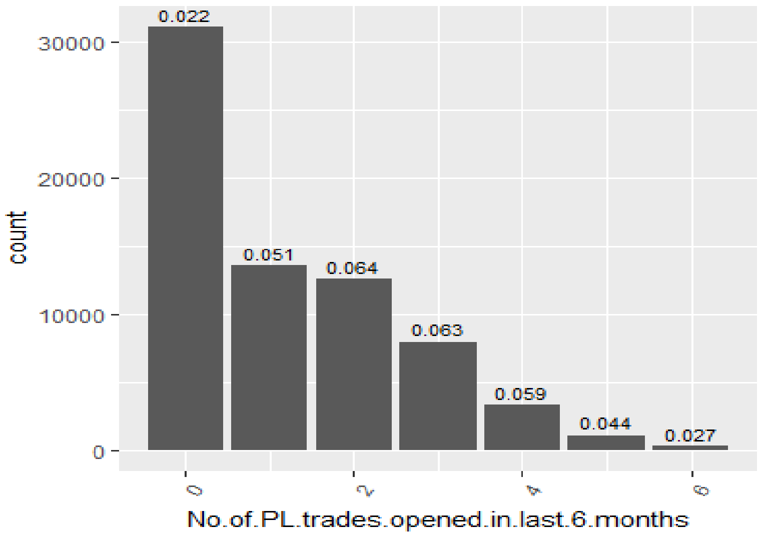


WOE- No.of.times.60.DPD.or.worse.in.last.12.months

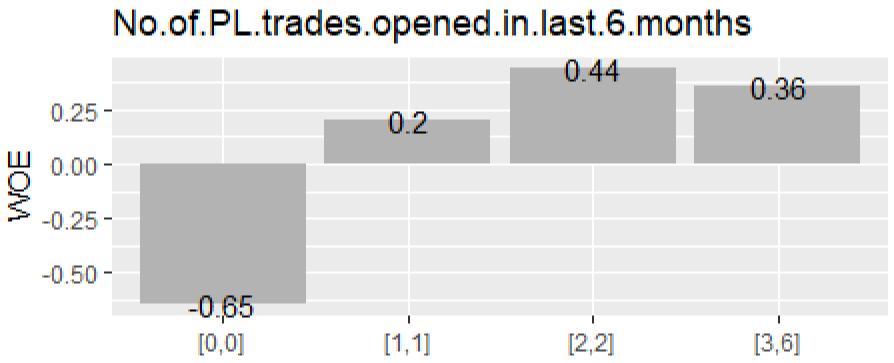


Categorical – No of PL Trades in 6m & No of times 90 DPD in 6m

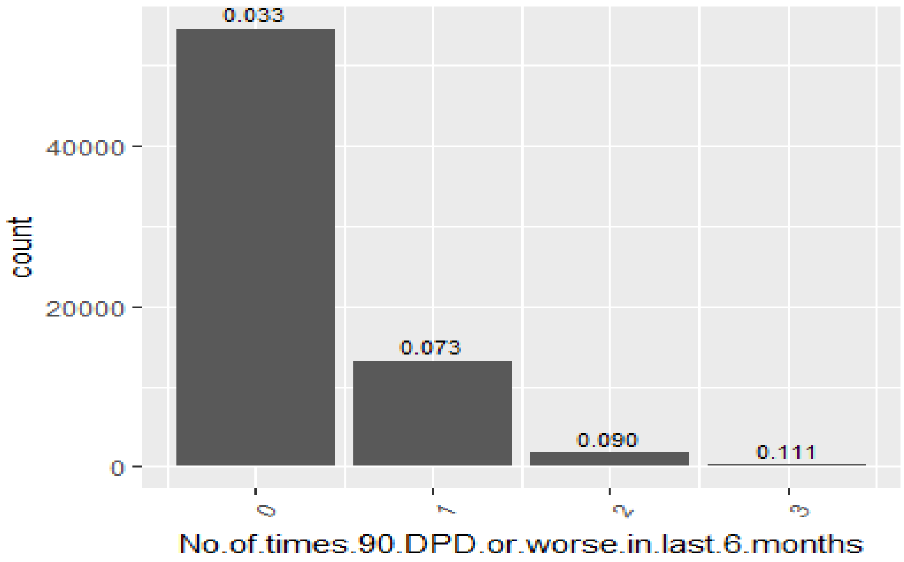
Bar Chart - No.of.PL.trades.opened.in.last.6.months



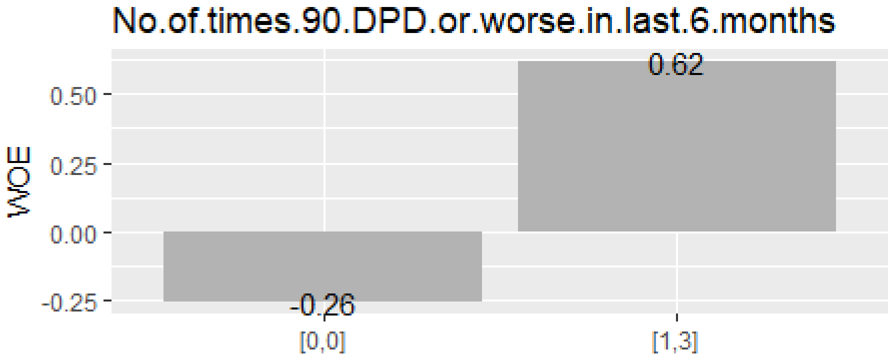
WOE- No.of.PL.trades.opened.in.last.6.months



Bar Chart - No.of.times.90.DPD.or.worse.in.last.6.months

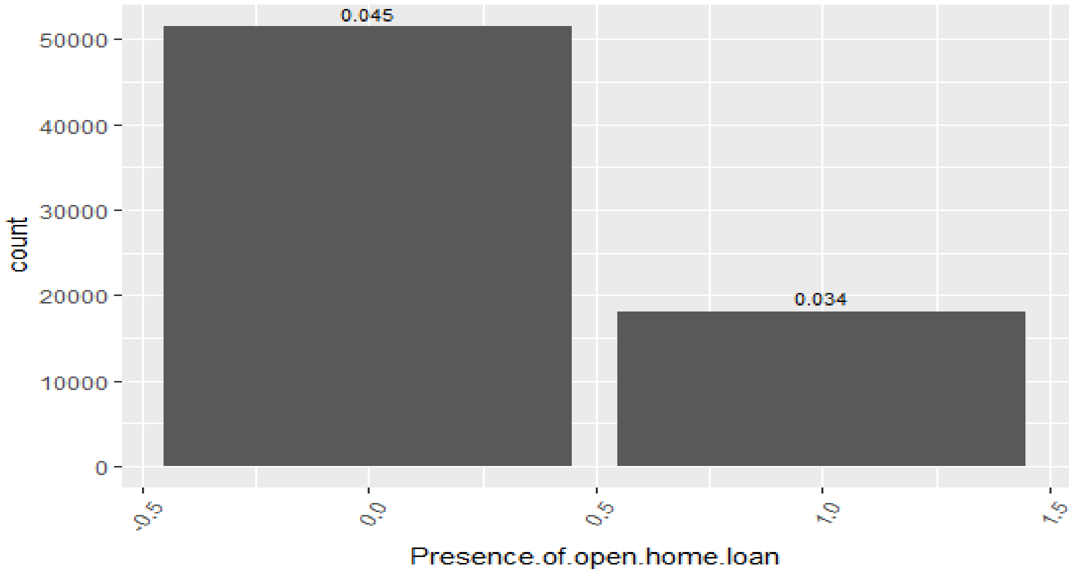


WOE- No.of.times.90.DPD.or.worse.in.last.6.months

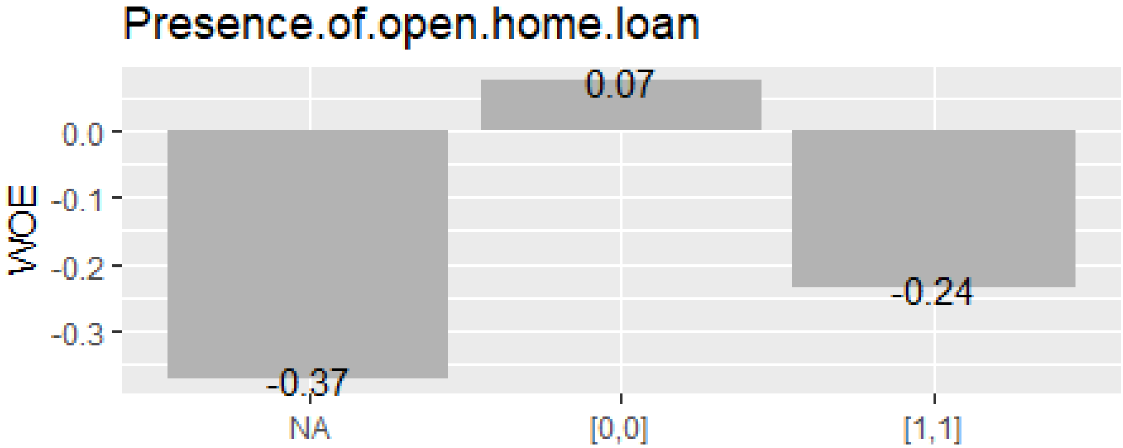


Categorical – Presence of open home loan & No of PL trades in 12m

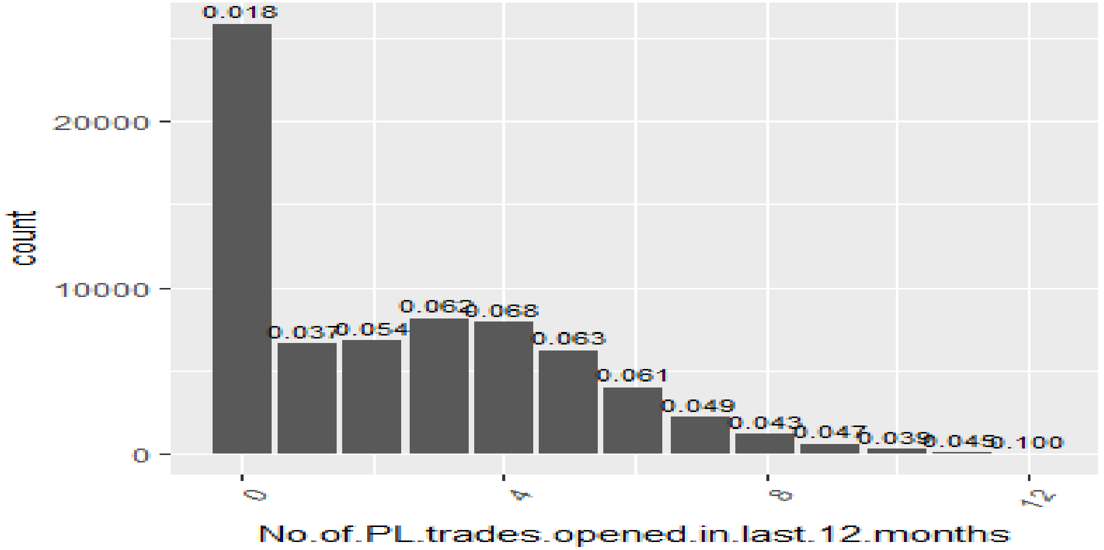
Bar Chart - Presence.of.open.home.loan



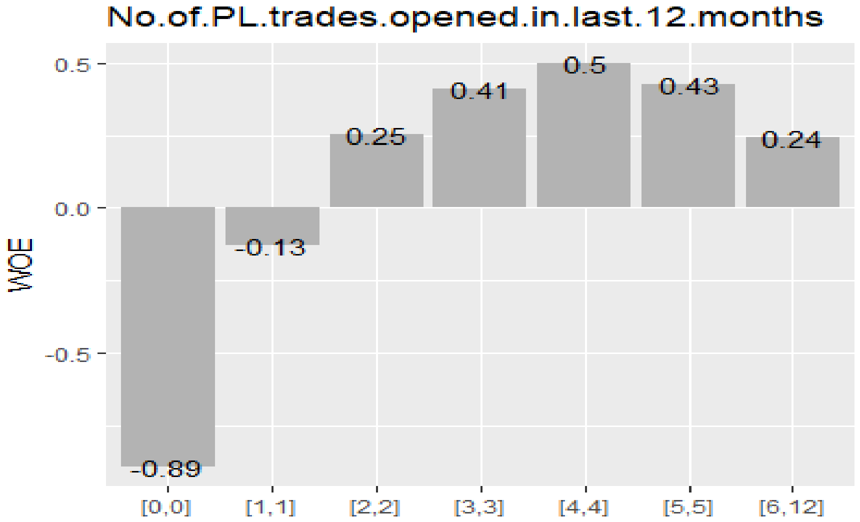
WOE- Presence.of.open.home.loan



Bar Chart - No.of.PL.trades.opened.in.last.12.months



WOE- No.of.PL.trades.opened.in.last.12.months



Categorical – No of 60 DPD+ in 6m & No of 90+ DPD in 12m