預測十年後是否發生冠心病

迴歸期末報告

碩一 107354029 許振榆

目錄

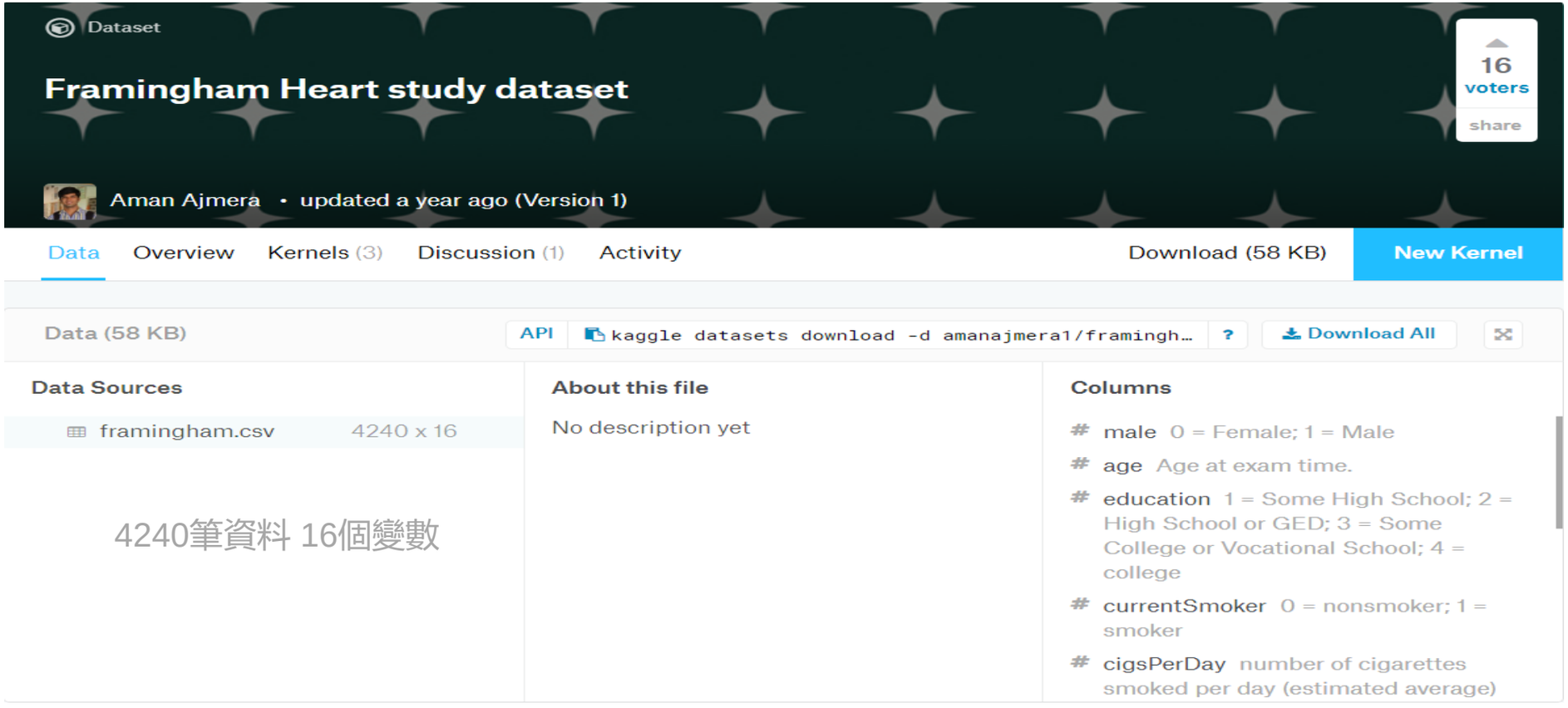
CONTENTS

01 資料來源

Kaggle

預測十年後是否會有冠狀動脈心臟疾病？

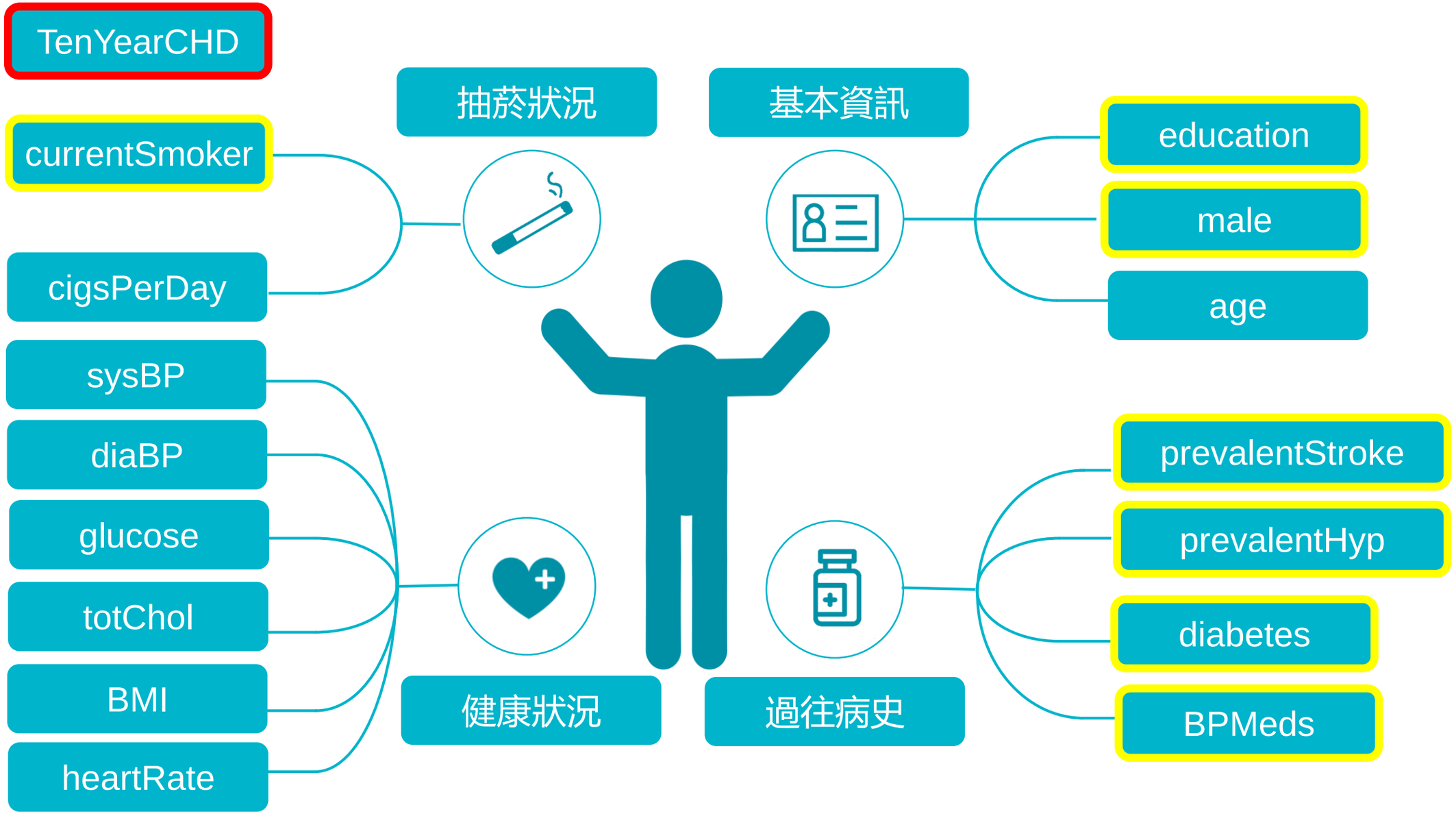
心肌梗塞、狹心症 、心臟衰竭、心律不整



4240筆資料 16個變數

02 變數介紹

變數介紹


TenYearCHD
抽菸狀況
基本資訊
currentSmoker
cigsPerDay
education
male
age
sysBP
diaBP
glucose
totChol
BMI
heartRate
prevalentStroke
prevalentHyp
diabetes
BPMeds
健康狀況
過往病史

變數介紹

```
> head(data)
  male age education currentSmoker cigsPerDay BPMeds prevalentStroke prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
1    1  39         4             0          0      0                0            0        0     195 106.0    70 26.97        80      77          0
2    0  46         2             0          0      0                0            0        0     250 121.0    81 28.73        95      76          0
3    1  48         1             1         20      0                0            0        0     245 127.5    80 25.34        75      70          0
4    0  61         3             1         30      0                0            1        0     225 150.0    95 28.58        65     103          1
5    0  46         3             1         23      0                0            0        0     285 130.0    84 23.10        85      85          0
6    0  43         2             0          0      0                0            1        0     228 180.0   110 30.30        77      99          0
```

```
> str(data)
'data.frame':   4240 obs. of  16 variables:
 $ male           : int  1 0 1 0 0 0 0 0 1 1 ...
 $ age            : int  39 46 48 61 46 43 63 45 52 43 ...
 $ education      : num  4 2 1 3 3 2 1 2 1 1 ...
 $ currentSmoker  : int  0 0 1 1 1 0 0 1 0 1 ...
 $ cigsPerDay     : num  0 0 20 30 23 0 0 20 0 30 ...
 $ BPMeds         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentHyp   : int  0 0 0 1 0 1 0 0 1 1 ...
 $ diabetes       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ totChol        : num  195 250 245 225 285 228 205 313 260 225 ...
 $ sysBP          : num  106 121 128 150 130 ...
 $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
 $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
 $ heartRate      : num  80 95 75 65 85 77 60 79 76 93 ...
 $ glucose        : num  77 76 70 103 85 99 85 78 79 88 ...
 $ TenYearCHD     : int  0 0 0 1 0 0 1 0 0 0 ...
```

03 建模分析

建模分析

- Lasso
- Forward / backward stepwise
- Test statistic

- Logistic Regression

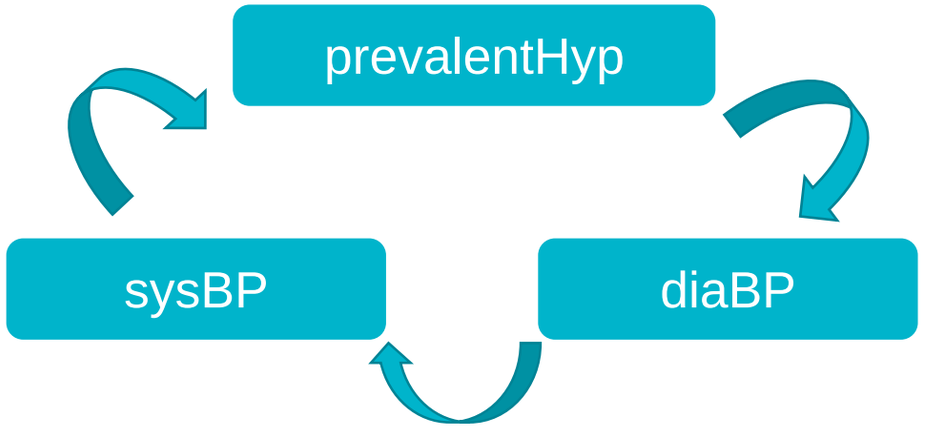
- AUC

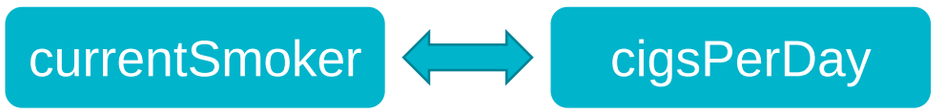
- Accuracy
- Sentivity
- Specificity

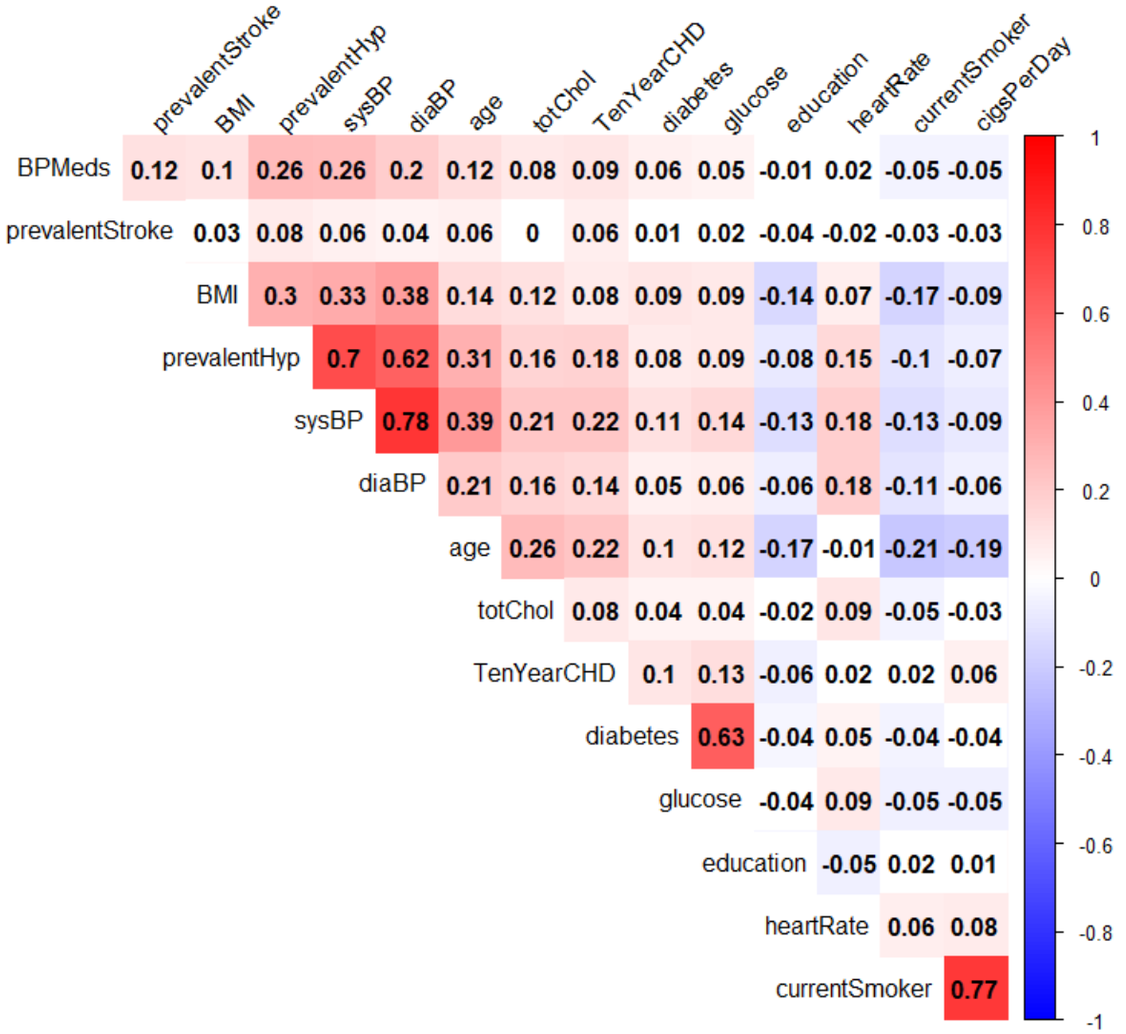
Wald Test

```
Call:
glm(formula = TenYearCHD ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7525  -0.5951  -0.4364  -0.2968   2.7892

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.325385   0.793896 -10.487  < 2e-16 ***
male             0.425641   0.120613   3.529 0.000417 ***
age              0.060514   0.007413   8.163 3.27e-16 ***
education       -0.015936   0.055539  -0.287 0.774170
currentSmoker   -0.005236   0.174273  -0.030 0.976030
cigsPerDay       0.024084   0.006873   3.504 0.000458 ***
BPMeds           0.122051   0.275717   0.443 0.658007
prevalentStroke  1.178688   0.590027   1.998 0.045751 *
prevalentHyp     0.167137   0.154159   1.084 0.278280
diabetes         0.155596   0.367419   0.423 0.671941
totChol          0.001836   0.001241   1.479 0.139076
sysBP            0.014063   0.004164   3.377 0.000732 ***
diaBP           -0.002438   0.007137  -0.342 0.732703
BMI              0.012123   0.013975   0.867 0.385690
heartRate       -0.003311   0.004695  -0.705 0.480716
glucose          0.009600   0.002748   3.494 0.000477 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2542.9  on 2967  degrees of freedom
Residual deviance: 2256.4  on 2952  degrees of freedom
AIC: 2288.4
```

Correlation Plot

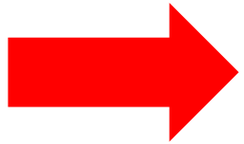






Lasso

```
16 x 1 sparse Matrix of class "dgCMatrix"
                          1
(Intercept)     -8.033955814
male             0.370864624
age              0.058069247
education        .
currentSmoker    .
cigsPerDay       0.020831820
BPMeds           0.031510653
prevalentStroke  0.951605323
prevalentHyp     0.133175738
diabetes         0.130100121
totChol          0.001134616
sysBP            0.012865415
diaBP            .
BMI              .
heartRate        .
glucose          0.008792177
```



刪除 education, currentSmoker
diaBP, heartRate

Stepwise

Forward

```
Step:  AIC=2275.61
TenYearCHD ~ age + sysBP + cigsPerDay + glucose + male + prevalentStroke +
    totChol

                Df Deviance    AIC
<none>               2259.6 2275.6
+ prevalentHyp   1   2258.3 2276.3
+ BMI            1   2258.7 2276.7
+ heartRate      1   2259.2 2277.2
+ BPMeds         1   2259.2 2277.2
+ diabetes       1   2259.3 2277.3
+ education      1   2259.5 2277.5
+ currentSmoker  1   2259.6 2277.6
+ diaBP          1   2259.6 2277.6
```

Backward

```
Step:  AIC=2275.61
TenYearCHD ~ male + age + cigsPerDay + prevalentStroke + totChol +
    sysBP + glucose

                   Df Deviance    AIC
<none>                  2259.6 2275.6
- totChol           1   2261.7 2275.7
- prevalentStroke   1   2263.8 2277.8
- male              1   2273.8 2287.8
- cigsPerDay        1   2284.3 2298.3
- glucose           1   2285.4 2299.4
- sysBP             1   2304.6 2318.6
- age               1   2337.8 2351.8
```

建模分析 - 變數篩選

-currentSmoker

-diaBP

-BPMeds

-heartRate

+age

+male

+BMI

+cigsPerDay

+sysBP

+glucose

+prevalentStroke

+age

+male

+cigsPerDay

+sysBP

+glucose

+prevalentStroke

+age

+male

+BMI

+cigsPerDay

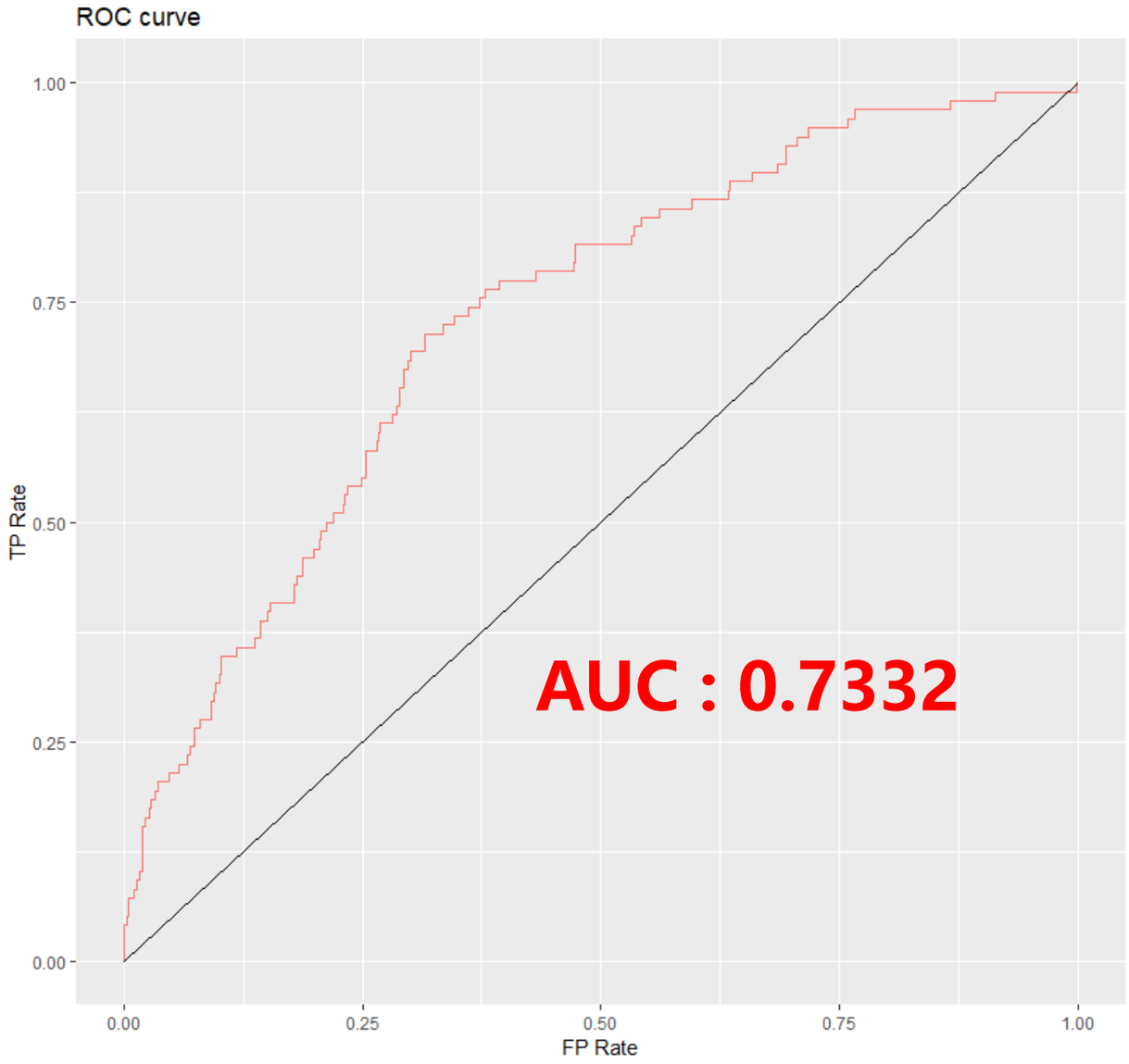
+sysBP

+glucose

+prevalentStroke

建模分析 – 模型評估

```
> model_select <- glm(TenYearCHD~age+male+BMI+cigsPerDay+sysBP+glucose
+prevalentStroke,family="binomial",data = train)
> summary(model_select)

Call:
glm(formula = TenYearCHD ~ age + male + BMI + cigsPerDay + sysBP +
    glucose + prevalentStroke, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7893  -0.5918  -0.4390  -0.3029   2.7695

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -8.686669   0.528827 -16.426  < 2e-16 ***
age               0.063212   0.007047   8.969  < 2e-16 ***
male1             0.411612   0.117021   3.517 0.000436 ***
BMI               0.012118   0.013376   0.906 0.364988
cigsPerDay        0.023812   0.004616   5.159 2.49e-07 ***
sysBP             0.015763   0.002477   6.365 1.96e-10 ***
glucose           0.009936   0.002143   4.637 3.54e-06 ***
prevalentStroke1  1.204161   0.582714   2.066 0.038784 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2542.9  on 2967  degrees of freedom
Residual deviance: 2263.3  on 2960  degrees of freedom
AIC: 2279.3
```



Confusion Matrix

	預測得冠心病	預測未發病	總共
實際有冠發病	13	176	189
實際未發病	10	1073	1083
總共	23	1249	1272

Accuracy : 85.38% (=1086/1272)

Sensitivity : 6.88% (=13/189) 高漏診率

Specificity : 99.08%(=1073/1083) 低誤診率

04 結論

結論

Model : $\text{logit}(\hat{\pi}(x)) = -8.687 + 0.063X_{age} + 0.412X_{male} + 0.012X_{BMI} + 0.024X_{cigsPerDay} + 0.016X_{sysBP} + 0.01X_{glucose} + 1.204X_{prevalentStroke}$

- 每增加1歲，冠心病發的勝算為未增加前1.07倍
- 男性冠心病發的勝算為女性的1.51倍

	勝算比乘數效果	解釋
age	e^0.063 = 1.07	隨著年紀上升，動脈彈性越差，血液循環不佳
male	e^0.412 = 1.51	女性荷爾蒙控制「低密度膽固醇」
BMI	e^0.012 = 1.01	肥胖者有血脂異常、高血壓及高血糖等問題
sysBP	e^0.016 = 1.016	高血壓增加動脈及心臟血管負擔，加速動脈硬化
glucose	e^0.01 = 1.01	冠狀動脈粥樣硬化，引發心血管疾病
cigsPerDay	e^0.024 = 1.024	脂肪累積、血管壁塞，導致血管壁變厚
prevalentStroke	e^1.204 = 3.33	缺血性中風由血管阻塞引起

感謝聆聽