CredX_Credit_Risk_Analytics

November 1, 2018

BFSI CAPSTONE PROJECT Credit Risk Analytics

Objective: The main objective is to identify the right customers using predictive models.
1) Determine the factors affecting credit risk using past data of the
bank's applicants
2) Create strategies to mitigate the acquisition risk
3) Assess the financial benefit of this project

Business understanding

Check and Import required libraries

```
library(tidyverse)
library(cowplot)
library(formattable)
library(corrplot)
library(Information)
library(caret)
library(caTools)
library(MASS)
library(car)
library(e1071)
library(ROCR)
library(fuzzyjoin)
```

Import Datasets

```
demographics_raw <- demographics  <-
  read.csv("Demographic data.csv",stringsAsFactors = F, na.strings = c("", "NA"))
creditbu_raw     <- creditbu      <-
  read.csv("Credit Bureau data.csv",stringsAsFactors = F, na.strings = c("", "NA"))
```

Data Understanding & Cleaning

```
dim(demographics)

## [1] 71295    12

# 71295    12
dim(creditbu)

## [1] 71295    19

# 71295    19
# Both the datasets have the same number of observations
str(demographics)

## 'data.frame':    71295 obs. of  12 variables:
##  $ Application.ID                             : int  954457215 432830445 941387308 39216167
7 182011211 312196805 532217204 74788849 782743811 96964957 ...
```

```
## $ Age : int 48 31 32 43 35 20 42 34 30 22 ...
## $ Gender : chr "F" "M" "M" "M" ...
## $ Marital.Status..at.the.time.of.application.: chr "Married" "Married" "Single" "Married"
...
## $ No.of.dependents : int 2 4 2 1 5 1 2 2 3 1 ...
## $ Income : num 40 55 46 53 44 39 55 49 48 38 ...
## $ Education : chr "Bachelor" "Professional" "Bachelor" "
Bachelor" ...
## $ Profession : chr "SAL" "SE_PROF" "SE_PROF" "SE" ...
## $ Type.of.residence : chr "Rented" "Rented" "Rented" "Rented" ..
.
## $ No.of.months.in.current.residence : int 113 112 104 94 112 116 104 108 115 111
...
## $ No.of.months.in.current.company : int 56 46 49 53 43 52 41 40 58 57 ...
## $ Performance.Tag : int 0 0 0 0 0 0 0 0 0 0 ...
```

Rename lengthy column names for convenience

```
demographics %>%
rename(Marital.Status = Marital.Status..at.the.time.of.application.,
       Curr.rsdnc.months = No.of.months.in.current.residence,
       Curr.cmpny.months = No.of.months.in.current.company) -> demographics

str(creditbu)
```

```
## 'data.frame': 71295 obs. of 19 variables:
## $ Application.ID : int 954457215 43283044
5 941387308 392161677 182011211 312196805 532217204 74788849 782743811 96964957 ...
## $ No.of.times.90.DPD.or.worse.in.last.6.months : int 0 0 0 0 0 0 0 0 0
0 ...
## $ No.of.times.60.DPD.or.worse.in.last.6.months : int 0 0 0 0 0 0 0 0 0
0 ...
## $ No.of.times.30.DPD.or.worse.in.last.6.months : int 0 0 0 0 0 0 0 0 0
0 ...
## $ No.of.times.90.DPD.or.worse.in.last.12.months : int 0 0 0 0 0 0 0 0 0
0 ...
## $ No.of.times.60.DPD.or.worse.in.last.12.months : int 0 0 0 0 0 0 0 0 0
0 ...
## $ No.of.times.30.DPD.or.worse.in.last.12.months : int 0 0 0 0 0 0 0 0 1
0 ...
## $ Avgas.CC.Utilization.in.last.12.months : int 4 3 7 11 12 10 11
13 9 6 ...
## $ No.of.trades.opened.in.last.6.months : int 1 1 0 1 0 0 0 1 0
1 ...
## $ No.of.trades.opened.in.last.12.months : int 2 2 0 1 1 0 1 1 0
1 ...
## $ No.of.PL.trades.opened.in.last.6.months : int 0 0 0 0 0 0 0 0 0
0 ...
## $ No.of.PL.trades.opened.in.last.12.months : int 0 0 0 0 0 0 0 0 0
0 ...
## $ No.of.Inquiries.in.last.6.months..excluding.home...auto.loans. : int 0 0 0 0 0 0 0 0 0
0 ...
## $ No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.: int 0 0 0 0 0 0 0 0 0
0 ...
## $ Presence.of.open.home.loan : int 1 0 1 1 1 0 1 1 1
0 ...
## $ Outstanding.Balance : int 2999395 3078 30049
72 3355373 3014283 2569 3005535 3004790 3007428 170860 ...
## $ Total.No.of.Trades : int 4 5 2 4 4 1 4 3 2
1 ...
## $ Presence.of.open.auto.loan : int 0 0 0 1 0 0 0 0 0
```

```
1 ...
##  $ Performance.Tag                                        : int  0 0 0 0 0 0 0 0 0
0 ...
```

Rename lengthy column names for convenience.

```
creditbu %>%
rename(DPD90.6months      = No.of.times.90.DPD.or.worse.in.last.6.months,
       DPD60.6months      = No.of.times.60.DPD.or.worse.in.last.6.months,
       DPD30.6months      = No.of.times.30.DPD.or.worse.in.last.6.months,
       DPD90.12months     = No.of.times.90.DPD.or.worse.in.last.12.months,
       DPD60.12months     = No.of.times.60.DPD.or.worse.in.last.12.months,
       DPD30.12months     = No.of.times.30.DPD.or.worse.in.last.12.months,
       CC.utilization     = Avgas.CC.Utilization.in.last.12.months,
       Trades.6months     = No.of.trades.opened.in.last.6.months,
       Trades.12months    = No.of.trades.opened.in.last.12.months,
       PL.Trades.6months  = No.of.PL.trades.opened.in.last.6.months,
       PL.Trades.12months = No.of.PL.trades.opened.in.last.12.months,
       Inquiries.6months  = No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.,
       Inquiries.12months = No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.,
       Has.Home.loan      = Presence.of.open.home.loan,
       Has.Auto.loan      = Presence.of.open.auto.loan) -> creditbu

head(demographics)
```

```
##   Application.ID Age Gender Marital.Status No.of.dependents Income
## 1      954457215  48      F        Married                2     40
## 2      432830445  31      M        Married                4     55
## 3      941387308  32      M         Single                2     46
## 4      392161677  43      M        Married                1     53
## 5      182011211  35      F        Married                5     44
## 6      312196805  20      M        Married                1     39
##      Education Profession Type.of.residence Curr.rsdnc.months
## 1     Bachelor        SAL            Rented               113
## 2 Professional     SE_PROF            Rented               112
## 3     Bachelor     SE_PROF            Rented               104
## 4     Bachelor         SE            Rented                94
## 5 Professional        SAL            Rented               112
## 6     Bachelor        SAL              <NA>               116
##   Curr.cmpny.months Performance.Tag
## 1                56               0
## 2                46               0
## 3                49               0
## 4                53               0
## 5                43               0
## 6                52               0
```

```
tail(demographics)
```

```
##       Application.ID Age Gender Marital.Status No.of.dependents Income
## 71290      304125466  49      F        Married                5    7.0
## 71291      254036864  44      M        Married                3   15.0
## 71292      375231276  24      M         Single                1    4.5
## 71293       32481239  33      M        Married                4    6.0
## 71294      704812159  52      M        Married                3    4.5
## 71295       37493797  54      M        Married                3   42.0
##          Education Profession Type.of.residence Curr.rsdnc.months
## 71290      Masters         SE            Rented                10
## 71291 Professional        SAL            Rented                 6
## 71292     Bachelor        SAL             Owned                20
## 71293     Bachelor    SE_PROF            Rented                37
```

```
## 71294     Bachelor         SE            Rented                  76
## 71295     Bachelor         SE            Rented                  96
##       Curr.cmpny.months Performance.Tag
## 71290                71               1
## 71291                 3               0
## 71292                 7               1
## 71293                25               0
## 71294                57               0
## 71295                29               0
```

```r
head(creditbu)
```

```
##   Application.ID DPD90.6months DPD60.6months DPD30.6months DPD90.12months
## 1      954457215             0             0             0              0
## 2      432830445             0             0             0              0
## 3      941387308             0             0             0              0
## 4      392161677             0             0             0              0
## 5      182011211             0             0             0              0
## 6      312196805             0             0             0              0
##   DPD60.12months DPD30.12months CC.utilization Trades.6months
## 1              0              0              4              1
## 2              0              0              3              1
## 3              0              0              7              0
## 4              0              0             11              1
## 5              0              0             12              0
## 6              0              0             10              0
##   Trades.12months PL.Trades.6months PL.Trades.12months Inquiries.6months
## 1               2                 0                  0                 0
## 2               2                 0                  0                 0
## 3               0                 0                  0                 0
## 4               1                 0                  0                 0
## 5               1                 0                  0                 0
## 6               0                 0                  0                 0
##   Inquiries.12months Has.Home.loan Outstanding.Balance Total.No.of.Trades
## 1                  0             1             2999395                  4
## 2                  0             0                3078                  5
## 3                  0             1             3004972                  2
## 4                  0             1             3355373                  4
## 5                  0             1             3014283                  4
## 6                  0             0                2569                  1
##   Has.Auto.loan Performance.Tag
## 1             0               0
## 2             0               0
## 3             0               0
## 4             1               0
## 5             0               0
## 6             0               0
```

```r
tail(creditbu)
```

```
##       Application.ID DPD90.6months DPD60.6months DPD30.6months
## 71290      304125466             1             1             2
## 71291      254036864             1             2             4
## 71292      375231276             0             1             2
## 71293       32481239             0             1             2
## 71294      704812159             2             2             4
## 71295       37493797             2             3             4
##       DPD90.12months DPD60.12months DPD30.12months CC.utilization
## 71290              2              2              3             NA
## 71291              1              3              6             NA
## 71292              0              1              2             NA
```

```
## 71293              1              3              2              NA
## 71294              3              4              5              62
## 71295              3              4              5              33
##       Trades.6months Trades.12months PL.Trades.6months PL.Trades.12months
## 71290              0               3                 0                  2
## 71291              3               9                 3                  5
## 71292              4              11                 3                  6
## 71293              1               8                 1                  5
## 71294              3              10                 3                  5
## 71295              2               5                 2                  3
##       Inquiries.6months Inquiries.12months Has.Home.loan
## 71290                 1                  5             0
## 71291                 4                  6             0
## 71292                 2                  4             1
## 71293                 2                  4             1
## 71294                 4                  6             1
## 71295                 3                  5             1
##       Outstanding.Balance Total.No.of.Trades Has.Auto.loan Performance.Tag
## 71290              396536                  2             0               1
## 71291             1028144                  8             0               0
## 71292             3564911                  9             0               1
## 71293             3386883                  7             0               0
## 71294             3475822                  9             0               0
## 71295             3088029                  4             0               0
```

Store the column names from the demographics dataset for building our 1st model later.

```
demographics_cols <- colnames(demographics)
```

Merging datasets:

Lets check if there are there any differences in the Application IDs (Key field for merging the files)

```
length(setdiff(demographics$Application.ID,creditbu$Application.ID))

## [1] 0
```

0 -> Matches with the number of No Hits. i.e whatever IDs are present in Demographics dataset has a matching ID in the Credit Bureau dataset

But lets also check for any duplicate Application IDs

```
dim(demographics)[1] - length(unique(demographics$Application.ID))

## [1] 3

# 3 duplicates
dim(creditbu)[1] - length(unique(creditbu$Application.ID))

## [1] 3

# 3 duplicates

sum(duplicated(demographics))

## [1] 0

sum(duplicated(creditbu))

## [1] 0
```

0 -> Indicates that eventhough the application ID is duplicate, the other values corresponding to these duplicate rows are different.

2 ways to handle this:

- Ignore the duplicates since the proportion is very less,
- or fix it.

We will choose to fix these observations.

Lets check which Application IDs are duplicate and if the duplicates are the same in both the files

```
demographics %>%
  mutate(rownum = row_number()) %>%
  group_by(Application.ID) %>%
  filter(n()>1) %>%
  arrange(Application.ID) %>%
  dplyr::select(Application.ID,rownum)

## # A tibble: 6 x 2
## # Groups:   Application.ID [3]
##   Application.ID rownum
##            <int>  <int>
## 1      653287861   5244
## 2      653287861  42638
## 3      671989187  48603
## 4      671989187  59023
## 5      765011468  24387
## 6      765011468  27587

creditbu %>%
  mutate(rownum = row_number()) %>%
  group_by(Application.ID) %>%
  filter(n()>1) %>%
  arrange(Application.ID) %>%
  dplyr::select(Application.ID,rownum)

## # A tibble: 6 x 2
## # Groups:   Application.ID [3]
##   Application.ID rownum
##            <int>  <int>
## 1      653287861   5244
## 2      653287861  42638
## 3      671989187  48603
## 4      671989187  59023
## 5      765011468  24387
## 6      765011468  27587
```

The duplicates are the same and are present at the same exact position (as indicated by the generated rownumbers) in both the files

Fix the Application IDs at row numbers 42638, 59023 & 27587 in both the files We will assign the next available Application IDs to these.

```
demographics$Application.ID[42638] = max(demographics$Application.ID) + 1
demographics$Application.ID[59023] = max(demographics$Application.ID) + 2
demographics$Application.ID[27587] = max(demographics$Application.ID) + 3

creditbu$Application.ID[42638] = max(creditbu$Application.ID) + 1
creditbu$Application.ID[59023] = max(creditbu$Application.ID) + 2
creditbu$Application.ID[27587] = max(creditbu$Application.ID) + 3
```

No hit cases in credit bureau: The cases where all the variables in the credit bureau data are zero and credit card utilisation is missing, represent cases in which there is a no-hit in the credit bureau.

```
sum((rowSums((creditbu[,2:18]), na.rm=T) == 0))

## [1] 566
```

566 cases where there is no-hit in credit bureau Lets remove these rows from creditbu dataset

```
creditbu %>% filter((rowSums((creditbu[,2:18]), na.rm=T) > 0)) -> creditbu
```

The cases with the credit card utilisation missing, represent cases in which the applicant does not have any other credit card.

```
sum(is.na(creditbu$CC.utilization))

## [1] 492
```

After removing No hit cases, 492 cases doesn't have any other credit card.

Merge both the files using Application ID & Performance.Tag as the key

```
master <-
  demographics %>%
  merge(creditbu, by=c("Application.ID", "Performance.Tag"))
dim(master)

## [1] 70729    29
```

All the data from both the files are now matched and merged. - 70729 observations - 29 features (17 demographics cols + 10 creditbu cols + 1 common Application.ID + 1 common Performance.Tag)

```
summary(master)

##  Application.ID      Performance.Tag        Age             Gender
##  Min.   :1.004e+05   Min.   :0.0000   Min.   :-3.00   Length:70729
##  1st Qu.:2.484e+08   1st Qu.:0.0000   1st Qu.:37.00   Class :character
##  Median :4.976e+08   Median :0.0000   Median :45.00   Mode  :character
##  Mean   :4.990e+08   Mean   :0.0421   Mean   :44.95
##  3rd Qu.:7.499e+08   3rd Qu.:0.0000   3rd Qu.:53.00
##  Max.   :1.000e+09   Max.   :1.0000   Max.   :65.00
##                      NA's   :1425
##  Marital.Status     No.of.dependents     Income       Education
##  Length:70729       Min.   :1.000    Min.   :-0.5   Length:70729
##  Class :character   1st Qu.:2.000    1st Qu.:14.0   Class :character
##  Mode  :character   Median :3.000    Median :27.0   Mode  :character
##                     Mean   :2.865    Mean   :27.2
##                     3rd Qu.:4.000    3rd Qu.:40.0
##                     Max.   :5.000    Max.   :60.0
##                     NA's   :3
##   Profession        Type.of.residence  Curr.rsdnc.months Curr.cmpny.months
##  Length:70729       Length:70729       Min.   :  6.00    Min.   :  3.00
##  Class :character   Class :character   1st Qu.:  6.00    1st Qu.: 16.00
##  Mode  :character   Mode  :character   Median : 11.00    Median : 34.00
##                                        Mean   : 34.54    Mean   : 33.99
##                                        3rd Qu.: 60.00    3rd Qu.: 51.00
##                                        Max.   :126.00    Max.   :133.00
##
##  DPD90.6months     DPD60.6months    DPD30.6months     DPD90.12months
##  Min.   :0.0000    Min.   :0.000    Min.   :0.0000    Min.   :0.0000
##  1st Qu.:0.0000    1st Qu.:0.000    1st Qu.:0.0000    1st Qu.:0.0000
```

```
##  Median :0.0000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.2725   Mean   :0.434   Mean   :0.5818   Mean   :0.4539
##  3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :3.0000   Max.   :5.000   Max.   :7.0000   Max.   :5.0000
##
##  DPD60.12months   DPD30.12months   CC.utilization  Trades.6months
##  Min.   :0.0000   Min.   :0.0000   Min.   :  0.0   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  8.0   1st Qu.: 1.000
##  Median :0.0000   Median :0.0000   Median : 15.0   Median : 2.000
##  Mean   :0.6607   Mean   :0.8073   Mean   : 29.7   Mean   : 2.316
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 46.0   3rd Qu.: 3.000
##  Max.   :7.0000   Max.   :9.0000   Max.   :113.0   Max.   :12.000
##                                    NA's   :492
##  Trades.12months  PL.Trades.6months PL.Trades.12months Inquiries.6months
##  Min.   : 0.000   Min.   :0.000     Min.   : 0.000     Min.   : 0.000
##  1st Qu.: 2.000   1st Qu.:0.000     1st Qu.: 0.000     1st Qu.: 0.000
##  Median : 5.000   Median :1.000     Median : 2.000     Median : 1.000
##  Mean   : 5.874   Mean   :1.217     Mean   : 2.417     Mean   : 1.778
##  3rd Qu.: 9.000   3rd Qu.:2.000     3rd Qu.: 4.000     3rd Qu.: 3.000
##  Max.   :28.000   Max.   :6.000     Max.   :12.000     Max.   :10.000
##
##  Inquiries.12months Has.Home.loan    Outstanding.Balance
##  Min.   : 0.000     Min.   :0.0000   Min.   :      0
##  1st Qu.: 0.000     1st Qu.:0.0000   1st Qu.: 216248
##  Median : 3.000     Median :0.0000   Median : 777745
##  Mean   : 3.564     Mean   :0.2585   Mean   :1259198
##  3rd Qu.: 5.000     3rd Qu.:1.0000   3rd Qu.:2924409
##  Max.   :20.000     Max.   :1.0000   Max.   :5218801
##                     NA's   :272      NA's   :272
##  Total.No.of.Trades Has.Auto.loan
##  Min.   : 0.000     Min.   :0.0000
##  1st Qu.: 3.000     1st Qu.:0.0000
##  Median : 6.000     Median :0.0000
##  Mean   : 8.252     Mean   :0.0853
##  3rd Qu.:10.000     3rd Qu.:0.0000
##  Max.   :44.000     Max.   :1.0000
##
```

Age has -ve numbers. Lets check which one of those:

```
master[which(master$Age <= 0),] %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    20
```

20 observations where age is either 0 or -ve. We will replace these with NA and handle them during the WOE analysis

```
master$Age[which(master$Age <= 0)] <- NA
```

Income has -ve numbers. Lets check which one of those:

```
master[which(master$Income < 0),] %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    79
```

79 observations where Income is -ve (We will leave the incomes with 0 as is). We will replace the -ve's with NA and handle them during the WOE analysis

```
master$Income[which(master$Income < 0)] <- NA
```

We also have quite a few other missing/NA values. Lets check them.

```
master %>%
  summarise_all(funs(sum(is.na(.))/n())) %>%
  gather(key='Variable',value = 'Missing') %>%
  filter(Missing > 0) %>%
  arrange(desc(Missing)) %>%
  mutate(Missing = percent(Missing, 3))

##               Variable Missing
## 1      Performance.Tag  2.015%
## 2       CC.utilization  0.696%
## 3        Has.Home.loan  0.385%
## 4  Outstanding.Balance  0.385%
## 5            Education  0.168%
## 6               Income  0.112%
## 7                  Age  0.028%
## 8           Profession  0.020%
## 9     Type.of.residence  0.011%
## 10      Marital.Status  0.008%
## 11     No.of.dependents  0.004%
## 12              Gender  0.003%
```

Remove Application ID column as it would be no longer required The applicants who were not given the credit card in the first place(Rejected Applicants) have NAs in the Performance.Tag. So these rows can be removed. The Rejected applicants will be used further in the score card verification.

```
master %>%
  dplyr::select(-c(Application.ID)) %>%
  drop_na(Performance.Tag) -> master

dim(master)

## [1] 69304    28
```

69304 Observations & 28 features

EDA

Common Functions

Setting the theme of plots

```
plot_theme <- theme_classic() +
  theme(plot.title = element_text(hjust = 0.5, size = 12,face = 'bold'),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        axis.text.x  = element_text(size = 10),
        axis.text.y  = element_text(size = 10))
```

Continuous Univariate plots

```
ContUnivar <- function(yfeature, ylabel) {
  ggplot(master, aes(x = "", y = yfeature)) +
    geom_boxplot(fill = "#F8766D", outlier.colour = "red", outlier.shape = 1) +
    stat_boxplot(geom = "errorbar", width = 0.5) +
```

```
    labs( y = ylabel, title = paste(ylabel, "Distribution")) +
    plot_theme
}
```

Bivariate plots

```
CatBivar <-  function(xfeature, yfeature, xlabel, ylabel) {
  as.data.frame(percent(prop.table(table(yfeature, xfeature), 2))) %>%
    ggplot(aes(x = xfeature, y = Freq,  fill = yfeature)) +
    geom_col( position = "fill" ) +
    geom_text(aes(label = Freq),
              position = position_fill(vjust = .5),
              size = 2.5) +
    labs(x = xlabel, y = "Performance Proportion",
         title = paste(ylabel,"Proportion by", xlabel), fill = "Performance") +
    plot_theme +
    theme(legend.position = 'none')
}
```

Bivariate plots

```
ContCatBivar <- function(xfeature, yfeature, xlabel, ylabel) {
  ggplot(woe_data, aes(x = xfeature, y = yfeature, fill = xfeature)) +
    geom_boxplot(outlier.colour = "red", outlier.shape = 1, show.legend = F) +
    stat_boxplot(geom = "errorbar", width = 0.5) +
    labs(x = xlabel, y = ylabel, title = paste(ylabel, "vs", xlabel)) +
    plot_theme
}
```

Treating outliers

```
treatoutlier <- function(x) {
  x[which(x %in% boxplot.stats(x)$out)] <-
    boxplot.stats(x)$stats[5]
  return(x)
}
```

Univariate Analyis

Weight of Evidence WOE/Information value

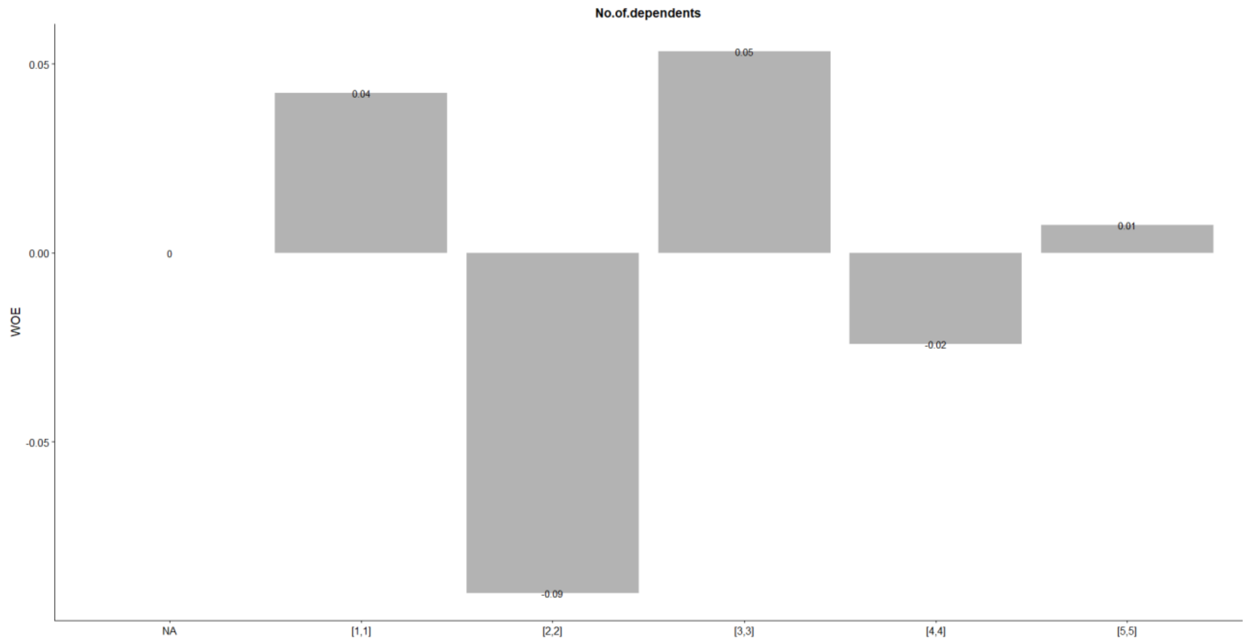
Lets identify the important variables using WOE/IV While doing so we will use WOE to fix the missing values.

```
infoTables <- create_infotables(data = master,
                                y = "Performance.Tag",
                                bins = 10,
                                parallel = T)
```

Treating NA/Missing values

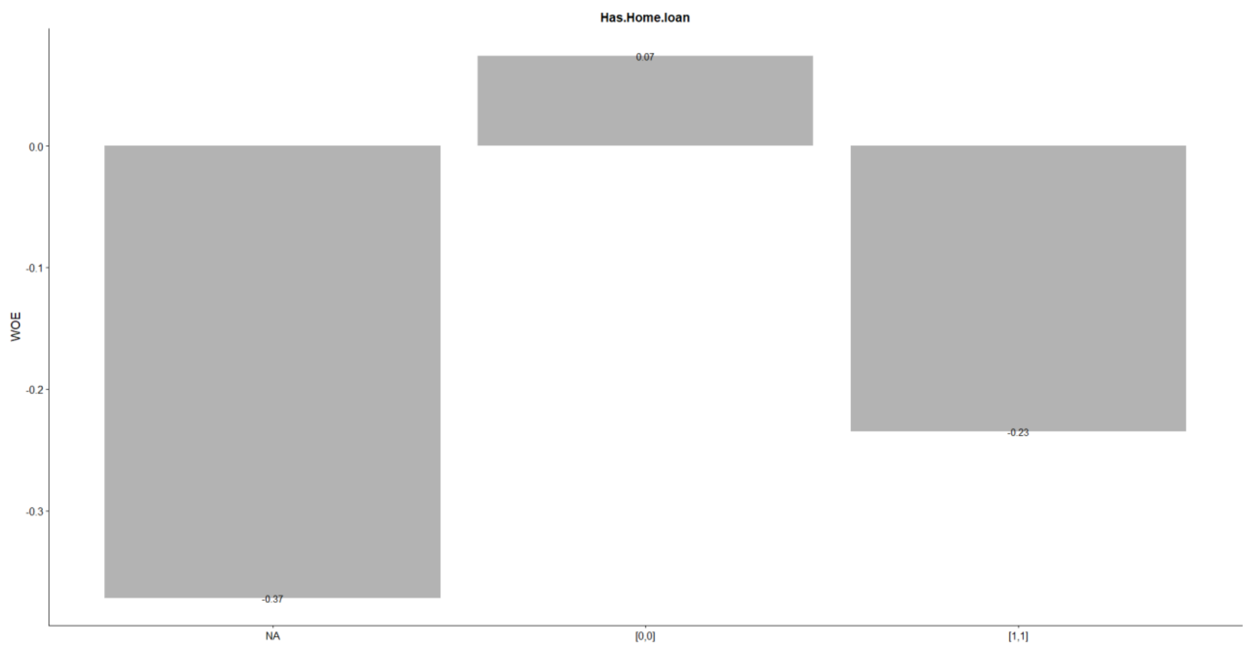
Check the WOE values under each bucket and replace NA value bucket to the closest possible bucket

```
plot_infotables(infoTables, "No.of.dependents", show_values=TRUE)
```



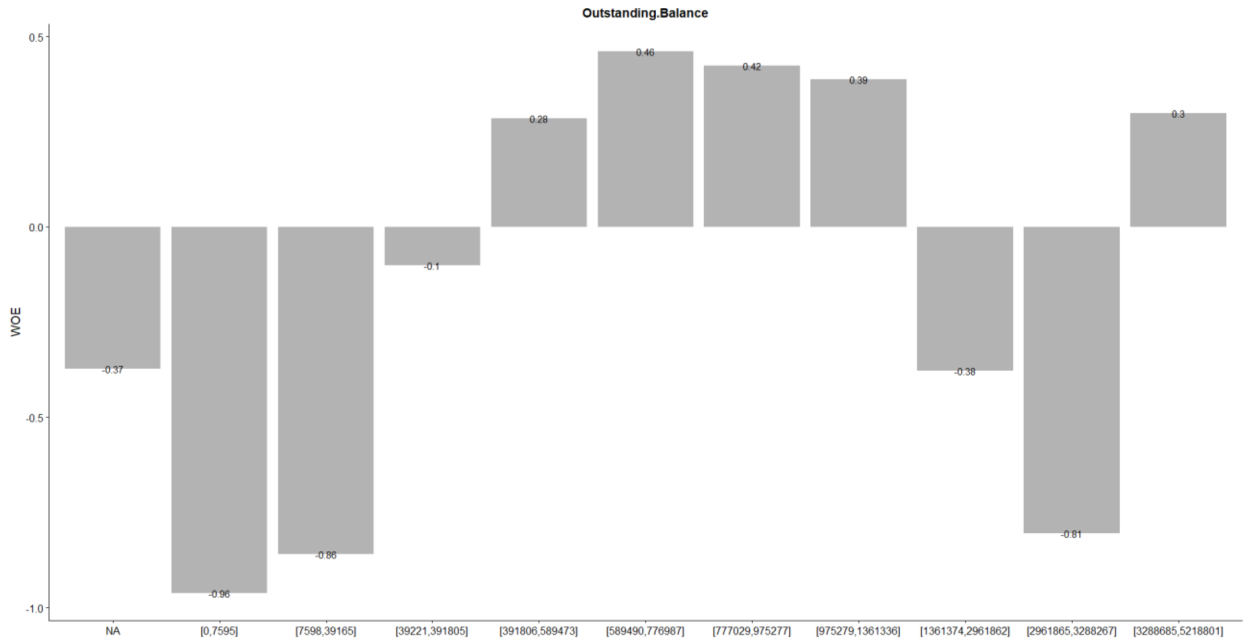

```
# Here NA bucket WOE is close to WOE of bucket 5. Hence replace NAs with 5.
master$No.of.dependents[which(is.na(master$No.of.dependents))] <- 5

plot_infotables(infoTables, "Has.Home.loan", show_values=TRUE)
```



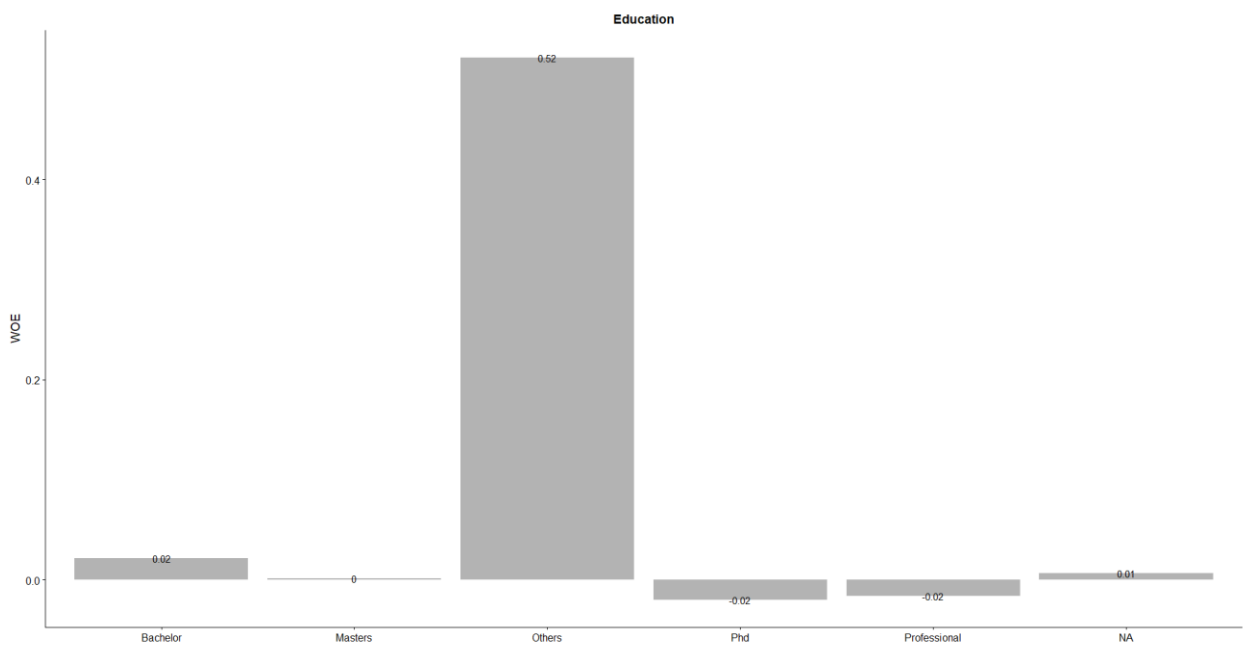

```
# NA bucket WOE is close to WOE of bucket 1. Hence replace with 1.
master$Has.Home.loan[which(is.na(master$Has.Home.loan))] <- 1

plot_infotables(infoTables, "Outstanding.Balance", show_values=TRUE)
```



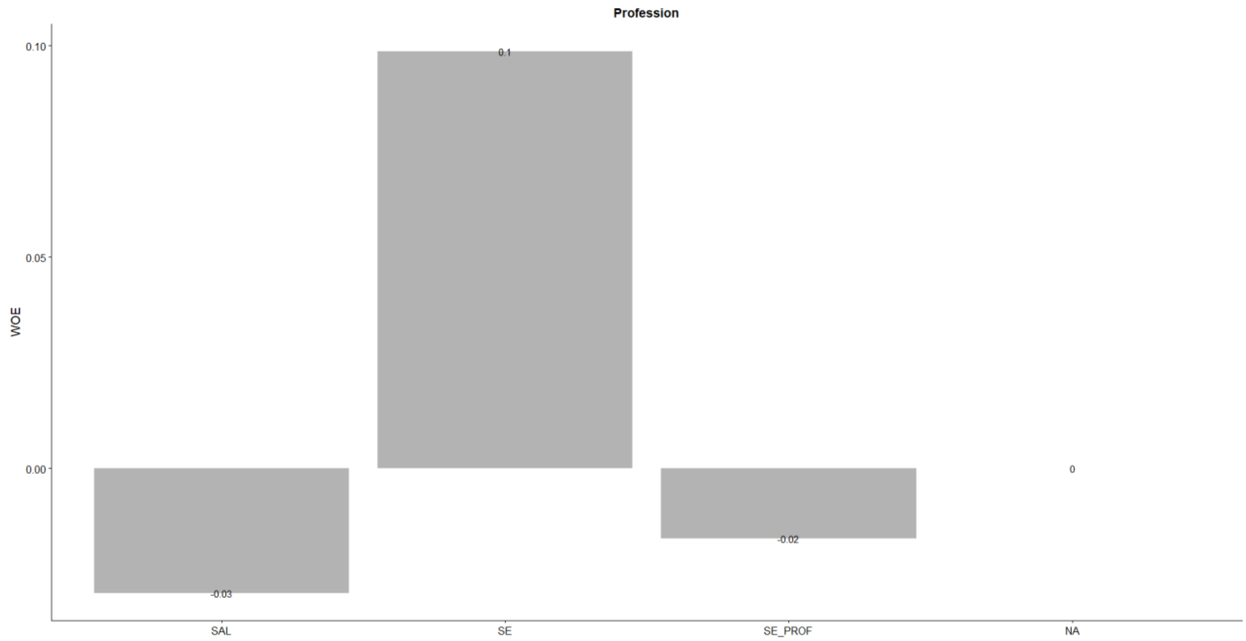

```
# NA bucket WOE is close to WOE of bucket [1357399:2960994]. Hence replace with random values
in this bucket.
master$Outstanding.Balance[which(is.na(master$Outstanding.Balance))] <- sample(1357399:2960994
, 272,replace=T)

plot_infotables(infoTables, "Education", show_values=TRUE)
```



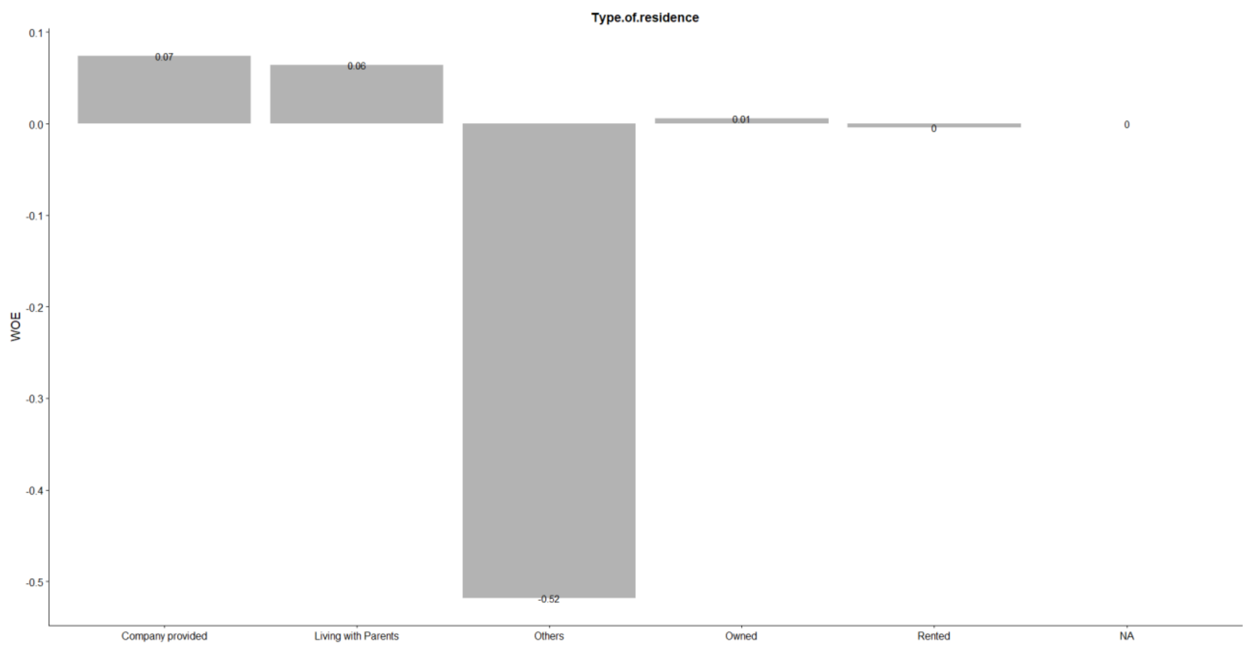

```
# NA bucket WOE is close to WOE of bucket "Masters". Hence replace it with "Masters".
master$Education[which(is.na(master$Education))] <- "Masters"

plot_infotables(infoTables, "Profession", show_values=TRUE)
```



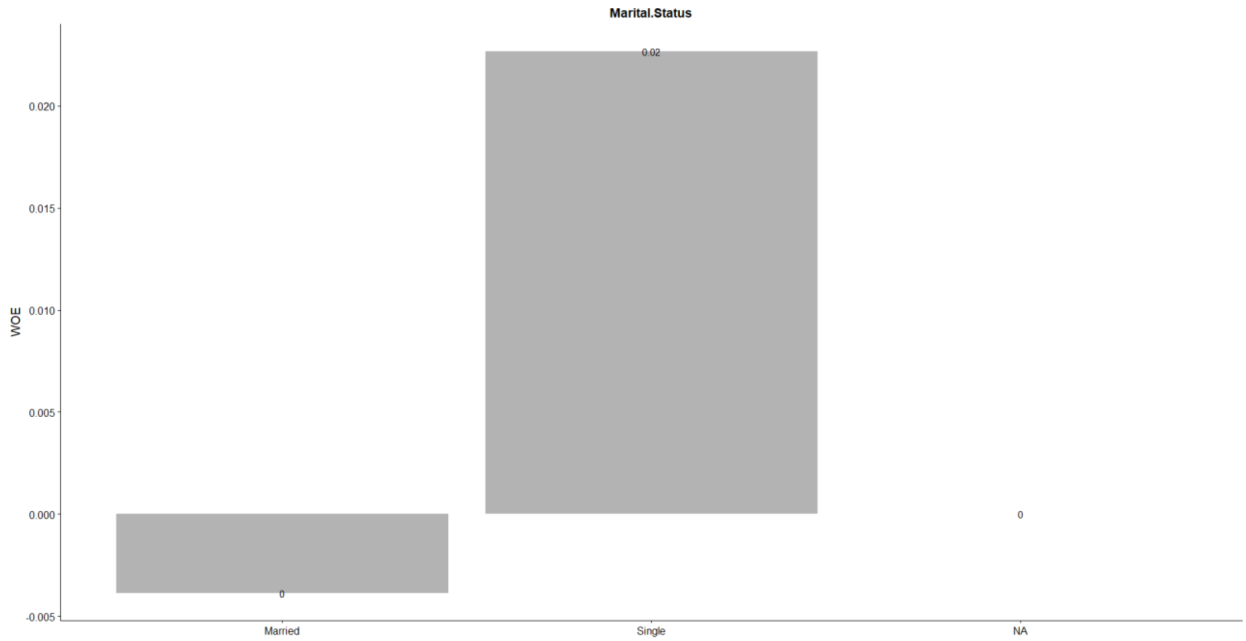

```r
# NA bucket WOE is close to WOE of bucket "SE_PROF". Hence replace it with "SE_PROF".
master$Profession[which(is.na(master$Profession))] <- "SE_PROF"

plot_infotables(infoTables, "Type.of.residence", show_values=TRUE)
```



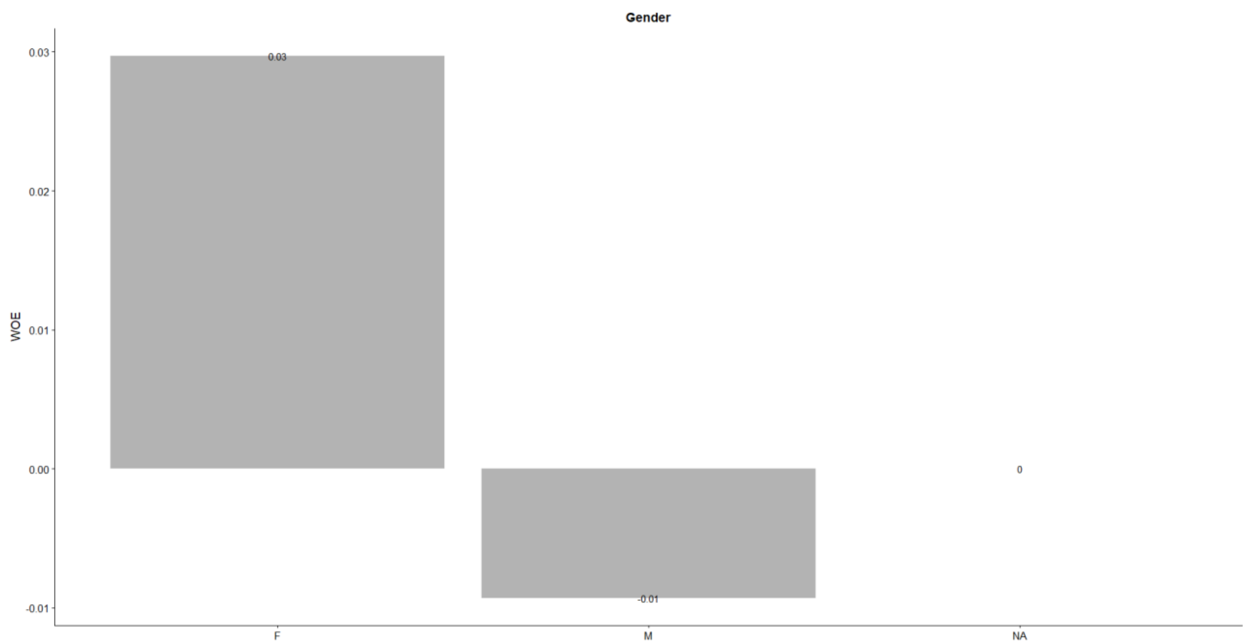

```r
# NA bucket WOE is close to WOE of bucket "Rented". Hence replace it with "Rented".
master$Type.of.residence[which(is.na(master$Type.of.residence))] <- "Rented"

plot_infotables(infoTables, "Marital.Status", show_values=TRUE)
```



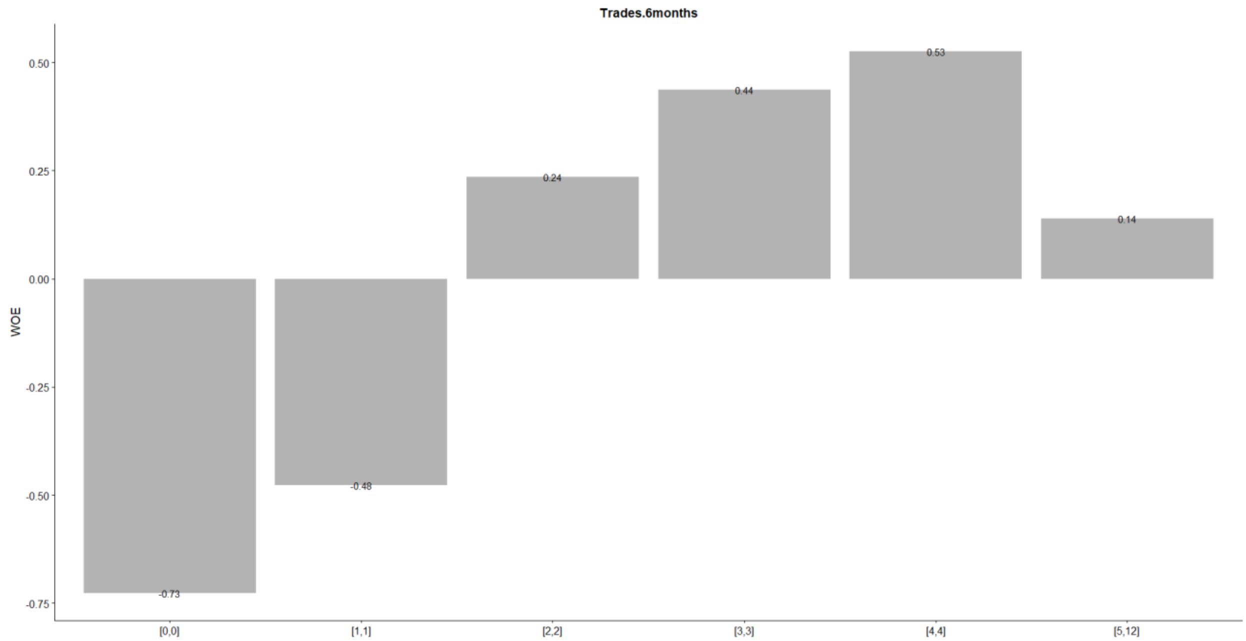

```r
# NA bucket WOE is close to WOE of bucket "Married". Hence replace it with "Married".
master$Marital.Status[which(is.na(master$Marital.Status))] <- "Married"

plot_infotables(infoTables, "Gender", show_values=TRUE)
```



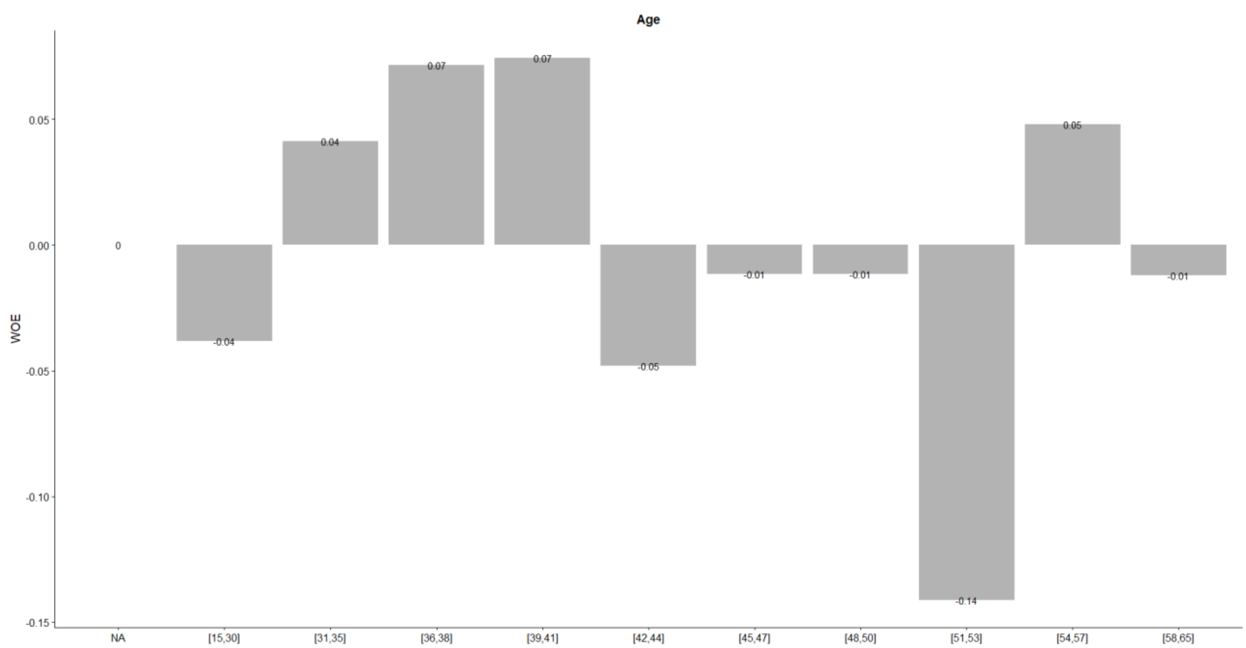

```r
# NA bucket WOE is close to WOE of bucket "M". Hence replace it with "M".
master$Gender[which(is.na(master$Gender))] <- "M"

plot_infotables(infoTables, "Trades.6months", show_values=TRUE)
```



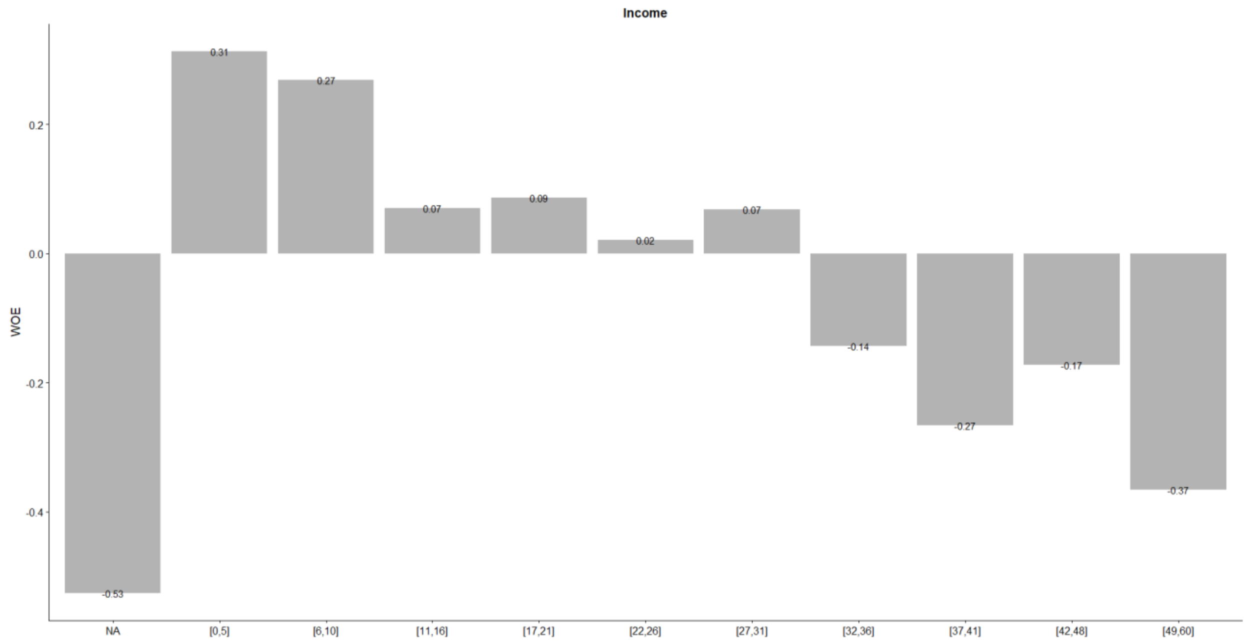

```
# NA bucket WOE is close to WOE of bucket [5,12]. Hence replace it with random values in this bucket.
master$Trades.6months[which(is.na(master$Trades.6months))] <- sample(5:12,1,replace=T)

plot_infotables(infoTables, "Age", show_values=TRUE)
```



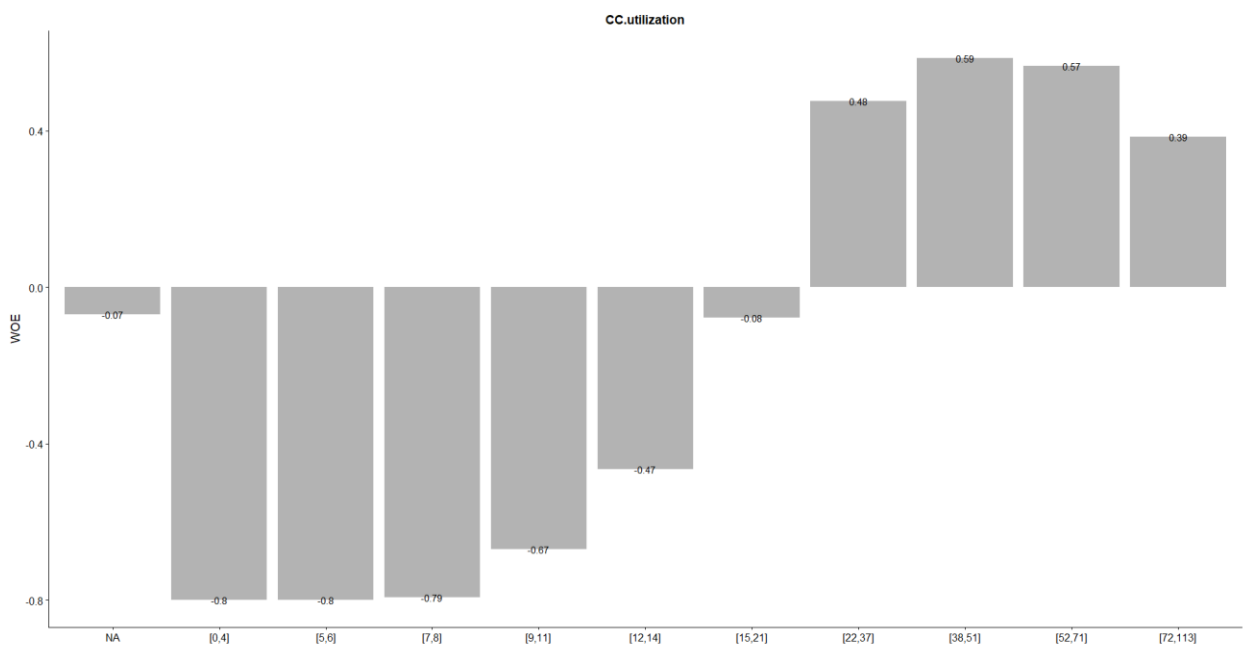

```
# NA bucket WOE is close to WOE of bucket [45,47],[48,50] and [58,65].
# Hence replace it with random values in this bucket.
master$Age[which(is.na(master$Age))] <- sample(c(45:50,58:65),20,replace=T)

plot_infotables(infoTables, "Income", show_values=TRUE)
```




```r
# NA bucket WOE is close to WOE of bucket [49,60].
# Hence replace it with random values in this bucket.
master$Income[which(is.na(master$Income))] <- sample(c(49:60),79,replace=T)

plot_infotables(infoTables, "CC.utilization", show_values=TRUE)
```




```r
# Unlike other missing variables, we will fix the CC.utilization such that the missing values
# are considered as a separate bucket
```

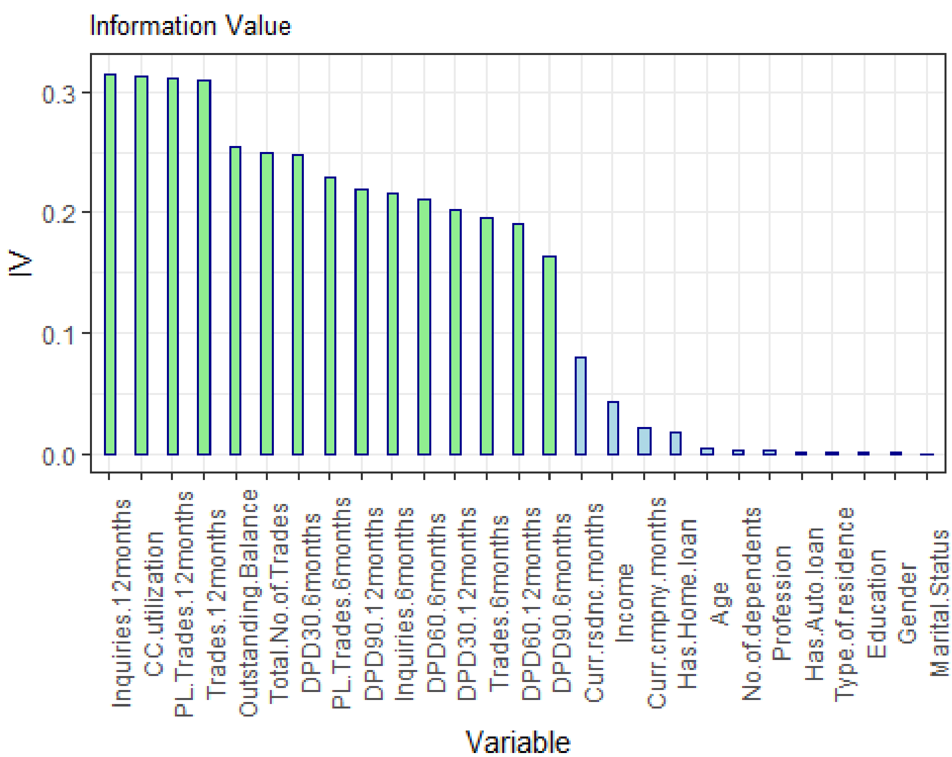
Create infotables again since the values have changed.

```r
infoTables <- create_infotables(data = master,
                                y = "Performance.Tag",
                                bins = 10,
                                parallel = T)

plot_infotables(infoTables, infoTables$Summary$Variable, same_scales=TRUE)
```

```r
plotFrame <- infoTables$Summary[order(-infoTables$Summary$IV), ]
plotFrame$Variable <- factor(plotFrame$Variable,
                             levels = plotFrame$Variable[order(-plotFrame$IV)])

ggplot(plotFrame, aes(x = Variable, y = IV)) +
  geom_bar(width = .35, stat = "identity", color = "darkblue",
           fill = "lightblue") +
  geom_bar(data  = filter(plotFrame, IV >= 0.10),
           width = .35, stat = "identity", color = "darkblue",
           fill  = "lightgreen") +
  ggtitle("Information Value") +
  theme_bw() +
  theme(plot.title  = element_text(size = 10)) +
  theme(axis.text.x = element_text(angle = 90))
```



Fix CC utilization with NA values so that it falls under a separate bucket

```r
infoTables$Tables$CC.utilization$CC.utilization[1] <- "[-1,-1]"
master$CC.utilization[which(is.na(master$CC.utilization))] <- -1
```

The below function will parse the Infotables and replace the WOE value for the corresponding variable value in the Master dataframe

```r
woe_replace <- function(df, IV) {
  df_clmtyp  <- data.frame(clmtyp = sapply(df, class))
  df_col_typ <- data.frame(clmnm = colnames(df), clmtyp = df_clmtyp$clmtyp)
  for (rownm in 1:nrow(df_col_typ)) {
    colmn_nm <- toString(df_col_typ$clmnm[rownm])
    if(colmn_nm %in% names(IV$Tables)){
      column_woe_df <- cbind(data.frame(IV$Tables[[toString(df_col_typ$clmnm[rownm])]]))
      if (df_col_typ$clmtyp[rownm] == "character") {
        df <- dplyr::inner_join(df, column_woe_df[,c(colmn_nm,"WOE")], by = colmn_nm,
                                type = "inner", match = "all")
        df[colmn_nm] <- NULL
        colnames(df)[colnames(df)=="WOE"] <- colmn_nm
      }
      else if (df_col_typ$clmtyp[rownm] == "numeric" | df_col_typ$clmtyp[rownm] == "integer")
{
        column_woe_df$lv<-as.numeric(str_sub(column_woe_df[,colmn_nm],
                                             regexpr("\\[", column_woe_df[,colmn_nm]) + 1,
                                             regexpr(",", column_woe_df[,colmn_nm]) - 1))
        column_woe_df$uv<-as.numeric(str_sub(column_woe_df[,colmn_nm],
                                             regexpr(",", column_woe_df[,colmn_nm]) + 1,
                                             regexpr("\\]", column_woe_df[,colmn_nm]) - 1))
        column_woe_df[colmn_nm] <- NULL
        column_woe_df <- column_woe_df[,c("lv","uv","WOE")]
        colnames(df)[colnames(df)==colmn_nm]<-"WOE_temp"
        df <-
          fuzzy_inner_join( df, column_woe_df[,c("lv","uv","WOE")],
                            by = c("WOE_temp"="lv","WOE_temp"="uv"),
                            match_fun=list(`>=`,`<=`))
        df["WOE_temp"]<-NULL
        df["lv"]<-NULL
        df["uv"]<-NULL
        colnames(df)[colnames(df)=="WOE"]<-colmn_nm
      }
    }
  }
  return(df)
}

woe_data <-  woe_replace(master, infoTables)
glimpse(woe_data)

## Observations: 69,304
## Variables: 28
## $ Performance.Tag     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ...
## $ Age                 <dbl> -0.01183112, -0.14120233, 0.07442647, -0.0...
## $ Gender              <dbl> -0.009350361, -0.009350361, -0.009350361, ...
## $ Marital.Status      <dbl> -0.003984326, -0.003984326, -0.003984326, ...
## $ No.of.dependents    <dbl> 0.007116783, -0.024128351, 0.042234398, -0...
## $ Income              <dbl> 0.02097236, -0.17246051, 0.07042667, -0.17...
## $ Education           <dbl> 0.001147737, -0.015965398, 0.001147737, 0....
## $ Profession          <dbl> 0.09871947, -0.02954263, 0.09871947, -0.02...
## $ Type.of.residence   <dbl> -0.004566674, -0.004566674, -0.004566674, ...
## $ Curr.rsdnc.months   <dbl> -0.27334573, -0.27334573, -0.27334573, -0....
## $ Curr.cmpny.months   <dbl> 0.03962433, -0.17095973, 0.20289770, 0.202...
## $ DPD90.6months       <dbl> 0.6251995, -0.2655128, -0.2655128, -0.2655...
## $ DPD60.6months       <dbl> 0.6250858, -0.3429386, -0.3429386, -0.3429...
## $ DPD30.6months       <dbl> 0.7446160, -0.3946889, -0.3946889, -0.3946...
## $ DPD90.12months      <dbl> 0.7254189, -0.3638612, -0.3638612, -0.3638...
## $ DPD60.12months      <dbl> 0.6969692, -0.3599566, -0.3599566, -0.3599...
## $ DPD30.12months      <dbl> 0.2834616, -0.3852153, -0.3852153, -0.3852...
## $ CC.utilization      <dbl> 0.58646657, -0.79988086, -0.79962482, -0.4...
```

```
## $ Trades.6months      <dbl> 0.4368879, -0.4776888, -0.7267022, -0.4776...
## $ Trades.12months     <dbl> 0.449730351, -0.814591270, -0.849424242, -...
## $ PL.Trades.6months   <dbl> 0.4403132, -0.6734343, -0.6734343, -0.6734...
## $ PL.Trades.12months  <dbl> 0.5018172, -0.9385167, -0.9385167, -0.9385...
## $ Inquiries.6months   <dbl> 0.21789213, -0.75288131, -0.75288131, -0.7...
## $ Inquiries.12months  <dbl> 0.48588941, -1.14169574, -1.14169574, -1.1...
## $ Has.Home.loan       <dbl> 0.07398413, -0.23669058, -0.23669058, 0.07...
## $ Outstanding.Balance <dbl> 0.46682662, -0.37102816, -0.37102816, -0.9...
## $ Total.No.of.Trades  <dbl> 0.38112947, -0.70014717, -0.70014717, -1.0...
## $ Has.Auto.loan       <dbl> 0.01193732, 0.01193732, 0.01193732, 0.0119...
```

All 69304 observations and 27 (excluding Performance.Tag) variables from the Master dataframe has been replaced with WOE values.

Convert to Factors:

```
master$Performance.Tag <- as.factor(master$Performance.Tag)
master$Has.Home.loan <- as.factor(master$Has.Home.loan)
master$Has.Auto.loan <- as.factor(master$Has.Auto.loan)
master <- master %>% mutate_if(is.character,as.factor)
```

Lets create Categorical & Continuous variable vectors

```
catvarnames  <- names(Filter(is.factor, master))
contvarnames <- names(Filter(is.numeric, master))
```

Lets look at summary once more:

```
sapply(master[catvarnames], table)

## $Performance.Tag
## 
##     0     1 
## 66386  2918 
## 
## $Gender
## 
##     F     M 
## 16367 52937 
## 
## $Marital.Status
## 
## Married  Single 
##   59055   10249 
## 
## $Education
## 
##     Bachelor      Masters       Others          Phd Professional 
##        17152        23416          116         4431        24189 
## 
## $Profession
## 
##     SAL      SE SE_PROF 
##   39361   13810   16133 
## 
## $Type.of.residence
## 
##    Company provided Living with Parents              Others 
##                1593                1765                 196 
##               Owned              Rented 
##               13890               51860
```

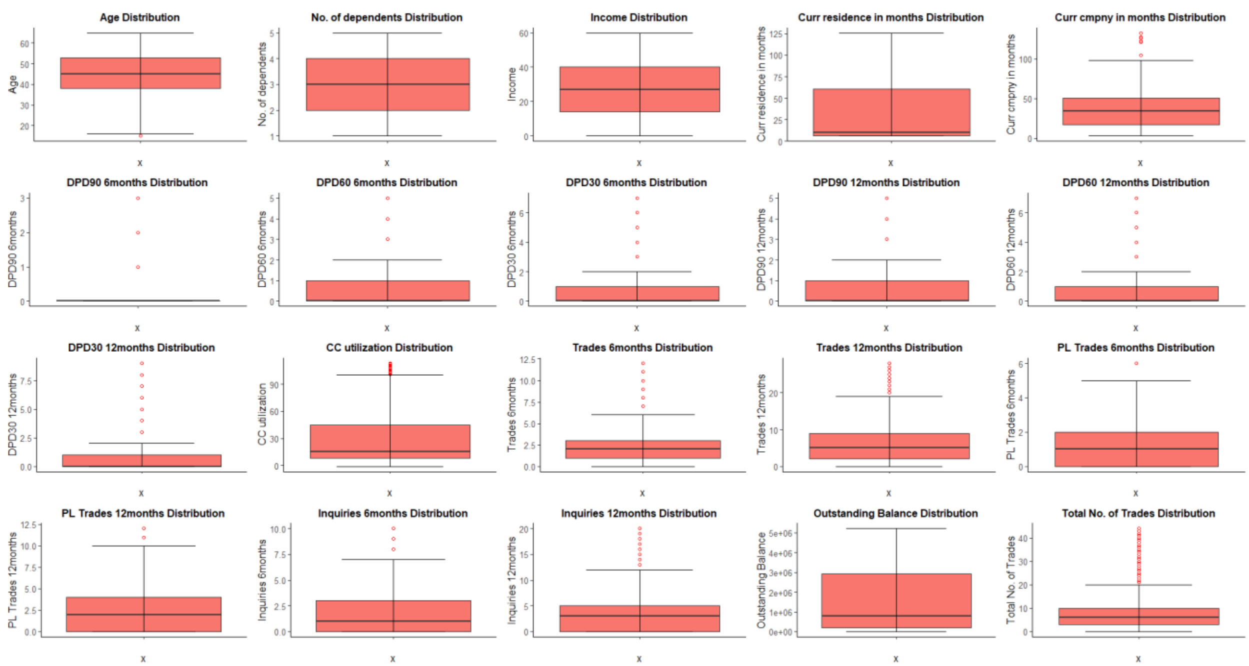
```
##
## $Has.Home.loan
##
##     0     1
## 50961 18343
##
## $Has.Auto.loan
##
##     0     1
## 63374  5930
```

```r
sapply(master[contvarnames], summary)
```

```
##              Age No.of.dependents   Income Curr.rsdnc.months
## Min.    15.00000         1.000000  0.00000           6.00000
## 1st Qu. 38.00000         2.000000 14.00000           6.00000
## Median  45.00000         3.000000 27.00000          10.00000
## Mean    45.01955         2.859085 27.47454          34.58255
## 3rd Qu. 53.00000         4.000000 40.00000          61.00000
## Max.    65.00000         5.000000 60.00000         126.00000
##         Curr.cmpny.months DPD90.6months DPD60.6months DPD30.6months
## Min.              3.00000       0.00000     0.0000000     0.0000000
## 1st Qu.          17.00000       0.00000     0.0000000     0.0000000
## Median           34.00000       0.00000     0.0000000     0.0000000
## Mean             34.23146       0.25101     0.3948834     0.5277906
## 3rd Qu.          51.00000       0.00000     1.0000000     1.0000000
## Max.            133.00000       3.00000     5.0000000     7.0000000
##         DPD90.12months DPD60.12months DPD30.12months CC.utilization
## Min.         0.0000000      0.0000000      0.0000000       -1.00000
## 1st Qu.      0.0000000      0.0000000      0.0000000        8.00000
## Median       0.0000000      0.0000000      0.0000000       15.00000
## Mean         0.4182298      0.6083343      0.7398996       29.06564
## 3rd Qu.      1.0000000      1.0000000      1.0000000       45.00000
## Max.         5.0000000      7.0000000      9.0000000      113.00000
##         Trades.6months Trades.12months PL.Trades.6months
## Min.          0.000000        0.000000          0.000000
## 1st Qu.       1.000000        2.000000          0.000000
## Median        2.000000        5.000000          1.000000
## Mean          2.303734        5.832189          1.199281
## 3rd Qu.       3.000000        9.000000          2.000000
## Max.         12.000000       28.000000          6.000000
##         PL.Trades.12months Inquiries.6months Inquiries.12months
## Min.              0.000000          0.000000           0.000000
## 1st Qu.           0.000000          0.000000           0.000000
## Median            2.000000          1.000000           3.000000
## Mean              2.382763          1.772336           3.553821
## 3rd Qu.           4.000000          3.000000           5.000000
## Max.             12.000000         10.000000          20.000000
##         Outstanding.Balance Total.No.of.Trades
## Min.                    0.0           0.000000
## 1st Qu.            213134.8           3.000000
## Median             778658.5           6.000000
## Mean              1267146.0           8.241458
## 3rd Qu.           2927999.0          10.000000
## Max.              5218801.0          44.000000
```

```r
plot_grid(ContUnivar(master$Age, "Age"),
          ContUnivar(master$No.of.dependents, "No. of dependents"),
          ContUnivar(master$Income, "Income"),
          ContUnivar(master$Curr.rsdnc.months, "Curr residence in months"),
          ContUnivar(master$Curr.cmpny.months, "Curr cmpny in months"),
```

```
          ContUnivar(master$DPD90.6months, "DPD90 6months"),
          ContUnivar(master$DPD60.6months, "DPD60 6months"),
          ContUnivar(master$DPD30.6months, "DPD30 6months"),
          ContUnivar(master$DPD90.12months, "DPD90 12months"),
          ContUnivar(master$DPD60.12months, "DPD60 12months"),
          ContUnivar(master$DPD30.12months, "DPD30 12months"),
          ContUnivar(master$CC.utilization, "CC utilization"),
          ContUnivar(master$Trades.6months, "Trades 6months"),
          ContUnivar(master$Trades.12months, "Trades 12months"),
          ContUnivar(master$PL.Trades.6months, "PL Trades 6months"),
          ContUnivar(master$PL.Trades.12months, "PL Trades 12months"),
          ContUnivar(master$Inquiries.6months, "Inquiries 6months"),
          ContUnivar(master$Inquiries.12months, "Inquiries 12months"),
          ContUnivar(master$Outstanding.Balance, "Outstanding Balance"),
          ContUnivar(master$Total.No.of.Trades, "Total No. of Trades"))
```



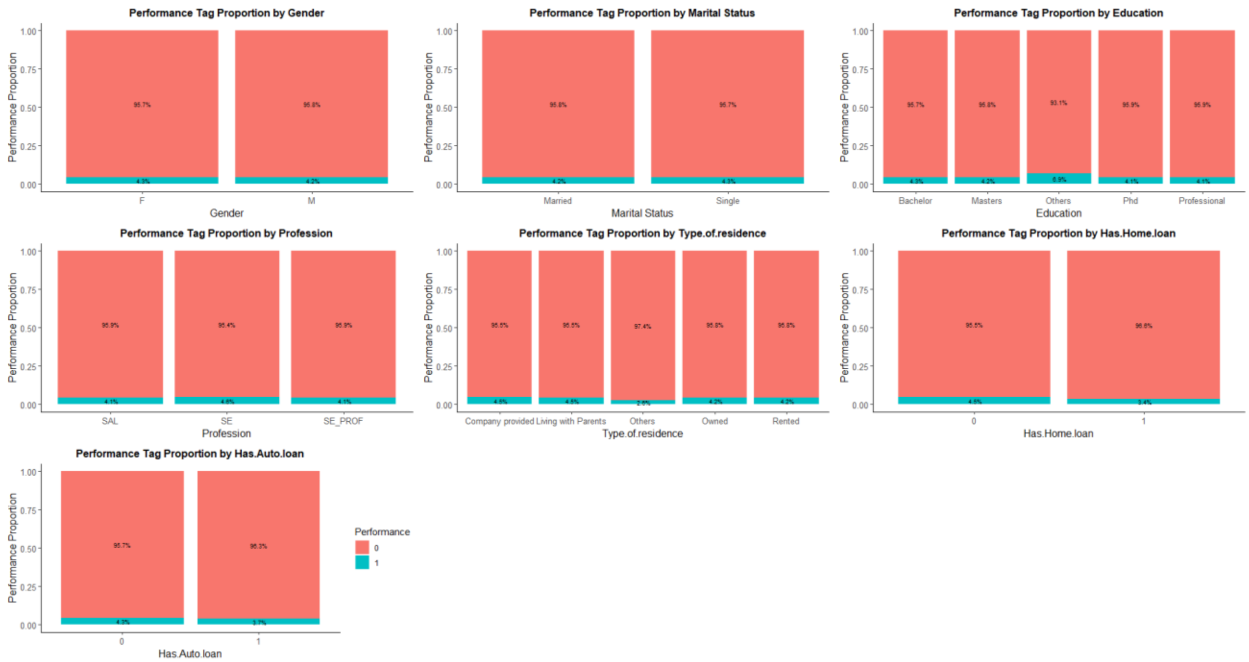

```
# treating outliers
# master$CC.utilization <- treatoutlier(master$CC.utilization)
#
#
# To be done if required
#
#
#
#

# Summary of observations:
```

Multivariate Analysis(Categorical Variables)

```
plot_grid(CatBivar(master$Gender, master$Performance.Tag,
                    "Gender", "Performance Tag"),
          CatBivar(master$Marital.Status, master$Performance.Tag,
                    "Marital Status",  "Performance Tag"),
          CatBivar(master$Education, master$Performance.Tag,
                   "Education", "Performance Tag"),
          CatBivar(master$Profession, master$Performance.Tag,
```

```
                    "Profession", "Performance Tag"),
          CatBivar(master$Type.of.residence, master$Performance.Tag,
                   "Type.of.residence", "Performance Tag"),
          CatBivar(master$Has.Home.loan, master$Performance.Tag,
                   "Has.Home.loan", "Performance Tag"),
          CatBivar(master$Has.Auto.loan, master$Performance.Tag,
                   "Has.Auto.loan", "Performance Tag") + theme(legend.position = 'right'))
```

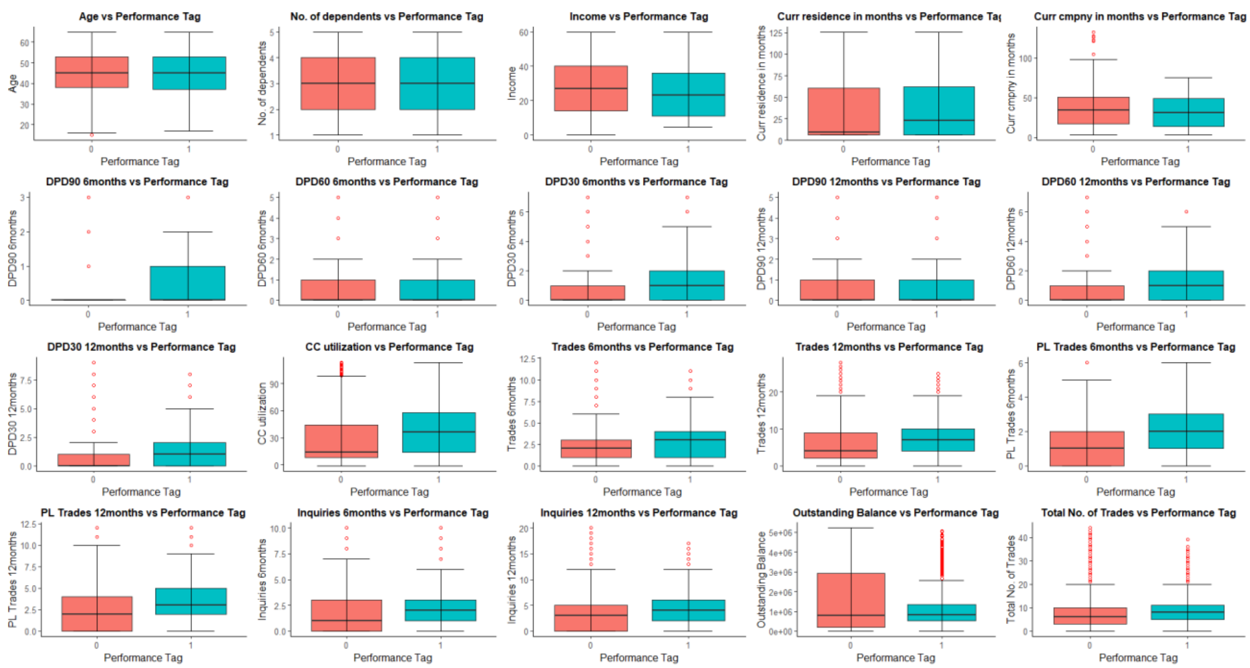


Summary of observations: The categorical variables doesn't seem to much impact on the Target variable. This confirms our analysis using IV values where all the categorical variables had very minimal IV values.

Multivariate Analysis(Performance vs Cont Variables)

```
plot_grid(ContCatBivar(master$Performance.Tag, master$Age,"Performance Tag", "Age"),
          ContCatBivar(master$Performance.Tag, master$No.of.dependents,"Performance Tag", "No.
of dependents"),
          ContCatBivar(master$Performance.Tag, master$Income,"Performance Tag", "Income"),
          ContCatBivar(master$Performance.Tag, master$Curr.rsdnc.months,"Performance Tag", "Cu
rr residence in months"),
          ContCatBivar(master$Performance.Tag, master$Curr.cmpny.months,"Performance Tag", "Cu
rr cmpny in months"),
          ContCatBivar(master$Performance.Tag, master$DPD90.6months,"Performance Tag", "DPD90
6months"),
          ContCatBivar(master$Performance.Tag, master$DPD60.6months,"Performance Tag", "DPD60
6months"),
          ContCatBivar(master$Performance.Tag, master$DPD30.6months,"Performance Tag", "DPD30
6months"),
          ContCatBivar(master$Performance.Tag, master$DPD90.12months,"Performance Tag", "DPD90
12months"),
          ContCatBivar(master$Performance.Tag, master$DPD60.12months,"Performance Tag", "DPD60
12months"),
          ContCatBivar(master$Performance.Tag, master$DPD30.12months,"Performance Tag", "DPD30
12months"),
          ContCatBivar(master$Performance.Tag, master$CC.utilization,"Performance Tag", "CC ut
ilization"),
          ContCatBivar(master$Performance.Tag, master$Trades.6months,"Performance Tag", "Trade
s 6months"),
```

```
        ContCatBivar(master$Performance.Tag, master$Trades.12months,"Performance Tag", "Trades 12months"),
        ContCatBivar(master$Performance.Tag, master$PL.Trades.6months,"Performance Tag", "PL Trades 6months"),
        ContCatBivar(master$Performance.Tag, master$PL.Trades.12months,"Performance Tag","PL Trades 12months"),
        ContCatBivar(master$Performance.Tag, master$Inquiries.6months,"Performance Tag", "Inquiries 6months"),
        ContCatBivar(master$Performance.Tag, master$Inquiries.12months,"Performance Tag", "Inquiries 12months"),
        ContCatBivar(master$Performance.Tag, master$Outstanding.Balance,"Performance Tag", "Outstanding Balance"),
        ContCatBivar(master$Performance.Tag, master$Total.No.of.Trades,"Performance Tag", "Total No. of Trades"))
```



Summary of observations: Performance.Tag is poor/defaulting rate is higher for Applicants with:

- Lower Incomes
- Higher DPD30 6months
- Higher DPD60 6months
- Higher DPD90 6months
- Higher DPD30 12months
- Higher DPD60 12months
- Higher Credit card utilization
- Higher trades (both normal & PL trades)
- Higher Inquiries

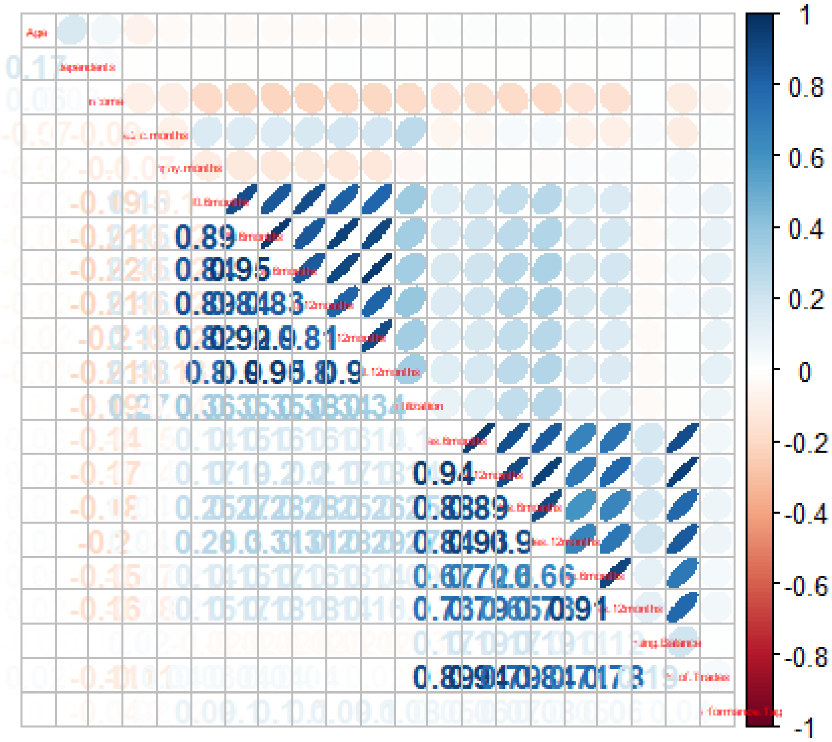
Age, No. of Dependents, No. of months in Cuurent residence or current company or Outsanding balance doesn't seem to be good predictors of the target variable.

This again confirms our analysis using IV values where we saw majority of the continuous variables are good predictors of the target variable.

Multivariate Analysis(Continuous Variables)

```r
master$Performance.Tag <- as.numeric(master$Performance.Tag)

corrplot.mixed(cor(cbind(master[contvarnames], master$Performance.Tag)),
               upper = "ellipse", tl.cex = 0.40, tl.pos = 'd')
```



Summary of observations:

There seems to be a number of dependant variables that are multicollinear.

- As expected response variable show come correlation with DPD, trades & Credit card utilization variables
- DPD variables seem seem to have strong multicollinearity within their group.
- Similarly varaibles related to Trades seem to have strong multicollinearity within the group
- Outstanding balance shows some +ve collinerality with Trades.
- All DPD variables show come collinerality with the Trades.
- Interestingly there is also some +ve correlation between No. of months in current resindence with CC utilization and DPD variables.

Model Building

Remove Application ID

```r
demographics_cols <- demographics_cols[2:12]
```

Subset only the demographics variables from the Master to build model based on demographics.

```r
demographic_model <- master %>% select_( .dots = demographics_cols)
```