Project Report - Code

Matthew Yau, ZiYing(Sophie) Chen, Xinyu Dong

2023-03-09

Dataset Introduction and

There are 14 categorical and 4 numerical variables in the dataset, and our target variable is "Heart-Disease". This is a clean dataset without any missing data. Among the 319,795 observations, we removed 18,078 duplicates. Therefore, the following explanatory data analysis would only perform on 301,717 observations.

Data Processed

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

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

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

Decision Tree

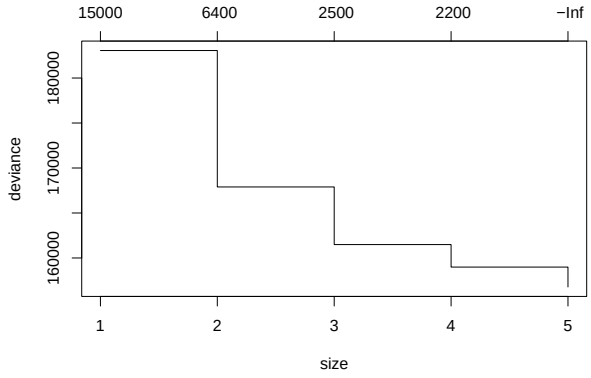
Through the explanatory analysis, we found that our target variable *HeartDisease* is unbalanced which most of the cases do not have heart disease.

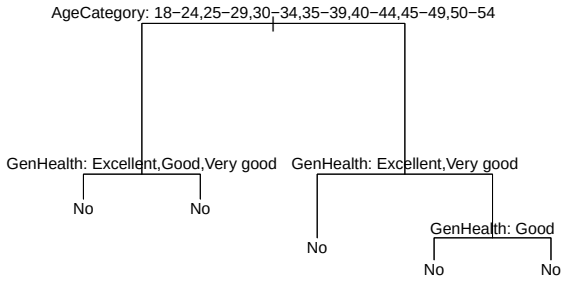
Fiting and Tunning

This is a seemingly strange but interesting decision tree. We can see from the results of the decision tree that the most certain thing about the decision tree is that if you are **under 54 years** of age and **in good overall health**, then the decision tree will assume that you will not develop heart disease. Other factors such as mental health, race, physical activity, etc., have little impact on whether or not you will develop heart disease if you meet the age and overall health criteria.

Now we have identified the low risk group: *they are under 45 years of age and their overall health is good or above*. So let's go a step further and identify those who are not in this range. We will

remove the people who meet the age under 45 and overall overall health. The conclusion is that the decision tree is still trying to identify people who do not have heart disease by their age and overall health status. Therefore, we directly remove the union set that satisfies both categories.





After removing those who passed age threshold and were in good overall health, we selected a total sample of 31,837 (only 10.55% of the total sample). The proportion of people suffering from heart disease in this sample rose to 29.78%. Compared to only 9.0% for the entire sample frame, it is already a significant and encouraging improvement - don't forget that we are only referring to the simple conditions of age 54+ and overall health below health.

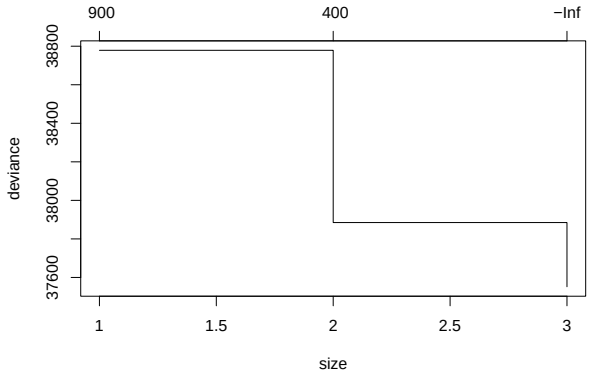
```
## [1] 31837    18
```

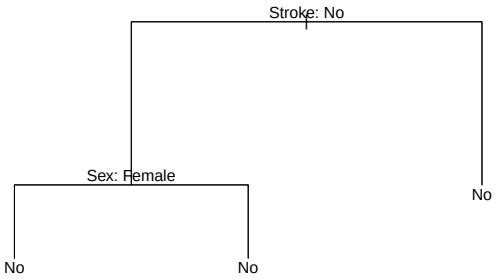
```
## [1] 0.1055194
```

```
##
##        No       Yes
## 0.7022018 0.2977982
```

```
##
##         No        Yes
## 0.90964712 0.09035288
```

Taking the screened non-low-risk people to the next step of decision tree regression, we identified another important signal: **whether or not the person had a stroke.** If there was no stroke the decision tree would assume that the person would not have had a heart attack. In fact if we pick out the people who are already in our risk population who also had a stroke, we can see that the risk of having heart disease if they had a stroke increased from 29.78% to 49.65%. This is also a significant increase. Such a result is not difficult to explain. Stroke is often associated with hardening and blockage of blood vessels, and this often indicates that the patient has a worse blood circulation, which is also an indicator of heart disease. Now that we have greatly identified our high-risk group by age, overall health status, and whether or not we have had a stroke. Let's go one step further and see if there are other factors that can help us determine this. We'll use the data from the further targeted high-risk group in a decision tree regression.





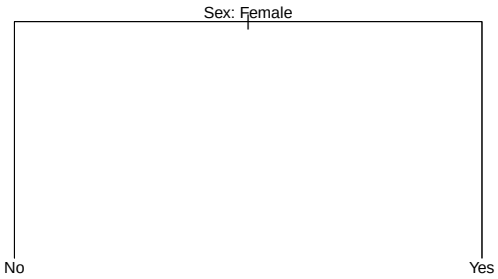

```
## [1] 4399   18
```

```
## [1] 0.01457989
```

```
##
##        No       Yes
## 0.5035235 0.4964765
```

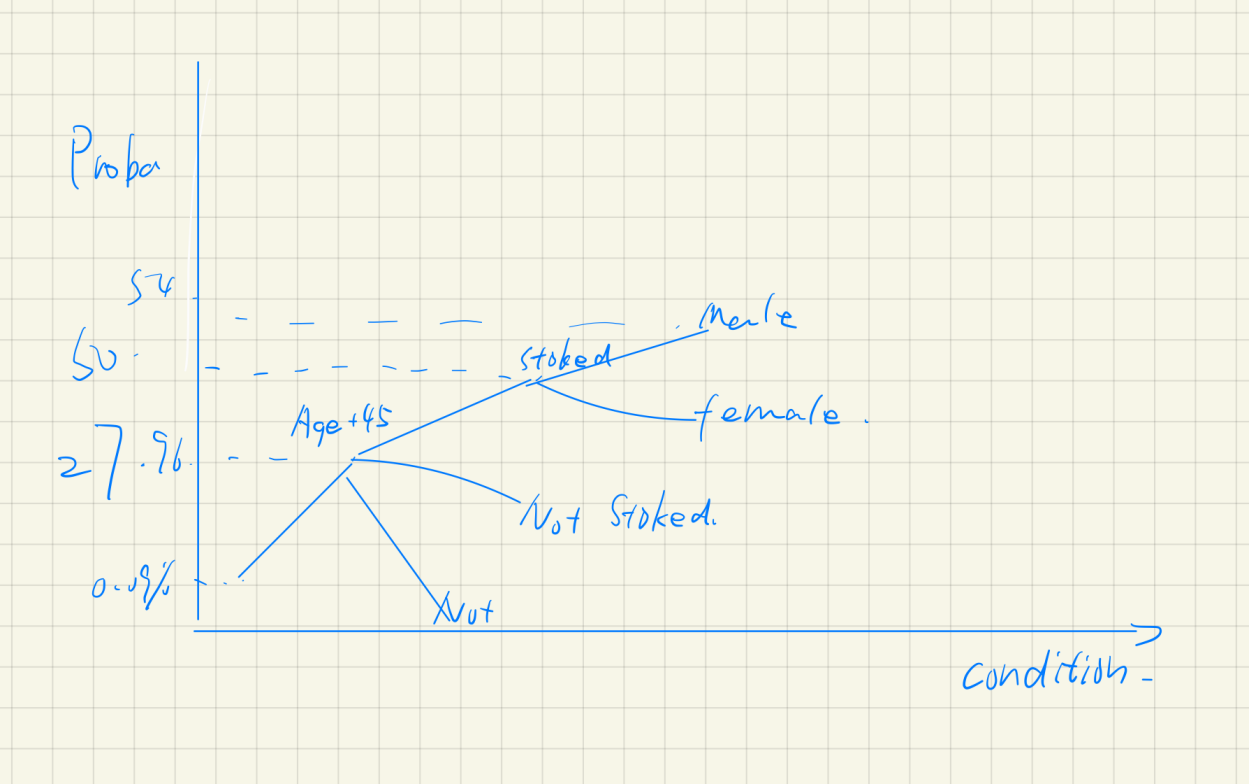
```
##
##         No        Yes
## 0.90964712 0.09035288
```

We were given the simplest decision tree, whether it was male or female. What it tells us is that for these people who are more prone to heart disease, men are at higher risk than women (56.28% for men and 43.46% for women).




```
## $Female
## x
##        No       Yes
## 0.5627907 0.4372093
##
## $Male
## x
##        No       Yes
## 0.4346116 0.5653884
```

The ramification plot is



Model Interpreatation

To be organized

logistic regression

Logistic regression is a statistical method used to analyze and model relationships between a binary dependent variable (i.e., one that takes on only two values, such as 0 or 1) and one or more independent variables (also known as predictors or explanatory variables). It is a type of regression analysis that is used to predict the probability of an event occurring based on the values of the independent variables.

Fiting and Tunning

```
##
## Call:
## glm(formula = HeartDisease ~ ., family = "binomial", data = heart_2020)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1047  -0.4293  -0.2540  -0.1326   3.5881
##
## Coefficients:
```

```
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -6.4830608  0.1034943 -62.642  < 2e-16 ***
## BMI                              0.0081347  0.0011373   7.153 8.52e-13 ***
## SmokingYes                       0.3482008  0.0143796  24.215  < 2e-16 ***
## AlcoholDrinkingYes              -0.2716103  0.0334590  -8.118 4.75e-16 ***
## StrokeYes                        1.0304371  0.0225307  45.735  < 2e-16 ***
## PhysicalHealth                   0.0029129  0.0008598   3.388 0.000704 ***
## MentalHealth                     0.0040747  0.0008799   4.631 3.64e-06 ***
## DiffWalkingYes                   0.2074626  0.0180683  11.482  < 2e-16 ***
## SexMale                          0.7083199  0.0145780  48.588  < 2e-16 ***
## AgeCategory25-29                 0.1236543  0.1241782   0.996 0.319357
## AgeCategory30-34                 0.4917089  0.1110833   4.426 9.58e-06 ***
## AgeCategory35-39                 0.6084139  0.1063690   5.720 1.07e-08 ***
## AgeCategory40-44                 1.0164925  0.1000598  10.159  < 2e-16 ***
## AgeCategory45-49                 1.3409679  0.0964953  13.897  < 2e-16 ***
## AgeCategory50-54                 1.7561326  0.0931489  18.853  < 2e-16 ***
## AgeCategory55-59                 1.9948168  0.0916947  21.755  < 2e-16 ***
## AgeCategory60-64                 2.2575566  0.0908500  24.849  < 2e-16 ***
## AgeCategory65-69                 2.4930843  0.0905818  27.523  < 2e-16 ***
## AgeCategory70-74                 2.7692245  0.0905100  30.596  < 2e-16 ***
## AgeCategory75-79                 2.9576256  0.0910387  32.488  < 2e-16 ***
## AgeCategory80 or older           3.2136495  0.0907803  35.400  < 2e-16 ***
## RaceWhite                        0.2009655  0.0185868  10.812  < 2e-16 ***
## DiabeticNo, borderline diabetes  0.0967987  0.0416679   2.323 0.020174 *
## DiabeticYes                      0.4549081  0.0166775  27.277  < 2e-16 ***
## DiabeticYes (during pregnancy)   0.0927621  0.1047582   0.885 0.375894
## PhysicalActivityYes              0.0354841  0.0160041   2.217 0.026610 *
## GenHealthFair                    1.4481710  0.0328406  44.097  < 2e-16 ***
## GenHealthGood                    0.9747177  0.0296685  32.854  < 2e-16 ***
## GenHealthPoor                    1.8457914  0.0408931  45.137  < 2e-16 ***
## GenHealthVery good               0.4479533  0.0305362  14.670  < 2e-16 ***
## SleepTime                       -0.0234114  0.0043126  -5.429 5.68e-08 ***
## AsthmaYes                        0.2596453  0.0191388  13.566  < 2e-16 ***
## KidneyDiseaseYes                 0.5572447  0.0243079  22.924  < 2e-16 ***
## SkinCancerYes                    0.0931997  0.0194867   4.783 1.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 183054  on 301716  degrees of freedom
## Residual deviance: 143259  on 301683  degrees of freedom
## AIC: 143327
##
## Number of Fisher Scoring iterations: 7

##
## Call:
```

```
## glm(formula = HeartDisease ~ BMI + Smoking + AlcoholDrinking +
##     Stroke + MentalHealth + DiffWalking + Sex + AgeCategory +
##     Race + Diabetic + GenHealth + SleepTime + Asthma + KidneyDisease +
##     SkinCancer, family = "binomial", data = heart_2020)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1016  -0.4292  -0.2541  -0.1326   3.5897
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -6.4460717  0.1020135 -63.188  < 2e-16 ***
## BMI                             0.0078765  0.0011329   6.953 3.58e-12 ***
## SmokingYes                      0.3471071  0.0143650  24.163  < 2e-16 ***
## AlcoholDrinkingYes             -0.2725509  0.0334583  -8.146 3.76e-16 ***
## StrokeYes                       1.0311016  0.0225283  45.769  < 2e-16 ***
## MentalHealth                    0.0045352  0.0008652   5.242 1.59e-07 ***
## DiffWalkingYes                  0.2134403  0.0174816  12.209  < 2e-16 ***
## SexMale                         0.7098993  0.0145599  48.757  < 2e-16 ***
## AgeCategory25-29                0.1247187  0.1241780   1.004   0.3152
## AgeCategory30-34                0.4936093  0.1110826   4.444 8.85e-06 ***
## AgeCategory35-39                0.6106418  0.1063674   5.741 9.42e-09 ***
## AgeCategory40-44                1.0191842  0.1000559  10.186  < 2e-16 ***
## AgeCategory45-49                1.3437634  0.0964909  13.926  < 2e-16 ***
## AgeCategory50-54                1.7592049  0.0931398  18.888  < 2e-16 ***
## AgeCategory55-59                1.9982142  0.0916847  21.794  < 2e-16 ***
## AgeCategory60-64                2.2610619  0.0908384  24.891  < 2e-16 ***
## AgeCategory65-69                2.4955103  0.0905758  27.552  < 2e-16 ***
## AgeCategory70-74                2.7705523  0.0905053  30.612  < 2e-16 ***
## AgeCategory75-79                2.9580492  0.0910300  32.495  < 2e-16 ***
## AgeCategory80 or older          3.2107211  0.0907598  35.376  < 2e-16 ***
## RaceWhite                       0.2042664  0.0185683  11.001  < 2e-16 ***
## DiabeticNo, borderline diabetes 0.0965672  0.0416666   2.318   0.0205 *
## DiabeticYes                     0.4532550  0.0166695  27.191  < 2e-16 ***
## DiabeticYes (during pregnancy)  0.0920104  0.1047763   0.878   0.3799
## GenHealthFair                   1.4630667  0.0321238  45.545  < 2e-16 ***
## GenHealthGood                   0.9754897  0.0296131  32.941  < 2e-16 ***
## GenHealthPoor                   1.8869713  0.0377813  49.945  < 2e-16 ***
## GenHealthVery good              0.4481023  0.0305333  14.676  < 2e-16 ***
## SleepTime                      -0.0238686  0.0043118  -5.536 3.10e-08 ***
## AsthmaYes                       0.2615269  0.0191319  13.670  < 2e-16 ***
## KidneyDiseaseYes                0.5587476  0.0242988  22.995  < 2e-16 ***
## SkinCancerYes                   0.0955836  0.0194694   4.909 9.13e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 183054  on 301716  degrees of freedom
## Residual deviance: 143275  on 301685  degrees of freedom
## AIC: 143339
##
## Number of Fisher Scoring iterations: 7
```

```
## 'log Lik.' 0.2173929 (df=34)
```

```
## 'log Lik.' 0.2173082 (df=32)
```

```
##      predicted.classes
##           No    Yes
##   No  156522 117934
##   Yes   2506  24755
```

```
## [1] 0.9080738
```

Add interaction terms

The R squared of these two are similar and both poor, so we decided to try adding interaction terms to see if we can predict heart disease better. Interaction term shows that one's effect on response variable depends on the other, with only one of the variable might not have predictive power, but combine them together, we can predict. In case other predictors depends on *PhysicalActivity* (which removed by stepwise process), we keep it to build the model with all interaction terms.

According to our EDA, there are no strong correlation among numerical variables, so we would not include interaction terms of them.

```
##
## Call:
## glm(formula = formula, family = "binomial", data = heart_2020)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1274  -0.4255  -0.2517  -0.1340   3.5680
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -5.9553171  0.1794494 -33.187  < 2e-16 ***
## BMI                             0.0078005  0.0011380   6.855 7.15e-12 ***
## SmokingYes                      0.3484604  0.0144340  24.142  < 2e-16 ***
## AlcoholDrinkingYes             -0.2694730  0.0335226  -8.039 9.09e-16 ***
## StrokeYes                       1.0317565  0.0225446  45.765  < 2e-16 ***
## MentalHealth                    0.0045738  0.0008678   5.271 1.36e-07 ***
## DiffWalkingYes                  0.2234664  0.0177403  12.597  < 2e-16 ***
## SexMale                         0.3366418  0.1811524   1.858 0.063121 .
```

```
## AgeCategory25-29                  0.3701813  0.2313194   1.600 0.109531
## AgeCategory30-34                  0.4432789  0.2156171   2.056 0.039796 *
## AgeCategory35-39                  0.5331098  0.2078047   2.565 0.010305 *
## AgeCategory40-44                  0.9785622  0.1937049   5.052 4.38e-07 ***
## AgeCategory45-49                  1.1344194  0.1884093   6.021 1.73e-09 ***
## AgeCategory50-54                  1.4715475  0.1822586   8.074 6.81e-16 ***
## AgeCategory55-59                  1.6179942  0.1799440   8.992  < 2e-16 ***
## AgeCategory60-64                  1.7573319  0.1782596   9.858  < 2e-16 ***
## AgeCategory65-69                  1.8710372  0.1780723  10.507  < 2e-16 ***
## AgeCategory70-74                  2.0149156  0.1785382  11.286  < 2e-16 ***
## AgeCategory75-79                  2.2200972  0.1812322  12.250  < 2e-16 ***
## AgeCategory80 or older            2.5652124  0.1799880  14.252  < 2e-16 ***
## RaceWhite                        -0.2375369  0.1774088  -1.339 0.180595
## DiabeticNo, borderline diabetes   0.0928287  0.0416823   2.227 0.025944 *
## DiabeticYes                       0.4563343  0.0166941  27.335  < 2e-16 ***
## DiabeticYes (during pregnancy)    0.0372118  0.1043681   0.357 0.721433
## GenHealthFair                     1.4688656  0.0322778  45.507  < 2e-16 ***
## GenHealthGood                     0.9777956  0.0296735  32.952  < 2e-16 ***
## GenHealthPoor                     1.8971121  0.0381010  49.792  < 2e-16 ***
## GenHealthVery good                0.4492436  0.0305664  14.697  < 2e-16 ***
## SleepTime                        -0.0246392  0.0043190  -5.705 1.16e-08 ***
## AsthmaYes                         0.2628819  0.0191611  13.720  < 2e-16 ***
## KidneyDiseaseYes                  0.5588551  0.0243178  22.981  < 2e-16 ***
## SkinCancerYes                     0.0751522  0.0195927   3.836 0.000125 ***
## PhysicalActivityYes               0.0324824  0.0159815   2.032 0.042103 *
## AgeCategory25-29:RaceWhite       -0.1106557  0.2496704  -0.443 0.657616
## AgeCategory30-34:RaceWhite        0.0981903  0.2248687   0.437 0.662361
## AgeCategory35-39:RaceWhite        0.2534951  0.2172044   1.167 0.243178
## AgeCategory40-44:RaceWhite        0.1553904  0.2023210   0.768 0.442464
## AgeCategory45-49:RaceWhite        0.1264382  0.1947562   0.649 0.516201
## AgeCategory50-54:RaceWhite        0.2683943  0.1881933   1.426 0.153821
## AgeCategory55-59:RaceWhite        0.3361309  0.1854915   1.812 0.069969 .
## AgeCategory60-64:RaceWhite        0.3963129  0.1836103   2.158 0.030893 *
## AgeCategory65-69:RaceWhite        0.4937586  0.1834706   2.691 0.007119 **
## AgeCategory70-74:RaceWhite        0.6266512  0.1839657   3.406 0.000658 ***
## AgeCategory75-79:RaceWhite        0.6283280  0.1866886   3.366 0.000764 ***
## AgeCategory80 or older:RaceWhite  0.6492878  0.1855970   3.498 0.000468 ***
## SexMale:AgeCategory25-29         -0.3343677  0.2517183  -1.328 0.184065
## SexMale:AgeCategory30-34         -0.0204413  0.2263218  -0.090 0.928033
## SexMale:AgeCategory35-39         -0.1690083  0.2162751  -0.781 0.434537
## SexMale:AgeCategory40-44         -0.1290879  0.2037688  -0.634 0.526406
## SexMale:AgeCategory45-49          0.2206847  0.1969496   1.121 0.262495
## SexMale:AgeCategory50-54          0.1990793  0.1901882   1.047 0.295215
## SexMale:AgeCategory55-59          0.2845198  0.1871957   1.520 0.128535
## SexMale:AgeCategory60-64          0.4201838  0.1854541   2.266 0.023470 *
## SexMale:AgeCategory65-69          0.4921193  0.1848275   2.663 0.007754 **
## SexMale:AgeCategory70-74          0.5286793  0.1844027   2.867 0.004144 **
## SexMale:AgeCategory75-79          0.4979006  0.1851314   2.689 0.007157 **
```

```
## SexMale:AgeCategory80 or older    0.2797246  0.1842405   1.518 0.128949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 183054  on 301716  degrees of freedom
## Residual deviance: 143020  on 301660  degrees of freedom
## AIC: 143134
##
## Number of Fisher Scoring iterations: 7

## 'log Lik.' 0.2187004 (df=57)
```

Analysis of the model

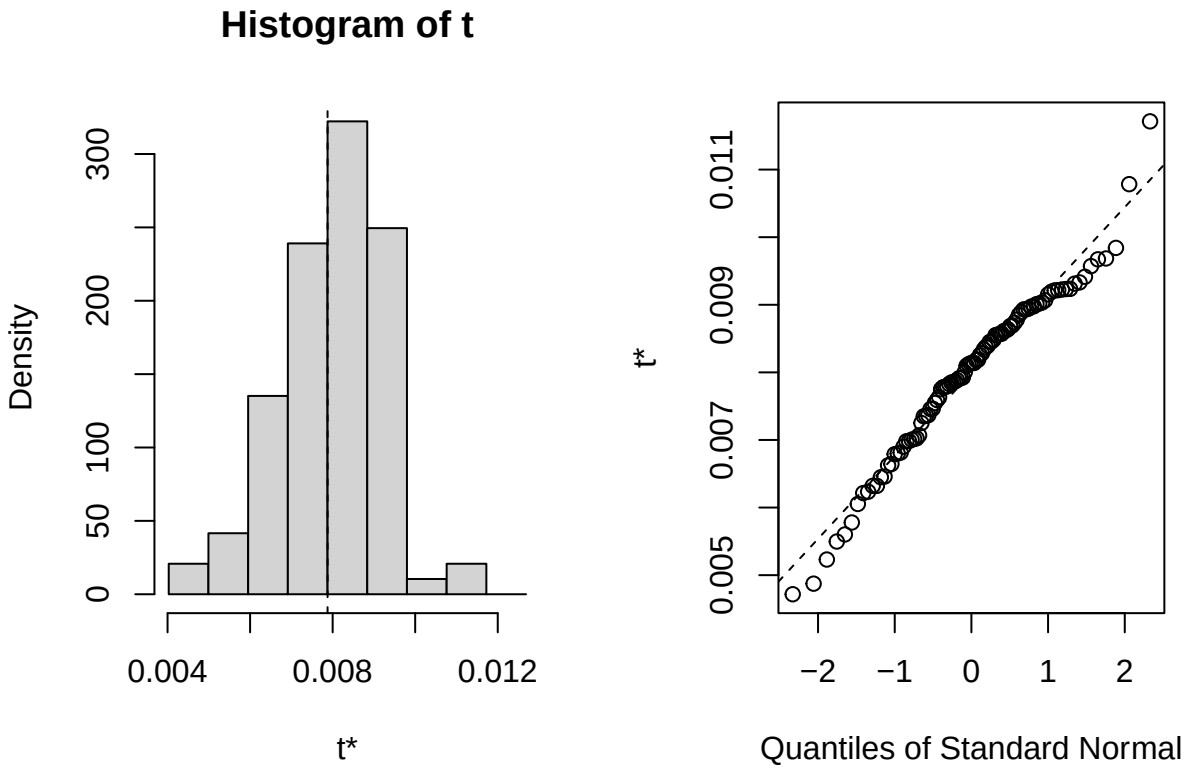
After stepwise selection, the predictors *PhysicalHealth* and *PhysicalActivity* were removed from the heart.logistic model, we can apply F-test to see if the reduced model is statistically better. Since the p-value 0.0004292 is pretty small at 0.05 significant level, hence, the reduced model is significantly better than the full model.

```
## Analysis of Deviance Table
##
## Model 1: HeartDisease ~ BMI + Smoking + AlcoholDrinking + Stroke + MentalHealth +
##     DiffWalking + Sex + AgeCategory + Race + Diabetic + GenHealth +
##     SleepTime + Asthma + KidneyDisease + SkinCancer
## Model 2: HeartDisease ~ BMI + Smoking + AlcoholDrinking + Stroke + PhysicalHealth +
##     MentalHealth + DiffWalking + Sex + AgeCategory + Race + Diabetic +
##     PhysicalActivity + GenHealth + SleepTime + Asthma + KidneyDisease +
##     SkinCancer
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1    301685     143275
## 2    301683     143259  2   15.507 0.0004292 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 0.0007187478

## HeartDisease ~ BMI + Smoking + AlcoholDrinking + Stroke + MentalHealth +
##     DiffWalking + Sex + AgeCategory + Race + Diabetic + GenHealth +
##     SleepTime + Asthma + KidneyDisease + SkinCancer
```

Model Checking

Applying bootstrap to check the model, each term in the log-likelihood sum should be reasonably large in the model. Any terms that are very small indicates the failure to be modeled properly.

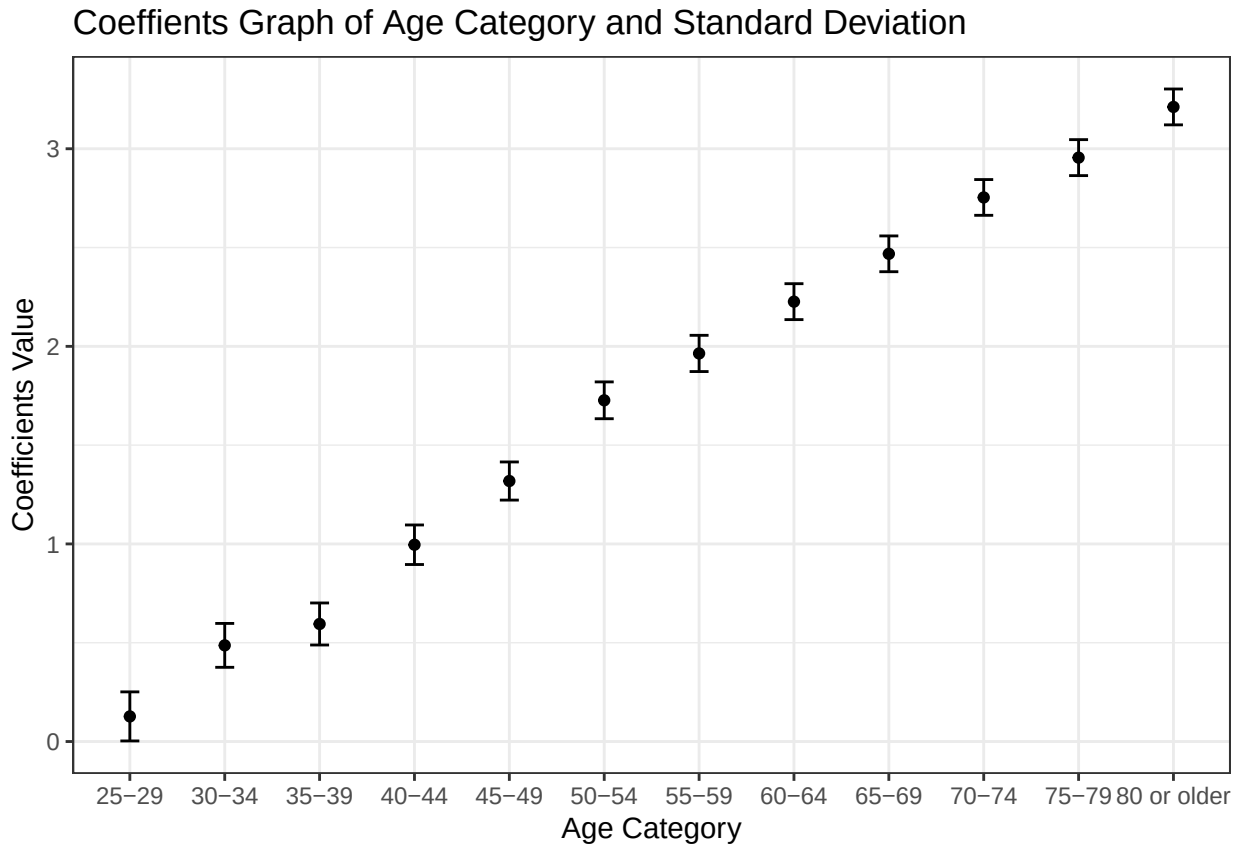
```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = heart_2020, statistic = logit_test, R = 100)
##
##
## Bootstrap Statistics :
##          original        bias     std. error
## t1*  -6.446071740 -2.091603e-02 0.1145323735
## t2*   0.007876473  1.139442e-04 0.0012266150
## t3*   0.347107149 -1.011339e-03 0.0151733748
## t4*  -0.272550891 -1.745688e-04 0.0299487071
## t5*   1.031101597 -9.460360e-04 0.0214199910
## t6*   0.004535173 -6.211811e-06 0.0008911005
## t7*   0.213440306 -2.409625e-04 0.0211502810
## t8*   0.709899346  2.748347e-03 0.0141254727
## t9*   0.124718718  5.114783e-03 0.1383294916
## t10*  0.493609263  1.124711e-02 0.1181623018
## t11*  0.610641794 -6.137073e-05 0.1125165388
## t12*  1.019184237  1.164150e-02 0.1087608119
## t13*  1.343763412  1.548137e-02 0.1047256085
## t14*  1.759204937  8.776048e-03 0.1063702963
## t15*  1.998214155  1.178647e-02 0.0987604665
## t16*  2.261061879  1.181663e-02 0.0992172239
## t17*  2.495510329  1.518921e-02 0.0976727037
## t18*  2.770552283  1.197332e-02 0.0989283986
## t19*  2.958049208  1.440908e-02 0.0986428377
## t20*  3.210721100  1.029761e-02 0.0966182311
## t21*  0.204266386 -2.821392e-03 0.0202840028
## t22*  0.096567209  1.561594e-04 0.0487821231
## t23*  0.453255037 -2.404704e-03 0.0167870240
## t24*  0.092010359 -7.312978e-03 0.1171018504
## t25*  1.463066694  4.623866e-04 0.0304348863
## t26*  0.975489723 -1.627737e-04 0.0286363731
## t27*  1.886971317 -1.639648e-04 0.0394831868
## t28*  0.448102259  1.076328e-03 0.0280590972
## t29* -0.023868629  7.537616e-04 0.0046628508
## t30*  0.261526894  2.099826e-03 0.0175411373
## t31*  0.558747560  1.615355e-03 0.0272860465
## t32*  0.095583605 -9.602778e-04 0.0186617831
```



- Based on this model we can hava a quantitative understanding of this model. Considering the AgeCategory which is considered as the most dominant factor in our decision tree. Their coeffiecnts can be show as

```
##                  (Intercept)                          BMI
##                -6.446071740                  0.007876473
##                   SmokingYes            AlcoholDrinkingYes
##                  0.347107149                 -0.272550891
##                    StrokeYes                MentalHealth
##                  1.031101597                  0.004535173
##                 DiffWalkingYes                      SexMale
##                  0.213440306                  0.709899346
##              AgeCategory25-29              AgeCategory30-34
##                  0.124718718                  0.493609263
##              AgeCategory35-39              AgeCategory40-44
##                  0.610641794                  1.019184237
##              AgeCategory45-49              AgeCategory50-54
##                  1.343763412                  1.759204937
##              AgeCategory55-59              AgeCategory60-64
##                  1.998214155                  2.261061879
##              AgeCategory65-69              AgeCategory70-74
##                  2.495510329                  2.770552283
```

```
##                AgeCategory75-79          AgeCategory80 or older
##                     2.958049208                     3.210721100
##                       RaceWhite DiabeticNo, borderline diabetes
##                     0.204266386                     0.096567209
##                      DiabeticYes  DiabeticYes (during pregnancy)
##                     0.453255037                     0.092010359
##                    GenHealthFair                   GenHealthGood
##                     1.463066694                     0.975489723
##                    GenHealthPoor               GenHealthVery good
##                     1.886971317                     0.448102259
##                       SleepTime                       AsthmaYes
##                    -0.023868629                     0.261526894
##                KidneyDiseaseYes                   SkinCancerYes
##                     0.558747560                     0.095583605
```



Model Interpreatation

GAM model

```
## Loading required package: nlme
```

```
##
```

```
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.8-41. For overview type 'help("mgcv-package")'.

##
## Family: binomial
## Link function: logit
##
## Formula:
## HeartDisease ~ s(BMI) + Smoking + AlcoholDrinking + Stroke +
##     s(MentalHealth) + DiffWalking + Sex + Race + Diabetic + GenHealth +
##     s(SleepTime) + Asthma + KidneyDisease + SkinCancer + PhysicalActivity +
##     s(PhysicalHealth)
##
## Parametric coefficients:
##                                  Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                      -4.63278    0.03480 -133.111  < 2e-16 ***
## SmokingYes                        0.39143    0.01415   27.664  < 2e-16 ***
## AlcoholDrinkingYes               -0.45070    0.03287  -13.711  < 2e-16 ***
## StrokeYes                         1.20850    0.02257   53.555  < 2e-16 ***
## DiffWalkingYes                    0.49802    0.01793   27.782  < 2e-16 ***
## SexMale                           0.54950    0.01441   38.144  < 2e-16 ***
## RaceWhite                         0.47379    0.01801   26.306  < 2e-16 ***
## DiabeticNo, borderline diabetes   0.31728    0.04132    7.678 1.62e-14 ***
## DiabeticYes                       0.68266    0.01653   41.296  < 2e-16 ***
## DiabeticYes (during pregnancy)   -0.19932    0.10257   -1.943  0.05198 .
## GenHealthFair                     1.67682    0.03278   51.146  < 2e-16 ***
## GenHealthGood                     1.18741    0.02942   40.363  < 2e-16 ***
## GenHealthPoor                     2.07128    0.04054   51.090  < 2e-16 ***
## GenHealthVery good                0.60778    0.03018   20.138  < 2e-16 ***
## AsthmaYes                         0.10200    0.01872    5.450 5.05e-08 ***
## KidneyDiseaseYes                  0.68914    0.02435   28.296  < 2e-16 ***
## SkinCancerYes                     0.49803    0.01918   25.964  < 2e-16 ***
## PhysicalActivityYes              -0.05072    0.01579   -3.213  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                      edf Ref.df Chi.sq p-value
## s(BMI)             5.096  6.144 206.26  <2e-16 ***
## s(MentalHealth)    6.233  7.094 659.29  <2e-16 ***
## s(SleepTime)       6.648  7.457 133.18  <2e-16 ***
## s(PhysicalHealth) 4.624  5.523  69.61  <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.13   Deviance explained = 16.9%
## UBRE = -0.49576  Scale est. = 1         n = 301717

##
## Family: binomial
## Link function: logit
##
## Formula:
## HeartDisease ~ BMI + Smoking + AlcoholDrinking + Stroke + MentalHealth +
##     DiffWalking + Sex + AgeCategory + Race + Diabetic + GenHealth +
##     SleepTime + Asthma + KidneyDisease + SkinCancer
##
## Parametric coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -6.4460719  0.1020348 -63.175  < 2e-16 ***
## BMI                             0.0078765  0.0011329   6.953 3.58e-12 ***
## SmokingYes                      0.3471071  0.0143651  24.163  < 2e-16 ***
## AlcoholDrinkingYes             -0.2725509  0.0334585  -8.146 3.76e-16 ***
## StrokeYes                       1.0311016  0.0225283  45.769  < 2e-16 ***
## MentalHealth                    0.0045352  0.0008652   5.242 1.59e-07 ***
## DiffWalkingYes                  0.2134403  0.0174816  12.209  < 2e-16 ***
## SexMale                         0.7098993  0.0145600  48.757  < 2e-16 ***
## AgeCategory25-29                0.1247189  0.1241997   1.004   0.3153
## AgeCategory30-34                0.4936094  0.1111021   4.443 8.88e-06 ***
## AgeCategory35-39                0.6106420  0.1063877   5.740 9.48e-09 ***
## AgeCategory40-44                1.0191844  0.1000775  10.184  < 2e-16 ***
## AgeCategory45-49                1.3437636  0.0965133  13.923  < 2e-16 ***
## AgeCategory50-54                1.7592051  0.0931630  18.883  < 2e-16 ***
## AgeCategory55-59                1.9982143  0.0917082  21.789  < 2e-16 ***
## AgeCategory60-64                2.2610620  0.0908621  24.885  < 2e-16 ***
## AgeCategory65-69                2.4955105  0.0905996  27.544  < 2e-16 ***
## AgeCategory70-74                2.7705524  0.0905291  30.604  < 2e-16 ***
## AgeCategory75-79                2.9580494  0.0910537  32.487  < 2e-16 ***
## AgeCategory80 or older          3.2107213  0.0907836  35.367  < 2e-16 ***
## RaceWhite                       0.2042664  0.0185684  11.001  < 2e-16 ***
## DiabeticNo, borderline diabetes 0.0965672  0.0416667   2.318   0.0205 *
## DiabeticYes                     0.4532550  0.0166695  27.191  < 2e-16 ***
## DiabeticYes (during pregnancy)  0.0920104  0.1047765   0.878   0.3799
## GenHealthFair                   1.4630667  0.0321239  45.544  < 2e-16 ***
## GenHealthGood                   0.9754897  0.0296133  32.941  < 2e-16 ***
## GenHealthPoor                   1.8869713  0.0377815  49.944  < 2e-16 ***
## GenHealthVery good              0.4481023  0.0305334  14.676  < 2e-16 ***
## SleepTime                      -0.0238686  0.0043118  -5.536 3.10e-08 ***
## AsthmaYes                       0.2615269  0.0191320  13.670  < 2e-16 ***
## KidneyDiseaseYes                0.5587476  0.0242988  22.995  < 2e-16 ***
```

```
## SkinCancerYes                    0.0955836  0.0194694   4.909 9.13e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.159   Deviance explained = 21.7%
## UBRE = -0.52492  Scale est. = 1         n = 301717
```

```
## [1] 152138.9
```

```
## [1] 143338.7
```