G5291 Final PROJECT

Lingjia Zhang

10/19/2020

```
# Load libraries
library(tidyr)
library(tidyverse)
```

Feature Selection

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v dplyr   1.0.2
## v tibble  3.0.4     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
## v purrr   0.3.4

## Warning: package 'tibble' was built under R version 4.0.3

## Warning: package 'readr' was built under R version 4.0.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.0.3
```

```
library(ggplot2)
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(bnstruct)
```

```
## Warning: package 'bnstruct' was built under R version 4.0.3

## Loading required package: bitops

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loading required package: igraph

## Warning: package 'igraph' was built under R version 4.0.3

##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union

## The following objects are masked from 'package:purrr':
##
##     compose, simplify

## The following object is masked from 'package:tibble':
##
##     as_data_frame

## The following object is masked from 'package:tidyr':
##
##     crossing

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union

##
## Attaching package: 'bnstruct'

## The following object is masked from 'package:tidyr':
##
##     complete
```

```
library(glmnet)
```

```
## Loaded glmnet 4.0-2
```

```
# Load the dataset
df <- read.csv("../data/NFWBS_PUF_2016_data.csv")
head(df)
```

```
##   PUF_ID sample fpl SWB_1 SWB_2 SWB_3 FWBscore FWB1_1 FWB1_2 FWB1_3 FWB1_4
## 1  10350      2   3     5     5     6       55      3      3      3      3
## 2   7740      1   3     6     6     6       51      2      2      3      3
## 3  13699      1   3     4     3     4       49      3      3      3      3
## 4   7267      1   3     6     6     6       49      3      3      3      3
## 5   7375      1   3     4     4     4       49      3      3      3      3
## 6  10910      1   3     5     7     5       67      5      1      1      1
##   FWB1_5 FWB1_6 FWB2_1 FWB2_2 FWB2_3 FWB2_4 FSscore FS1_1 FS1_2 FS1_3 FS1_4
## 1      2      3      2      3      2      4      44     3     3     4     3
## 2      3      4      2      2      2      3      43     3     3     3     3
## 3      3      3      3      3      3      3      42     3     3     3     3
## 4      3      3      3      3      3      3      42     3     3     3     3
```

```
## 5      3      3      3      3      3      3      42     3     3     3     3
## 6      1      1      2      5      2      2      57     4     4     4     4
##   FS1_5 FS1_6 FS1_7 FS2_1 FS2_2 FS2_3 SUBKNOWL1 ACT1_1 ACT1_2 FINGOALS
## 1     3     3     4     4     3     4         5      4      3        1
## 2     4     3     2     4     3     2         5      4      3        0
## 3     3     3     3     3     3     3         5      3      3        1
## 4     3     3     3     3     3     3        -1     -1     -1       -1
## 5     3     3     3     3     3     3         4      3      3        1
## 6     3     4     4     4     4     1         6      5      4        1
##   PROPPLAN_1 PROPPLAN_2 PROPPLAN_3 PROPPLAN_4 MANAGE1_1 MANAGE1_2 MANAGE1_3
## 1          5          4          4          3         4         4         2
## 2          3          2          2          1         4         4         1
## 3          4          4          4          4         3         3         3
## 4          3          3          3          3         4         4         2
## 5          3          3          3          3         3         3         3
## 6          5          4          3          4         5         3         5
##   MANAGE1_4 SAVEHABIT FRUGALITY AUTOMATED_1 AUTOMATED_2 ASK1_1 ASK1_2
## 1         4         4         6           0           0      4      3
## 2         4         1         5           0           0      3      2
## 3         3         5         5           1           1      4      4
## 4         4         4         6          -1          -1     -1     -1
## 5         3         4         4           0           1      3      3
## 6         5         4         5           1           1      5      3
##   SUBNUMERACY2 SUBNUMERACY1 CHANGEABLE GOALCONF LMscore FINKNOWL1 FINKNOWL2
## 1            3            3          4        3       3         1         3
## 2            5            5          2        3       3         1         3
## 3            4            4          6        3       3         1         3
## 4           -1           -1         -1       -1       2         1         1
## 5            4            4          4        3       1         2         2
## 6            5            6          1        4       3         1         3
##   FINKNOWL3 FK1correct FK2correct FK3correct KHscore KHKNOWL1 KHKNOWL2 KHKNOWL3
## 1         2          1          1          1   1.267        3        3        2
## 2         2          1          1          1  -0.570        2        3        3
## 3         2          1          1          1  -0.188        3        3        1
## 4         2          1          0          1  -1.485        2        2        3
## 5         2          0          0          1  -1.900        1        1        3
## 6         2          1          1          1   0.242        3        3        2
##   KHKNOWL4 KHKNOWL5 KHKNOWL6 KHKNOWL7 KHKNOWL8 KHKNOWL9 KH1correct KH2correct
## 1        1        1        2        4        2        1          1          1
## 2        1        1        2        2        3        1          0          1
## 3        1        2        2        2        2        1          1          1
## 4        1        2        2        2       -1        1          0          0
## 5        2        2        1        3        2        2          0          0
## 6        1        1        2        3        4        1          1          1
##   KH3correct KH4correct KH5correct KH6correct KH7correct KH8correct KH9correct
## 1          1          1          1          1          1          1          1
## 2          0          1          1          1          0          0          1
## 3          0          1          0          1          0          1          1
## 4          0          1          0          1          0          0          1
## 5          0          0          0          0          0          1          0
## 6          1          1          1          1          0          0          1
##   ENDSMEET HOUSING LIVINGARRANGEMENT HOUSERANGES IMPUTATION_FLAG VALUERANGES
## 1        2       1                 1           4               0           2
## 2        2       1                 2           4               0           2
```

```
## 3        1       1                2           3              0           3
## 4       -1      -1               -1          99              0          -2
## 5        2       2                3           2              0          -2
## 6        1       1                2           4              0           1
##   MORTGAGE SAVINGSRANGES PRODHAVE_1 PRODHAVE_2 PRODHAVE_3 PRODHAVE_4 PRODHAVE_5
## 1        2             6          1          1          1          1          1
## 2        2             2          1          0          1          0          0
## 3        2             4          1          1          0          1          1
## 4       -2            -1          0          0          0          0          0
## 5       -2            98          1          0          0          1          0
## 6        2             5          1          1          1          1          0
##   PRODHAVE_6 PRODHAVE_7 PRODHAVE_8 PRODHAVE_9 PRODUSE_1 PRODUSE_2 PRODUSE_3
## 1          0          0          0          0         0         0         0
## 2          0          0          0          0         0         0         0
## 3          0          0          0          0         0         0         1
## 4          0          0          0          1         0         0         0
## 5          0          0          0          0         0         1         0
## 6          0          0          1          0         0         0         0
##   PRODUSE_4 PRODUSE_5 PRODUSE_6 CONSPROTECT1 CONSPROTECT2 CONSPROTECT3 EARNERS
## 1         0         0         1            3            2            1       1
## 2         0         0         1            2            1            0       2
## 3         1         1         0            3            3            1       2
## 4         0         0         1           -1           -1           -1      -1
## 5         0         0         0            3            2            0       2
## 6         0         0         1            2            2            1       1
##   VOLATILITY SNAP MATHARDSHIP_1 MATHARDSHIP_2 MATHARDSHIP_3 MATHARDSHIP_4
## 1          2    0             1             1             1             1
## 2          2    0             1             1             1             1
## 3          3    0             1             1             1             1
## 4         -1   -1             2             2             2             2
## 5          2    8             2             2             2             2
## 6          1    0             1             1             1             1
##   MATHARDSHIP_5 MATHARDSHIP_6 COLLECT REJECTED_1 REJECTED_2 ABSORBSHOCK
## 1             1             1       1          0          0           4
## 2             1             1       1          0          0           3
## 3             1             1       0          0          0           4
## 4             2             2      -1          0          1           8
## 5             2             2       8          0          0           2
## 6             1             1       0          0          0           4
##   BENEFITS_1 BENEFITS_2 BENEFITS_3 BENEFITS_4 BENEFITS_5 FRAUD2 COVERCOSTS
## 1          0          0          1          0          0      8          1
## 2          1          0          0          0          1      0          3
## 3          1          0          0          0          0      0         -1
## 4          0          0          0          0          0      8         -1
## 5          1          1          0          0          0      1          1
## 6          1          1          0          1          1      1          2
##   BORROW_1 BORROW_2 SHOCKS_1 SHOCKS_2 SHOCKS_3 SHOCKS_4 SHOCKS_5 SHOCKS_6
## 1        0        1        0        0        0        0        1        0
## 2        1        1        0        0        0        0        0        0
## 3       -1       -1        0        0        0        0        0        0
## 4       -1       -1        0        0        0        0        0        0
## 5        0        1        0        0        0        0        0        0
## 6        0        1        0        0        0        0        0        0
##   SHOCKS_7 SHOCKS_8 SHOCKS_9 SHOCKS_10 SHOCKS_11 SHOCKS_12 MANAGE2 PAIDHELP
```

```
## 1        0        1        0         0         0         0       3       -2
## 2        0        0        0         0         0         1       1        0
## 3        0        0        0         0         0         1       3       -2
## 4        0        0        0         0         0         1      -1       -2
## 5        0        0        0         1         0         0       2        1
## 6        0        0        1         0         0         0       2        0
##   HSLOC PAREDUC FINSOC2_1 FINSOC2_2 FINSOC2_3 FINSOC2_4 FINSOC2_5 FINSOC2_6
## 1    -1       4         0         1         1         1         1         1
## 2     1       2         0         0         0         0         1         0
## 3     1       3         0         0         0         1         0         0
## 4     1       2         0         1         1         1         1         1
## 5    -1       2         1         0         1         1         0         0
## 6     1       4         1         1         0         1         1         0
##   FINSOC2_7 OBJNUMERACY1 ON2correct ON1correct MATERIALISM_1 MATERIALISM_2
## 1         1            3          1          0            3             5
## 2         0            2          1          1            3             4
## 3         1            2          0          1            4             4
## 4         1           -1          0          0           -1            -1
## 5         1            2          1          1            3             3
## 6         1            1          1          0            3             3
##   MATERIALISM_3 CONNECT HEALTH SCFHORIZON DISCOUNT MEMLOSS DISTRESS
## 1             4      80      2          3        2       0        4
## 2             3      95      3          3        1       0        3
## 3             3      50      3          4        2       0        2
## 4            -1      -1     -1         -1       -1      -1       -1
## 5             3       0      3          3        2       1        3
## 6             2      80      5          1        1       0        4
##   SELFCONTROL_1 SELFCONTROL_2 SELFCONTROL_3 OUTLOOK_1 OUTLOOK_2
## 1             2             3             3         3         2
## 2             2             4             3         2         5
## 3             3             3             3         4         4
## 4            -1            -1            -1        -1        -1
## 5             3             3             3         3         3
## 6             1             3             3         3         5
##   INTERCONNECTIONS_1 INTERCONNECTIONS_2 INTERCONNECTIONS_3 INTERCONNECTIONS_4
## 1                  0                  0                  0                  0
## 2                  0                  1                  0                  0
## 3                  0                  1                  0                  0
## 4                  0                  0                  0                  0
## 5                  0                  0                  0                  0
## 6                  0                  1                  0                  0
##   INTERCONNECTIONS_5 INTERCONNECTIONS_6 INTERCONNECTIONS_7 INTERCONNECTIONS_8
## 1                  1                  0                  1                  1
## 2                  0                  0                  0                  0
## 3                  0                  0                  0                  0
## 4                  0                  0                  0                  0
## 5                  0                  0                  0                  1
## 6                  0                  1                  0                  0
##   INTERCONNECTIONS_9 INTERCONNECTIONS_10 PEM HOUSESAT SOCSEC1 SOCSEC2 SOCSEC3
## 1                  0                   0   3        4       1      62      -2
## 2                  0                   0   4        3      -2      -2      66
## 3                  0                   0   6        3      -2      -2      68
## 4                  0                   1  -1       -1      -2      -2      -1
## 5                  0                   0   4        3      -2      -2      65
```

```
## 6                  0                    0   7        4      -2      -2      71
##   LIFEEXPECT HHEDUC KIDS_NoChildren KIDS_1 KIDS_2 KIDS_3 KIDS_4 EMPLOY
## 1         -2      4              -1      0      0      0      0      8
## 2         90      2               1      0      0      0      0      2
## 3         78      3               0      0      0      0      1      2
## 4         -1     -1              -1      0      0      0      0     99
## 5         75      2               1      0      0      0      0      2
## 6         10      4               1      0      0      0      0      2
##   EMPLOY1_1 EMPLOY1_2 EMPLOY1_3 EMPLOY1_4 EMPLOY1_5 EMPLOY1_6 EMPLOY1_7
## 1         0         0         0         0         0         0         0
## 2         0         1         0         0         0         0         0
## 3         0         1         0         0         0         0         0
## 4         0         0         0         0         0         0         0
## 5         0         1         0         1         0         0         0
## 6         0         1         0         0         0         0         0
##   EMPLOY1_8 EMPLOY1_9 RETIRE MILITARY Military_Status agecat generation PPEDUC
## 1         1         0      1        0               5      8          1      4
## 2         0         0     -2        0               5      3          3      2
## 3         0         0     -2        0               5      3          3      3
## 4         0         1     -2       -1              -1      3          3      2
## 5         0         0     -2        0               5      2          4      2
## 6         0         0     -2        1               3      2          4      4
##   PPETHM PPGENDER PPHHSIZE PPINCIMP PPMARIT PPMSACAT PPREG4 PPREG9 PPT01 PPT25
## 1      1        1        1        7       3        1      4      8     0     0
## 2      1        1        2        6       3        1      2      3     0     0
## 3      2        1        3        6       3        1      4      9     0     0
## 4      1        1        1        8       3        1      3      7     0     0
## 5      3        1        5        7       1        1      2      4     0     0
## 6      1        1        2        7       1        1      2      3     0     0
##   PPT612 PPT1317 PPT18OV PCTLT200FPL   finalwt
## 1      0       0       1           0 0.3672919
## 2      0       0       2           0 1.3275607
## 3      0       1       2           1 0.8351558
## 4      0       0       1           0 1.4108710
## 5      1       0       4           1 4.2606681
## 6      0       0       2           0 0.7600609
```

```
# Show dimension of the dataframe
dim(df)
```

```
## [1] 6394  217
```

```
# Set PUF_ID as index
df <- df %>%
  remove_rownames %>%
  column_to_rownames(var="PUF_ID")

# notice that negative values are invalid entries,
# so replacing them with NA
for (i in 1:nrow(df)){
  for (j in 1:ncol(df)){
    if (df[i,j] < 0){
      df[i,j] = NA
```

```
    }
  }
}
```

```
# use knn impute to resolve NA problem
df = knn.impute(as.matrix(df)) %>%
  as.data.frame()

head(df)
```

```
##       sample fpl SWB_1 SWB_2 SWB_3 FWBscore FWB1_1 FWB1_2 FWB1_3 FWB1_4 FWB1_5
## 10350      2   3     5     5     6       55      3      3      3      3      2
## 7740       1   3     6     6     6       51      2      2      3      3      3
## 13699      1   3     4     3     4       49      3      3      3      3      3
## 7267       1   3     6     6     6       49      3      3      3      3      3
## 7375       1   3     4     4     4       49      3      3      3      3      3
## 10910      1   3     5     7     5       67      5      1      1      1      1
##       FWB1_6 FWB2_1 FWB2_2 FWB2_3 FWB2_4 FSscore FS1_1 FS1_2 FS1_3 FS1_4 FS1_5
## 10350      3      2      3      2      4      44     3     3     4     3     3
## 7740       4      2      2      2      3      43     3     3     3     3     4
## 13699      3      3      3      3      3      42     3     3     3     3     3
## 7267       3      3      3      3      3      42     3     3     3     3     3
## 7375       3      3      3      3      3      42     3     3     3     3     3
## 10910      1      2      5      2      2      57     4     4     4     4     3
##       FS1_6 FS1_7 FS2_1 FS2_2 FS2_3 SUBKNOWL1 ACT1_1 ACT1_2 FINGOALS PROPPLAN_1
## 10350     3     4     4     3     4         5      4      3        1          5
## 7740      3     2     4     3     2         5      4      3        0          3
## 13699     3     3     3     3     3         5      3      3        1          4
## 7267      3     3     3     3     3         4      3      3        0          3
## 7375      3     3     3     3     3         4      3      3        1          3
## 10910     4     4     4     4     1         6      5      4        1          5
##       PROPPLAN_2 PROPPLAN_3 PROPPLAN_4 MANAGE1_1 MANAGE1_2 MANAGE1_3 MANAGE1_4
## 10350          4          4          3         4         4         2         4
## 7740           2          2          1         4         4         1         4
## 13699          4          4          4         3         3         3         3
## 7267           3          3          3         4         4         2         4
## 7375           3          3          3         3         3         3         3
## 10910          4          3          4         5         3         5         5
##       SAVEHABIT FRUGALITY AUTOMATED_1 AUTOMATED_2 ASK1_1 ASK1_2 SUBNUMERACY2
## 10350         4         6           0           0      4      3            3
## 7740          1         5           0           0      3      2            5
## 13699         5         5           1           1      4      4            4
## 7267          4         6           7           7      3      3            3
## 7375          4         4           0           1      3      3            4
## 10910         4         5           1           1      5      3            5
##       SUBNUMERACY1 CHANGEABLE GOALCONF LMscore FINKNOWL1 FINKNOWL2 FINKNOWL3
## 10350            3          4        3       3         1         3         2
## 7740             5          2        3       3         1         3         2
## 13699            4          6        3       3         1         3         2
## 7267             4          4        3       2         1         1         2
## 7375             4          4        3       1         2         2         2
## 10910            6          1        4       3         1         3         2
##       FK1correct FK2correct FK3correct KHscore KHKNOWL1 KHKNOWL2 KHKNOWL3
```

```
## 10350          1          1          1   1.267        3        3        2
## 7740           1          1          1   0.242        2        3        3
## 13699          1          1          1   0.242        3        3        1
## 7267           1          0          1   0.242        2        2        3
## 7375           0          0          1   0.242        1        1        3
## 10910          1          1          1   0.242        3        3        2
##       KHKNOWL4 KHKNOWL5 KHKNOWL6 KHKNOWL7 KHKNOWL8 KHKNOWL9 KH1correct
## 10350        1        1        2        4        2        1          1
## 7740         1        1        2        2        3        1          0
## 13699        1        2        2        2        2        1          1
## 7267         1        2        2        2        2        1          0
## 7375         2        2        1        3        2        2          0
## 10910        1        1        2        3        4        1          1
##       KH2correct KH3correct KH4correct KH5correct KH6correct KH7correct
## 10350          1          1          1          1          1          1
## 7740           1          0          1          1          1          0
## 13699          1          0          1          0          1          0
## 7267           0          0          1          0          1          0
## 7375           0          0          0          0          0          0
## 10910          1          1          1          1          1          0
##       KH8correct KH9correct ENDSMEET HOUSING LIVINGARRANGEMENT HOUSERANGES
## 10350          1          1        2       1                 1           4
## 7740           0          1        2       1                 2           4
## 13699          1          1        1       1                 2           3
## 7267           0          1        2       2                 2          99
## 7375           1          0        2       2                 3           2
## 10910          0          1        1       1                 2           4
##       IMPUTATION_FLAG VALUERANGES MORTGAGE SAVINGSRANGES PRODHAVE_1 PRODHAVE_2
## 10350               0           2        2             6          1          1
## 7740                0           2        2             2          1          0
## 13699               0           3        2             4          1          1
## 7267                0           1        2            99          0          0
## 7375                0           1        2            98          1          0
## 10910               0           1        2             5          1          1
##       PRODHAVE_3 PRODHAVE_4 PRODHAVE_5 PRODHAVE_6 PRODHAVE_7 PRODHAVE_8
## 10350          1          1          1          0          0          0
## 7740           1          0          0          0          0          0
## 13699          0          1          1          0          0          0
## 7267           0          0          0          0          0          0
## 7375           0          1          0          0          0          0
## 10910          1          1          0          0          0          1
##       PRODHAVE_9 PRODUSE_1 PRODUSE_2 PRODUSE_3 PRODUSE_4 PRODUSE_5 PRODUSE_6
## 10350          0         0         0         0         0         0         1
## 7740           0         0         0         0         0         0         1
## 13699          0         0         0         1         1         1         0
## 7267           1         0         0         0         0         0         1
## 7375           0         0         1         0         0         0         0
## 10910          0         0         0         0         0         0         1
##       CONSPROTECT1 CONSPROTECT2 CONSPROTECT3 EARNERS VOLATILITY SNAP
## 10350            3            2            1       1          2    0
## 7740             2            1            0       2          2    0
## 13699            3            3            1       2          3    0
## 7267             3            2            0       1          1    0
## 7375             3            2            0       2          2    8
```

```
## 10910            2            2            1       1          1    0
##       MATHARDSHIP_1 MATHARDSHIP_2 MATHARDSHIP_3 MATHARDSHIP_4 MATHARDSHIP_5
## 10350             1             1             1             1             1
## 7740              1             1             1             1             1
## 13699             1             1             1             1             1
## 7267              2             2             2             2             2
## 7375              2             2             2             2             2
## 10910             1             1             1             1             1
##       MATHARDSHIP_6 COLLECT REJECTED_1 REJECTED_2 ABSORBSHOCK BENEFITS_1
## 10350             1       1          0          0           4          0
## 7740              1       1          0          0           3          1
## 13699             1       0          0          0           4          1
## 7267              2       0          0          1           8          0
## 7375              2       8          0          0           2          1
## 10910             1       0          0          0           4          1
##       BENEFITS_2 BENEFITS_3 BENEFITS_4 BENEFITS_5 FRAUD2 COVERCOSTS BORROW_1
## 10350          0          1          0          0      8          1        0
## 7740           0          0          0          1      0          3        1
## 13699          0          0          0          0      0          2        1
## 7267           0          0          0          0      8          2        0
## 7375           1          0          0          0      1          1        0
## 10910          1          0          1          1      1          2        0
##       BORROW_2 SHOCKS_1 SHOCKS_2 SHOCKS_3 SHOCKS_4 SHOCKS_5 SHOCKS_6 SHOCKS_7
## 10350        1        0        0        0        0        1        0        0
## 7740         1        0        0        0        0        0        0        0
## 13699        0        0        0        0        0        0        0        0
## 7267         0        0        0        0        0        0        0        0
## 7375         1        0        0        0        0        0        0        0
## 10910        1        0        0        0        0        0        0        0
##       SHOCKS_8 SHOCKS_9 SHOCKS_10 SHOCKS_11 SHOCKS_12 MANAGE2 PAIDHELP HSLOC
## 10350        1        0         0         0         0       3        0     1
## 7740         0        0         0         0         1       1        0     1
## 13699        0        0         0         0         1       3        0     1
## 7267         0        0         0         0         1       3        0     1
## 7375         0        0         1         0         0       2        1     1
## 10910        0        1         0         0         0       2        0     1
##       PAREDUC FINSOC2_1 FINSOC2_2 FINSOC2_3 FINSOC2_4 FINSOC2_5 FINSOC2_6
## 10350       4         0         1         1         1         1         1
## 7740        2         0         0         0         0         1         0
## 13699       3         0         0         0         1         0         0
## 7267        2         0         1         1         1         1         1
## 7375        2         1         0         1         1         0         0
## 10910       4         1         1         0         1         1         0
##       FINSOC2_7 OBJNUMERACY1 ON2correct ON1correct MATERIALISM_1 MATERIALISM_2
## 10350         1            3          1          0            3            5
## 7740          0            2          1          1            3            4
## 13699         1            2          0          1            4            4
## 7267          1            2          0          0            3            3
## 7375          1            2          1          1            3            3
## 10910         1            1          1          0            3            3
##       MATERIALISM_3 CONNECT HEALTH SCFHORIZON DISCOUNT MEMLOSS DISTRESS
## 10350             4      80      2          3        2       0        4
## 7740              3      95      3          3        1       0        3
## 13699             3      50      3          4        2       0        2
```

```
## 7267              3     100      3          3        1       0        3
## 7375              3       0      3          3        2       1        3
## 10910             2      80      5          1        1       0        4
##       SELFCONTROL_1 SELFCONTROL_2 SELFCONTROL_3 OUTLOOK_1 OUTLOOK_2
## 10350             2             3             3         3         2
## 7740              2             4             3         2         5
## 13699             3             3             3         4         4
## 7267              3             3             3         3         3
## 7375              3             3             3         3         3
## 10910             1             3             3         3         5
##       INTERCONNECTIONS_1 INTERCONNECTIONS_2 INTERCONNECTIONS_3
## 10350                  0                  0                  0
## 7740                   0                  1                  0
## 13699                  0                  1                  0
## 7267                   0                  0                  0
## 7375                   0                  0                  0
## 10910                  0                  1                  0
##       INTERCONNECTIONS_4 INTERCONNECTIONS_5 INTERCONNECTIONS_6
## 10350                  0                  1                  0
## 7740                   0                  0                  0
## 13699                  0                  0                  0
## 7267                   0                  0                  0
## 7375                   0                  0                  0
## 10910                  0                  0                  1
##       INTERCONNECTIONS_7 INTERCONNECTIONS_8 INTERCONNECTIONS_9
## 10350                  1                  1                  0
## 7740                   0                  0                  0
## 13699                  0                  0                  0
## 7267                   0                  0                  0
## 7375                   0                  1                  0
## 10910                  0                  0                  0
##       INTERCONNECTIONS_10 PEM HOUSESAT SOCSEC1 SOCSEC2 SOCSEC3 LIFEEXPECT
## 10350                   0   3        4       1      62      70         90
## 7740                    0   4        3       0      62      66         90
## 13699                   0   6        3       1      62      68         78
## 7267                    1   4        3       1      65      70         50
## 7375                    0   4        3       1      65      65         75
## 10910                   0   7        4       1      66      71         10
##       HHEDUC KIDS_NoChildren KIDS_1 KIDS_2 KIDS_3 KIDS_4 EMPLOY EMPLOY1_1
## 10350      4               1      0      0      0      0      8         0
## 7740       2               1      0      0      0      0      2         0
## 13699      3               0      0      0      0      1      2         0
## 7267       3               1      0      0      0      0     99         0
## 7375       2               1      0      0      0      0      2         0
## 10910      4               1      0      0      0      0      2         0
##       EMPLOY1_2 EMPLOY1_3 EMPLOY1_4 EMPLOY1_5 EMPLOY1_6 EMPLOY1_7 EMPLOY1_8
## 10350         0         0         0         0         0         0         1
## 7740          1         0         0         0         0         0         0
## 13699         1         0         0         0         0         0         0
## 7267          0         0         0         0         0         0         0
## 7375          1         0         1         0         0         0         0
## 10910         1         0         0         0         0         0         0
##       EMPLOY1_9 RETIRE MILITARY Military_Status agecat generation PPEDUC PPETHM
## 10350         0      1        0               5      8          1      4      1
```

```
## 7740          0      2        0               5      3          3      2      1
## 13699         0      2        0               5      3          3      3      2
## 7267          1      2        0               5      3          3      2      1
## 7375          0      2        0               5      2          4      2      3
## 10910         0      2        1               3      2          4      4      1
##       PPGENDER PPHHSIZE PPINCIMP PPMARIT PPMSACAT PPREG4 PPREG9 PPT01 PPT25
## 10350        1        1        7       3        1      4      8     0     0
## 7740         1        2        6       3        1      2      3     0     0
## 13699        1        3        6       3        1      4      9     0     0
## 7267         1        1        8       3        1      3      7     0     0
## 7375         1        5        7       1        1      2      4     0     0
## 10910        1        2        7       1        1      2      3     0     0
##       PPT612 PPT1317 PPT18OV PCTLT200FPL   finalwt
## 10350      0       0       1           0 0.3672919
## 7740       0       0       2           0 1.3275607
## 13699      0       1       2           1 0.8351558
## 7267       0       0       1           0 1.4108710
## 7375       1       0       4           1 4.2606681
## 10910      0       0       2           0 0.7600609
```

```
# Check NA values
colSums(is.na(df))
```

```
##             sample                fpl              SWB_1              SWB_2
##                  0                  0                  0                  0
##              SWB_3           FWBscore              FWB1_1             FWB1_2
##                  0                  0                  0                  0
##             FWB1_3             FWB1_4             FWB1_5             FWB1_6
##                  0                  0                  0                  0
##             FWB2_1             FWB2_2             FWB2_3             FWB2_4
##                  0                  0                  0                  0
##            FSscore              FS1_1              FS1_2              FS1_3
##                  0                  0                  0                  0
##              FS1_4              FS1_5              FS1_6              FS1_7
##                  0                  0                  0                  0
##              FS2_1              FS2_2              FS2_3          SUBKNOWL1
##                  0                  0                  0                  0
##             ACT1_1             ACT1_2           FINGOALS         PROPPLAN_1
##                  0                  0                  0                  0
##         PROPPLAN_2         PROPPLAN_3         PROPPLAN_4          MANAGE1_1
##                  0                  0                  0                  0
##          MANAGE1_2          MANAGE1_3          MANAGE1_4          SAVEHABIT
##                  0                  0                  0                  0
##          FRUGALITY        AUTOMATED_1        AUTOMATED_2             ASK1_1
##                  0                  0                  0                  0
##             ASK1_2        SUBNUMERACY2       SUBNUMERACY1         CHANGEABLE
##                  0                  0                  0                  0
##           GOALCONF            LMscore          FINKNOWL1          FINKNOWL2
##                  0                  0                  0                  0
##          FINKNOWL3         FK1correct         FK2correct         FK3correct
##                  0                  0                  0                  0
##            KHscore            KHKNOWL1           KHKNOWL2           KHKNOWL3
##                  0                  0                  0                  0
##           KHKNOWL4           KHKNOWL5           KHKNOWL6           KHKNOWL7
```

```
##                 0                 0                 0                 0
##         KHKNOWL8          KHKNOWL9       KH1correct       KH2correct
##                 0                 0                 0                 0
##       KH3correct        KH4correct       KH5correct       KH6correct
##                 0                 0                 0                 0
##       KH7correct        KH8correct       KH9correct          ENDSMEET
##                 0                 0                 0                 0
##          HOUSING  LIVINGARRANGEMENT      HOUSERANGES  IMPUTATION_FLAG
##                 0                 0                 0                 0
##      VALUERANGES          MORTGAGE    SAVINGSRANGES       PRODHAVE_1
##                 0                 0                 0                 0
##       PRODHAVE_2        PRODHAVE_3       PRODHAVE_4       PRODHAVE_5
##                 0                 0                 0                 0
##       PRODHAVE_6        PRODHAVE_7       PRODHAVE_8       PRODHAVE_9
##                 0                 0                 0                 0
##        PRODUSE_1         PRODUSE_2         PRODUSE_3         PRODUSE_4
##                 0                 0                 0                 0
##        PRODUSE_5         PRODUSE_6      CONSPROTECT1      CONSPROTECT2
##                 0                 0                 0                 0
##     CONSPROTECT3           EARNERS        VOLATILITY              SNAP
##                 0                 0                 0                 0
##    MATHARDSHIP_1     MATHARDSHIP_2     MATHARDSHIP_3     MATHARDSHIP_4
##                 0                 0                 0                 0
##    MATHARDSHIP_5     MATHARDSHIP_6           COLLECT       REJECTED_1
##                 0                 0                 0                 0
##       REJECTED_2       ABSORBSHOCK       BENEFITS_1       BENEFITS_2
##                 0                 0                 0                 0
##       BENEFITS_3        BENEFITS_4       BENEFITS_5           FRAUD2
##                 0                 0                 0                 0
##       COVERCOSTS          BORROW_1          BORROW_2          SHOCKS_1
##                 0                 0                 0                 0
##         SHOCKS_2          SHOCKS_3          SHOCKS_4          SHOCKS_5
##                 0                 0                 0                 0
##         SHOCKS_6          SHOCKS_7          SHOCKS_8          SHOCKS_9
##                 0                 0                 0                 0
##        SHOCKS_10         SHOCKS_11         SHOCKS_12           MANAGE2
##                 0                 0                 0                 0
##         PAIDHELP             HSLOC          PAREDUC         FINSOC2_1
##                 0                 0                 0                 0
##       FINSOC2_2         FINSOC2_3         FINSOC2_4         FINSOC2_5
##                 0                 0                 0                 0
##       FINSOC2_6         FINSOC2_7      OBJNUMERACY1       ON2correct
##                 0                 0                 0                 0
##       ON1correct     MATERIALISM_1     MATERIALISM_2     MATERIALISM_3
##                 0                 0                 0                 0
##          CONNECT            HEALTH       SCFHORIZON          DISCOUNT
##                 0                 0                 0                 0
##          MEMLOSS          DISTRESS     SELFCONTROL_1     SELFCONTROL_2
##                 0                 0                 0                 0
##    SELFCONTROL_3         OUTLOOK_1         OUTLOOK_2 INTERCONNECTIONS_1
##                 0                 0                 0                 0
## INTERCONNECTIONS_2 INTERCONNECTIONS_3 INTERCONNECTIONS_4 INTERCONNECTIONS_5
##                 0                 0                 0                 0
## INTERCONNECTIONS_6 INTERCONNECTIONS_7 INTERCONNECTIONS_8 INTERCONNECTIONS_9
```

```
##                   0                   0                   0                   0
## INTERCONNECTIONS_10                 PEM           HOUSESAT             SOCSEC1
##                   0                   0                   0                   0
##             SOCSEC2             SOCSEC3          LIFEEXPECT             HHEDUC
##                   0                   0                   0                   0
##     KIDS_NoChildren              KIDS_1              KIDS_2              KIDS_3
##                   0                   0                   0                   0
##              KIDS_4              EMPLOY          EMPLOY1_1           EMPLOY1_2
##                   0                   0                   0                   0
##           EMPLOY1_3           EMPLOY1_4           EMPLOY1_5           EMPLOY1_6
##                   0                   0                   0                   0
##           EMPLOY1_7           EMPLOY1_8           EMPLOY1_9              RETIRE
##                   0                   0                   0                   0
##            MILITARY     Military_Status              agecat          generation
##                   0                   0                   0                   0
##              PPEDUC              PPETHM           PPGENDER            PPHHSIZE
##                   0                   0                   0                   0
##            PPINCIMP             PPMARIT           PPMSACAT             PPREG4
##                   0                   0                   0                   0
##              PPREG9               PPT01               PPT25              PPT612
##                   0                   0                   0                   0
##             PPT1317             PPT18OV         PCTLT200FPL             finalwt
##                   0                   0                   0                   0
```

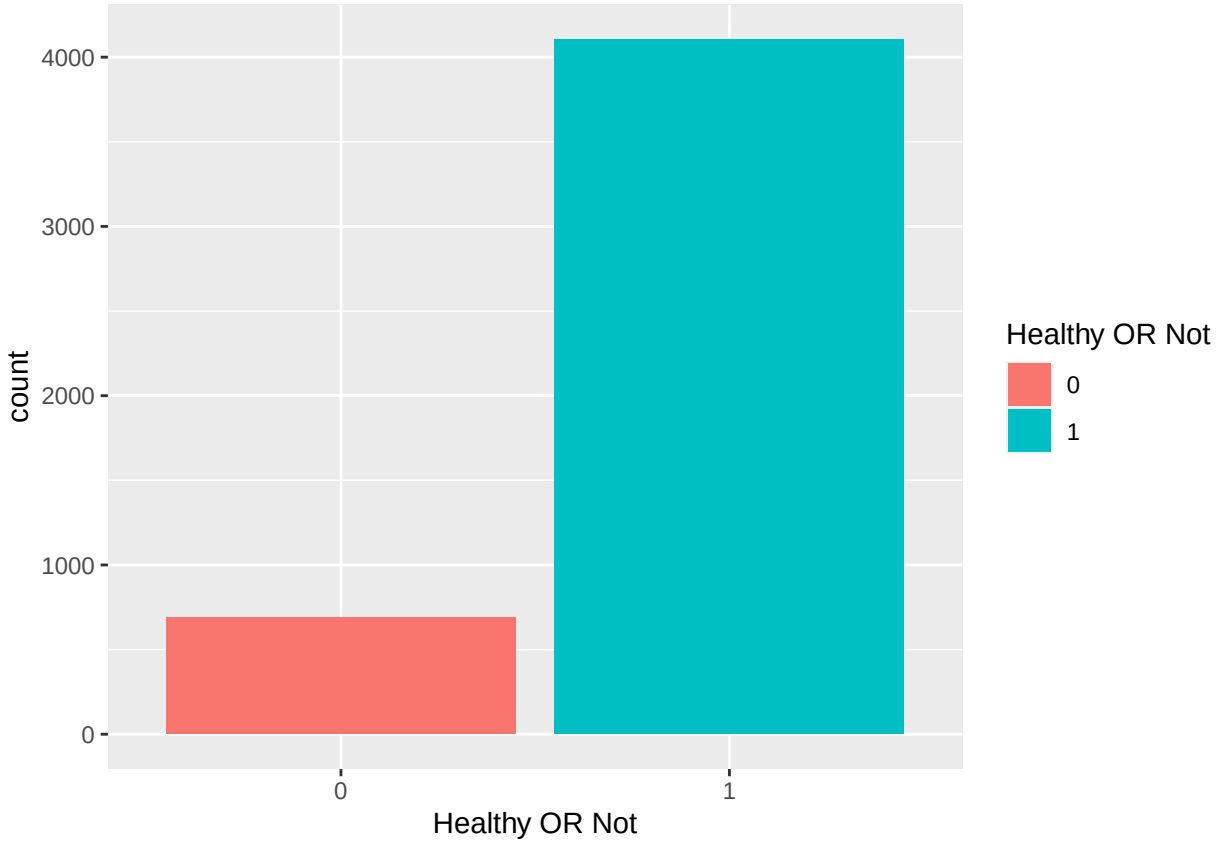
```
df <- df %>%
  mutate(HEALTH = ifelse(HEALTH == -1 | HEALTH == 1 | HEALTH == 2, 0, 1))
```

```
df <- df %>%
  mutate(HEALTH = as.factor(HEALTH))

# Train / Test dataset split

## 75% of the sample size
smp_size <- floor(0.75 * nrow(df))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(df)), size = smp_size)

train_df <- df[train_ind, ]
test_df <- df[-train_ind, ]

# Export to csv
write.csv(train_df, "../data/train_data.csv", row.names = FALSE)
write.csv(test_df, "../data/test_data.csv", row.names = FALSE)
```

```
# Imbalance Data

as.data.frame(table(train_df$HEALTH))
```

```
##   Var1 Freq
## 1    0  690
## 2    1 4105
```

```
ggplot(train_df , aes(x = factor(HEALTH), fill = factor(HEALTH))) +
  geom_bar() +
  xlab("Healthy OR Not") +
  labs(fill="Healthy OR Not")
```




```
# SMOTE : : Synthetic Minority Oversampling Technique To Handle Class Imbalancy In Binary Classificatio
library(DMwR)
```

```
## Warning: package 'DMwR' was built under R version 4.0.3
```

```
## Loading required package: grid
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

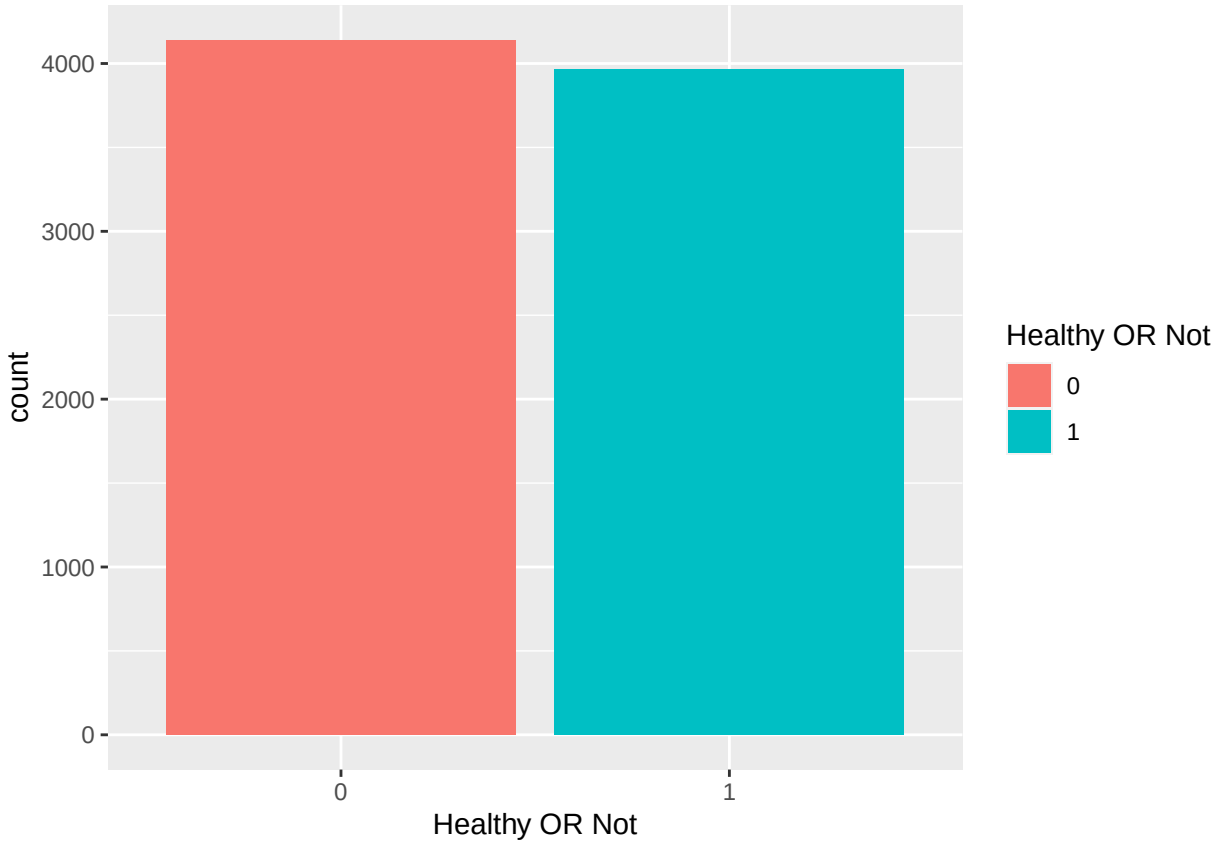
```
##
## Attaching package: 'DMwR'
```

```
## The following object is masked from 'package:bnstruct':
##
##     bootstrap
```

```
train_balanced_df <- SMOTE(HEALTH~., train_df, perc.over = 500, perc.under = 115, k = 5)

as.data.frame(table(train_balanced_df$HEALTH))
```

```
##   Var1 Freq
## 1    0 4140
## 2    1 3967
```

```
ggplot(train_balanced_df , aes(x = factor(HEALTH), fill = factor(HEALTH))) +
  geom_bar() +
  xlab("Healthy OR Not") +
  labs(fill="Healthy OR Not")
```




```
# Export to csv
write.csv(train_balanced_df, "../data/train_balanced_data.csv", row.names = FALSE)
```