成年人高血压关联因素探索

吴沛豪

2023-05-16

目录

1 环境准备

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Chinese (Simplified)_China.utf8
## [2] LC_CTYPE=Chinese (Simplified)_China.utf8
## [3] LC_MONETARY=Chinese (Simplified)_China.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=Chinese (Simplified)_China.utf8
##
## attached base packages:
## [1] stats4    grid      stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] ggpubr_0.6.0         runway_0.0.0.9000    pROC_1.18.0
##  [4] tensorflow_2.11.0    keras_2.11.1         party_1.3-13
##  [7] strucchange_1.5-3    sandwich_3.0-2       zoo_1.8-11
## [10] modeltools_0.2-23    mvtnorm_1.1-3        RWeka_0.4-46
## [13] rattle_5.5.1         bitops_1.0-7         xgboost_1.7.5.1
## [16] glmnet_4.1-7         Matrix_1.5-4         e1071_1.7-13
## [19] MASS_7.3-58.3        caret_6.0-94         lattice_0.20-45
## [22] randomForest_4.7-1.1 rpart.plot_3.1.1     rpart_4.1.19
## [25] mice_3.15.0          yardstick_1.1.0      workflowsets_1.0.1
## [28] workflows_1.1.3      tune_1.1.0           rsample_1.1.1
## [31] recipes_1.0.5        parsnip_1.0.4        modeldata_1.1.0
## [34] infer_1.0.4          dials_1.2.0          scales_1.2.1
## [37] broom_1.0.4          tidymodels_1.0.0     VIM_6.2.2
## [40] colorspace_2.1-0     patchwork_1.1.2      qgraph_1.9.4
## [43] reshape2_1.4.4       lubridate_1.9.2      forcats_1.0.0
## [46] stringr_1.5.0        dplyr_1.1.1          purrr_1.0.1
## [49] readr_2.1.4          tidyr_1.3.0          tibble_3.2.1
## [52] ggplot2_3.4.2        tidyverse_2.0.0      igraph_1.4.1
##
```

```
## loaded via a namespace (and not attached):
##   [1] backports_1.4.1      Hmisc_5.0-1          plyr_1.8.8
##   [4] sp_1.6-0             splines_4.2.2        listenv_0.9.0
##   [7] tfruns_1.5.1         TH.data_1.1-1        digest_0.6.31
##  [10] foreach_1.5.2        htmltools_0.5.4      fansi_1.0.4
##  [13] magrittr_2.0.3       checkmate_2.1.0      cluster_2.1.4
##  [16] tzdb_0.3.0           globals_0.16.2       gower_1.0.1
##  [19] matrixStats_0.63.0   hardhat_1.3.0        timechange_0.2.0
##  [22] jpeg_0.1-10          xfun_0.37            libcoin_1.0-9
##  [25] jsonlite_1.8.4       zeallot_0.1.0        survival_3.5-5
##  [28] iterators_1.0.14     glue_1.6.2           gtable_0.3.3
##  [31] ipred_0.9-14         car_3.1-2            shape_1.4.6
##  [34] future.apply_1.10.0  DEoptimR_1.0-12      abind_1.4-5
##  [37] rstatix_0.7.2        Rcpp_1.0.10          laeken_0.5.2
##  [40] htmlTable_2.4.1      reticulate_1.28      GPfit_1.0-8
##  [43] foreign_0.8-84       proxy_0.4-27         Formula_1.2-5
##  [46] lava_1.7.2.1         prodlim_2023.03.31   vcd_1.4-11
##  [49] htmlwidgets_1.6.2    lavaan_0.6-15        rJava_1.0-6
##  [52] pkgconfig_2.0.3      nnet_7.3-18          utf8_1.2.3
##  [55] tidyselect_1.2.0     rlang_1.1.0          DiceDesign_1.9
##  [58] munsell_0.5.0        tools_4.2.2          cli_3.6.0
##  [61] generics_0.1.3       ranger_0.15.1        fdrtool_1.2.17
##  [64] evaluate_0.20        fastmap_1.1.1        yaml_2.3.7
##  [67] rticles_0.24         ModelMetrics_1.2.2.2 knitr_1.42
##  [70] robustbase_0.95-1    coin_1.4-2           glasso_1.11
##  [73] pbapply_1.7-0        future_1.32.0        nlme_3.1-162
##  [76] whisker_0.4.1        compiler_4.2.2       rstudioapi_0.14
##  [79] png_0.1-8            ggsignif_0.6.4       lhs_1.1.6
##  [82] pbivnorm_0.6.0       stringi_1.7.12       psych_2.3.3
##  [85] RWekajars_3.9.3-2    vctrs_0.6.1          pillar_1.9.0
##  [88] lifecycle_1.0.3      furrr_0.3.1          lmtest_0.9-40
##  [91] data.table_1.14.8    corpcor_1.6.10       R6_2.5.1
##  [94] gridExtra_2.3        parallelly_1.35.0    codetools_0.2-19
##  [97] boot_1.3-28.1        gtools_3.9.4         withr_2.5.0
## [100] mnormt_2.1.1         multcomp_1.4-23      parallel_4.2.2
## [103] hms_1.1.3            quadprog_1.5-8       timeDate_4022.108
## [106] class_7.3-21         rmarkdown_2.21       carData_3.0-5
## [109] base64enc_0.1-3
```

```
if (length(tf$config$list_physical_devices("GPU"))) {
  message("TensorFlow **IS** using the GPU")
} else {
```

```
  message("TensorFlow **IS NOT** using the GPU")
}
```

2 数据准备

2.1 运行 DataClean.Rmd 脚本以获取清洗后数据；

2.2 运行 DataClean2.Rmd 脚本重编码变量：

2.3 运行 DataClean3.Rmd 脚本将有序分类变量及无序二分类变量转换成 numeric

2.4 运行 data_use2.R 脚本合并数据

```
load(file = paste0(getwd(), "/data_use/data_use_4.RData"))
```

2.5 检查数据

```
str(YData)
```

```
## 'data.frame':    4335 obs. of  4 variables:
##  $ SEQN    : num  93705 93706 93708 93711 93712 ...
##  $ BPXSY   : num  200 111 142 101 113 ...
##  $ BPXDI   : num  68 73.3 76 66.7 70 ...
##  $ response: num  1 0 1 0 0 0 1 0 0 0 ...
```

```
str(XData)
```

```
## 'data.frame':    4335 obs. of  140 variables:
##  $ SEQN    : num  93705 93706 93708 93711 93712 ...
##  $ RIAGENDR: num  0 1 0 1 1 1 0 1 1 1 ...
##  $ RIDRETH3: Factor w/ 6 levels "Mexican American",..: 4 5 5 5 1 3 4 6 5 3 ...
##  $ DMDBORN4: num  1 1 0 0 0 1 1 1 0 1 ...
##  $ DMDCITZN: num  1 1 1 1 0 1 1 1 1 1 ...
##  $ DMDHHSIZ: num  1 5 2 3 4 1 3 5 3 2 ...
##  $ DMDFMSIZ: num  1 5 2 3 4 1 3 5 3 1 ...
##  $ DMDHHSZA: num  0 0 0 0 0 0 0 1 0 0 ...
##  $ DMDHHSZB: num  0 0 0 0 2 0 1 2 0 0 ...
##  $ DMDHHSZE: num  1 1 2 0 0 1 0 1 1 0 ...
##  $ DMDHRGND: num  0 1 1 1 0 1 1 1 1 1 ...
##  $ DMDHRAGZ: num  4 4 4 3 3 4 3 4 4 2 ...
##  $ DMDHREDZ: num  1 3 1 3 1 2 2 2 3 1 ...
##  $ DMDHRMAZ: Factor w/ 3 levels "Married/Living with partner",..: 2 1 1 1 2 2 1 1 1 3 ...
##  $ INDHHIN2: num  1 2 2 3 1 2 2 2 3 1 ...
##  $ INDFMIN2: num  1 2 2 3 1 2 2 2 3 1 ...
```

```
##  $ AGE     : num  66 18 66 56 18 67 54 71 61 22 ...
##  $ DMDEDUC : num  1 2 1 3 1 2 3 2 3 2 ...
##  $ WTDRD1  : num  7186 6464 10826 9098 60947 ...
##  $ WTDR2D  : num  5640 6464 22482 8230 89066 ...
##  $ DBQ095Z : num  1 3 1 3 3 3 3 3 3 3 ...
##  $ DRQSPREP: num  3 3 4 2 3 1 4 4 3 3 ...
##  $ DRQSDIET: num  0 0 0 1 0 0 0 0 0 0 ...
##  $ DR1TNUMF: int  17 8 14 27 12 17 16 9 18 18 ...
##  $ DR1TKCAL: int  1202 1987 1251 2840 2045 2040 2493 1287 2917 3151 ...
##  $ DR1TPROT: num  20 94.2 51 101.3 99.7 ...
##  $ DR1TCARB: num  157.4 89.8 123.7 339.6 268.2 ...
##  $ DR1TSUGR: num  91.5 14.7 49.8 148.2 125 ...
##  $ DR1TFIBE: num  8.4 7.1 16.6 44.5 22.3 14.6 11.1 2.4 31.4 18 ...
##  $ DR1TTFAT: num  57 137.4 65.5 124.2 63.9 ...
##  $ DR1TSFAT: num  16.4 35.2 17.4 41.3 15.9 ...
##  $ DR1TMFAT: num  16.4 45.8 29 39.6 24.2 ...
##  $ DR1TPFAT: num  19.8 49.9 14.8 31.3 19 ...
##  $ DR1TCHOL: int  14 462 71 546 216 176 965 470 300 384 ...
##  $ DR1TATOC: num  5.66 10.02 6.2 14.27 7.05 ...
##  $ DR1TATOA: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DR1TRET : int  32 198 35 691 23 212 584 280 384 472 ...
##  $ DR1TVARA: int  436 431 236 1012 46 577 608 300 1222 886 ...
##  $ DR1TACAR: int  1551 872 323 414 51 1095 9 0 3314 287 ...
##  $ DR1TBCAR: int  4096 2363 2245 3639 171 3736 265 181 7461 4736 ...
##  $ DR1TCRYP: int  2 2 26 31 156 200 34 65 1774 167 ...
##  $ DR1TLYCO: int  1573 4605 0 23074 618 548 25 342 0 6435 ...
##  $ DR1TLZ  : int  1645 313 2148 5629 316 1745 928 628 2054 3932 ...
##  $ DR1TVB1 : num  0.589 1.152 1.143 1.79 1.619 ...
##  $ DR1TVB2 : num  1.24 1.03 0.84 3.22 1.51 ...
##  $ DR1TNIAC: num  7.58 26.83 15.37 17.38 31.34 ...
##  $ DR1TVB6 : num  0.458 1.821 1.096 2.177 2.59 ...
##  $ DR1TFOLA: int  179 267 260 609 437 262 203 120 655 519 ...
##  $ DR1TFA  : int  32 125 74 84 76 55 33 45 328 256 ...
##  $ DR1TFF  : int  146 139 185 526 361 206 169 75 327 263 ...
##  $ DR1TFDFE: int  202 354 311 669 490 300 224 152 888 696 ...
##  $ DR1TCHL : num  95 368 176 546 373 ...
##  $ DR1TVB12: num  0.33 2.3 1.09 3.62 3.62 2.55 3.72 3.24 3.22 3.02 ...
##  $ DR1TB12A: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DR1TVC  : num  21.4 9.7 146.4 124 182.1 ...
##  $ DR1TVD  : num  0.2 0.7 0.8 4.7 1.3 0.6 7 6.7 4.2 7.6 ...
##  $ DR1TVK  : num  156 138 137 277 49 ...
```

```
##  $ DR1TCALC: int  314 869 412 1635 391 583 981 623 972 1959 ...
##  $ DR1TPHOS: int  466 1025 635 2141 1256 950 1908 839 1638 2027 ...
##  $ DR1TMAGN: int  162 187 248 541 260 210 276 119 451 282 ...
##  $ DR1TIRON: num  8.8 8.52 11.49 17 12.07 ...
##  $ DR1TZINC: num  2.93 8.05 6.45 13.25 15 ...
##  $ DR1TCOPP: num  0.689 0.614 1.049 1.983 1.256 ...
##  $ DR1TSODI: int  3574 3657 2135 4382 3753 2456 5000 1430 4831 6470 ...
##  $ DR1TPOTA: int  1640 1247 1631 4457 3358 2488 2449 1634 4190 3089 ...
##  $ DR1TSELE: num  22.1 118.5 54.3 129.7 109.7 ...
##  $ DR1TCAFF: int  361 0 33 347 0 385 60 432 95 70 ...
##  $ DR1TTHEO: int  120 0 69 68 0 125 161 0 0 25 ...
##  $ DR1TALCO: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DR1TMOIS: num  1774 3405 2822 4345 3217 ...
##  $ DR1TS040: num  0.156 0.263 0.07 0.88 0.033 0.543 0.982 0.252 0.491 0.936 ...
##  $ DR1TS060: num  0.077 0.203 0.044 0.594 0.027 0.368 0.624 0.198 0.376 0.706 ...
##  $ DR1TS080: num  0.058 0.14 0.027 0.459 0.02 0.253 0.457 0.136 0.267 0.457 ...
##  $ DR1TS100: num  0.122 0.377 0.091 1.022 0.07 ...
##  $ DR1TS120: num  0.145 0.459 0.097 1.601 0.07 ...
##  $ DR1TS140: num  0.447 1.816 0.499 3.742 0.633 ...
##  $ DR1TS160: num  8.95 23.15 9.44 23.52 10.38 ...
##  $ DR1TS180: num  5.98 7.75 6.13 8.38 4.24 ...
##  $ DR1TM161: num  0.118 3.387 0.446 1.027 1.085 ...
##  $ DR1TM181: num  16 41.6 28.2 38.2 22.7 ...
##  $ DR1TM201: num  0.101 0.524 0.31 0.285 0.248 0.312 0.564 0.186 0.306 0.586 ...
##  $ DR1TM221: num  0.014 0.011 0.003 0.004 0.001 0.017 0.078 0.038 0 0.131 ...
##  $ DR1TP182: num  17.8 44.1 13.9 21.7 17.1 ...
##  $ DR1TP183: num  1.943 5.074 0.804 9.337 1.522 ...
##  $ DR1TP184: num  0 0.016 0 0 0 0 0.005 0.002 0 0.008 ...
##  $ DR1TP204: num  0.014 0.308 0.038 0.254 0.161 0.102 0.49 0.231 0.137 0.249 ...
##  $ DR1TP205: num  0.001 0.021 0.001 0.004 0.003 0.006 0.007 0.007 0.015 0.014 ...
##  $ DR1TP225: num  0.001 0.044 0.004 0.02 0.03 0.009 0.033 0.023 0.015 0.039 ...
##  $ DR1TP226: num  0.001 0.021 0 0.062 0.001 0.002 0.097 0.059 0.018 0.018 ...
##  $ DR1_300 : int  2 3 2 2 3 1 1 1 2 2 ...
##  $ DR1_320Z: num  315 3042 2160 1902 1014 ...
##  $ DR1_330Z: num  315 0 720 1902 0 ...
##  $ DR1BWATZ: num  0 3042 1440 0 1014 ...
##  $ DRD360  : num  0 1 1 0 1 0 1 1 1 1 ...
##  $ BMXWT   : num  79.5 66.3 53.5 62.1 58.9 74.9 87.1 65.6 77.7 74.4 ...
##  $ BMXHT   : num  158 176 150 171 173 ...
##  $ BMXBMI  : num  31.7 21.5 23.7 21.3 19.7 23.5 39.9 22.5 30.7 24.5 ...
##  $ BMXLEG  : num  37 46.6 31.8 40.1 44.5 39.1 26 42 37.4 44 ...
```

```
##  $ BMXARML : num  36 38.8 30.6 37.2 37.2 41.4 32 39.3 32.6 41.4 ...
##   [list output truncated]
```

3 数据标准化、虚拟化

```
source("std.r")
```

4 方差选择法

选择方差较大的特征。如果一个特征的方差很小，那么它对预测结果的影响也很小

```
df_var2 <- nearZeroVar(Data_SD[-c(1, 2, 3)], saveMetrics = TRUE)
Data_SD <- Data_SD[, !df_var2$nzv]
```

```
# 重命名变量，防止作图显示问题
Data_SD <- Data_SD %>%
  dplyr::rename(RIDRETH3MA = RIDRETH3Mexican.American,
                RIDRETH3OH = RIDRETH3Other.Hispanic,
                RIDRETH3NHW = RIDRETH3Non.Hispanic.White,
                RIDRETH3NHB = RIDRETH3Non.Hispanic.Black,
                RIDRETH3NHA = RIDRETH3Non.Hispanic.Asian,
                RIDRETH3OR = RIDRETH3Other.Race...Including.Multi.Rac,
                DMDHRMAZWDS = DMDHRMAZWidowed.Divorced.Separated,
                DMDHRMAZNM = DMDHRMAZNever.Married,
                PHDSESNA = PHDSESNafternoon,
                PHDSESNE = PHDSESNevening)
```

5 相关性计算

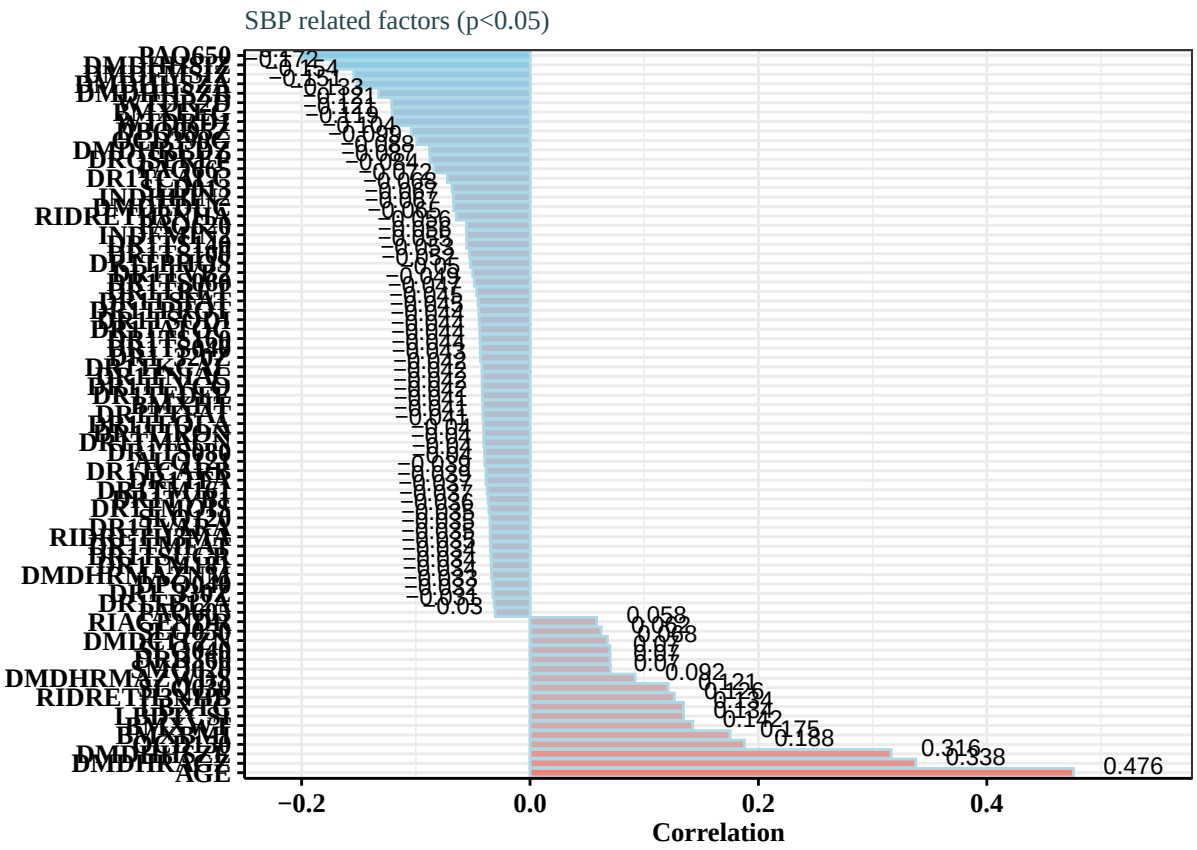
5.1 收缩压关联因素

```
cor.f <- function(m, n = 4, data = Data_SD) {
  cor <- NULL
  t_value <- NULL
  p_value <- NULL
  for (i in n:ncol(data)) {
    temp <- cor.test(data[, m], data[, i])
    cor <- cor %>% append(temp$estimate)
    t_value <- t_value %>% append(temp$statistic)
    p_value <- p_value %>% append(temp$p.value)
  }
  opt <- data.frame(
    cor = cor,
    t_value = t_value,
    p_value = p_value
  )
  rownames(opt) <- colnames(data)[n:ncol(data)]
  return(opt)
}
```

```
cor_S <- cor.f(1, data = Data_SD)
cor_S <- cor_S %>%
  dplyr::filter(p_value < 0.05) %>%
  arrange(-cor)
```

```
pA <- ggplot(data = cor_S, aes(y = cor, x = reorder(rownames(cor_S), -cor))) +
  geom_col(aes(fill = cor), col = "lightblue") +
  scale_fill_gradient(low = "skyblue", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   size = 12,
                                   hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "Correlation") +
  ggtitle("SBP related factors (p<0.05)") +
  scale_y_continuous(expand = c(0,0),
                     limits = c(-0.25, 0.58)) +
  geom_text(aes(label = round(cor, 3)),
```

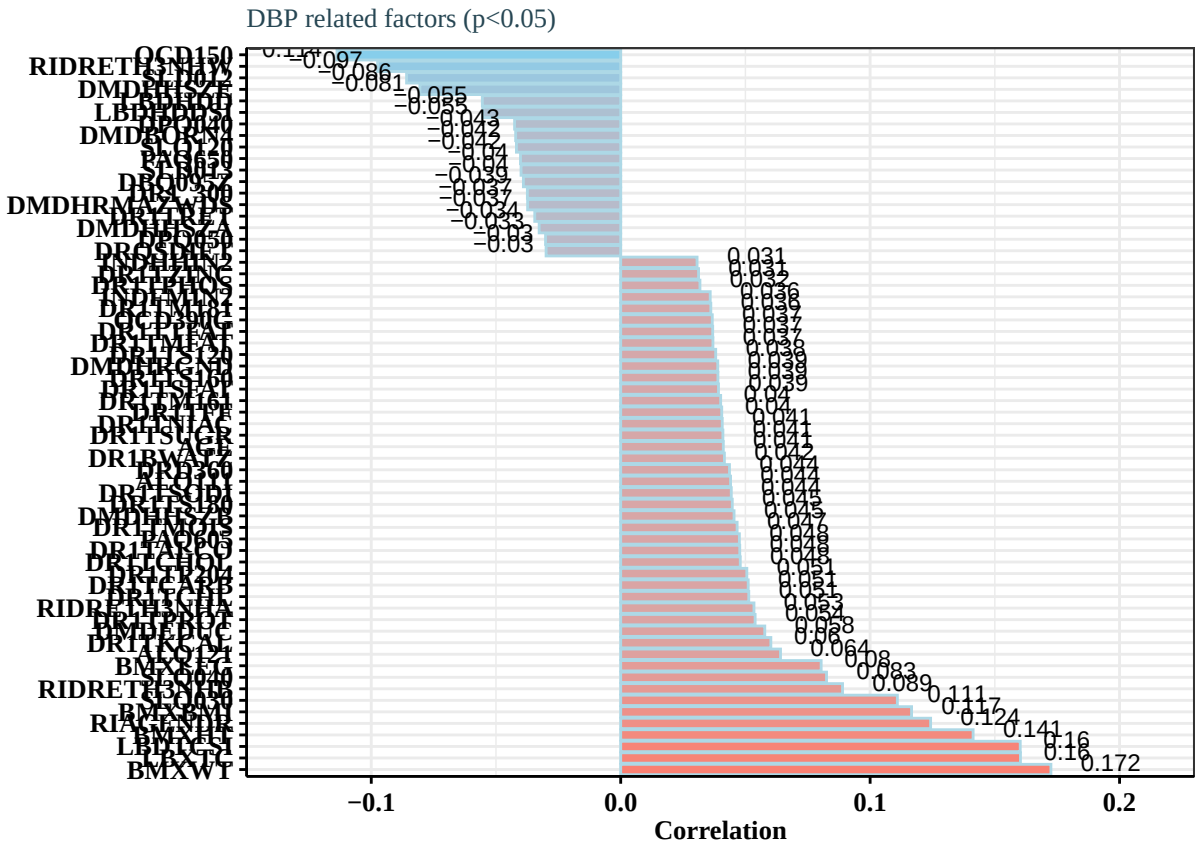
```
            vjust = 0.1, hjust = if_else(cor_S$cor >0,-0.5,1.2),
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
        axis.ticks = element_line(color = "black"),# 设置刻度线颜色
        axis.line = element_line(size = 0.5,
                                 colour = "black"),# 设置边框线颜色
        axis.title = element_text(colour = "black",
                                  size = 10,
                                  face = "bold"),# 设置标题字体
        axis.text = element_text(colour = "black",
                                 size = 10,
                                 face = "bold"),# 设置 x,y 轴标签字体
        axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
        text = element_text(size = 8,
                            color = "#264653",
                            family = "serif"))+# 设置文本字体
  guides(fill=FALSE)
pA
```



5.2 舒张压关联因素

```
cor_D <- cor.f(2, data = Data_SD)
cor_D <- cor_D %>%
  dplyr::filter(p_value < 0.05) %>%
  arrange(-cor)
```

```
pB <- ggplot(data = cor_D, aes(y = cor, x = reorder(rownames(cor_D), -cor))) +
  geom_col(aes(fill = cor), col = "lightblue") +
  scale_fill_gradient(low = "skyblue", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   size = 12,
                                   hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "Correlation") +
  ggtitle("DBP related factors (p<0.05)") +
  scale_y_continuous(expand = c(0, 0),
                     limits = c(-0.15, 0.23)) +
  geom_text(aes(label = round(cor, 3)),
            vjust = 0.1, hjust = if_else(cor_D$cor >0,-0.5,1.2),
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
          axis.ticks = element_line(color = "black"),# 设置刻度线颜色
          axis.line = element_line(size = 0.5,
                                   colour = "black"),# 设置边框线颜色
          axis.title = element_text(colour = "black",
                                    size = 10,
                                    face = "bold"),# 设置标题字体
          axis.text = element_text(colour = "black",
                                   size = 10,
                                   face = "bold"),# 设置 x,y 轴标签字体
          axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
          text = element_text(size = 8,
                              color = "#264653",
                              family = "serif"))+# 设置文本字体
  guides(fill=FALSE)
pB
```

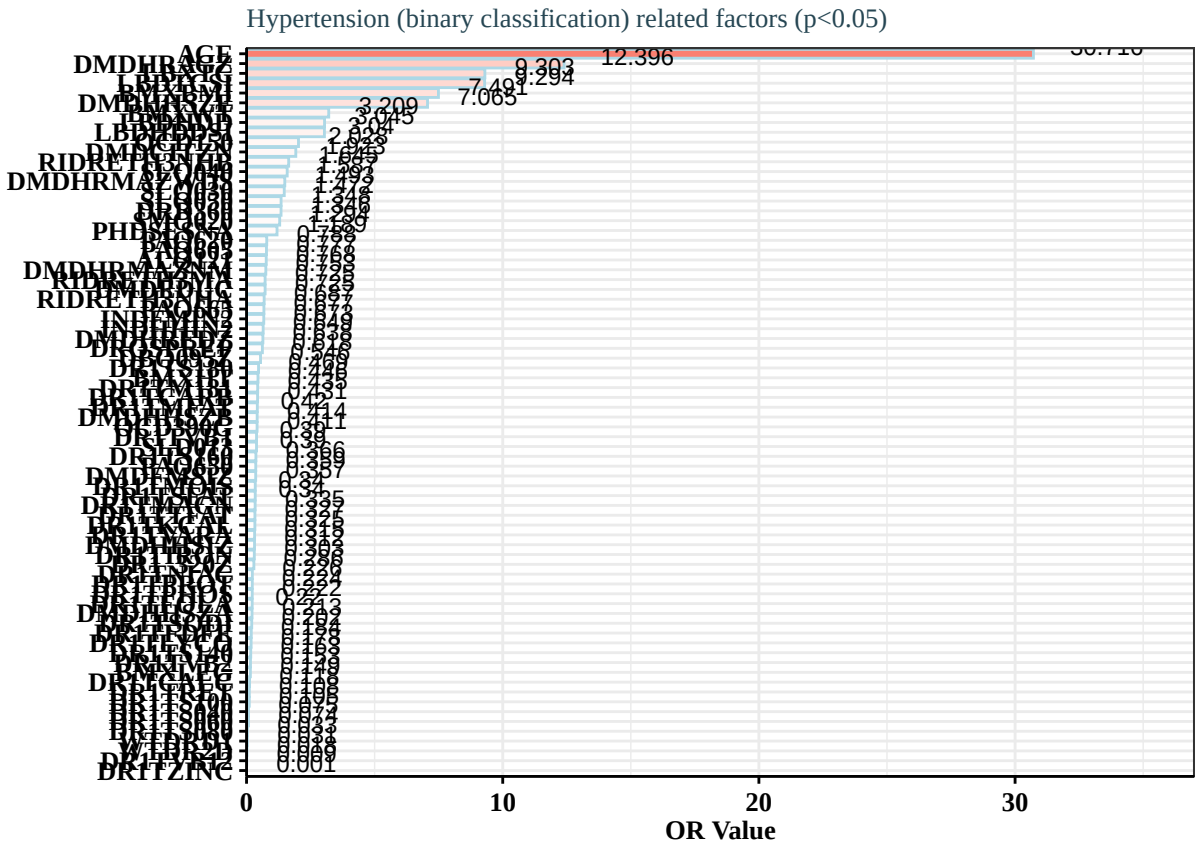


5.3 高血压关联因素（logistic）

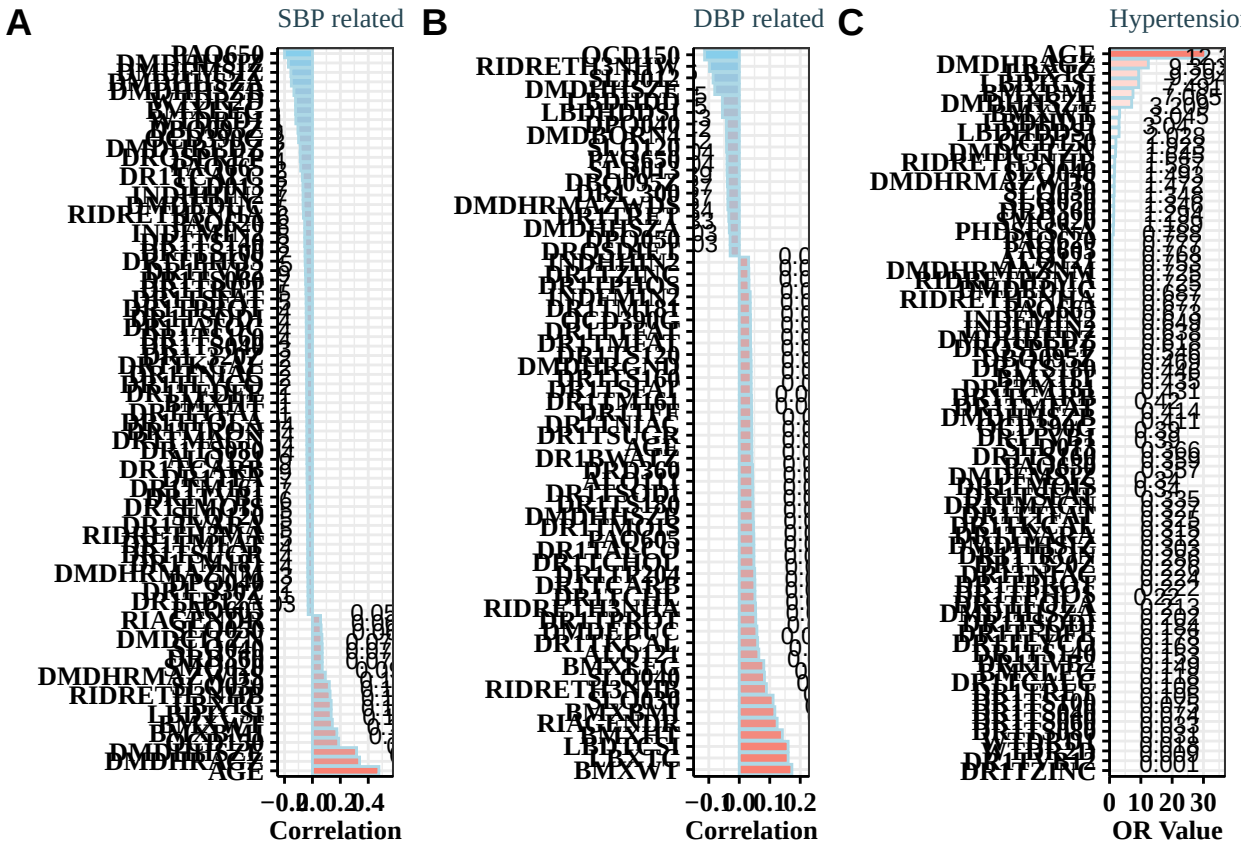
```
beta <- NULL
P_value <- NULL
for (i in 4:139) {
  temp.fit <- glm(response ~ Data_SD[, i],
    data = Data_SD,
    family = binomial(link = "logit")
  )
  temp <- summary(temp.fit)
  beta <- beta %>% append(temp$coefficients[2, 1])
  P_value <- P_value %>% append(temp$coefficients[2, 4])
}
opt <- data.frame(
  beta = beta,
  P_value = P_value
)
rownames(opt) <- colnames(Data_SD[4:139])
opt$OR_value <- exp(opt$beta)
```

```
opt$Q_value <- sprintf("%0.3f",
                       p.adjust(opt$P_value,
                                method = "BH", nrow(opt)))
opt <- opt %>%
  dplyr::filter(P_value < 0.05) %>%
  arrange(-OR_value, P_value)
```

```
pC <- ggplot(data = opt, aes(y = OR_value, x = reorder(rownames(opt), OR_value))) +
  geom_col(aes(fill = OR_value), col = "lightblue") +
  scale_fill_gradient(low = "white", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   size = 12,
                                   hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "OR Value") +
  ggtitle("Hypertension (binary classification) related factors (p<0.05)") +
  scale_y_continuous(expand = c(0, 0),
                     limits = c(0, 37)) +
  geom_text(aes(label = round(OR_value, 3)),
            vjust = 0.1, hjust = -0.5,
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
          axis.ticks = element_line(color = "black"),# 设置刻度线颜色
          axis.line = element_line(size = 0.5,
                                   colour = "black"),# 设置边框线颜色
          axis.title = element_text(colour = "black",
                                    size = 10,
                                    face = "bold"),# 设置标题字体
          axis.text = element_text(colour = "black",
                                   size = 10,
                                   face = "bold"),# 设置 x,y 轴标签字体
          axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
          text = element_text(size = 8,
                              color = "#264653",
                              family = "serif"))+# 设置文本字体
  guides(fill=FALSE)
pC
```




```
ggarrange(pA, pB,pC,
          labels = c("A","B","C"), # 添加标签
          ncol = 3, nrow = 1,
          font.label = list(size = 14, face = "bold")) # 设置标签字体样式
```


A
SBP related
Correlation
B
DBP related
Correlation
C
Hypertensio
OR Value

6 选择变量

6.1 特征筛选

6.1.1 合并上述相关性计算的有效变量

```
namesc <- unique(c(rownames(cor_D),
                   rownames(cor_S),
                   row.names(opt)))
namesc
```

```
##  [1] "BMXWT"       "LBXTC"       "LBDTCSI"     "BMXHT"       "RIAGENDR"
##  [6] "BMXBMI"      "SLQ030"      "RIDRETH3NHB" "SLQ040"      "BMXLEG"
## [11] "ALQ121"      "DR1TKCAL"    "DMDEDUC"     "DR1TPROT"    "RIDRETH3NHA"
## [16] "DR1TCHL"     "DR1TCARB"    "DR1TP204"    "DR1TCHOL"    "DR1TALCO"
## [21] "PAQ605"      "DR1TMOIS"    "DMDHHSZB"    "DR1TS180"    "DR1TSODI"
## [26] "ALQ111"      "DRD360"      "DR1BWATZ"    "AGE"         "DR1TSUGR"
## [31] "DR1TNIAC"    "DR1TFF"      "DR1TM161"    "DR1TSFAT"    "DR1TS160"
## [36] "DMDHRGND"    "DR1TS120"    "DR1TMFAT"    "DR1TTFAT"    "OCD390G"
## [41] "DR1TM181"    "INDFMIN2"    "DR1TPHOS"    "DR1TZINC"    "INDHHIN2"
## [46] "DRQSDIET"    "DPQ050"      "DMDHHSZA"    "DR1TRET"     "DMDHRMAZWDS"
## [51] "DR1_300"     "DBQ095Z"     "SLD013"      "PAQ650"      "SLQ120"
## [56] "DMDBORN4"    "DPQ040"      "LBDHDDSI"    "LBDHDD"      "DMDHHSZE"
## [61] "SLD012"      "RIDRETH3NHW" "OCD150"      "DMDHRAGZ"    "SMQ020"
## [66] "DMDCITZN"    "SLQ050"      "DR1TB12A"    "DR1_330Z"    "DMDHRMAZNM"
## [71] "RIDRETH3MA"  "DR1TVARA"    "DR1TVB1"     "DR1TFA"      "DR1TS080"
## [76] "DR1TMAGN"    "DR1TIRON"    "DR1TFOLA"    "DR1TFDFE"    "DR1TLYCO"
## [81] "DR1_320Z"    "DR1TS040"    "DR1TATOC"    "DR1TS060"    "DR1TVB2"
## [86] "DR1TS100"    "DR1TS140"    "PAQ620"      "DR1TCALC"    "PAQ665"
## [91] "DRQSPREP"    "DMDHREDZ"    "WTDRD1"      "WTDR2D"      "DMDFMSIZ"
## [96] "DMDHHSIZ"    "PHDSESNA"    "DR1TVB12"
```

```
Data_SD2 <- Data_SD[c("response", namesc)]
X <- Data_SD2[-1]
Y <- Data_SD2[, 1]
```

6.1.2 计算皮尔逊相关

```
library(FeatureSelection)
Featurepearson <- func_correlation(
  data = Data_SD2,
```

```
  target = "response",
  use_obs = "all.obs",
  correlation_thresh = 0.001,
  correlation_method = "pearson"
)
Featurepearson
```

```
##                 response
## AGE          0.351673348
## DMDHRAGZ     0.248565986
## DMDHHSZE     0.234009708
## OCD150       0.141394147
## LBDTCSI      0.107547674
## LBXTC        0.107496019
## BMXBMI       0.093533063
## RIDRETH3NHB 0.091567008
## DMDCITZN     0.075180172
## DMDHRMAZWDS 0.073919379
## BMXWT        0.062299134
## SLQ030       0.061856924
## SLQ050       0.057644617
## SLQ040       0.057529530
## DRD360       0.055218704
## SMQ020       0.052846255
## LBDHDD       0.043546480
## LBDHDDSI     0.043525317
## PHDSESNA     0.034982058
## DMDBORN4     0.027614263
## DR1TALCO     0.015007906
## DMDHRGND     0.007152334
## RIAGENDR     0.002989241
```

6.1.3 基于集成学习选择变量

```
params_glmnet <- list(
  alpha = 1,
  family = "gaussian",
  nfolds = 3,
  parallel = TRUE
)
params_xgboost <- list(
```

```
  params = list(
    "objective" = "reg:linear",
    "bst:eta" = 0.001,
    "subsample" = 0.75,
    "max_depth" = 5,
    "colsample_bytree" = 0.75,
    "nthread" = 6
  ),
  nrounds = 1000,
  print.every.n = 250,
  maximize = FALSE
)



params_ranger <- list(
  dependent.variable.name = "y",
  probability = FALSE, num.trees = 1000,
  verbose = TRUE, mtry = 5,
  min.node.size = 10, num.threads = 6,
  classification = FALSE, importance = "permutation"
)



params_features <- list(keep_number_feat = NULL, union = TRUE)
```

```
params_barplot <- list(keep_features = 96,
                       horiz = TRUE,
                       cex.names = 1.0)
barplot_feat_select(feat, params_barplot, xgb_sort = "Cover")
```

```
feat_lasso <- cbind.data.frame(Feature = feat$all_feat$`glmnet-lasso`$Feature,
                               coef = feat$all_feat$`glmnet-lasso`$coefficients)
feat_xgboost <- cbind.data.frame(Feature = feat$all_feat$xgboost$Feature,
                               coef = feat$all_feat$xgboost$Cover)
feat_ranger <- cbind.data.frame(Feature = feat$all_feat$ranger$Feature,
                               coef = feat$all_feat$ranger$permutation)
feat_union <- cbind.data.frame(Feature = feat$union_feat$feature,
                               coef = feat$union_feat$importance)
```

```
plasso <- ggplot(data = feat_lasso, aes(y = coef,
                                        x = reorder(Feature, coef))) +
  geom_col(aes(fill = coef), col = "lightblue") +
```

```
  scale_fill_gradient(low = "skyblue", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   size = 12,
                                   hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "Varible Importance") +
  ggtitle("Glmnet Lasso") +
  scale_y_continuous(expand = c(0, 0),
                     limits = c(-0.19, 0.55)) +
  geom_text(aes(label = round(coef, 3)),
            vjust = 0.1, hjust = if_else(feat_lasso$coef >0,-0.5,1.2),
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
          axis.ticks = element_line(color = "black"),# 设置刻度线颜色
          axis.line = element_line(size = 0.5,
                                   colour = "black"),# 设置边框线颜色
          axis.title = element_text(colour = "black",
                                    size = 10,
                                    face = "bold"),# 设置标题字体
          axis.text = element_text(colour = "black",
                                   size = 10,
                                   face = "bold"),# 设置 x,y 轴标签字体
          axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
          text = element_text(size = 8,
                              color = "#264653",
                              family = "serif"))+# 设置文本字体
  guides(fill=FALSE)+
  theme(axis.ticks.length.y = unit(-0.1, 'cm'),
        axis.ticks.length.x = unit(-0.1, 'cm'))
```

```
feat_xgboost <- feat_xgboost %>%
  dplyr::filter(coef > 0.005)
pxgb <- ggplot(data = feat_xgboost, aes(y = coef,
                                        x = reorder(Feature, coef))) +
  geom_col(aes(fill = coef), col = "lightblue") +
  scale_fill_gradient(low = "white", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   size = 12,
```

```
                                  hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "Varible Importance") +
  ggtitle("XGBoost") +
  scale_y_continuous(expand = c(0, 0),
                     limits = c(0, 0.032)) +
  geom_text(aes(label = round(coef, 3)),
            vjust = 0.1, hjust = if_else(feat_xgboost$coef >0,-0.5,1.2),
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
          axis.ticks = element_line(color = "black"),# 设置刻度线颜色
          axis.line = element_line(size = 0.5,
                                   colour = "black"),# 设置边框线颜色
          axis.title = element_text(colour = "black",
                                    size = 10,
                                    face = "bold"),# 设置标题字体
          axis.text = element_text(colour = "black",
                                   size = 10,
                                   face = "bold"),# 设置 x,y 轴标签字体
          axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
          text = element_text(size = 8,
                              color = "#264653",
                              family = "serif"))+# 设置文本字体
  guides(fill=FALSE)+
  theme(axis.ticks.length.y = unit(-0.1, 'cm'),
        axis.ticks.length.x = unit(-0.1, 'cm'))
```

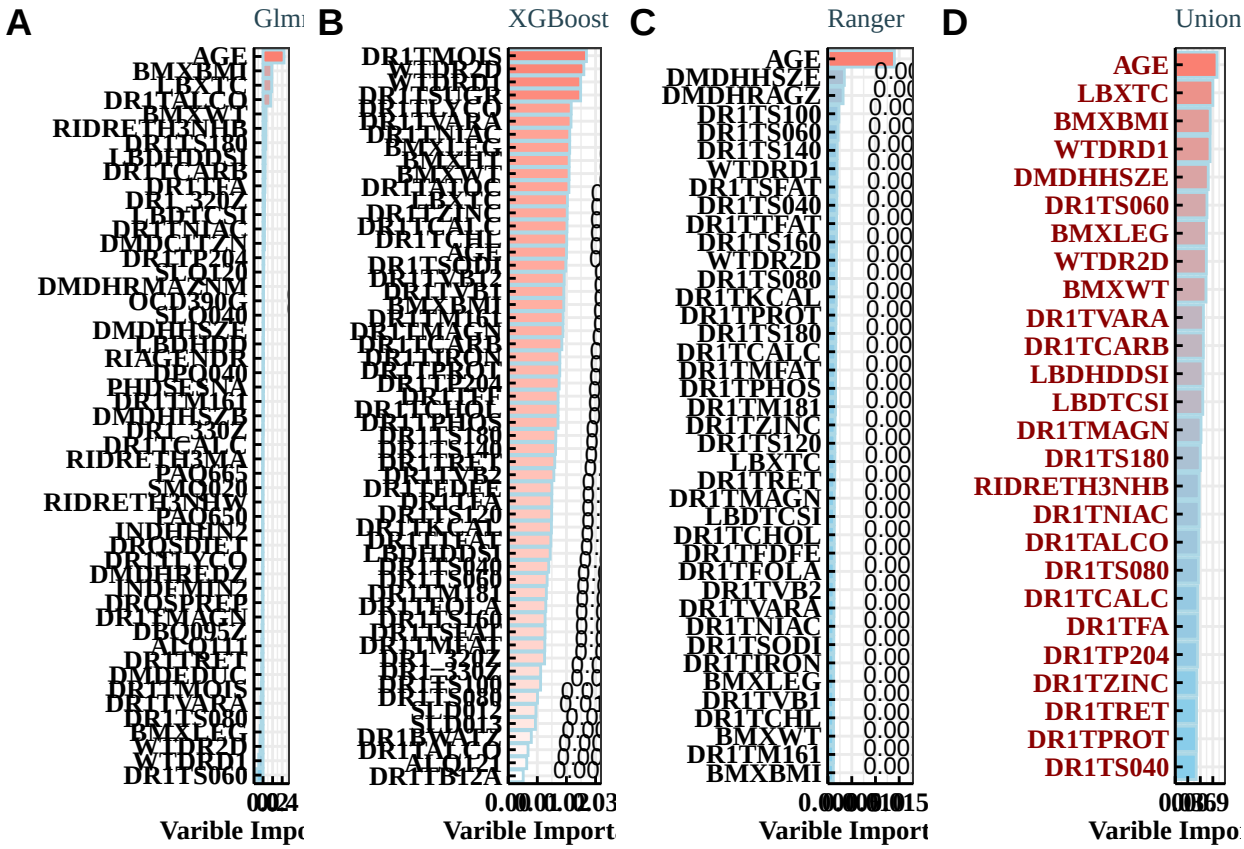
```
feat_ranger <- feat_ranger %>% dplyr::filter(coef > 0.001)
pranger <- ggplot(data = feat_ranger, aes(y = coef,
                                          x = reorder(Feature, coef))) +
  geom_col(aes(fill = coef), col = "lightblue") +
  scale_fill_gradient(low = "skyblue", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   size = 12,
                                   hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "Varible Importance") +
  ggtitle("Ranger") +
  scale_y_continuous(expand = c(0, 0),
```

```
                    limits = c(-0.0001, 0.018)) +
  geom_text(aes(label = round(coef, 3)),
            vjust = 0.1, hjust = if_else(feat_ranger$coef >0,-0.5,1.2),
            size = 3) +
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
          axis.ticks = element_line(color = "black"),# 设置刻度线颜色
          axis.line = element_line(size = 0.5,
                                   colour = "black"),# 设置边框线颜色
          axis.title = element_text(colour = "black",
                                    size = 10,
                                    face = "bold"),# 设置标题字体
          axis.text = element_text(colour = "black",
                                   size = 10,
                                   face = "bold"),# 设置 x,y 轴标签字体
          axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
          text = element_text(size = 8,
                              color = "#264653",
                              family = "serif"))+# 设置文本字体
  guides(fill=FALSE)+
  theme(axis.ticks.length.y = unit(-0.1, 'cm'),
        axis.ticks.length.x = unit(-0.1, 'cm'))
```

```
feat_union <- feat_union %>%
  dplyr::filter(coef > 0.5)
punion <- ggplot(data = feat_union, aes(y = coef,
                                        x = reorder(Feature, coef))) +
  geom_col(aes(fill = coef), col = "lightblue") +
  scale_fill_gradient(low = "skyblue", high = "#FA8072") +
  theme(axis.text.x = element_text(angle = 45,
                                   vjust = 1,
                                   size = 12,
                                   hjust = 1)) +
  coord_flip() +
  labs(x = "", y = "Varible Importance") +
  ggtitle("Union") +
  scale_y_continuous(expand = c(0, 0),
                     limits = c(0, 1.2)) +
  geom_text(aes(label = round(coef, 3)),
            vjust = 0.1, hjust = if_else(feat_union$coef >0,-0.5,1.2),
            size = 3) +
```

```
  theme_bw()+
  theme(panel.background = element_rect(fill = "transparent"),# 设置背景透明
          axis.ticks = element_line(color = "black"),# 设置刻度线颜色
          axis.line = element_line(size = 0.5,
                                   colour = "black"),# 设置边框线颜色
          axis.title = element_text(colour = "black",
                                    size = 10,
                                    face = "bold"),# 设置标题字体
          axis.text = element_text(colour = "black",
                                   size = 10,
                                   face = "bold"),# 设置 x,y 轴标签字体
          axis.text.x = element_text(angle = 0,hjust = 0.5,vjust = 0.5),
          axis.text.y = element_text(colour = 'darkred'),
          text = element_text(size = 8,
                              color = "#264653",
                              family = "serif"))+# 设置文本字体
  guides(fill=FALSE)+
  theme(axis.ticks.length.y = unit(-0.1, 'cm'),
        axis.ticks.length.x = unit(-0.1, 'cm'))
```

```
ggarrange(plasso, pxgb,pranger,punion,
          labels = c("A","B","C",'D'), # 添加标签
          ncol = 4, nrow = 1,
          font.label = list(size = 14, face = "bold")) # 设置标签字体样式
```




```
names_ts <- data.frame(
  imp = feat$union_feat$importance,
  feat = feat$union_feat$feature
)
names_fin <- names_ts %>%
  dplyr::filter(imp > 0.5)
names_final <- names_fin[, 2]
```

6.1.4 手动去除强共线性

```
match(names_final, colnames(Data_SD2))
```

```
##  [1] 30  3  7 94 61 85 11 95  2 73 18 59  4 77 25  9 32 21 76 90 75 19 45 50 15
## [26] 83
```

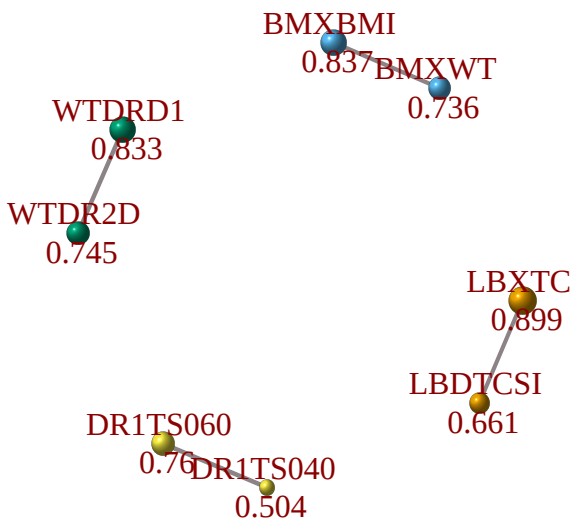
```
Data3 <- data.frame(
  response = Y,
  Data_SD2[, names_final]
)
```

```r
# Calculate the correlation matrix
cor_matrix <- cor(Data3)
```

```r
# Find the pairs of variables with correlation greater than 0.75
high_cor_pairs0 <- which(cor_matrix > 0.75 & cor_matrix != 1,
                         arr.ind = TRUE)
# Create adjacency matrix of highly correlated variables
adj_matrix0 <- cor_matrix[unique(high_cor_pairs0[, 1]),
                          unique(high_cor_pairs0[, 1])]
for (i in 1:length(adj_matrix0)) {
  if (adj_matrix0[i] < 0.75 | adj_matrix0[i] == 1) {
    adj_matrix0[i] <- 0
  }
}
```

```r
set.seed(1234)
g <- graph.adjacency(adj_matrix0, mode = "undirected", weighted = TRUE)
clusters <- cluster_louvain(g)
p <- plot(g, layout = layout_with_fr(g, area = nrow(adj_matrix0)^2),
     edge.color = '#8B8386', edge.width = E(g)/15+2,
     vertex.size = feat$union_feat$importance[match(colnames(adj_matrix0),
                                    feat$union_feat$feature)]*14, vertex.label = paste(
    colnames(adj_matrix0), "\n",
    feat$union_feat$importance[match(colnames(adj_matrix0),
                                    feat$union_feat$feature)] %>% round(3)
  ), vertex.label.cex = 1, vertex.label.dist = 0, vertex.label.color = "darkred",
  edge.label.family = 'Times',
  vertex.color = clusters$membership,
  vertex.shape = 'sphere',
  frame = T)
```

6.1.4.1 共线性图




```
names_ex <- c(
  "WTDR2D", "BMXWT", "DR1TS040",
  "LBDTCSI"
)

Data5 <- Data3[-match(names_ex, colnames(Data3))]
```

6.1.4.2 相关性大于 0.8

6.1.4.3 相关性大于 0.75

由于去除相关性 >0.8 的变量后，模型仍存在共线性。

6.2 过采样平衡结局变量

```
set.seed(1234)
Data5 <- Data5 %>%
  mutate(response = factor(response, levels = c(0, 1)))
Data6 <- DMwR::SMOTE(response ~ ., Data5, perc.over = 200)
```

```
set.seed(1234)
# 使用 downSample 函数进行负采样
Data_ds <- caret::downSample(x = Data5[, -1], y = Data5$response, yname = "response", list = FALSE)
Data_ds <- Data_ds %>%
  dplyr::select(response,everything())
```

```
table(Data5$response)
```

```
##
##    0    1
## 3390  945
```

```
table(Data6$response)
```

```
##
##    0    1
## 3780 2835
```

```
table(Data_ds$response)
```

```
##
##   0   1
## 945 945
```

6.3 交互作用

```
glm.fit <- glm(response ~ .,
  data = Data6,
  family = binomial(link = "logit")
)
```

```
summary(glm.fit)
```

```
##
## Call:
## glm(formula = response ~ ., family = binomial(link = "logit"),
##     data = Data6)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2658  -0.9330  -0.3489   0.9685   2.5874
##
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.14379    0.23799 -17.412  < 2e-16 ***
## AGE          4.47199    0.17385  25.723  < 2e-16 ***
## LBXTC        2.33381    0.27074   8.620  < 2e-16 ***
## BMXBMI       3.01751    0.30841   9.784  < 2e-16 ***
## WTDRD1      -1.86123    0.37701  -4.937 7.94e-07 ***
## DMDHHSZE    -0.71982    0.14097  -5.106 3.29e-07 ***
## DR1TS060    -1.98462    0.97754  -2.030  0.04233 *
## BMXLEG      -0.40410    0.23095  -1.750  0.08016 .
## DR1TVARA    -0.66717    0.55863  -1.194  0.23236
## DR1TCARB     2.73228    0.49624   5.506 3.67e-08 ***
## LBDHDDSI     0.63531    0.35645   1.782  0.07470 .
## DR1TMAGN    -2.32521    0.56254  -4.133 3.57e-05 ***
## DR1TS180    -0.07908    0.55627  -0.142  0.88695
## RIDRETH3NHB  0.39313    0.07452   5.275 1.33e-07 ***
## DR1TNIAC     0.45553    0.78354   0.581  0.56098
## DR1TALCO     1.55565    0.48271   3.223  0.00127 **
## DR1TS080    -1.38773    1.16427  -1.192  0.23329
## DR1TCALC     0.59778    0.62574   0.955  0.33941
## DR1TFA       0.82411    0.86876   0.949  0.34282
## DR1TP204     0.08978    0.54252   0.165  0.86857
## DR1TZINC     8.78519    3.98366   2.205  0.02743 *
## DR1TRET     -2.87563    1.09776  -2.620  0.00880 **
## DR1TPROT    -0.15812    1.01626  -0.156  0.87635
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9034.9  on 6614  degrees of freedom
## Residual deviance: 7285.3  on 6592  degrees of freedom
## AIC: 7331.3
##
## Number of Fisher Scoring iterations: 4
```

```r
# create a new model with interaction terms
glm.fit_int <- glm(response ~ . + (AGE + LBXTC + BMXBMI + WTDRD1 + DMDHHSZE + DR1TS060 + DR1TCARB + DR1TMA
  data = Data6,
  family = binomial(link = "logit")
)
```

```
summary(glm.fit_int)
```

```
##
## Call:
## glm(formula = response ~ . + (AGE + LBXTC + BMXBMI + WTDRD1 +
##     DMDHHSZE + DR1TS060 + DR1TCARB + DR1TMAGN + RIDRETH3NHB +
##     DR1TALCO + DR1TZINC + DR1TRET) * (AGE + LBXTC + BMXBMI +
##     WTDRD1 + DMDHHSZE + DR1TS060 + DR1TCARB + DR1TMAGN + RIDRETH3NHB +
##     DR1TALCO + DR1TZINC + DR1TRET), family = binomial(link = "logit"),
##     data = Data6)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3154  -0.9173  -0.2809   0.9394   2.7199
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -6.31053    0.69416  -9.091  < 2e-16 ***
## AGE                    5.95371    0.83875   7.098 1.26e-12 ***
## LBXTC                  5.89105    1.47086   4.005 6.20e-05 ***
## BMXBMI                 8.51304    1.40859   6.044 1.51e-09 ***
## WTDRD1                -6.06514    2.95951  -2.049 0.040425 *
## DMDHHSZE               2.07296    0.86392   2.399 0.016419 *
## DR1TS060              10.34977    4.44933   2.326 0.020011 *
## BMXLEG                -0.64338    0.24108  -2.669 0.007614 **
## DR1TVARA              -0.50965    0.57866  -0.881 0.378462
## DR1TCARB               4.93225    2.52500   1.953 0.050776 .
## LBDHDDSI               0.52643    0.37399   1.408 0.159245
## DR1TMAGN              -8.48246    3.05100  -2.780 0.005432 **
## DR1TS180               0.08402    0.57476   0.146 0.883780
## RIDRETH3NHB            1.11409    0.43725   2.548 0.010836 *
## DR1TNIAC               0.52508    0.84511   0.621 0.534394
## DR1TALCO               8.32379    2.93889   2.832 0.004622 **
## DR1TS080              -0.80703    1.12873  -0.715 0.474616
## DR1TCALC               1.01452    0.67282   1.508 0.131592
## DR1TFA                 0.72175    0.87634   0.824 0.410168
## DR1TP204               0.26825    0.56615   0.474 0.635639
## DR1TZINC              72.81911   18.81401   3.870 0.000109 ***
## DR1TRET              -18.39564    6.00411  -3.064 0.002185 **
## DR1TPROT              -0.81250    1.10964  -0.732 0.464033
## AGE:LBXTC             -1.18611    1.63988  -0.723 0.469498
```

```
## AGE:BMXBMI             -6.13689    1.59041  -3.859 0.000114 ***
## AGE:WTDRD1              5.44208    3.00174   1.813 0.069836 .
## AGE:DMDHHSZE           -1.45124    0.64910  -2.236 0.025367 *
## AGE:DR1TS060          -12.43652    4.17001  -2.982 0.002860 **
## AGE:DR1TCARB            7.41218    2.64649   2.801 0.005098 **
## AGE:DR1TMAGN           -0.19764    3.00513  -0.066 0.947564
## AGE:RIDRETH3NHB         0.24122    0.43412   0.556 0.578442
## AGE:DR1TALCO            1.99807    2.88313   0.693 0.488298
## AGE:DR1TZINC          -55.55697   17.89386  -3.105 0.001904 **
## AGE:DR1TRET            14.55230    5.19048   2.804 0.005053 **
## LBXTC:BMXBMI           -6.47960    2.99703  -2.162 0.030618 *
## LBXTC:WTDRD1            1.49962    3.79929   0.395 0.693055
## LBXTC:DMDHHSZE         -0.76929    1.33046  -0.578 0.563118
## LBXTC:DR1TS060         -1.54163    6.12037  -0.252 0.801130
## LBXTC:DR1TCARB        -11.61428    4.22576  -2.748 0.005988 **
## LBXTC:DR1TMAGN         13.03111    4.66720   2.792 0.005237 **
## LBXTC:RIDRETH3NHB      -0.63211    0.64569  -0.979 0.327600
## LBXTC:DR1TALCO        -15.82850    5.05352  -3.132 0.001735 **
## LBXTC:DR1TZINC        -59.68079   28.49253  -2.095 0.036206 *
## LBXTC:DR1TRET           6.29349    7.54675   0.834 0.404318
## BMXBMI:WTDRD1           4.65168    4.83644   0.962 0.336151
## BMXBMI:DMDHHSZE         0.53171    1.46902   0.362 0.717387
## BMXBMI:DR1TS060       -25.14999    6.98432  -3.601 0.000317 ***
## BMXBMI:DR1TCARB        -4.82758    4.21759  -1.145 0.252362
## BMXBMI:DR1TMAGN         5.70114    5.00192   1.140 0.254374
## BMXBMI:RIDRETH3NHB     -0.74177    0.69070  -1.074 0.282852
## BMXBMI:DR1TALCO         3.06027    5.85972   0.522 0.601493
## BMXBMI:DR1TZINC         6.76501   36.21652   0.187 0.851823
## BMXBMI:DR1TRET         16.45492    9.45964   1.739 0.081949 .
## WTDRD1:DMDHHSZE        -5.90973    1.96956  -3.001 0.002695 **
## WTDRD1:DR1TS060       -11.87340    8.51832  -1.394 0.163358
## WTDRD1:DR1TCARB         0.38760    6.30842   0.061 0.951007
## WTDRD1:DR1TMAGN        -7.29552    7.02854  -1.038 0.299277
## WTDRD1:RIDRETH3NHB     -3.41061    2.82874  -1.206 0.227933
## WTDRD1:DR1TALCO         4.05523    7.32180   0.554 0.579677
## WTDRD1:DR1TZINC        34.88979   48.48734   0.720 0.471793
## WTDRD1:DR1TRET         37.55001   11.57529   3.244 0.001179 **
## DMDHHSZE:DR1TS060       9.35863    3.55740   2.631 0.008520 **
## DMDHHSZE:DR1TCARB      -7.52961    2.26315  -3.327 0.000878 ***
## DMDHHSZE:DR1TMAGN       2.01531    2.53194   0.796 0.426057
## DMDHHSZE:RIDRETH3NHB   -1.36848    0.35063  -3.903 9.50e-05 ***
```

```
## DMDHHSZE:DR1TALCO      -8.70778    2.95001  -2.952 0.003159 **
## DMDHHSZE:DR1TZINC      -1.26894   17.19300  -0.074 0.941165
## DMDHHSZE:DR1TRET       -7.79706    4.43802  -1.757 0.078939 .
## DR1TS060:DR1TCARB     -18.21141    7.66211  -2.377 0.017463 *
## DR1TS060:DR1TMAGN      41.66529   10.29304   4.048 5.17e-05 ***
## DR1TS060:RIDRETH3NHB    3.61876    1.78481   2.028 0.042608 *
## DR1TS060:DR1TALCO      -3.18063   12.83870  -0.248 0.804337
## DR1TS060:DR1TZINC    -292.23497   66.28630  -4.409 1.04e-05 ***
## DR1TS060:DR1TRET        8.31508   10.71905   0.776 0.437909
## DR1TCARB:DR1TMAGN      -4.96548    4.56990  -1.087 0.277230
## DR1TCARB:RIDRETH3NHB   -0.43110    1.13893  -0.379 0.705047
## DR1TCARB:DR1TALCO      -5.93978    5.45843  -1.088 0.276513
## DR1TCARB:DR1TZINC       9.05068   37.29637   0.243 0.808262
## DR1TCARB:DR1TRET       37.70975   11.25859   3.349 0.000810 ***
## DR1TMAGN:RIDRETH3NHB    2.04128    1.23571   1.652 0.098554 .
## DR1TMAGN:DR1TALCO       1.61067    4.99530   0.322 0.747122
## DR1TMAGN:DR1TZINC      29.94616   36.19441   0.827 0.408028
## DR1TMAGN:DR1TRET      -31.59975   11.87930  -2.660 0.007812 **
## RIDRETH3NHB:DR1TALCO   -0.65875    1.35185  -0.487 0.626050
## RIDRETH3NHB:DR1TZINC  -12.22900    8.77881  -1.393 0.163616
## RIDRETH3NHB:DR1TRET    -3.53897    2.25176  -1.572 0.116034
## DR1TALCO:DR1TZINC     -17.50477   48.26236  -0.363 0.716829
## DR1TALCO:DR1TRET       -8.47277   18.12476  -0.467 0.640164
## DR1TZINC:DR1TRET       -0.83677   66.12216  -0.013 0.989903
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9034.9  on 6614  degrees of freedom
## Residual deviance: 7071.4  on 6526  degrees of freedom
## AIC: 7249.4
##
## Number of Fisher Scoring iterations: 6
```

```
glm.fit_int <- glm(
  response ~ . +
    AGE:BMXBMI +
    AGE:DMDHHSZE +
    AGE:DR1TS060 +
    AGE:DR1TCARB +
    AGE:DR1TZINC +
```

```
    AGE:DR1TRET +
    LBXTC:BMXBMI +
    LBXTC:DR1TCARB +
    LBXTC:DR1TMAGN +
    LBXTC:DR1TZINC +
    LBXTC:DR1TCARB +
    LBXTC:DR1TMAGN +
    LBXTC:DR1TALCO +
    LBXTC:DR1TZINC +
    BMXBMI:DR1TS060 +
    WTDRD1:DMDHHSZE +
    WTDRD1:DR1TRET +
    DMDHHSZE:DR1TS060 +
    DMDHHSZE:DR1TCARB +
    DMDHHSZE:RIDRETH3NHB +
    DMDHHSZE:DR1TALCO +
    DR1TS060:DR1TCARB +
    DR1TS060:DR1TMAGN +
    DR1TS060:DR1TZINC +
    DR1TCARB:DR1TRET +
    DR1TMAGN:DR1TRET,
  data = Data6,
  family = binomial(link = "logit")
)
```

```
summary(glm.fit_int)
```

```
##
## Call:
## glm(formula = response ~ . + AGE:BMXBMI + AGE:DMDHHSZE + AGE:DR1TS060 +
##     AGE:DR1TCARB + AGE:DR1TZINC + AGE:DR1TRET + LBXTC:BMXBMI +
##     LBXTC:DR1TCARB + LBXTC:DR1TMAGN + LBXTC:DR1TZINC + LBXTC:DR1TCARB +
##     LBXTC:DR1TMAGN + LBXTC:DR1TALCO + LBXTC:DR1TZINC + BMXBMI:DR1TS060 +
##     WTDRD1:DMDHHSZE + WTDRD1:DR1TRET + DMDHHSZE:DR1TS060 + DMDHHSZE:DR1TCARB +
##     DMDHHSZE:RIDRETH3NHB + DMDHHSZE:DR1TALCO + DR1TS060:DR1TCARB +
##     DR1TS060:DR1TMAGN + DR1TS060:DR1TZINC + DR1TCARB:DR1TRET +
##     DR1TMAGN:DR1TRET, family = binomial(link = "logit"), data = Data6)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2447  -0.9142  -0.2997   0.9429   2.6561
##
```

```
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -6.0865     0.4596 -13.244  < 2e-16 ***
## AGE                     5.9505     0.4997  11.909  < 2e-16 ***
## LBXTC                   4.3718     0.9267   4.718 2.39e-06 ***
## BMXBMI                  8.4423     1.1088   7.614 2.67e-14 ***
## WTDRD1                 -2.4398     0.7073  -3.450 0.000561 ***
## DMDHHSZE                1.7783     0.5649   3.148 0.001644 **
## DR1TS060                6.1350     3.2630   1.880 0.060087 .
## BMXLEG                 -0.5530     0.2355  -2.348 0.018872 *
## DR1TVARA               -0.4732     0.5650  -0.838 0.402279
## DR1TCARB                2.4092     1.7396   1.385 0.166077
## LBDHDDSI                0.5076     0.3668   1.384 0.166445
## DR1TMAGN               -7.0301     1.6193  -4.341 1.42e-05 ***
## DR1TS180                0.1186     0.5650   0.210 0.833769
## RIDRETH3NHB             0.7275     0.1051   6.920 4.53e-12 ***
## DR1TNIAC                0.4393     0.8026   0.547 0.584168
## DR1TALCO                6.8802     1.4359   4.791 1.66e-06 ***
## DR1TS080               -0.8518     1.1371  -0.749 0.453783
## DR1TCALC                0.7872     0.6433   1.224 0.221092
## DR1TFA                  0.5072     0.8493   0.597 0.550372
## DR1TP204                0.1199     0.5571   0.215 0.829577
## DR1TZINC               75.7072    11.8158   6.407 1.48e-10 ***
## DR1TRET               -10.0553     3.2464  -3.097 0.001953 **
## DR1TPROT               -0.6203     1.0606  -0.585 0.558645
## AGE:BMXBMI             -5.4591     1.1418  -4.781 1.74e-06 ***
## AGE:DMDHHSZE           -1.6553     0.6250  -2.649 0.008081 **
## AGE:DR1TS060           -8.2871     3.6696  -2.258 0.023925 *
## AGE:DR1TCARB            7.4334     2.1149   3.515 0.000440 ***
## AGE:DR1TZINC          -52.4799    11.2199  -4.677 2.91e-06 ***
## AGE:DR1TRET             7.1615     3.2949   2.174 0.029740 *
## LBXTC:BMXBMI           -5.0951     2.8168  -1.809 0.070472 .
## LBXTC:DR1TCARB        -11.2148     3.9550  -2.836 0.004574 **
## LBXTC:DR1TMAGN         13.4644     4.3595   3.089 0.002012 **
## LBXTC:DR1TZINC        -55.2297    25.0934  -2.201 0.027739 *
## LBXTC:DR1TALCO        -13.4182     4.0650  -3.301 0.000964 ***
## BMXBMI:DR1TS060       -13.0708     5.2853  -2.473 0.013396 *
## WTDRD1:DMDHHSZE        -3.7936     1.3516  -2.807 0.005005 **
## WTDRD1:DR1TRET         27.8799     7.3185   3.810 0.000139 ***
## DMDHHSZE:DR1TS060       5.5904     2.7580   2.027 0.042663 *
## DMDHHSZE:DR1TCARB      -6.1558     1.7337  -3.551 0.000384 ***
```

```
## DMDHHSZE:RIDRETH3NHB   -1.0883     0.2472  -4.402 1.07e-05 ***
## DMDHHSZE:DR1TALCO      -6.5090     2.1279  -3.059 0.002221 **
## DR1TS060:DR1TCARB     -15.6499     7.3189  -2.138 0.032495 *
## DR1TS060:DR1TMAGN      35.5614     9.1995   3.866 0.000111 ***
## DR1TS060:DR1TZINC    -249.2111    52.4820  -4.749 2.05e-06 ***
## DR1TCARB:DR1TRET       31.2168     9.5432   3.271 0.001071 **
## DR1TMAGN:DR1TRET      -26.1131     9.0367  -2.890 0.003856 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9034.9  on 6614  degrees of freedom
## Residual deviance: 7117.4  on 6569  degrees of freedom
## AIC: 7209.4
##
## Number of Fisher Scoring iterations: 5
```

```
# compare the two models using anova()
anova(glm.fit, glm.fit_int, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: response ~ AGE + LBXTC + BMXBMI + WTDRD1 + DMDHHSZE + DR1TS060 +
##     BMXLEG + DR1TVARA + DR1TCARB + LBDHDDSI + DR1TMAGN + DR1TS180 +
##     RIDRETH3NHB + DR1TNIAC + DR1TALCO + DR1TS080 + DR1TCALC +
##     DR1TFA + DR1TP204 + DR1TZINC + DR1TRET + DR1TPROT
## Model 2: response ~ AGE + LBXTC + BMXBMI + WTDRD1 + DMDHHSZE + DR1TS060 +
##     BMXLEG + DR1TVARA + DR1TCARB + LBDHDDSI + DR1TMAGN + DR1TS180 +
##     RIDRETH3NHB + DR1TNIAC + DR1TALCO + DR1TS080 + DR1TCALC +
##     DR1TFA + DR1TP204 + DR1TZINC + DR1TRET + DR1TPROT + AGE:BMXBMI +
##     AGE:DMDHHSZE + AGE:DR1TS060 + AGE:DR1TCARB + AGE:DR1TZINC +
##     AGE:DR1TRET + LBXTC:BMXBMI + LBXTC:DR1TCARB + LBXTC:DR1TMAGN +
##     LBXTC:DR1TZINC + LBXTC:DR1TCARB + LBXTC:DR1TMAGN + LBXTC:DR1TALCO +
##     LBXTC:DR1TZINC + BMXBMI:DR1TS060 + WTDRD1:DMDHHSZE + WTDRD1:DR1TRET +
##     DMDHHSZE:DR1TS060 + DMDHHSZE:DR1TCARB + DMDHHSZE:RIDRETH3NHB +
##     DMDHHSZE:DR1TALCO + DR1TS060:DR1TCARB + DR1TS060:DR1TMAGN +
##     DR1TS060:DR1TZINC + DR1TCARB:DR1TRET + DR1TMAGN:DR1TRET
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      6592     7285.3
## 2      6569     7117.4 23   167.86 < 2.2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(glm.fit_int)
```

```
##
## Call:
## glm(formula = response ~ . + AGE:BMXBMI + AGE:DMDHHSZE + AGE:DR1TS060 +
##     AGE:DR1TCARB + AGE:DR1TZINC + AGE:DR1TRET + LBXTC:BMXBMI +
##     LBXTC:DR1TCARB + LBXTC:DR1TMAGN + LBXTC:DR1TZINC + LBXTC:DR1TCARB +
##     LBXTC:DR1TMAGN + LBXTC:DR1TALCO + LBXTC:DR1TZINC + BMXBMI:DR1TS060 +
##     WTDRD1:DMDHHSZE + WTDRD1:DR1TRET + DMDHHSZE:DR1TS060 + DMDHHSZE:DR1TCARB +
##     DMDHHSZE:RIDRETH3NHB + DMDHHSZE:DR1TALCO + DR1TS060:DR1TCARB +
##     DR1TS060:DR1TMAGN + DR1TS060:DR1TZINC + DR1TCARB:DR1TRET +
##     DR1TMAGN:DR1TRET, family = binomial(link = "logit"), data = Data6)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2447  -0.9142  -0.2997   0.9429   2.6561
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -6.0865     0.4596 -13.244  < 2e-16 ***
## AGE                     5.9505     0.4997  11.909  < 2e-16 ***
## LBXTC                   4.3718     0.9267   4.718 2.39e-06 ***
## BMXBMI                  8.4423     1.1088   7.614 2.67e-14 ***
## WTDRD1                 -2.4398     0.7073  -3.450 0.000561 ***
## DMDHHSZE                1.7783     0.5649   3.148 0.001644 **
## DR1TS060                6.1350     3.2630   1.880 0.060087 .
## BMXLEG                 -0.5530     0.2355  -2.348 0.018872 *
## DR1TVARA               -0.4732     0.5650  -0.838 0.402279
## DR1TCARB                2.4092     1.7396   1.385 0.166077
## LBDHDDSI                0.5076     0.3668   1.384 0.166445
## DR1TMAGN               -7.0301     1.6193  -4.341 1.42e-05 ***
## DR1TS180                0.1186     0.5650   0.210 0.833769
## RIDRETH3NHB             0.7275     0.1051   6.920 4.53e-12 ***
## DR1TNIAC                0.4393     0.8026   0.547 0.584168
## DR1TALCO                6.8802     1.4359   4.791 1.66e-06 ***
## DR1TS080               -0.8518     1.1371  -0.749 0.453783
## DR1TCALC                0.7872     0.6433   1.224 0.221092
## DR1TFA                  0.5072     0.8493   0.597 0.550372
## DR1TP204                0.1199     0.5571   0.215 0.829577
## DR1TZINC               75.7072    11.8158   6.407 1.48e-10 ***
```

```
## DR1TRET               -10.0553     3.2464  -3.097 0.001953 **
## DR1TPROT               -0.6203     1.0606  -0.585 0.558645
## AGE:BMXBMI             -5.4591     1.1418  -4.781 1.74e-06 ***
## AGE:DMDHHSZE           -1.6553     0.6250  -2.649 0.008081 **
## AGE:DR1TS060           -8.2871     3.6696  -2.258 0.023925 *
## AGE:DR1TCARB            7.4334     2.1149   3.515 0.000440 ***
## AGE:DR1TZINC          -52.4799    11.2199  -4.677 2.91e-06 ***
## AGE:DR1TRET             7.1615     3.2949   2.174 0.029740 *
## LBXTC:BMXBMI           -5.0951     2.8168  -1.809 0.070472 .
## LBXTC:DR1TCARB        -11.2148     3.9550  -2.836 0.004574 **
## LBXTC:DR1TMAGN         13.4644     4.3595   3.089 0.002012 **
## LBXTC:DR1TZINC        -55.2297    25.0934  -2.201 0.027739 *
## LBXTC:DR1TALCO        -13.4182     4.0650  -3.301 0.000964 ***
## BMXBMI:DR1TS060       -13.0708     5.2853  -2.473 0.013396 *
## WTDRD1:DMDHHSZE        -3.7936     1.3516  -2.807 0.005005 **
## WTDRD1:DR1TRET         27.8799     7.3185   3.810 0.000139 ***
## DMDHHSZE:DR1TS060       5.5904     2.7580   2.027 0.042663 *
## DMDHHSZE:DR1TCARB      -6.1558     1.7337  -3.551 0.000384 ***
## DMDHHSZE:RIDRETH3NHB   -1.0883     0.2472  -4.402 1.07e-05 ***
## DMDHHSZE:DR1TALCO      -6.5090     2.1279  -3.059 0.002221 **
## DR1TS060:DR1TCARB     -15.6499     7.3189  -2.138 0.032495 *
## DR1TS060:DR1TMAGN      35.5614     9.1995   3.866 0.000111 ***
## DR1TS060:DR1TZINC    -249.2111    52.4820  -4.749 2.05e-06 ***
## DR1TCARB:DR1TRET       31.2168     9.5432   3.271 0.001071 **
## DR1TMAGN:DR1TRET      -26.1131     9.0367  -2.890 0.003856 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9034.9  on 6614  degrees of freedom
## Residual deviance: 7117.4  on 6569  degrees of freedom
## AIC: 7209.4
##
## Number of Fisher Scoring iterations: 5
```

6.3.1 交互作用网络图

```
in_var1 <- c('AGE','AGE','AGE','AGE','AGE','AGE',
             'LBXTC','LBXTC','LBXTC','LBXTC','LBXTC',
             'BMXBMI',
```

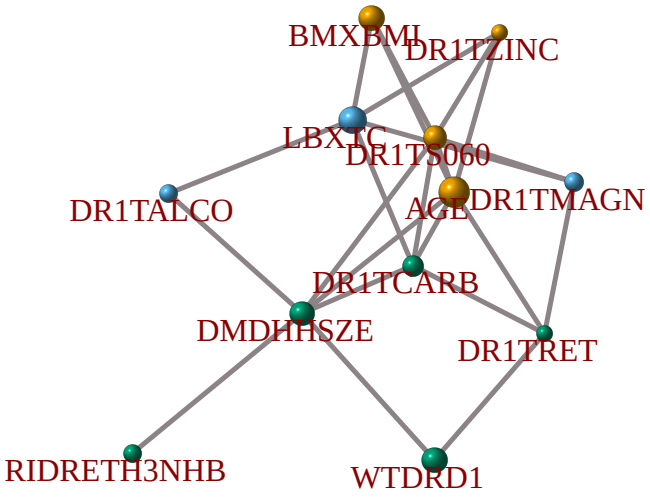
```
              'WTDRD1','WTDRD1',
              'DMDHHSZE','DMDHHSZE','DMDHHSZE','DMDHHSZE',
              'DR1TS060','DR1TS060','DR1TS060',
              'DR1TCARB',
              'DR1TMAGN')

in_var2 <- c('BMXBMI','DMDHHSZE','DR1TS060','DR1TCARB',
              'DR1TZINC', 'DR1TRET','BMXBMI','DR1TCARB','DR1TMAGN',
              'DR1TZINC','DR1TALCO','DR1TS060', 'DMDHHSZE','DR1TRET',
              'DR1TS060','DR1TCARB','RIDRETH3NHB' ,'DR1TALCO','DR1TCARB',
              'DR1TMAGN','DR1TZINC','DR1TRET','DR1TRET')

in_var <- unique(c(in_var1,in_var2))
```

```
in_ma <- Matrix(rep(0,12*12),
       ncol = 12)
colnames(in_ma) <- rownames(in_ma) <- in_var
for (i in 1:length(in_var1)){
  in_ma[match(in_var1[i],in_var),match(in_var2[i],in_var)] <- 1
  in_ma[match(in_var2[i],in_var),match(in_var1[i],in_var)] <- 1
}
```

```
set.seed(1234)
g2 <- graph.adjacency(in_ma, mode = "undirected", weighted = TRUE)
clusters <- cluster_louvain(g2)
plot(g2, layout = layout_with_fr(g2),
     edge.color = '#8B8386', edge.width = 2.5,
     vertex.size = feat$union_feat$importance[match(colnames(in_ma),
                                    feat$union_feat$feature)]*14, vertex.label = paste(
    colnames(in_ma)), vertex.label.cex = 1, vertex.label.dist = -1.5, vertex.label.color = "darkred",
  edge.label.family = 'Times',
  vertex.color = clusters$membership,
  vertex.shape = 'sphere',
  frame = T)
```



6.4 最终数据集

```
Data7 <- Data6 %>%
  mutate(
    `AGE_BMXBMI` = AGE*BMXBMI,
    `AGE_DMDHHSZE` = AGE*DMDHHSZE,
    `AGE_DR1TS060` = AGE*DR1TS060,
    `AGE_DR1TCARB`= AGE*DR1TCARB,
    `AGE_DR1TZINC`= AGE*DR1TZINC,
    `AGE_DR1TRET` = AGE*DR1TRET,
    `LBXTC_BMXBMI` = LBXTC*BMXBMI,
    `LBXTC_DR1TCARB` = LBXTC*DR1TCARB,
    `LBXTC_DR1TMAGN` = LBXTC*DR1TMAGN,
    `LBXTC_DR1TZINC` = LBXTC*DR1TZINC,
    `LBXTC_DR1TALCO` = LBXTC*DR1TALCO,
    `BMXBMI_DR1TS060` =  BMXBMI*DR1TS060,
    `WTDRD1_DMDHHSZE` = WTDRD1*DMDHHSZE,
    `WTDRD1_DR1TRET` = WTDRD1*DR1TRET,
    `DMDHHSZE_DR1TS060` = DMDHHSZE*DR1TS060,
```

```
    `DMDHHSZE_DR1TCARB` = DMDHHSZE*DR1TCARB,
    `DMDHHSZE_RIDRETH3NHB` = DMDHHSZE*RIDRETH3NHB,
    `DMDHHSZE_DR1TALCO` = DMDHHSZE*DR1TALCO,
    `DR1TS060_DR1TCARB` = DR1TS060*DR1TCARB,
    `DR1TS060_DR1TMAGN` = DR1TS060*DR1TMAGN,
    `DR1TS060_DR1TZINC` = DR1TS060*DR1TZINC,
    `DR1TCARB_DR1TRET` = DR1TCARB*DR1TRET,
    `DR1TMAGN_DR1TRET` = DR1TMAGN*DR1TRET
  )

Data_ds <- Data_ds %>%
  mutate(
    `AGE_BMXBMI` = AGE*BMXBMI,
    `AGE_DMDHHSZE` = AGE*DMDHHSZE,
    `AGE_DR1TS060` = AGE*DR1TS060,
    `AGE_DR1TCARB`= AGE*DR1TCARB,
    `AGE_DR1TZINC`= AGE*DR1TZINC,
    `AGE_DR1TRET` = AGE*DR1TRET,
    `LBXTC_BMXBMI` = LBXTC*BMXBMI,
    `LBXTC_DR1TCARB` = LBXTC*DR1TCARB,
    `LBXTC_DR1TMAGN` = LBXTC*DR1TMAGN,
    `LBXTC_DR1TZINC` = LBXTC*DR1TZINC,
    `LBXTC_DR1TALCO` = LBXTC*DR1TALCO,
    `BMXBMI_DR1TS060` =  BMXBMI*DR1TS060,
    `WTDRD1_DMDHHSZE` = WTDRD1*DMDHHSZE,
    `WTDRD1_DR1TRET` = WTDRD1*DR1TRET,
    `DMDHHSZE_DR1TS060` = DMDHHSZE*DR1TS060,
    `DMDHHSZE_DR1TCARB` = DMDHHSZE*DR1TCARB,
    `DMDHHSZE_RIDRETH3NHB` = DMDHHSZE*RIDRETH3NHB,
    `DMDHHSZE_DR1TALCO` = DMDHHSZE*DR1TALCO,
    `DR1TS060_DR1TCARB` = DR1TS060*DR1TCARB,
    `DR1TS060_DR1TMAGN` = DR1TS060*DR1TMAGN,
    `DR1TS060_DR1TZINC` = DR1TS060*DR1TZINC,
    `DR1TCARB_DR1TRET` = DR1TCARB*DR1TRET,
    `DR1TMAGN_DR1TRET` = DR1TMAGN*DR1TRET
  )
```

```
save(Data7,Data_ds,
     file = paste0(getwd(), "/data_use/featureS.RData"))
```