阜外心血管项目建模结果

推想科技 科研组

2017/10/7

数据预处理

原始文件情况

- 文件名：验前概率17-10-6.xls
- 文件大小：2,337,280 字节
- MD5：0025ecdb32846a9518d33f1fb267f923

数据读取

由于2017年10月6日数据修改，前文分析不再有效。经商议，我们不再重复进行数据清洗和描述统计，直接开始建模。

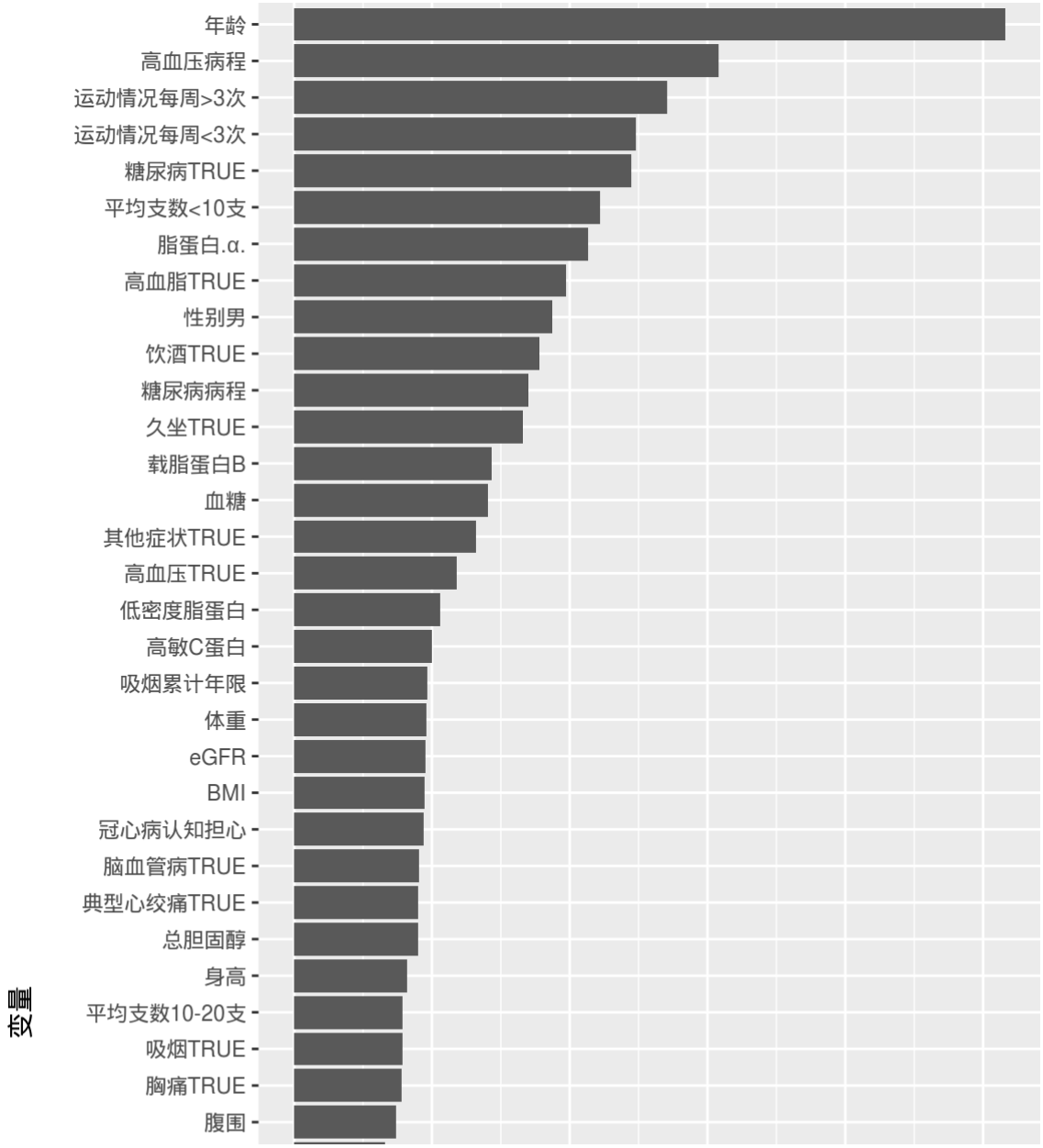
广义线性模型

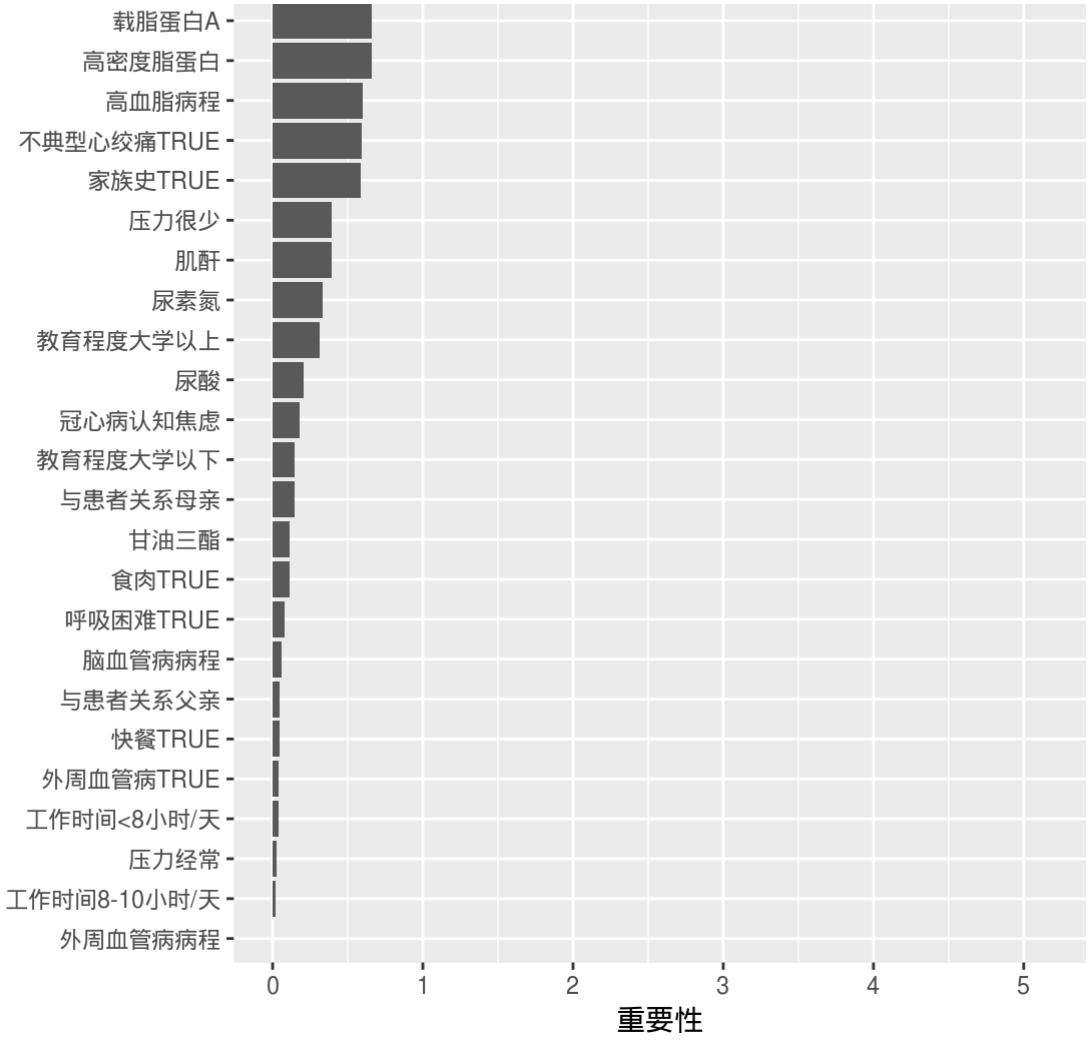
初步模型

首先对所有变量建立logistic模型，根据变量重要性与p值选择纳入最终模型的变量。

```
##
## Call:
## glm(formula = 结果 ~ ., family = "binomial", data = tbLab[,
##     c(1, 3:length(tbLab)), with = F])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3298  -1.0059   0.3785   1.0079   2.3423
##
## Coefficients: (2 not defined because of singularities)
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -7.346e-01  8.163e+00  -0.090  0.92829
## 年龄                 6.997e-02  1.356e-02   5.159 2.48e-07 ***
## 性别男               6.378e-01  3.401e-01   1.875  0.06075 .
## 身高                -3.546e+00  4.314e+00  -0.822  0.41104
## 体重                 4.805e-02  5.008e-02   0.959  0.33734
## BMI                 -1.312e-01  1.391e-01  -0.943  0.34545
## 腹围                -6.426e-03  8.707e-03  -0.738  0.46052
## 胸痛TRUE            -8.603e-02  1.099e-01  -0.783  0.43361
## 不典型心绞痛TRUE    -1.313e-01  2.221e-01  -0.591  0.55438
## 典型心绞痛TRUE      -6.300e-01  6.988e-01  -0.901  0.36733
## 呼吸困难TRUE         7.407e-03  9.621e-02   0.077  0.93863
## 其他症状TRUE        -1.459e-01  1.107e-01  -1.318  0.18759
## 高血压TRUE           1.431e-01  1.212e-01   1.181  0.23745
## 高血脂TRUE           2.463e-01  1.247e-01   1.976  0.04818 *
## 糖尿病TRUE           5.510e-01  2.254e-01   2.444  0.01451 *
## 外周血管病TRUE       1.457e+01  3.567e+02   0.041  0.96742
## 脑血管病TRUE         3.346e-01  3.694e-01   0.906  0.36495
## 家族史TRUE           1.059e-01  1.801e-01   0.588  0.55666
## 与患者关系母亲      -2.831e-02  1.971e-01  -0.144  0.88581
## 与患者关系父亲       1.018e-02  2.053e-01   0.050  0.96047
## 与患者关系无                NA         NA      NA       NA
## 吸烟TRUE             2.212e-01  2.811e-01   0.787  0.43144
## 吸烟累计年限         6.644e-03  6.887e-03   0.965  0.33471
## 平均支数10-20支     -1.717e-01  2.177e-01  -0.788  0.43041
## 平均支数<10支       -6.730e-01  3.027e-01  -2.224  0.02618 *
## 平均支数>20支               NA         NA      NA       NA
## 教育程度大学以上    -1.602e-01  5.149e-01  -0.311  0.75570
## 教育程度大学以下    -3.213e-02  2.229e-01  -0.144  0.88541
## 久坐TRUE            -2.841e-01  1.714e-01  -1.658  0.09736 .
## 工作时间8-10小时/天 -6.403e-03  3.355e-01  -0.019  0.98477
## 工作时间<8小时/天   -1.433e-02  3.619e-01  -0.040  0.96840
## 压力很少             9.276e-02  2.342e-01   0.396  0.69203
## 压力经常             6.108e-03  2.285e-01   0.027  0.97867
## 饮酒TRUE             2.453e-01  1.377e-01   1.782  0.07482 .
## 食肉TRUE            -1.132e-02  1.028e-01  -0.110  0.91233
## 快餐TRUE             1.727e-02  3.995e-01   0.043  0.96552
## 运动情况每周<3次     3.177e-01  1.280e-01   2.483  0.01302 *
## 运动情况每周>3次     3.638e-01  1.345e-01   2.705  0.00684 **
## 冠心病认知担心      -1.710e-01  1.824e-01  -0.938  0.34847
## 冠心病认知焦虑      -7.091e-02  3.995e-01  -0.178  0.85911
## 肌酐                 6.939e-03  1.779e-02   0.390  0.69659
## 尿素氮               1.198e-02  3.590e-02   0.334  0.73860
## 尿酸                -1.326e-04  6.510e-04  -0.204  0.83863
## 血糖                 5.569e-02  3.957e-02   1.407  0.15933
## 总胆固醇            -1.545e-01  1.716e-01  -0.900  0.36807
## 甘油三酯            -7.757e-03  6.872e-02  -0.113  0.91012
```

```
## 高密度脂蛋白         -1.911e-01  2.890e-01  -0.661  0.50847
## 低密度脂蛋白          1.964e-01  1.853e-01   1.060  0.28937
## 脂蛋白.α.             4.040e-04  1.896e-04   2.131  0.03310 *
## 载脂蛋白A            -1.966e-01  2.969e-01  -0.662  0.50791
## 载脂蛋白B             4.495e-01  3.139e-01   1.432  0.15216
## 高敏C蛋白             1.851e-02  1.848e-02   1.002  0.31651
## 高血压病程            2.968e-02  9.631e-03   3.082  0.00206 **
## 高血脂病程            1.043e-02  1.730e-02   0.603  0.54646
## 糖尿病病程           -3.286e-02  1.936e-02  -1.698  0.08959 .
## 外周血管病病程       -7.917e-02  1.039e+02  -0.001  0.99939
## 脑血管病病程          2.339e-03  3.938e-02   0.059  0.95264
## eGFR                  1.787e-02  1.872e-02   0.955  0.33974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3354.0  on 2419  degrees of freedom
## Residual deviance: 2895.9  on 2364  degrees of freedom
## AIC: 3007.9
##
## Number of Fisher Scoring iterations: 13
```

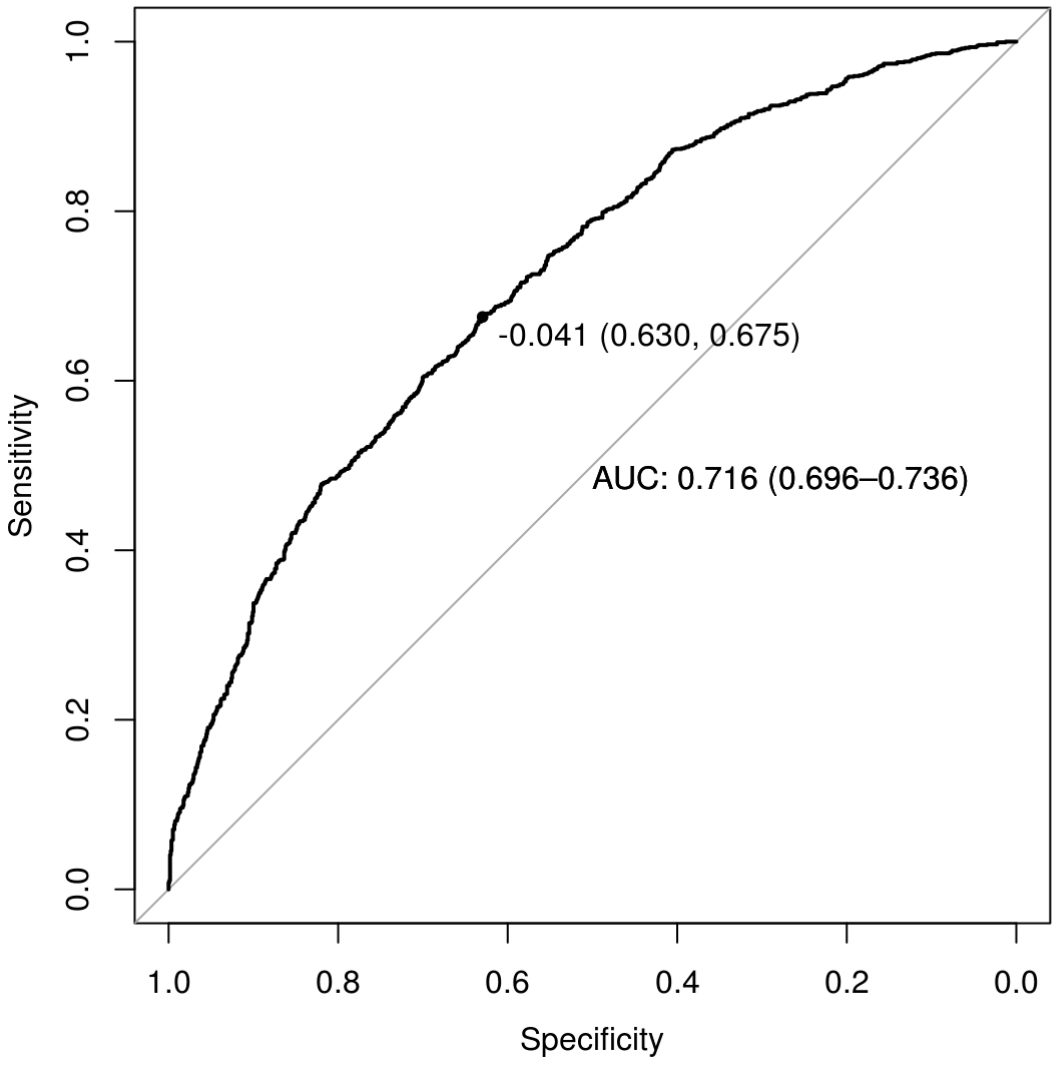




变量选择后模型

纳入变量 年龄 + 性别 + 高血脂 + 糖尿病 + 平均支数 + 久坐 + 运动情况 + 脂蛋白.α.

```
##
## Call:
## glm(formula = 结果 ~ 年龄 + 性别 + 高血脂 + 糖尿病 +
##     平均支数 + 久坐 + 运动情况 + 脂蛋白.α., family = "binomial",
##     data = tbLab[, c(1, 3:length(tbLab)), with = F])
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.197  -1.049   0.474   1.039   2.132
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.5238347  0.3248877 -13.924  < 2e-16 ***
## 年龄              0.0608410  0.0050657  12.010  < 2e-16 ***
## 性别男            0.8411443  0.1113923   7.551 4.31e-14 ***
## 高血脂TRUE        0.4233961  0.0913467   4.635 3.57e-06 ***
## 糖尿病TRUE        0.5457983  0.1315993   4.147 3.36e-05 ***
## 平均支数10-20支   0.2812621  0.1329858   2.115  0.03443 *
## 平均支数<10支    -0.1675538  0.2374874  -0.706  0.48048
## 平均支数>20支     0.5281942  0.2033042   2.598  0.00938 **
## 久坐TRUE         -0.3065212  0.1436040  -2.134  0.03280 *
## 运动情况每周<3次  0.2942912  0.1188070   2.477  0.01325 *
## 运动情况每周>3次  0.3143950  0.1290620   2.436  0.01485 *
## 脂蛋白.α.         0.0004680  0.0001811   2.584  0.00978 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3354.0  on 2419  degrees of freedom
## Residual deviance: 2979.8  on 2408  degrees of freedom
## AIC: 3003.8
##
## Number of Fisher Scoring iterations: 4
```



Cross Validation结果

```
## $cvAUC
## [1] 0.7093674
##
## $se
## [1] 0.0103827
##
## $ci
## [1] 0.6890177 0.7297171
##
## $confidence
## [1] 0.95
```

决策树类模型

C5.0决策树

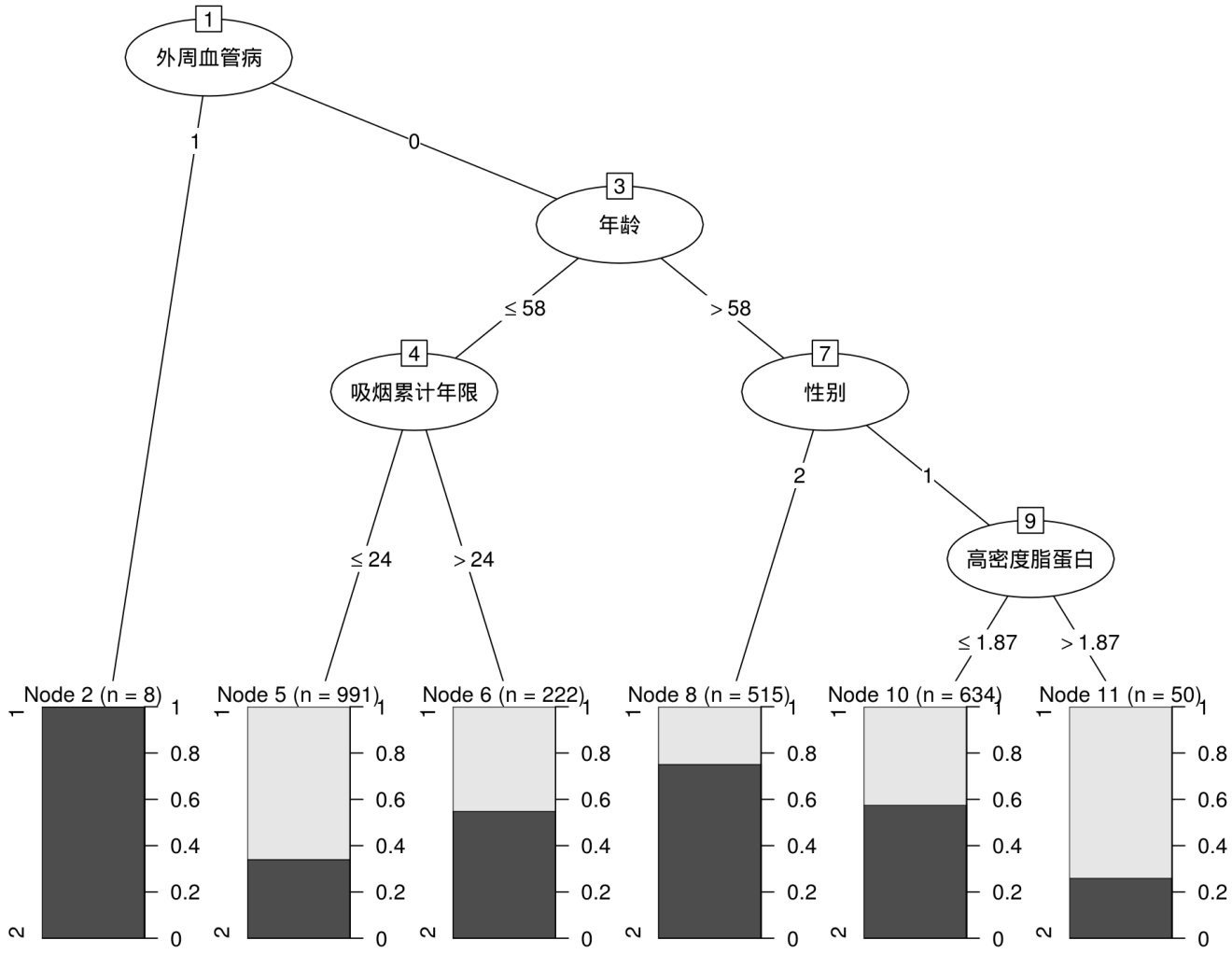
参数搜索

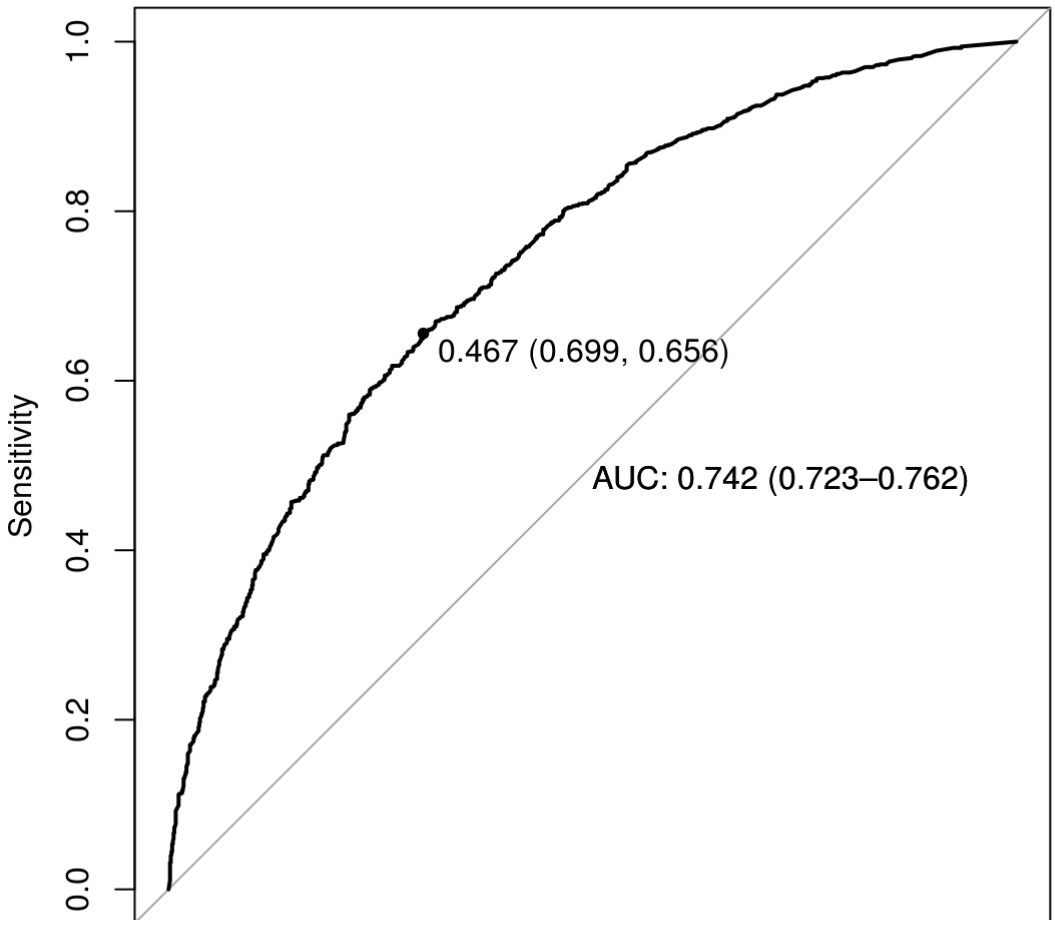
```
##
## Attaching package: 'plyr'
```

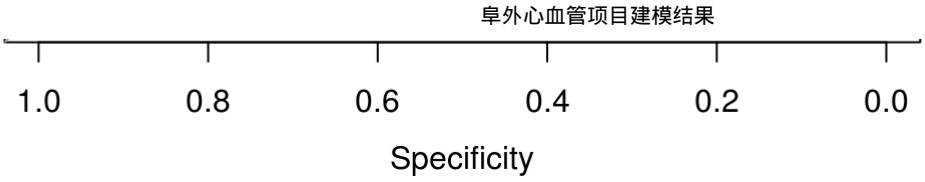
```
## The following objects are masked from 'package:Hmisc':
##
##     is.discrete, summarize
```

```
## C5.0
##
## 2420 samples
##   48 predictors
##    2 classes: 'L1', 'L2'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 2178, 2178, 2179, 2177, 2178, 2178, ...
## Resampling results across tuning parameters:
##
##   model  winnow  trials  ROC        Sens       Spec
##   rules  FALSE    1      0.6385256  0.5791468  0.6892460
##   rules  FALSE   10      0.7025694  0.6318074  0.6619552
##   rules  FALSE   20      0.7031770  0.6351745  0.6606543
##   rules   TRUE    1      0.6412250  0.5821635  0.6873026
##   rules   TRUE   10      0.6976476  0.6363324  0.6533832
##   rules   TRUE   20      0.6979252  0.6371728  0.6528980
##   tree   FALSE    1      0.6690475  0.5729184  0.6894007
##   tree   FALSE   10      0.7000789  0.6516550  0.6387136
##   tree   FALSE   20      0.7004864  0.6518160  0.6400144
##   tree    TRUE    1      0.6697499  0.5830095  0.6825911
##   tree    TRUE   10      0.6966902  0.6476043  0.6428455
##   tree    TRUE   20      0.6967430  0.6460917  0.6439785
##
## ROC was used to select the optimal model using  the largest value.
## The final values used for the model were trials = 20, model = rules
##  and winnow = FALSE.
```

最佳模型


1
外周血管病
1
0
3
年龄
≤ 58
> 58
4
吸烟累计年限
7
性别
≤ 24
> 24
2
1
9
高密度脂蛋白
≤ 1.87
> 1.87
Node 2 (n = 8)
Node 5 (n = 991)
Node 6 (n = 222)
Node 8 (n = 515)
Node 10 (n = 634)
Node 11 (n = 50)


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.467 (0.699, 0.656)
AUC: 0.742 (0.723–0.762)



Cross Validation结果

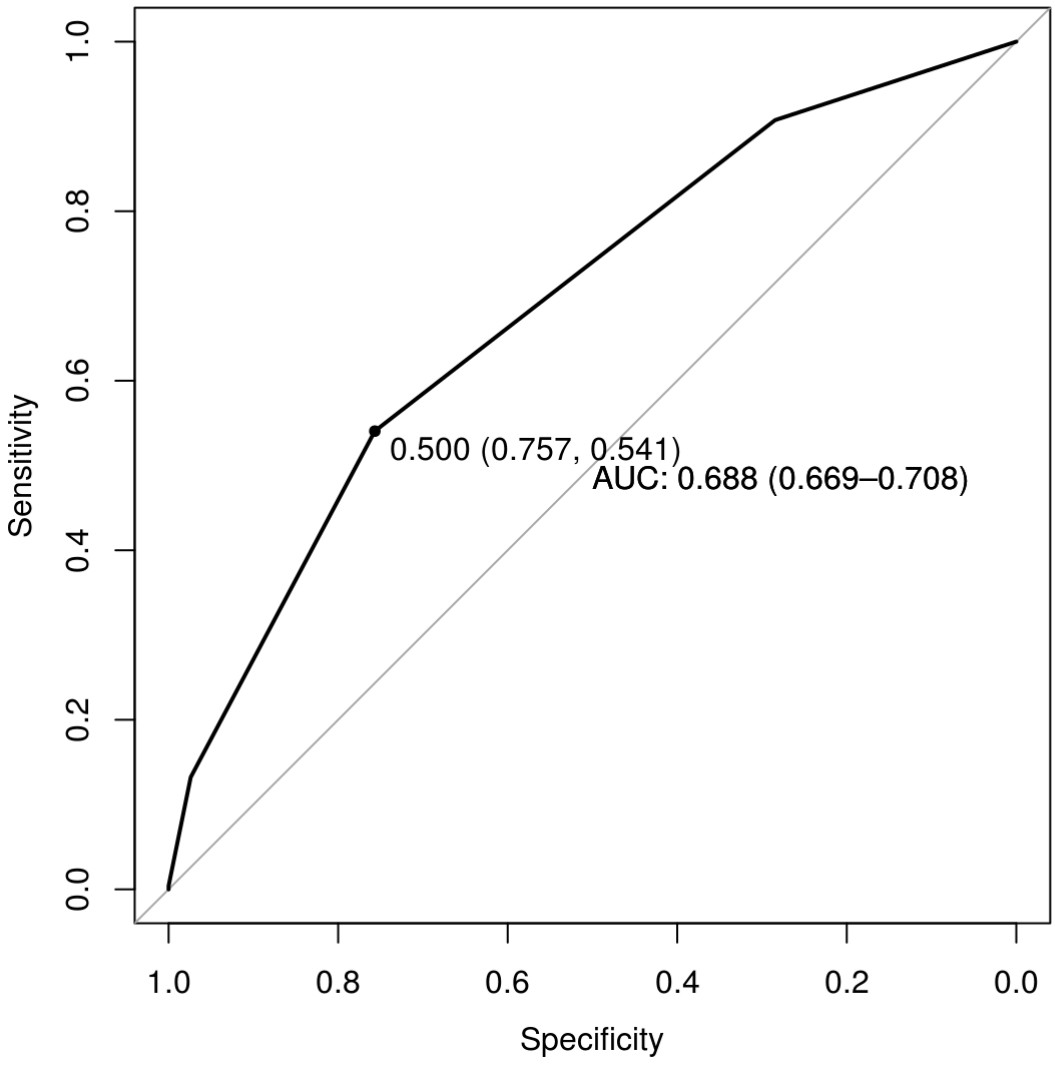
```
## $cvAUC
## [1] 0.6956447
##
## $se
## [1] 0.01074453
##
## $ci
## [1] 0.6745858 0.7167036
##
## $confidence
## [1] 0.95
```

LogitBoost模型

参数搜索

```
## Boosted Logistic Regression
##
## 2420 samples
##   48 predictors
##    2 classes: '无斑块', '有斑块'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 2178, 2178, 2179, 2177, 2178, 2178, ...
## Resampling results across tuning parameters:
##
##   nIter  ROC        Sens       Spec
##    1     0.6235307  0.6096544  0.6374071
##    2     0.6569054  0.7355814  0.7237471
##    3     0.6360508  0.5697764  0.6750131
##    4     0.6532635  0.7540306  0.6204796
##    5     0.6699564  0.6974814  0.5746744
##    6     0.6522436  0.6905045  0.7383984
##    7     0.6272668  0.6054503  0.6417344
##    8     0.5228714  0.1969906  0.8615404
##    9     0.5788328  0.6666097  0.4844392
##   10     0.6160084  0.6541642  0.6460850
##   12     0.6214437  0.6491591  0.7276971
##   15     0.6302670  0.5885653  0.6360237
##   18     0.5891469  0.5863018  0.6370583
##   20     0.5985739  0.6377842  0.6755777
##   23     0.6358876  0.5906044  0.6430326
##   25     0.6277598  0.5873166  0.6347386
##   30     0.6062167  0.5866449  0.6908333
##   35     0.6311502  0.5718606  0.6489990
##   40     0.6075154  0.5987069  0.6833460
##   50     0.6096202  0.5666579  0.6959648
##
## ROC was used to select the optimal model using  the largest value.
## The final value used for the model was nIter = 5.
```

最佳模型



Cross Validation结果

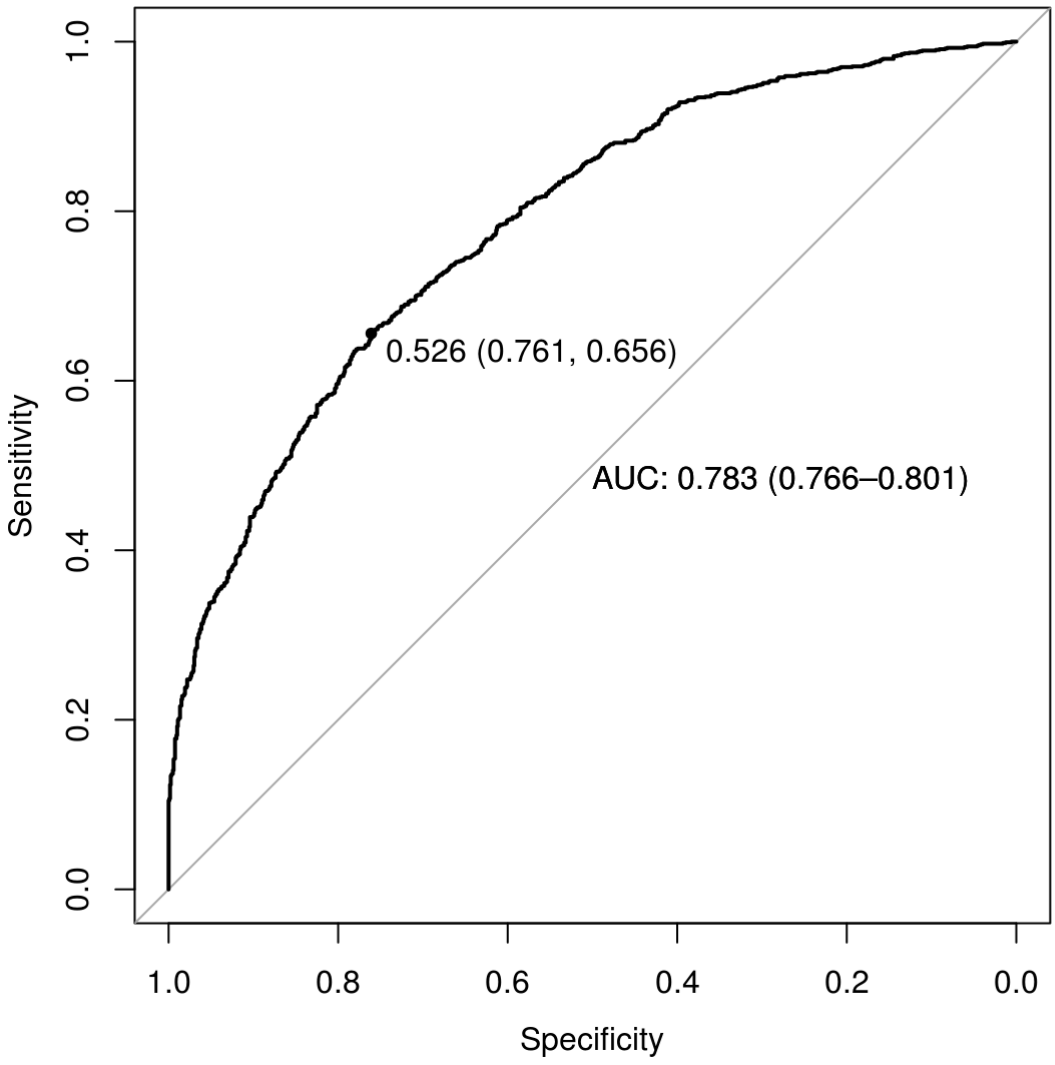
```
## $cvAUC
## [1] 0.6088343
##
## $se
## [1] 0.01514876
##
## $ci
## [1] 0.5791432 0.6385253
##
## $confidence
## [1] 0.95
```

xgBoost模型

参数搜索

```
## eXtreme Gradient Boosting
##
## 2420 samples
##   48 predictors
##    2 classes: '无斑块', '有斑块'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2178, 2178, 2179, 2177, 2178, 2178, ...
## Resampling results across tuning parameters:
##
##   eta    max_depth  ROC        Sens       Spec
##   1e-04  2          0.6776220  0.6093719  0.6460136
##   1e-04  3          0.6860777  0.6060960  0.6614805
##   1e-04  4          0.6893093  0.5925082  0.6752885
##   1e-04  6          0.6711668  0.6102621  0.6436664
##   1e-03  2          0.6945690  0.5765347  0.6972069
##   1e-03  3          0.6982398  0.5959621  0.6931812
##   1e-03  4          0.6996628  0.6102265  0.6680108
##   1e-03  6          0.6893473  0.6211936  0.6526292
##   1e-02  2          0.7019579  0.6262712  0.6387687
##   1e-02  3          0.6949641  0.6296325  0.6420076
##   1e-02  4          0.6879364  0.6270546  0.6460726
##   1e-02  6          0.6824971  0.6144495  0.6468922
##
## Tuning parameter 'nrounds' was held constant at a value of 1000
##  1
## Tuning parameter 'min_child_weight' was held constant at a value of
##  1
## Tuning parameter 'subsample' was held constant at a value of 1
## ROC was used to select the optimal model using  the largest value.
## The final values used for the model were nrounds = 1000, max_depth =
##  2, eta = 0.01, gamma = 1, colsample_bytree = 1, min_child_weight = 1
##  and subsample = 1.
```

最佳模型



Cross Validation结果

```
## $cvAUC
## [1] 0.7066013
##
## $se
## [1] 0.01044251
##
## $ci
## [1] 0.6861344 0.7270683
##
## $confidence
## [1] 0.95
```