Credit Scoring Project

Lương Hồng Nhung

2024-12-25

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
library(scales)
```

```
df <- read.csv("C:/Users/T480s/Downloads/hmeq.csv")
attach(df)
```

```
summary(df)
```

```
##       BAD              LOAN          MORTDUE          VALUE
##  Min.   :0.0000   Min.   : 1100   Min.   :  2063   Min.   :  8000
##  1st Qu.:0.0000   1st Qu.:11100   1st Qu.: 46276   1st Qu.: 66076
##  Median :0.0000   Median :16300   Median : 65019   Median : 89236
##  Mean   :0.1995   Mean   :18608   Mean   : 73761   Mean   :101776
##  3rd Qu.:0.0000   3rd Qu.:23300   3rd Qu.: 91488   3rd Qu.:119824
##  Max.   :1.0000   Max.   :89900   Max.   :399550   Max.   :855909
##                                   NA's   :518      NA's   :112
##     REASON              JOB                YOJ              DEROG
##  Length:5960        Length:5960        Min.   : 0.000   Min.   : 0.0000
##  Class :character   Class :character   1st Qu.: 3.000   1st Qu.: 0.0000
##  Mode  :character   Mode  :character   Median : 7.000   Median : 0.0000
```

```
##                                         Mean   : 8.922   Mean   : 0.2546
##                                         3rd Qu.:13.000   3rd Qu.: 0.0000
##                                         Max.   :41.000   Max.   :10.0000
##                                         NA's   :515      NA's   :708
##      DELINQ            CLAGE             NINQ            CLNO
##  Min.   : 0.0000   Min.   :   0.0   Min.   : 0.000   Min.   : 0.0
##  1st Qu.: 0.0000   1st Qu.: 115.1   1st Qu.: 0.000   1st Qu.:15.0
##  Median : 0.0000   Median : 173.5   Median : 1.000   Median :20.0
##  Mean   : 0.4494   Mean   : 179.8   Mean   : 1.186   Mean   :21.3
##  3rd Qu.: 0.0000   3rd Qu.: 231.6   3rd Qu.: 2.000   3rd Qu.:26.0
##  Max.   :15.0000   Max.   :1168.2   Max.   :17.000   Max.   :71.0
##  NA's   :580       NA's   :308      NA's   :510      NA's   :222
##     DEBTINC
##  Min.   :  0.5245
##  1st Qu.: 29.1400
##  Median : 34.8183
##  Mean   : 33.7799
##  3rd Qu.: 39.0031
##  Max.   :203.3121
##  NA's   :1267
```

Dữ liệu có nhiều quan sát NA

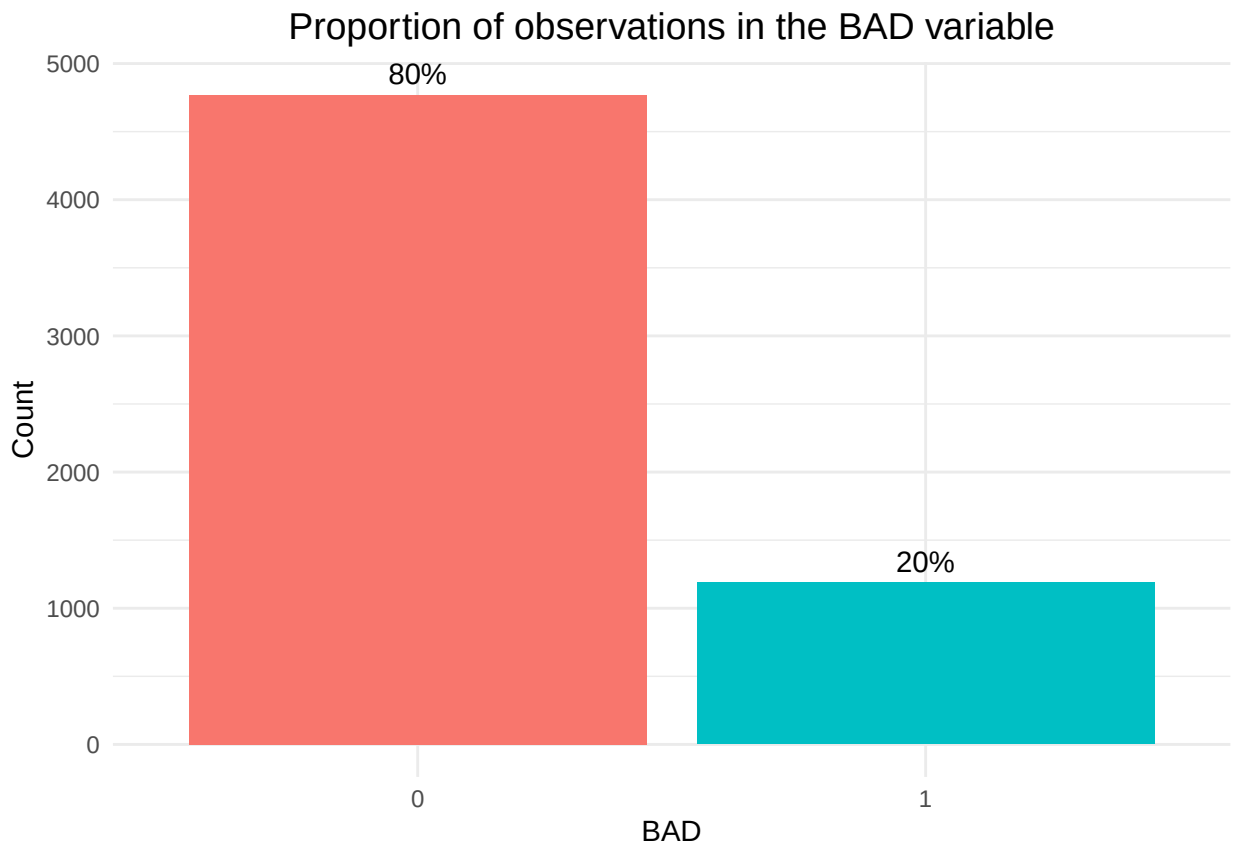
```
df$BAD <- as.factor(df$BAD)
```

Exploratory Data Analysis

Check imbalanced dataset

```
ggplot(df, aes(x = BAD, fill = BAD)) +
  geom_bar() +
  geom_text(stat = 'count', aes(label = scales::percent(..count.. / sum(..count..))), vjust = -0.5) +
  labs(x = "BAD", y = "Count", title = "Proportion of observations in the BAD variable") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 14), # Center the title
        legend.position = 'none')
```

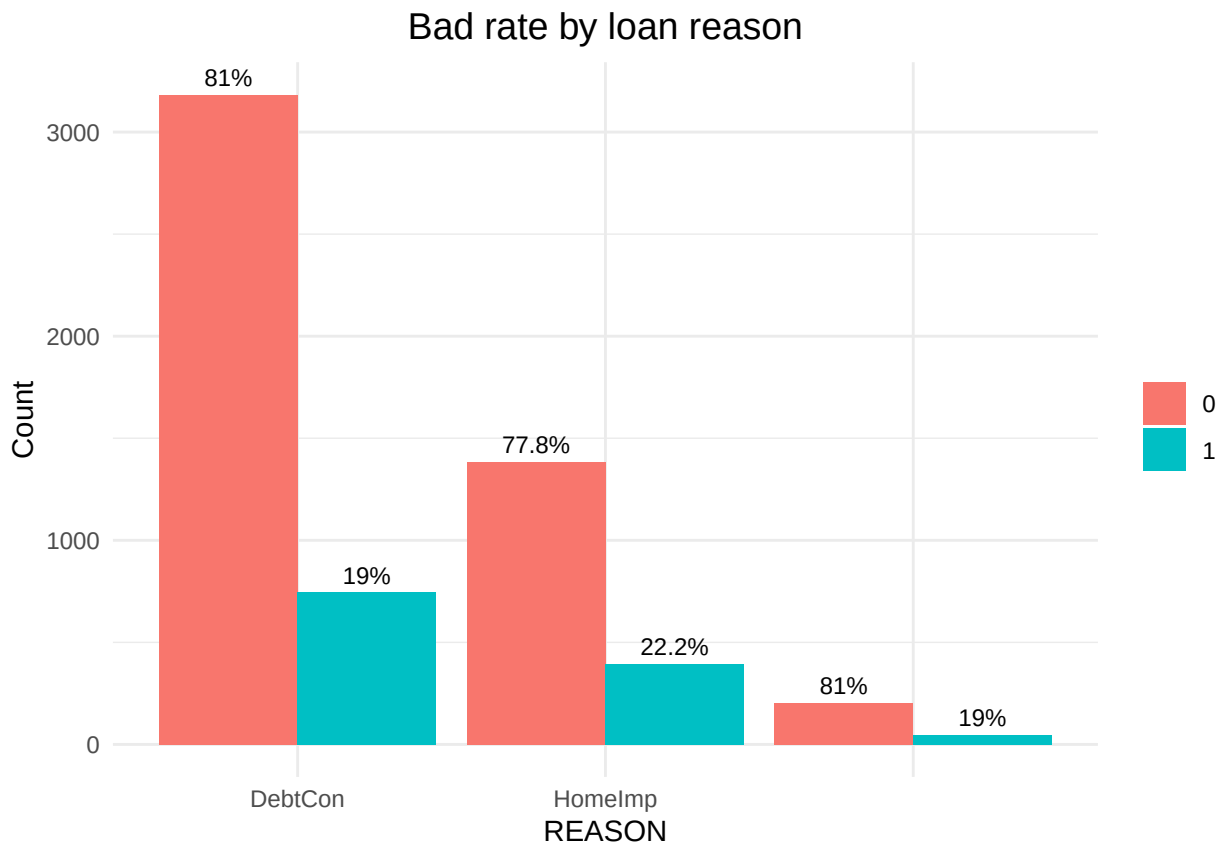
```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



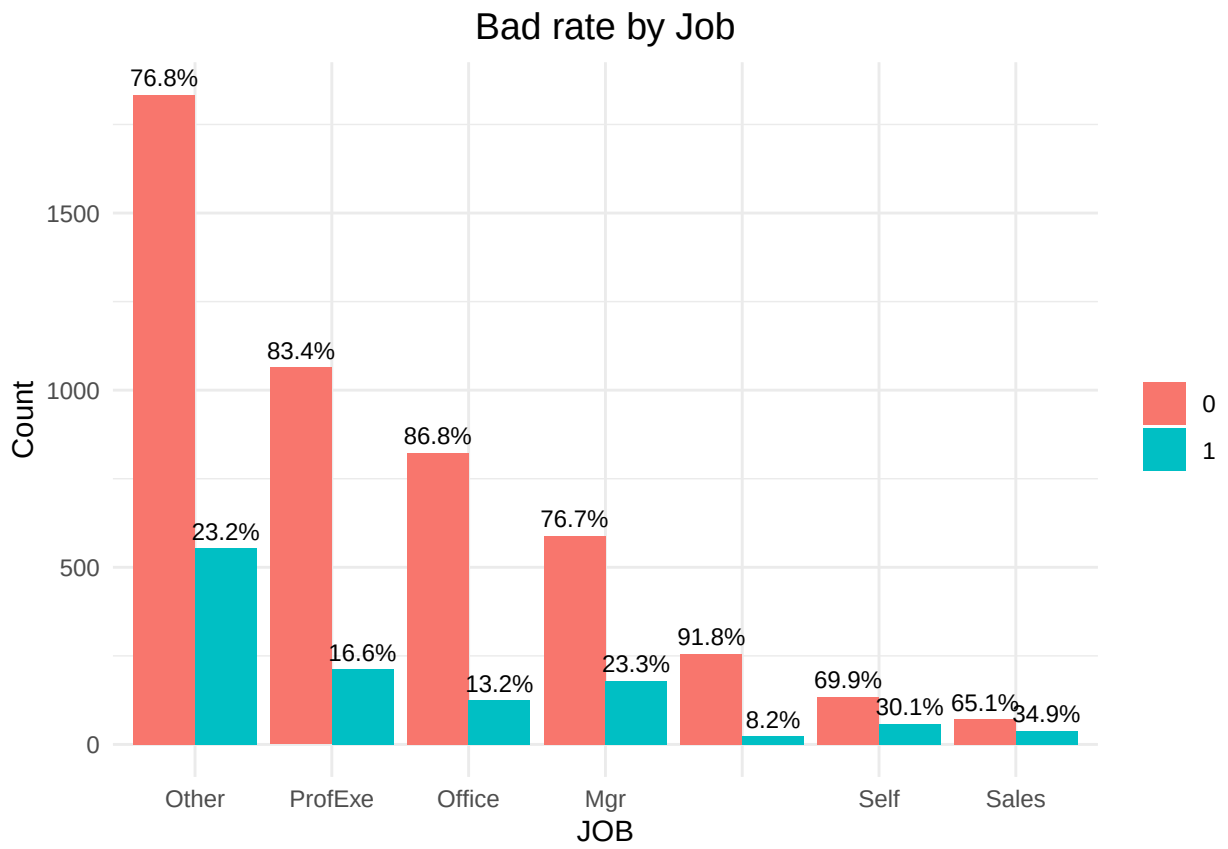
Bad rate > 15% nên dữ liệu không bị mất cân bằng

Data visualization

```
df %>%
  count(REASON, BAD) %>%
  group_by(REASON) %>%
  mutate(percent = n / sum(n) * 100) %>%
  ggplot(aes(x = reorder(REASON, -n), y = n, fill = BAD)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = paste0(round(percent, 1), "%")),
            position = position_dodge(width = 0.9),
            vjust = -0.5, size = 3) +
  labs(title = "Bad rate by loan reason",
       x = "REASON",
       y = "Count") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 14),
        legend.title = element_blank())
```




```
df %>%
  count(JOB, BAD) %>%
  group_by(JOB) %>%
  mutate(percent = n / sum(n) * 100) %>%
  ggplot(aes(x = reorder(JOB, -n), y = n, fill = BAD)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = paste0(round(percent, 1), "%")),
            position = position_dodge(width = 0.9),
            vjust = -0.5, size = 3) +
  labs(title = "Bad rate by Job",
       x = "JOB",
       y = "Count") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 14),
        legend.title = element_blank())
```



Handle missing value - Check missing value

```
df[df == ""] <- NA
colSums(is.na(df))
```

```
##     BAD    LOAN MORTDUE   VALUE  REASON     JOB     YOJ   DEROG  DELINQ   CLAGE
##       0       0     518     112     252     279     515     708     580     308
##    NINQ    CLNO DEBTINC
##     510     222    1267
```

Kết luận: Các phương pháp phổ biến để thay thế missing value thừng được sử dụng như thay thế bằng Mean hoặc Mode. Tuy nhiên, missing value có thể có ý nghĩa nên sẽ được giữ lại

```
# Chuyển đổi factor sang numeric
df$BAD <- as.numeric(as.character(df$BAD))
```

Chia tập dữ liệu train và test:

```
# train 70% - test 30%
set.seed(1230000)
ind <- sample(2, nrow(df), replace = TRUE, prob = c(0.7, 0.3))
train_data <- df [ind == 1, ]
test_data<- df [ind == 2, ]
```

Sử dụng mô hình Logit trong chấm điểm tín dụng:

- Tính chỉ số IV của các biến trên tập train_data

```
library('ROSE')
```

```
## Warning: package 'ROSE' was built under R version 4.2.3
```

```
## Loaded ROSE 0.0-4
```

```
IV <- Information::create_infotables(data = train_data, y = "BAD", parallel = FALSE)
print(IV$Summary)
```

```
##    Variable          IV
## 12  DEBTINC 1.902658214
## 8    DELINQ 0.505235083
## 3     VALUE 0.493648315
## 7     DEROG 0.327531151
## 9     CLAGE 0.215421107
## 10     NINQ 0.179189424
## 5       JOB 0.135148427
## 1      LOAN 0.125745413
## 6       YOJ 0.076973004
## 11     CLNO 0.071406374
## 2   MORTDUE 0.046687447
## 4    REASON 0.002455575
```

Kết luận: IV là chỉ số đo lường sức mạnh của từng biến trong mô hình. Biến nào có IV < 0.02 có nghĩa biến không có tác dụng trong việc phân loại Good/Bad và cần được loại bỏ

```
vars_removed <- IV$Summary %>% as.data.frame %>%
                                    subset(IV < 0.02) %>% pull(1)
vars_removed
```

```
## [1] "REASON"
```

```
train_data_removed<- train_data %>% dplyr::select(-all_of(vars_removed))
```

Bin các biến theo woe:

```
library("scorecard")
```

```
## Warning: package 'scorecard' was built under R version 4.2.3
```

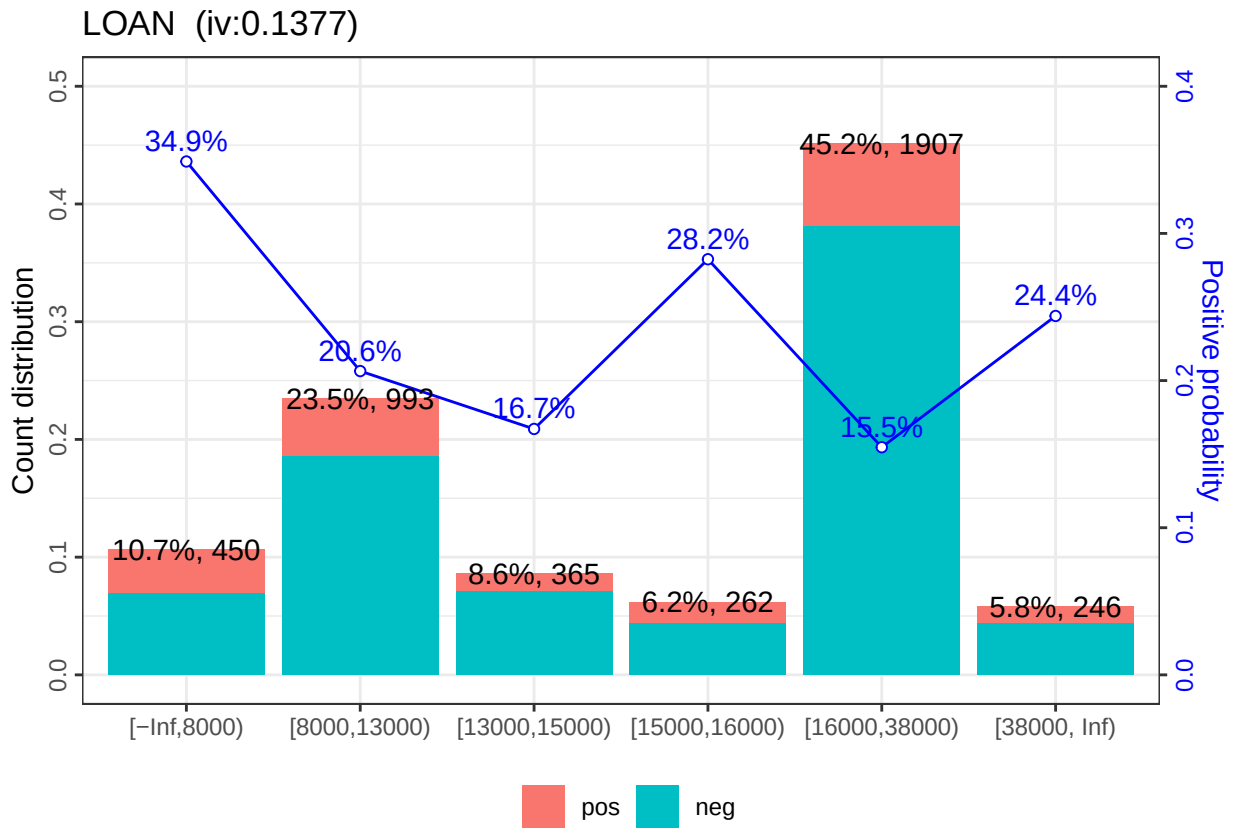
```
bins <- woebin(train_data_removed, y = "BAD")
```

```
## i Creating woe binning ...
```

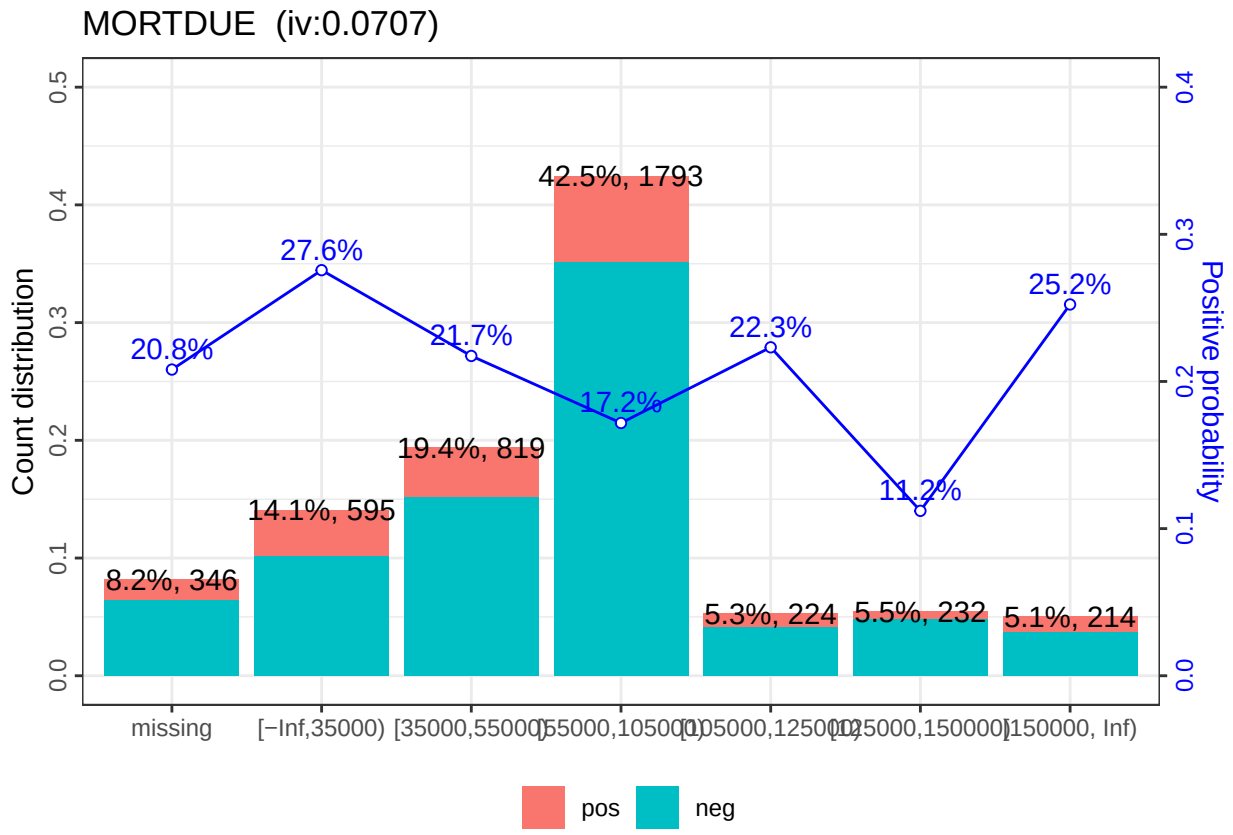
```
## v Binning on 4223 rows and 12 columns in 00:00:05
```

```
woebin_plot(bins)
```

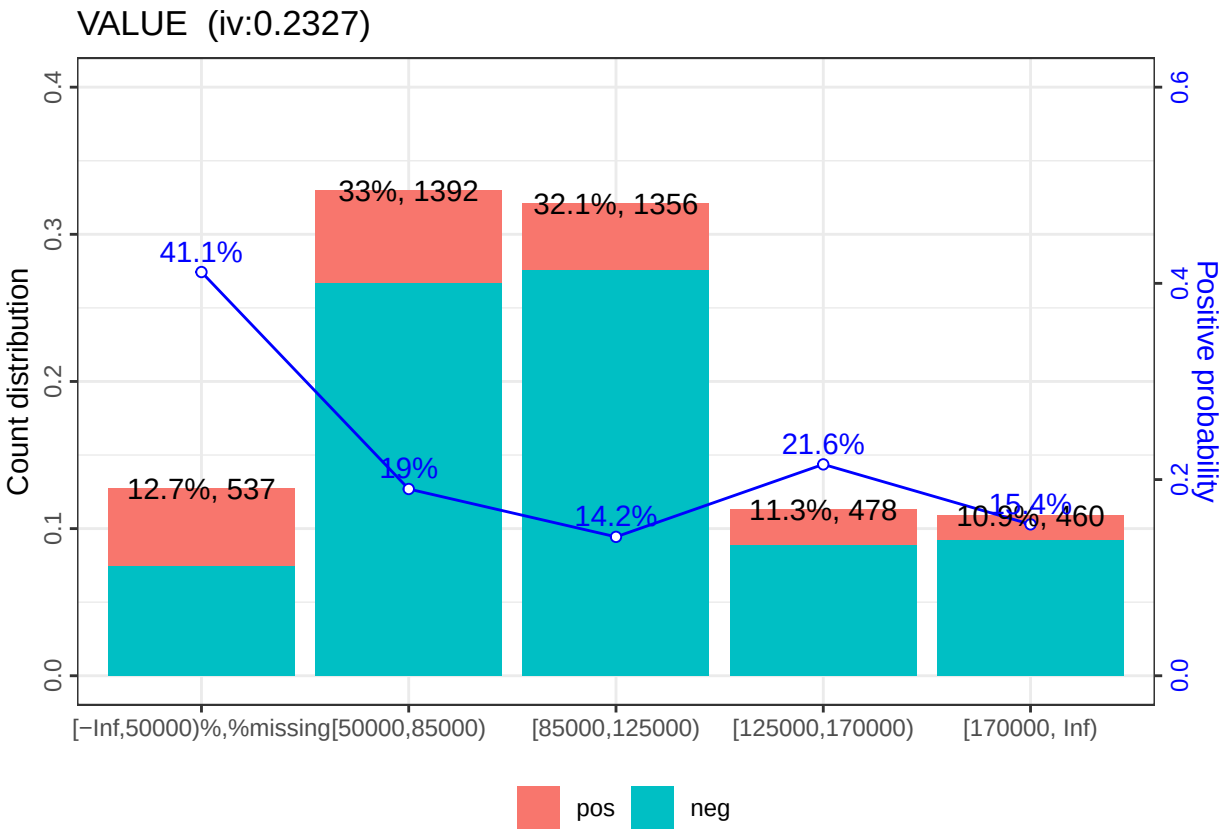
```
## $LOAN
```



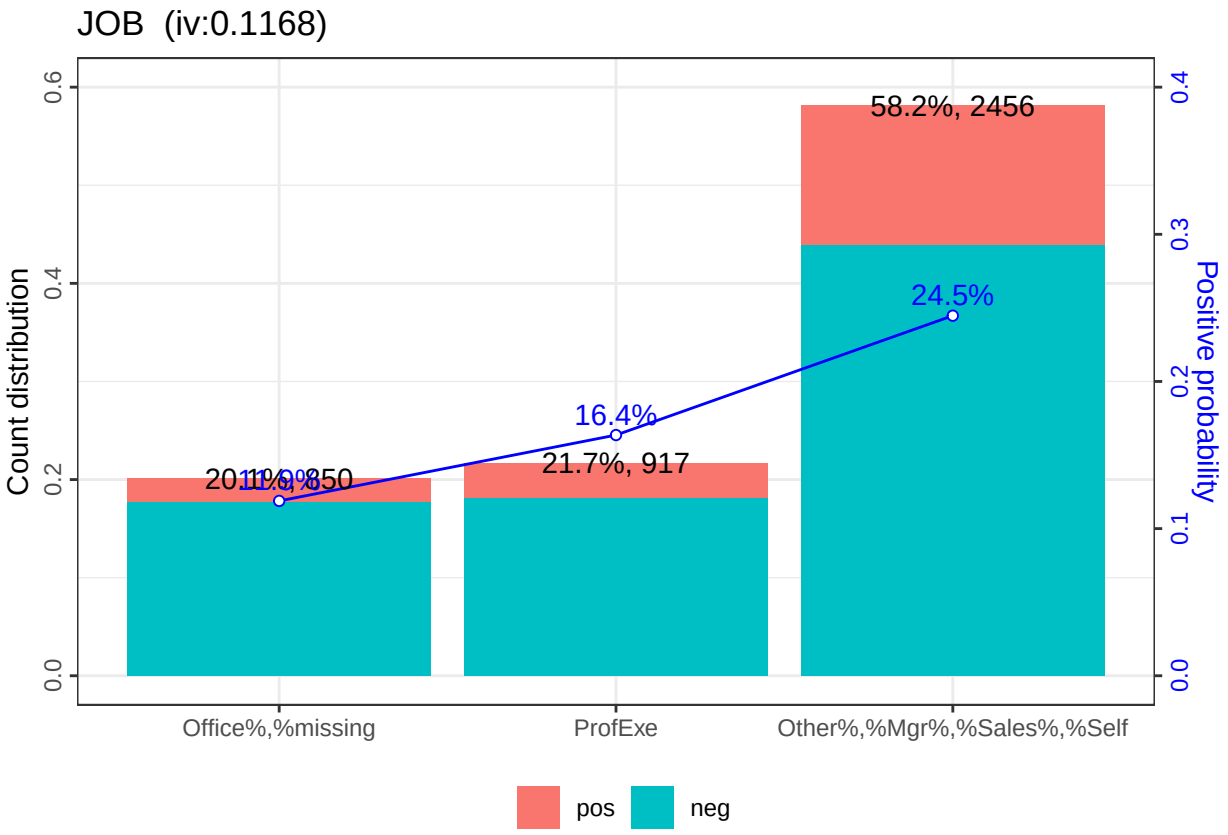

```
##
## $MORTDUE
```


MORTDUE (iv:0.0707)
Count distribution
Positive probability
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.1
0.2
0.3
0.4
8.2%, 346
14.1%, 595
19.4%, 819
42.5%, 1793
5.3%, 224
5.5%, 232
5.1%, 214
20.8%
27.6%
21.7%
17.2%
22.3%
11.2%
25.2%
missing
[−Inf,35000)
[35000,55000)
[55000,105000)
[105000,125000)
[125000,150000)
[150000, Inf)
pos
neg

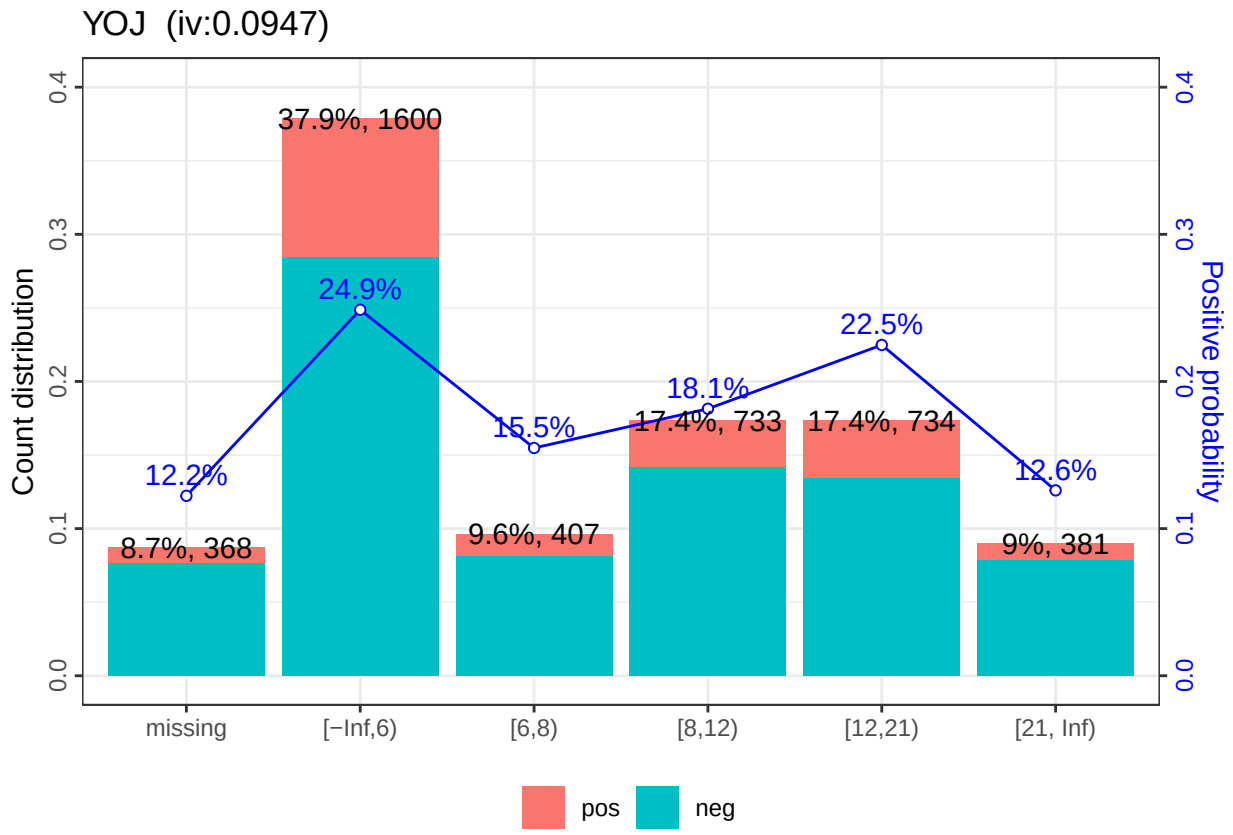
```
##
## $VALUE
```


VALUE (iv:0.2327)
Count distribution
Positive probability
0.0
0.1
0.2
0.3
0.4
0.6
12.7%, 537
33%, 1392
32.1%, 1356
11.3%, 478
10.9%, 460
41.1%
19%
14.2%
21.6%
15.4%
[−Inf,50000)%,%missing
[50000,85000)
[85000,125000)
[125000,170000)
[170000, Inf)
pos
neg

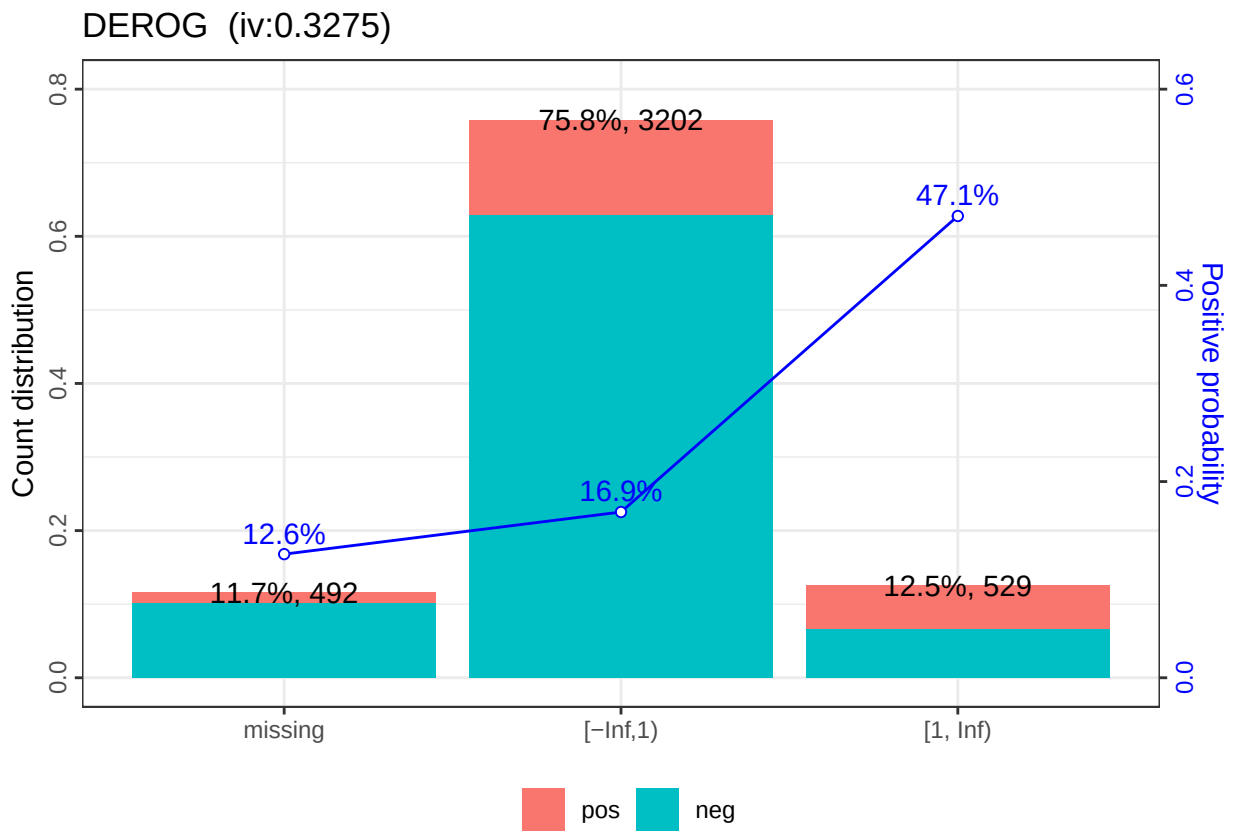
```
##
## $JOB
```


JOB (iv:0.1168)
Count distribution
Positive probability
0.0
0.2
0.4
0.6
0.1
0.3
11.0%, 850
16.4%
21.7%, 917
24.5%
58.2%, 2456
Office%,%missing
ProfExe
Other%,%Mgr%,%Sales%,%Self
pos
neg

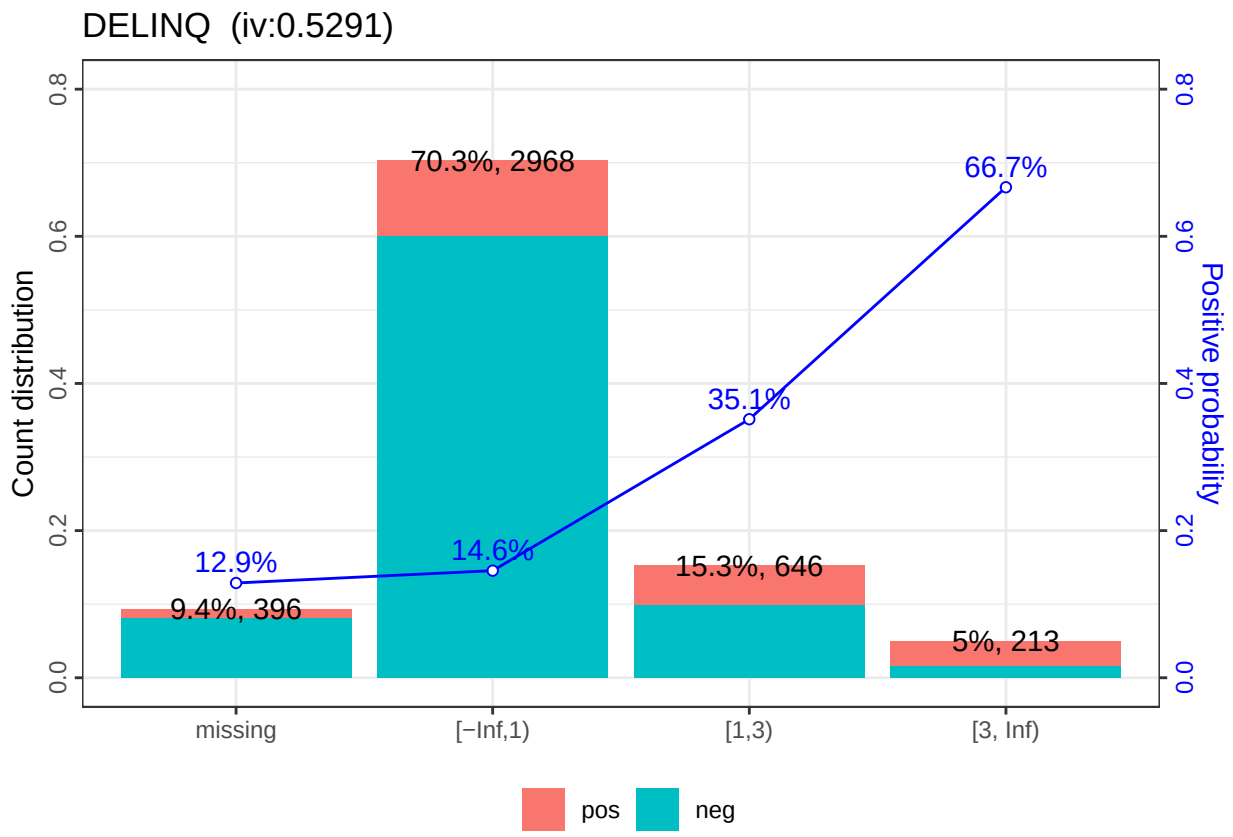
```
##
## $YOJ
```


YOJ (iv:0.0947)
Count distribution
Positive probability
0.0
0.1
0.2
0.3
0.4
37.9%, 1600
24.9%
12.2%
15.5%
18.1%
17.4%, 733
17.4%, 734
22.5%
12.6%
8.7%, 368
9.6%, 407
9%, 381
missing
[−Inf,6)
[6,8)
[8,12)
[12,21)
[21, Inf)
pos
neg

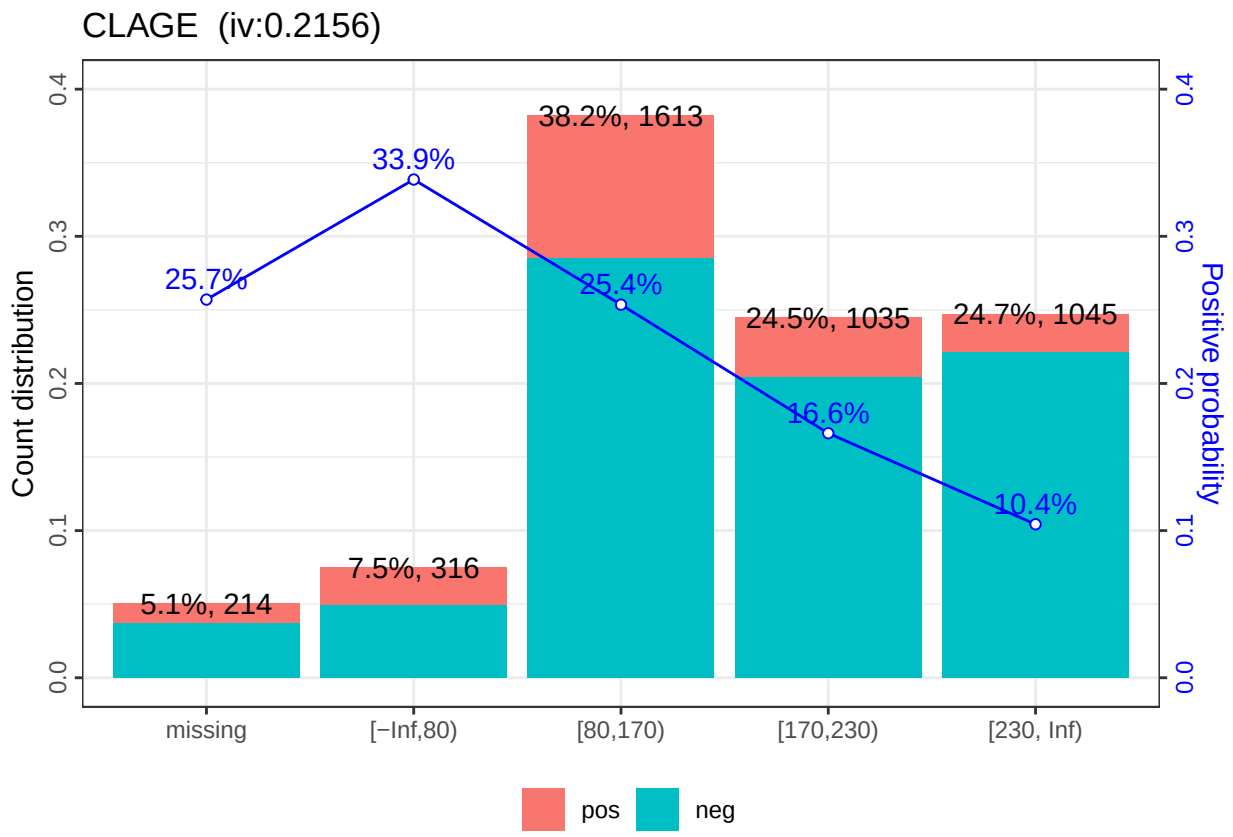
```
##
## $DEROG
```


DEROG (iv:0.3275)
Count distribution
Positive probability
0.0
0.2
0.4
0.6
0.8
0.0
0.2
0.4
0.6
75.8%, 3202
47.1%
16.9%
12.6%
11.7%, 492
12.5%, 529
missing
[−Inf,1)
[1, Inf)
pos
neg

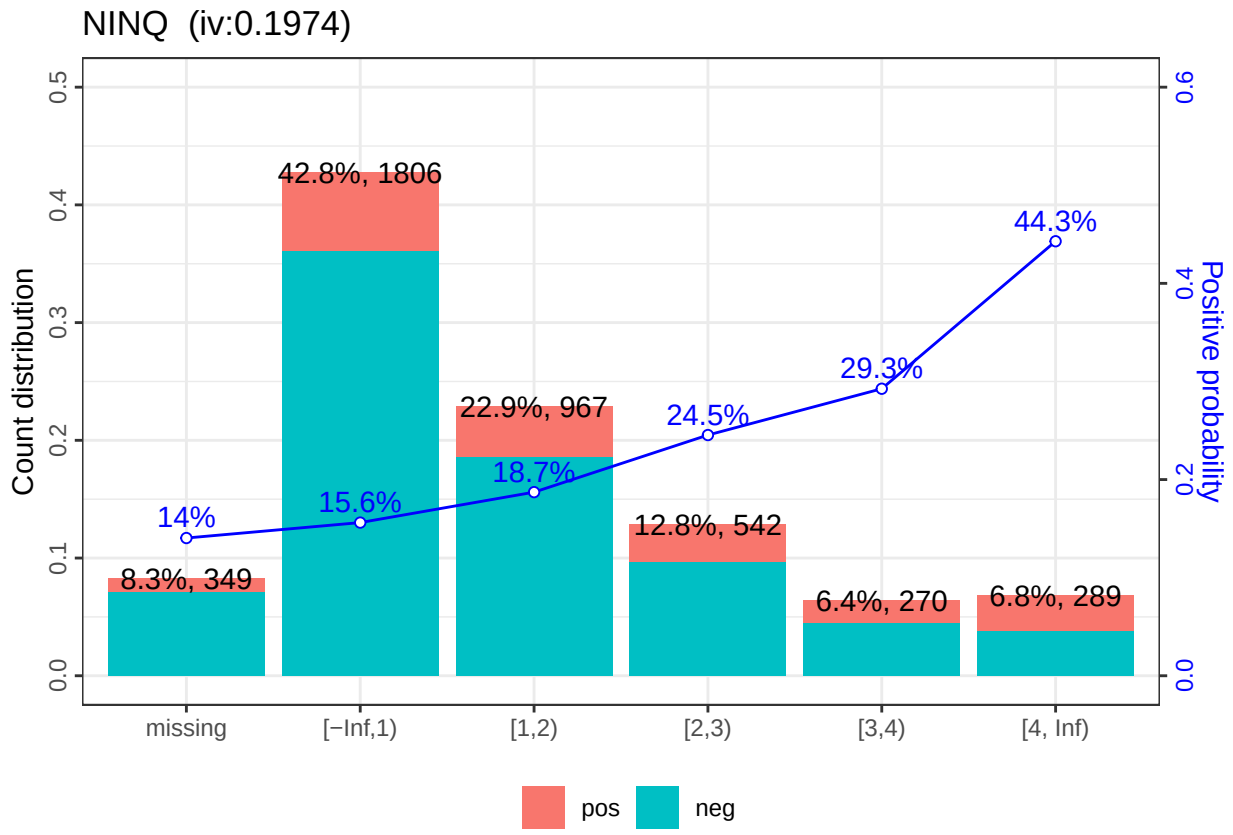
```
##
## $DELINQ
```


DELINQ (iv:0.5291)
Count distribution
Positive probability
0.0
0.2
0.4
0.6
0.8
70.3%, 2968
66.7%
35.1%
12.9%
14.6%
15.3%, 646
9.4%, 396
5%, 213
missing
[−Inf,1)
[1,3)
[3, Inf)
pos
neg

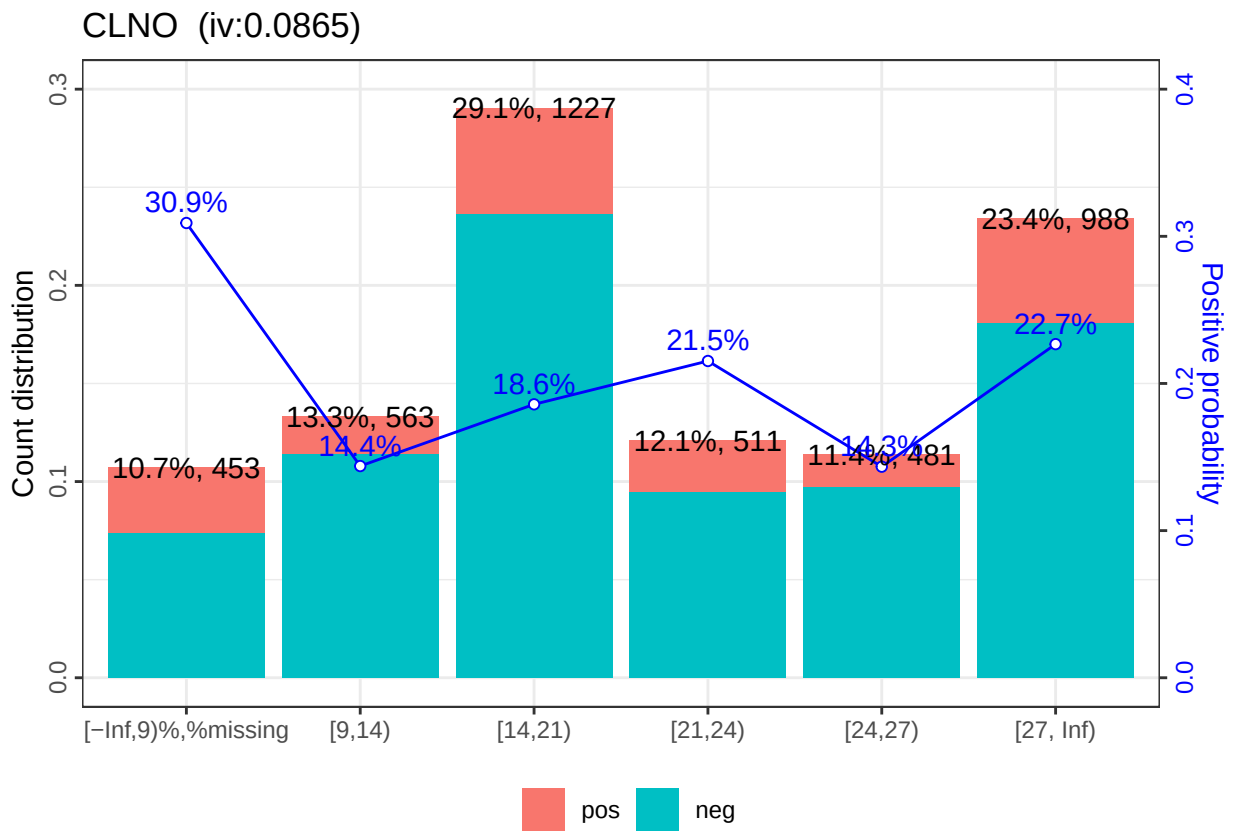
```
##
## $CLAGE
```


CLAGE (iv:0.2156)
Count distribution
Positive probability
0.0
0.1
0.2
0.3
0.4
5.1%, 214
7.5%, 316
38.2%, 1613
24.5%, 1035
24.7%, 1045
25.7%
33.9%
25.4%
16.6%
10.4%
missing
[−Inf,80)
[80,170)
[170,230)
[230, Inf)
pos
neg

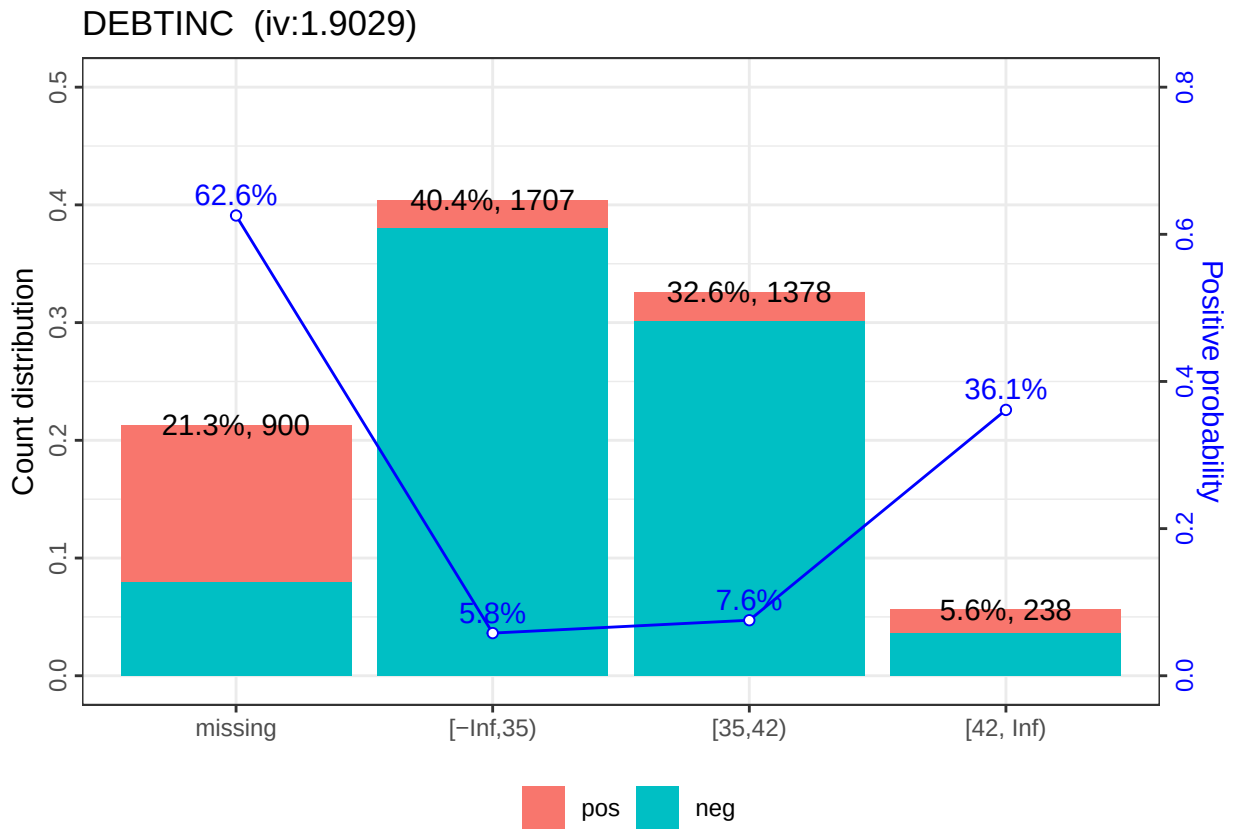
```
##
## $NINQ
```


NINQ (iv:0.1974)
Count distribution
Positive probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
42.8%, 1806
22.9%, 967
12.8%, 542
8.3%, 349
6.4%, 270
6.8%, 289
14%
15.6%
18.7%
24.5%
29.3%
44.3%
missing
[−Inf,1)
[1,2)
[2,3)
[3,4)
[4, Inf)
pos
neg

```
##
## $CLNO
```


CLNO (iv:0.0865)
Count distribution
Positive probability
0.0
0.1
0.2
0.3
0.4
30.9%
18.6%
21.5%
14.3%
22.7%
10.7%, 453
13.3%, 563
14.4%
29.1%, 1227
12.1%, 511
11.4%, 481
23.4%, 988
[−Inf,9)%,%missing
[9,14)
[14,21)
[21,24)
[24,27)
[27, Inf)
pos
neg

```
##
## $DEBTINC
```




```
bins
```

```
## $LOAN
##    variable           bin count count_distr  neg pos   posprob         woe
## 1:     LOAN   [-Inf,8000)   450  0.10655932  293 157 0.3488889  0.75145138
## 2:     LOAN  [8000,13000)   993  0.23514090  788 205 0.2064451  0.02889008
## 3:     LOAN [13000,15000)   365  0.08643145  304  61 0.1671233 -0.23077565
## 4:     LOAN [15000,16000)   262  0.06204120  188  74 0.2824427  0.44300132
## 5:     LOAN [16000,38000)  1907  0.45157471 1612 295 0.1546932 -0.32287738
## 6:     LOAN  [38000, Inf)   246  0.05825243  186  60 0.2439024  0.24397608
##         bin_iv  total_iv breaks is_special_values
## 1: 0.073157153 0.1377386   8000             FALSE
## 2: 0.000197949 0.1377386  13000             FALSE
## 3: 0.004288899 0.1377386  15000             FALSE
## 4: 0.013770545 0.1377386  16000             FALSE
## 5: 0.042604396 0.1377386  38000             FALSE
## 6: 0.003719667 0.1377386    Inf             FALSE
##
## $MORTDUE
##    variable             bin count count_distr  neg pos   posprob         woe
## 1:  MORTDUE         missing   346  0.08193228  274  72 0.2080925  0.03891620
## 2:  MORTDUE    [-Inf,35000)   595  0.14089510  431 164 0.2756303  0.40913653
## 3:  MORTDUE   [35000,55000)   819  0.19393796  641 178 0.2173382  0.09413228
## 4:  MORTDUE  [55000,105000)  1793  0.42457968 1485 308 0.1717791 -0.19769208
## 5:  MORTDUE [105000,125000)   224  0.05304286  174  50 0.2232143  0.12834589
## 6:  MORTDUE [125000,150000)   232  0.05493725  206  26 0.1120690 -0.69440144
```

```
## 7:  MORTDUE   [150000, Inf)   214  0.05067488  160  54 0.2523364  0.28918842
##          bin_iv   total_iv  breaks is_special_values
## 1: 0.0001255251 0.07071158 missing             TRUE
## 2: 0.0264437322 0.07071158   35000            FALSE
## 3: 0.0017667591 0.07071158   55000            FALSE
## 4: 0.0156215793 0.07071158  105000            FALSE
## 5: 0.0009072406 0.07071158  125000            FALSE
## 6: 0.0212438450 0.07071158  150000            FALSE
## 7: 0.0046029025 0.07071158     Inf            FALSE
##
## $VALUE
##    variable                    bin count count_distr  neg pos   posprob
## 1:    VALUE [-Inf,50000)%,%missing   537   0.1271608  316 221 0.4115456
## 2:    VALUE          [50000,85000)  1392   0.3296235 1127 265 0.1903736
## 3:    VALUE         [85000,125000)  1356   0.3210987 1164 192 0.1415929
## 4:    VALUE        [125000,170000)   478   0.1131897  375 103 0.2154812
## 5:    VALUE          [170000, Inf)   460   0.1089273  389  71 0.1543478
##            woe       bin_iv  total_iv          breaks is_special_values
## 1:  1.01779868 0.1685972735 0.2327049 50000%,%missing            FALSE
## 2: -0.07220650 0.0016816385 0.2327049           85000            FALSE
## 3: -0.42674407 0.0511862518 0.2327049          125000            FALSE
## 4:  0.08318115 0.0008026203 0.2327049          170000            FALSE
## 5: -0.32552128 0.0104370883 0.2327049             Inf            FALSE
##
## $JOB
##    variable                         bin count count_distr  neg pos   posprob
## 1:      JOB           Office%,%missing   850   0.2012787  749 101 0.1188235
## 2:      JOB                    ProfExe   917   0.2171442  767 150 0.1635769
## 3:      JOB Other%,%Mgr%,%Sales%,%Self  2456   0.5815771 1855 601 0.2447068
##           woe     bin_iv  total_iv                    breaks is_special_values
## 1: -0.6282403 0.06511375 0.1168358           Office%,%missing             FALSE
## 2: -0.2564733 0.01320134 0.1168358                    ProfExe             FALSE
## 3:  0.2483331 0.03852075 0.1168358 Other%,%Mgr%,%Sales%,%Self             FALSE
##
## $YOJ
##    variable       bin count count_distr  neg pos   posprob        woe
## 1:      YOJ   missing   368  0.08714184  323  45 0.1222826 -0.5956116
## 2:      YOJ  [-Inf,6)  1600  0.37887758 1202 398 0.2487500  0.2700881
## 3:      YOJ     [6,8)   407  0.09637698  344  63 0.1547912 -0.3221287
## 4:      YOJ    [8,12)   733  0.17357329  600 133 0.1814461 -0.1312023
## 5:      YOJ   [12,21)   734  0.17381009  569 165 0.2247956  0.1374432
## 6:      YOJ [21, Inf)   381  0.09022022  333  48 0.1259843 -0.5615633
##         bin_iv   total_iv  breaks is_special_values
## 1: 0.025611517 0.09465226 missing             TRUE
## 2: 0.029862396 0.09465226       6            FALSE
## 3: 0.009052851 0.09465226       8            FALSE
## 4: 0.002871421 0.09465226      12            FALSE
## 5: 0.003418121 0.09465226      21            FALSE
## 6: 0.023835956 0.09465226     Inf            FALSE
##
## $DEROG
##    variable      bin count count_distr  neg pos   posprob        woe     bin_iv
## 1:    DEROG  missing   492   0.1165049  430  62 0.1260163 -0.5612726 0.03075136
## 2:    DEROG [-Inf,1)  3202   0.7582287 2661 541 0.1689569 -0.2176598 0.03360743
```

```
## 3:    DEROG [1, Inf)   529   0.1252664  280 249 0.4706994  1.2580415 0.26317236
##     total_iv  breaks is_special_values
## 1: 0.3275312 missing             TRUE
## 2: 0.3275312       1            FALSE
## 3: 0.3275312     Inf            FALSE
##
## $DELINQ
##    variable      bin count count_distr  neg pos   posprob        woe     bin_iv
## 1:   DELINQ  missing   396  0.09377220  345  51 0.1287879 -0.5363406 0.02278609
## 2:   DELINQ [-Inf,1)  2968  0.70281790 2536 432 0.1455526 -0.3945395 0.09676349
## 3:   DELINQ    [1,3)   646  0.15297182  419 227 0.3513932  0.7624573 0.10837298
## 4:   DELINQ [3, Inf)   213  0.05043808   71 142 0.6666667  2.0685254 0.30118695
##     total_iv  breaks is_special_values
## 1: 0.5291095 missing             TRUE
## 2: 0.5291095       1            FALSE
## 3: 0.5291095       3            FALSE
## 4: 0.5291095     Inf            FALSE
##
## $CLAGE
##    variable        bin count count_distr  neg pos   posprob        woe
## 1:    CLAGE    missing   214  0.05067488  159  55 0.2570093  0.3138072
## 2:    CLAGE  [-Inf,80)   316  0.07482832  209 107 0.3386076  0.7058728
## 3:    CLAGE   [80,170)  1613  0.38195596 1204 409 0.2535648  0.2956887
## 4:    CLAGE  [170,230)  1035  0.24508643  863 172 0.1661836 -0.2375420
## 5:    CLAGE [230, Inf)  1045  0.24745442  936 109 0.1043062 -0.7748894
##         bin_iv  total_iv  breaks is_special_values
## 1: 0.005456158 0.2155566 missing             TRUE
## 2: 0.044884648 0.2155566      80            FALSE
## 3: 0.036335088 0.2155566     170            FALSE
## 4: 0.012857956 0.2155566     230            FALSE
## 5: 0.116022751 0.2155566     Inf            FALSE
##
## $NINQ
##    variable      bin count count_distr  neg pos   posprob         woe
## 1:     NINQ  missing   349  0.08264267  300  49 0.1404011 -0.43658399
## 2:     NINQ [-Inf,1)  1806  0.42765806 1524 282 0.1561462 -0.31180848
## 3:     NINQ    [1,2)   967  0.22898413  786 181 0.1871768 -0.09308157
## 4:     NINQ    [2,3)   542  0.12834478  409 133 0.2453875  0.25201216
## 5:     NINQ    [3,4)   270  0.06393559  191  79 0.2925926  0.49255261
## 6:     NINQ [4, Inf)   289  0.06843476  161 128 0.4429066  1.14600409
##         bin_iv  total_iv  breaks is_special_values
## 1: 0.013744811 0.1973983 missing             TRUE
## 2: 0.037761707 0.1973983       1            FALSE
## 3: 0.001929015 0.1973983       2            FALSE
## 4: 0.008763547 0.1973983       3            FALSE
## 5: 0.017763053 0.1973983       4            FALSE
## 6: 0.117436119 0.1973983     Inf            FALSE
##
## $CLNO
##    variable                bin count count_distr neg pos   posprob         woe
## 1:     CLNO [-Inf,9)%,%missing   453   0.1072697 313 140 0.3090508  0.57081742
## 2:     CLNO             [9,14)   563   0.1333175 482  81 0.1438721 -0.40811677
## 3:     CLNO            [14,21)  1227   0.2905517 999 228 0.1858191 -0.10203096
## 4:     CLNO            [21,24)   511   0.1210040 401 110 0.2152642  0.08189713
```

```
## 5:     CLNO            [24,27)   481   0.1139001 412  69 0.1434511 -0.41153866
## 6:     CLNO          [27, Inf)   988   0.2339569 764 224 0.2267206  0.14845645
##          bin_iv   total_iv      breaks is_special_values
## 1: 0.0407954435 0.08646803 9%,%missing             FALSE
## 2: 0.0195544478 0.08646803          14             FALSE
## 3: 0.0029329409 0.08646803          21             FALSE
## 4: 0.0008314345 0.08646803          24             FALSE
## 5: 0.0169689778 0.08646803          27             FALSE
## 6: 0.0053847855 0.08646803         Inf             FALSE
##
## $DEBTINC
##    variable       bin count count_distr  neg pos    posprob       woe
## 1:  DEBTINC   missing   900  0.21311864  337 563 0.62555556  1.888575
## 2:  DEBTINC [-Inf,35)  1707  0.40421501 1608  99 0.05799649 -1.412248
## 3:  DEBTINC   [35,42)  1378  0.32630831 1274 104 0.07547170 -1.130148
## 4:  DEBTINC [42, Inf)   238  0.05635804  152  86 0.36134454  0.805845
##        bin_iv total_iv  breaks is_special_values
## 1: 1.05916527 1.902891 missing              TRUE
## 2: 0.50955703 1.902891      35             FALSE
## 3: 0.28916375 1.902891      42             FALSE
## 4: 0.04500522 1.902891     Inf             FALSE
```

Kết luận: pos=1, neg=0: Data được lấy trên kaggle có thể không phải data thực nên một số mối quan hệ có thể không phù hợp với lý thuyết kinh tế

Dựa vào biểu đồ bins của các biến, có thể nhận thấy:

- **LOAN**: Xác suất vỡ nợ của khách hàng cao xảy ra ở nhóm có khoản vay rất nhỏ hoặc trung bình
- **MORTDUE**: Xác suất vỡ nợ có xu hướng giảm khi số tiền phải trả cho khoản vay thế chấp hiện tại tăng lên, nhưng với nhóm giá trị MORTDUE cao nhất, xác suất lại tăng trở lại
- **VALUE**: Nhìn chung giá trị tài sản hiện tại của khách hàng càng cao thì xác suất vỡ nợ càng giảm. Những khách hàng có tài sản thế chấp thấp thường có ít sự đảm bảo hơn và dễ gặp rủi ro tài chính. Tài sản thế chấp thấp không đủ để giảm bớt rủi ro tín dụng, dẫn đến xác suất vỡ nợ cao.
- **JOB**: Nhóm khách hàng làm việc ở văn phòng và là các chuyên gia có xác suất vỡ nợ thấp nhất, còn đối với các nhóm OTHER (có thể là lao động tự do, phổ thông) và người tự kinh doanh, bán hàng thì có xác suất vỡ nợ cao hơn - có thể do thu nhập không ổn định so với các nhóm khác
- **YOJ**: Nhìn chung, số năm kinh nghiệm làm việc càng ít (nhỏ hơn 6 năm) thì xác suất vỡ nợ là cao nhất
- **DEROG**: Nếu khách hàng có số lượng báo cáo xấu từ 1 trở lên thì xác suất vỡ nợ sẽ rất lớn
- **DELINQ**: Số hạn mức tín dụng quá hạn càng lớn thì khách hàng đó có xác suất vỡ nợ càng lớn
- **CLAGE**: Nhóm khách hàng có lịch sử tín dụng càng dài thì xác suất vỡ nợ càng giảm
- **NINQ**: Số lượng yêu cầu tín dụng gần đây càng nhiều thì xác xuất vỡ nợ càng lớn - có thể do khách hàng đang gặp khó khăn về tài chính và nhiều khoản vay dẫn đến rủi ro thanh khoản giảm
- **CLNO**: Khách hàng có số lượng tín dụng ít nhất hay để trống thông tin này thì xác xuất vỡ nợ là cao nhất - có thể do đây là nhóm ít thông tin về lịch sử tín dụng nên sẽ rủi ro hơn.
- **DEBTINC**: Nhìn chung, khi tỷ lệ nợ/thu nhập tăng, xác suất vỡ nợ cũng tăng, còn đối với những khách hàng không có thông tin về tỷ lệ này thì xác suất vỡ nợ gần như là khá lớn.

Chuyển data sang woe

```
train_data_woe <- woebin_ply(train_data_removed, bins)
```

```
## i Converting into woe values ...
```

```
## v Woe transformating on 4223 rows and 11 columns in 00:00:02
```

Chạy mô hình logit cho tập train cho dữ liệu đã binning theo WOE

```
logit.model_woe <- glm(BAD ~., family = binomial(link = 'logit'), data = train_data_woe)
summary(logit.model_woe)
```

```
##
## Call:
## glm(formula = BAD ~ ., family = binomial(link = "logit"), data = train_data_woe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7254  -0.4021  -0.2339  -0.1330   3.0139
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.36819    0.05518 -24.796  < 2e-16 ***
## LOAN_woe     0.42185    0.14210   2.969 0.002991 **
## MORTDUE_woe  0.22392    0.22867   0.979 0.327468
## VALUE_woe    0.74878    0.12069   6.204 5.50e-10 ***
## JOB_woe      0.85764    0.15801   5.428 5.70e-08 ***
## YOJ_woe      1.02967    0.18004   5.719 1.07e-08 ***
## DEROG_woe    0.70615    0.09046   7.806 5.90e-15 ***
## DELINQ_woe   0.95840    0.07321  13.092  < 2e-16 ***
## CLAGE_woe    1.06135    0.11995   8.848  < 2e-16 ***
## NINQ_woe     0.43985    0.11928   3.688 0.000226 ***
## CLNO_woe     1.26576    0.18225   6.945 3.78e-12 ***
## DEBTINC_woe  0.92976    0.03663  25.386  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4246.8  on 4222  degrees of freedom
## Residual deviance: 2388.0  on 4211  degrees of freedom
## AIC: 2412
##
## Number of Fisher Scoring iterations: 6
```

Loc biến theo stepwise

```r
logit.model.step_woe <- step(logit.model_woe, direction = "backward", trace = 0)
summary(logit.model.step_woe)
```

```
##
## Call:
## glm(formula = BAD ~ LOAN_woe + VALUE_woe + JOB_woe + YOJ_woe +
##     DEROG_woe + DELINQ_woe + CLAGE_woe + NINQ_woe + CLNO_woe +
##     DEBTINC_woe, family = binomial(link = "logit"), data = train_data_woe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7113  -0.4024  -0.2345  -0.1339   3.0414
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.36898    0.05517 -24.814  < 2e-16 ***
## LOAN_woe     0.42216    0.14197   2.974 0.002944 **
## VALUE_woe    0.80127    0.10835   7.395 1.41e-13 ***
## JOB_woe      0.86203    0.15795   5.458 4.83e-08 ***
## YOJ_woe      1.02025    0.17984   5.673 1.40e-08 ***
## DEROG_woe    0.70459    0.09038   7.796 6.40e-15 ***
## DELINQ_woe   0.95731    0.07312  13.091  < 2e-16 ***
## CLAGE_woe    1.06200    0.12000   8.850  < 2e-16 ***
## NINQ_woe     0.44370    0.11922   3.722 0.000198 ***
## CLNO_woe     1.26452    0.18240   6.933 4.13e-12 ***
## DEBTINC_woe  0.93084    0.03661  25.424  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4246.8  on 4222  degrees of freedom
## Residual deviance: 2388.9  on 4212  degrees of freedom
## AIC: 2410.9
##
## Number of Fisher Scoring iterations: 6
```

Kiểm tra mô hình trên tập train cho mô hình logit với dữ liệu đã binning theo WOE

```r
train.prob <- predict(logit.model.step_woe, type = "response")
train.pred <- as.factor(ifelse(train.prob > .5, "1", "0"))
table(train.pred, train_data_woe$BAD)
```

```
##
## train.pred    0    1
##          0 3209  316
##          1  162  536
```

```
train_data_woe$BAD <- as.factor(train_data_woe$BAD)
caret::confusionMatrix(train.pred, train_data_woe$BAD, positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 3209  316
##          1  162  536
##
##                Accuracy : 0.8868
##                  95% CI : (0.8769, 0.8962)
##     No Information Rate : 0.7982
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6231
##
##  Mcnemar's Test P-Value : 2.595e-12
##
##             Sensitivity : 0.6291
##             Specificity : 0.9519
##          Pos Pred Value : 0.7679
##          Neg Pred Value : 0.9104
##              Prevalence : 0.2018
##          Detection Rate : 0.1269
##    Detection Prevalence : 0.1653
##       Balanced Accuracy : 0.7905
##
##        'Positive' Class : 1
##
```

Kiểm tra mô hình trên tập test được binning theo WOE:

```
test.data_woe <- woebin_ply(test_data, bins)
```

```
## i Converting into woe values ...
```

```
## v Woe transformating on 1737 rows and 11 columns in 00:00:11
```

- Thực hiện mô hình trên tập test:

```
logit.pred.prob_woe <- predict(logit.model.step_woe, test.data_woe, type = 'response')
logit.pred_woe <- as.factor(ifelse(logit.pred.prob_woe > 0.5, 1, 0))
test.data_woe$BAD <- as.factor(test.data_woe$BAD)

caret::confusionMatrix(logit.pred_woe, test.data_woe$BAD, positive = "1")
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction    0    1
##          0 1335  128
##          1   65  209
##
##                Accuracy : 0.8889
##                  95% CI : (0.8732, 0.9033)
##     No Information Rate : 0.806
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6176
##
##  Mcnemar's Test P-Value : 8.087e-06
##
##             Sensitivity : 0.6202
##             Specificity : 0.9536
##          Pos Pred Value : 0.7628
##          Neg Pred Value : 0.9125
##              Prevalence : 0.1940
##          Detection Rate : 0.1203
##    Detection Prevalence : 0.1577
##       Balanced Accuracy : 0.7869
##
##        'Positive' Class : 1
##
```

Tính toán AUC và Gini trên tập train

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.2.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

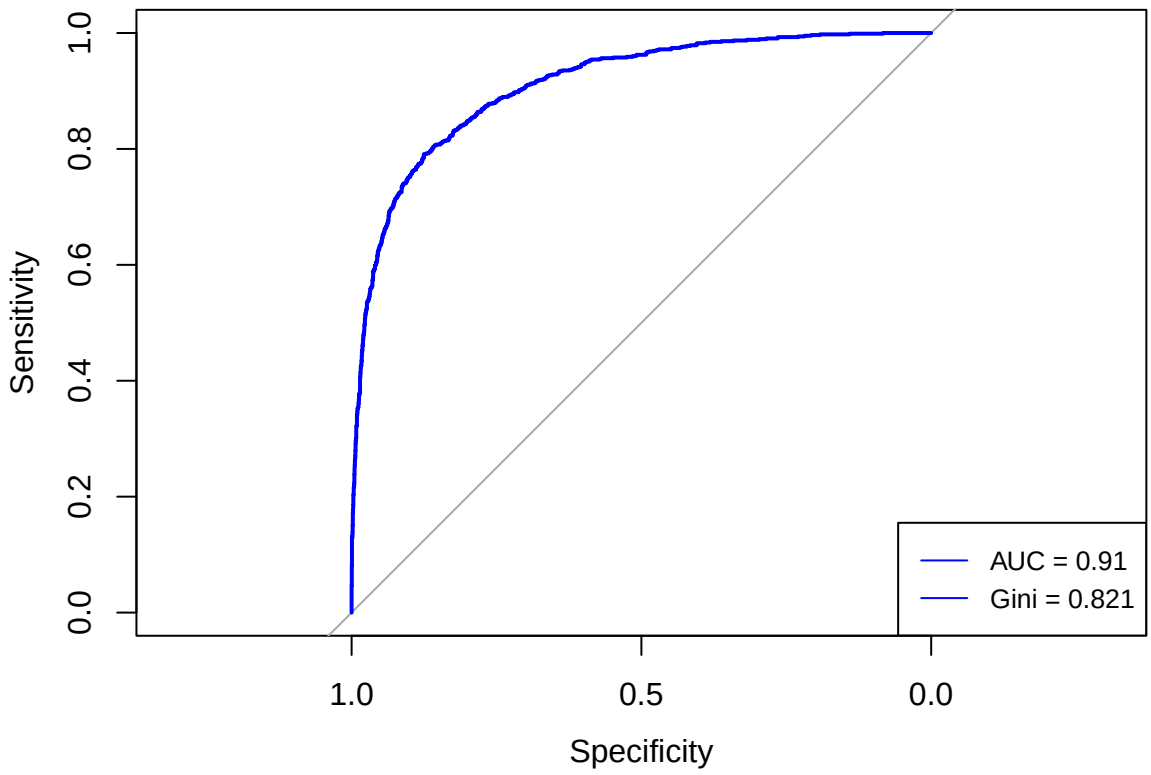
```
# Tính toán AUC
roc_curve <- roc(train_data_woe$BAD, train.prob)  # Đường ROC
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
auc_value <- auc(roc_curve)                          # Giá trị AUC
gini_value <- 2 * auc_value - 1                      # Giá trị Gini

# Vẽ đường cong ROC
plot(roc_curve, col = "blue")
legend("bottomright", legend = c(
  paste("AUC =", round(auc_value, 3)),
  paste("Gini =", round(gini_value, 3))
), col = c("blue"), lty = 1, cex = 0.8)
```



Tính toán AUC và Gini trên tập test

```r
roc_curve_test <- roc(test.data_woe$BAD, logit.pred.prob_woe)  # Đường ROC
```

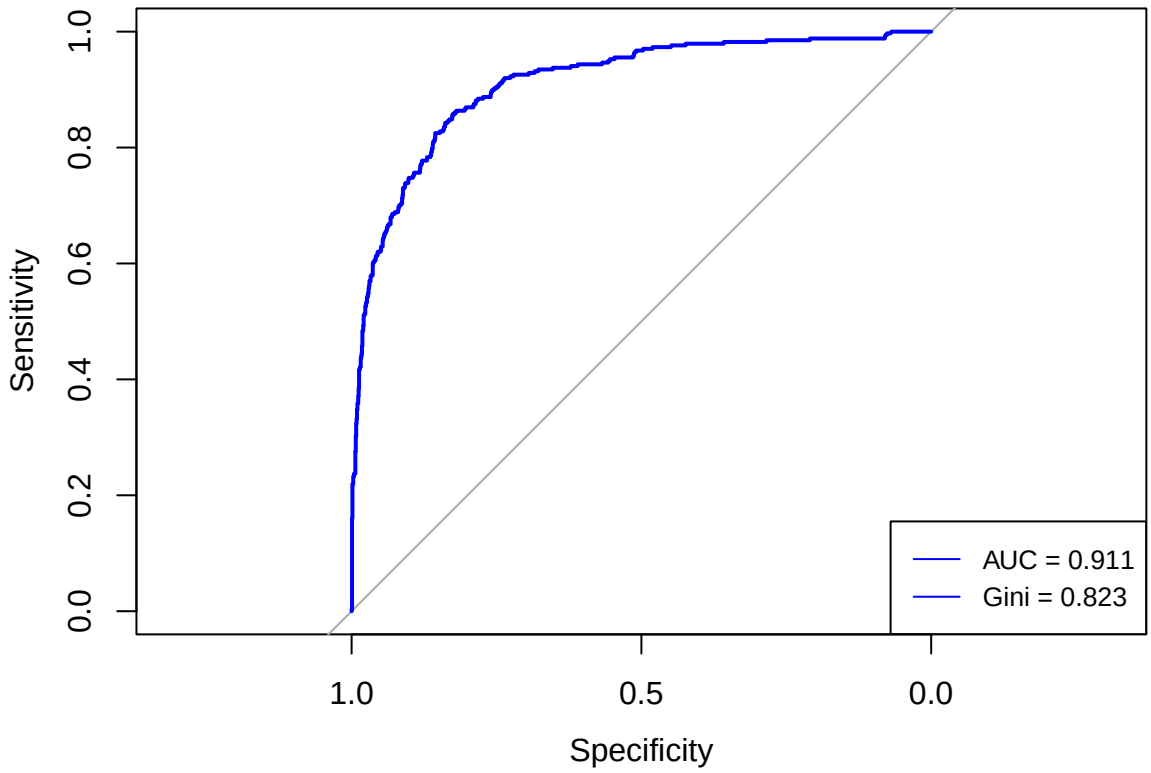
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
auc_value_test <- auc(roc_curve_test)                    # Giá trị AUC
gini_value_test <- 2 * auc_value_test - 1                # Giá trị Gini

# Vẽ đường cong ROC
plot(roc_curve_test, col = "blue")
legend("bottomright", legend = c(
  paste("AUC =", round(auc_value_test, 3)),
```

```
  paste("Gini =", round(gini_value_test, 3))
), col = c("blue"), lty = 1, cex = 0.8)
```



Kết luận: Mô hình tốt

Thực hiện tính score:

```
my_card <- scorecard(bins,logit.model.step_woe, points0 = 600, odds0 = 1/19, pdo = 50)
my_card
```

```
## $basepoints
##      variable bin woe points
## 1: basepoints  NA  NA    486
##
## $LOAN
##    variable           bin count count_distr  neg pos   posprob         woe
## 1:     LOAN   [-Inf,8000)   450  0.10655932  293 157 0.3488889  0.75145138
## 2:     LOAN  [8000,13000)   993  0.23514090  788 205 0.2064451  0.02889008
## 3:     LOAN [13000,15000)   365  0.08643145  304  61 0.1671233 -0.23077565
## 4:     LOAN [15000,16000)   262  0.06204120  188  74 0.2824427  0.44300132
## 5:     LOAN [16000,38000)  1907  0.45157471 1612 295 0.1546932 -0.32287738
## 6:     LOAN  [38000, Inf)   246  0.05825243  186  60 0.2439024  0.24397608
##         bin_iv  total_iv breaks is_special_values points
```

```
## 1: 0.073157153 0.1377386   8000             FALSE    -23
## 2: 0.000197949 0.1377386  13000             FALSE     -1
## 3: 0.004288899 0.1377386  15000             FALSE      7
## 4: 0.013770545 0.1377386  16000             FALSE    -13
## 5: 0.042604396 0.1377386  38000             FALSE     10
## 6: 0.003719667 0.1377386    Inf             FALSE     -7
##
## $VALUE
##    variable                    bin count count_distr  neg pos   posprob
## 1:    VALUE [-Inf,50000)%,%missing   537   0.1271608  316 221 0.4115456
## 2:    VALUE          [50000,85000)  1392   0.3296235 1127 265 0.1903736
## 3:    VALUE         [85000,125000)  1356   0.3210987 1164 192 0.1415929
## 4:    VALUE        [125000,170000)   478   0.1131897  375 103 0.2154812
## 5:    VALUE          [170000, Inf)   460   0.1089273  389  71 0.1543478
##           woe       bin_iv  total_iv          breaks is_special_values points
## 1:  1.01779868 0.1685972735 0.2327049 50000%,%missing             FALSE    -59
## 2: -0.07220650 0.0016816385 0.2327049           85000             FALSE      4
## 3: -0.42674407 0.0511862518 0.2327049          125000             FALSE     25
## 4:  0.08318115 0.0008026203 0.2327049          170000             FALSE     -5
## 5: -0.32552128 0.0104370883 0.2327049             Inf             FALSE     19
##
## $JOB
##    variable                         bin count count_distr  neg pos   posprob
## 1:      JOB           Office%,%missing   850   0.2012787  749 101 0.1188235
## 2:      JOB                     ProfExe   917   0.2171442  767 150 0.1635769
## 3:      JOB Other%,%Mgr%,%Sales%,%Self  2456   0.5815771 1855 601 0.2447068
##           woe     bin_iv  total_iv                    breaks is_special_values
## 1: -0.6282403 0.06511375 0.1168358           Office%,%missing             FALSE
## 2: -0.2564733 0.01320134 0.1168358                    ProfExe             FALSE
## 3:  0.2483331 0.03852075 0.1168358 Other%,%Mgr%,%Sales%,%Self             FALSE
##    points
## 1:     39
## 2:     16
## 3:    -15
##
## $YOJ
##    variable       bin count count_distr  neg pos   posprob        woe
## 1:      YOJ   missing   368  0.08714184  323  45 0.1222826 -0.5956116
## 2:      YOJ  [-Inf,6)  1600  0.37887758 1202 398 0.2487500  0.2700881
## 3:      YOJ     [6,8)   407  0.09637698  344  63 0.1547912 -0.3221287
## 4:      YOJ    [8,12)   733  0.17357329  600 133 0.1814461 -0.1312023
## 5:      YOJ   [12,21)   734  0.17381009  569 165 0.2247956  0.1374432
## 6:      YOJ [21, Inf)   381  0.09022022  333  48 0.1259843 -0.5615633
##         bin_iv   total_iv  breaks is_special_values points
## 1: 0.025611517 0.09465226 missing             TRUE     44
## 2: 0.029862396 0.09465226       6            FALSE    -20
## 3: 0.009052851 0.09465226       8            FALSE     24
## 4: 0.002871421 0.09465226      12            FALSE     10
## 5: 0.003418121 0.09465226      21            FALSE    -10
## 6: 0.023835956 0.09465226     Inf            FALSE     41
##
## $DEROG
##    variable      bin count count_distr  neg pos   posprob        woe     bin_iv
## 1:    DEROG  missing   492   0.1165049  430  62 0.1260163 -0.5612726 0.03075136
```

```
## 2:    DEROG [-Inf,1)  3202   0.7582287 2661 541 0.1689569 -0.2176598 0.03360743
## 3:    DEROG [1, Inf)   529   0.1252664  280 249 0.4706994  1.2580415 0.26317236
##     total_iv  breaks is_special_values points
## 1: 0.3275312 missing              TRUE     29
## 2: 0.3275312       1             FALSE     11
## 3: 0.3275312     Inf             FALSE    -64
##
## $DELINQ
##    variable      bin count count_distr  neg pos   posprob        woe     bin_iv
## 1:   DELINQ  missing   396  0.09377220  345  51 0.1287879 -0.5363406 0.02278609
## 2:   DELINQ [-Inf,1)  2968  0.70281790 2536 432 0.1455526 -0.3945395 0.09676349
## 3:   DELINQ    [1,3)   646  0.15297182  419 227 0.3513932  0.7624573 0.10837298
## 4:   DELINQ [3, Inf)   213  0.05043808   71 142 0.6666667  2.0685254 0.30118695
##     total_iv  breaks is_special_values points
## 1: 0.5291095 missing              TRUE     37
## 2: 0.5291095       1             FALSE     27
## 3: 0.5291095       3             FALSE    -53
## 4: 0.5291095     Inf             FALSE   -143
##
## $CLAGE
##    variable        bin count count_distr  neg pos   posprob        woe
## 1:    CLAGE    missing   214  0.05067488  159  55 0.2570093  0.3138072
## 2:    CLAGE  [-Inf,80)   316  0.07482832  209 107 0.3386076  0.7058728
## 3:    CLAGE   [80,170)  1613  0.38195596 1204 409 0.2535648  0.2956887
## 4:    CLAGE  [170,230)  1035  0.24508643  863 172 0.1661836 -0.2375420
## 5:    CLAGE [230, Inf)  1045  0.24745442  936 109 0.1043062 -0.7748894
##         bin_iv  total_iv  breaks is_special_values points
## 1: 0.005456158 0.2155566 missing              TRUE    -24
## 2: 0.044884648 0.2155566      80             FALSE    -54
## 3: 0.036335088 0.2155566     170             FALSE    -23
## 4: 0.012857956 0.2155566     230             FALSE     18
## 5: 0.116022751 0.2155566     Inf             FALSE     59
##
## $NINQ
##    variable      bin count count_distr  neg pos   posprob         woe
## 1:     NINQ  missing   349  0.08264267  300  49 0.1404011 -0.43658399
## 2:     NINQ [-Inf,1)  1806  0.42765806 1524 282 0.1561462 -0.31180848
## 3:     NINQ    [1,2)   967  0.22898413  786 181 0.1871768 -0.09308157
## 4:     NINQ    [2,3)   542  0.12834478  409 133 0.2453875  0.25201216
## 5:     NINQ    [3,4)   270  0.06393559  191  79 0.2925926  0.49255261
## 6:     NINQ [4, Inf)   289  0.06843476  161 128 0.4429066  1.14600409
##         bin_iv  total_iv  breaks is_special_values points
## 1: 0.013744811 0.1973983 missing              TRUE     14
## 2: 0.037761707 0.1973983       1             FALSE     10
## 3: 0.001929015 0.1973983       2             FALSE      3
## 4: 0.008763547 0.1973983       3             FALSE     -8
## 5: 0.017763053 0.1973983       4             FALSE    -16
## 6: 0.117436119 0.1973983     Inf             FALSE    -37
##
## $CLNO
##    variable                bin count count_distr neg pos   posprob         woe
## 1:     CLNO [-Inf,9)%,%missing   453   0.1072697 313 140 0.3090508  0.57081742
## 2:     CLNO             [9,14)   563   0.1333175 482  81 0.1438721 -0.40811677
## 3:     CLNO            [14,21)  1227   0.2905517 999 228 0.1858191 -0.10203096
```

```
## 4:     CLNO            [21,24)   511   0.1210040 401 110 0.2152642  0.08189713
## 5:     CLNO            [24,27)   481   0.1139001 412  69 0.1434511 -0.41153866
## 6:     CLNO          [27, Inf)   988   0.2339569 764 224 0.2267206  0.14845645
##          bin_iv   total_iv      breaks is_special_values points
## 1: 0.0407954435 0.08646803 9%,%missing            FALSE    -52
## 2: 0.0195544478 0.08646803          14            FALSE     37
## 3: 0.0029329409 0.08646803          21            FALSE      9
## 4: 0.0008314345 0.08646803          24            FALSE     -7
## 5: 0.0169689778 0.08646803          27            FALSE     38
## 6: 0.0053847855 0.08646803         Inf            FALSE    -14
##
## $DEBTINC
##    variable       bin count count_distr  neg pos    posprob       woe
## 1:  DEBTINC   missing   900  0.21311864  337 563 0.62555556  1.888575
## 2:  DEBTINC [-Inf,35)  1707  0.40421501 1608  99 0.05799649 -1.412248
## 3:  DEBTINC   [35,42)  1378  0.32630831 1274 104 0.07547170 -1.130148
## 4:  DEBTINC [42, Inf)   238  0.05635804  152  86 0.36134454  0.805845
##        bin_iv total_iv  breaks is_special_values points
## 1: 1.05916527 1.902891 missing             TRUE   -127
## 2: 0.50955703 1.902891      35            FALSE     95
## 3: 0.28916375 1.902891      42            FALSE     76
## 4: 0.04500522 1.902891     Inf            FALSE    -54
```