Project Report - Code

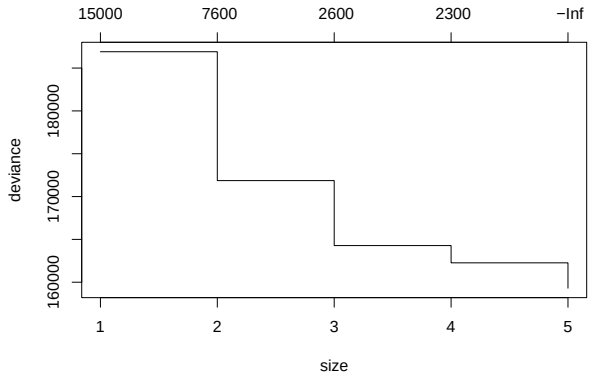
Xinyu Dong

2023-03-17

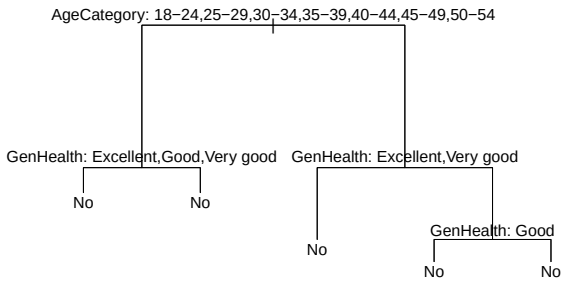
Decision Tree

Fiting and Tunning

```
# Load data
library(readr)
heart_2020 <- read.csv("heart_2020_cleaned.csv",stringsAsFactors = TRUE)

library(tree)


# Fit decision tree model
model.trees <- tree(HeartDisease ~ ., data = heart_2020)


heart.treecv <- cv.tree(model.trees) # Cross-validation
plot(heart.treecv)
```




```
# Tuning
heart.tree.prune <- prune.misclass(model.trees, best=5)
plot(heart.tree.prune)
text(heart.tree.prune,pretty = 0)
```



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.

```
obs_not_low_risk <- subset(heart_2020,!(AgeCategory %in% c('18-24',
                                                           '25-29',
                                                    '30-34',
                                                    '35-39',
                                                    '40-44',
                                                    '45-49',
                                                    '50-54') |

                                 GenHealth %in% c('Good','Very good','Excellent'))
                           )

#New data Investigation
dim(obs_not_low_risk)
```

```
## [1] 31857    18
```

```
nrow(obs_not_low_risk)/nrow(heart_2020)
```

```
## [1] 0.09961694
```

```
prop.table(table(obs_not_low_risk$HeartDisease))
```

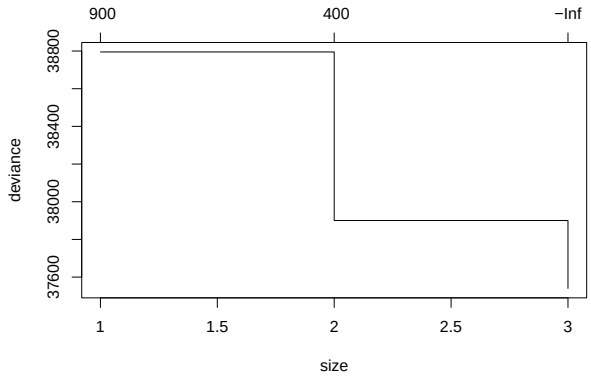
```
##
##        No       Yes
## 0.7023574 0.2976426
```

```
prop.table(table(heart_2020$HeartDisease))
```

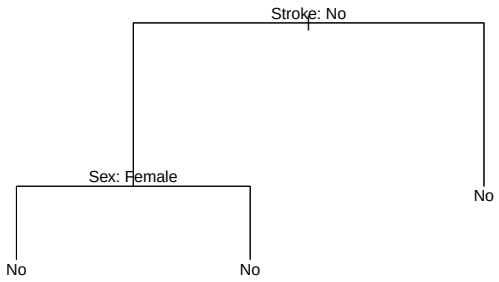
```
##
##         No        Yes
## 0.91440454 0.08559546
```

After removing those who passed age threshold and were in good overall health, we selected a total sample of 31,857 (only 9.96% of the total sample). The proportion of people suffering from heart disease in this sample rose to 29.76%. Compared to only 8.6% 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 45+ and overall health below health.

```
# Fit decision tree model
model.trees <- tree(HeartDisease ~ ., data = obs_not_low_risk)


heart.treecv <- cv.tree(model.trees) # Cross-validation
plot(heart.treecv)
```




```
# Tuning
heart.tree.prune <- prune.misclass(model.trees, best=3)
plot(heart.tree.prune)
text(heart.tree.prune,pretty = 0)
```




```
Stroked <- subset(obs_not_low_risk,Stroke == 'Yes')

#New data Investigation
dim(Stroked)
```

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

```
nrow(Stroked)/nrow(heart_2020)
```

```
## [1] 0.01375569
```

```
prop.table(table(Stroked$HeartDisease))
```

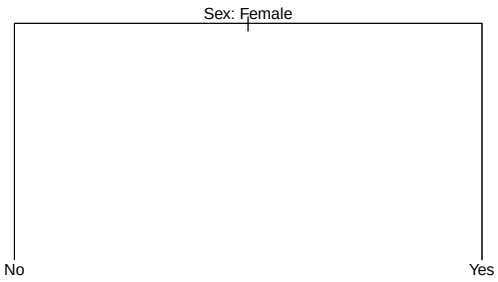
```
##
##        No       Yes
## 0.5035235 0.4964765
```

```
prop.table(table(heart_2020$HeartDisease))
```

```
##
##         No        Yes
## 0.91440454 0.08559546
```

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 29.76 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.

```
# Fit decision tree model
model.trees <- tree(HeartDisease ~ ., data = Stroked)

# Tuning
#heart.treecv <- cv.tree(model.trees) # Cross-validation
#plot(heart.treecv)


plot(model.trees)
text(model.trees,pretty = 0)
```




```
tapply(Stroked$HeartDisease, Stroked$Sex, function(x) prop.table(table(x)))
```

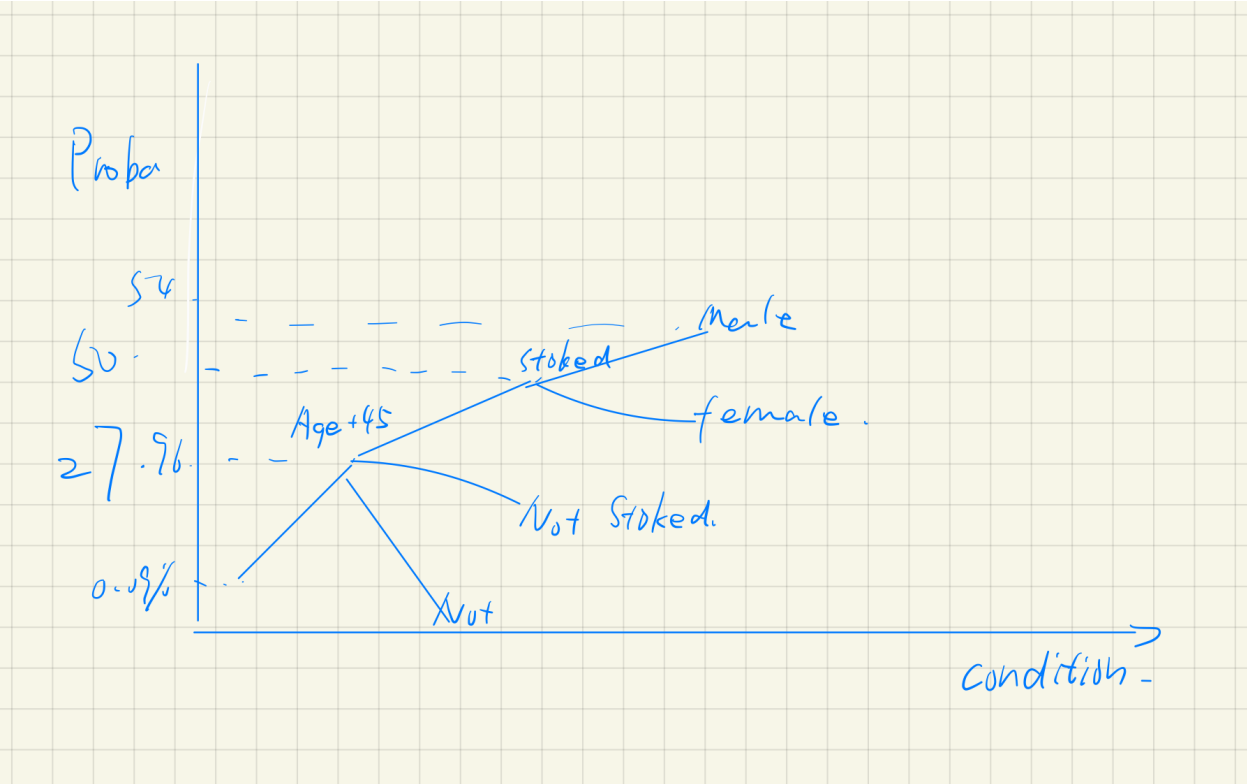
```
## $Female
## x
##        No       Yes
## 0.5627907 0.4372093
##
```

```
## $Male
## x
##        No       Yes
## 0.4346116 0.5653884
```

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 (56.54 for men and 43.72% for women).

The ramification plot is

```
# To be ploted
probability <- c(0.2976 ,0.4965,0.5654)
```



Model Interpreatation

To be organized

logistic regression

Fiting and Tunning

```
heart.logistic <- glm(HeartDisease~.,data = heart_2020,family = 'binomial')

#summary(heart.logistic)
stepwise_model <- step(heart.logistic, direction = "both", k = log(nrow(heart_2020)), trace = 0)

summary(stepwise_model)
```

```
##
## Call:
```

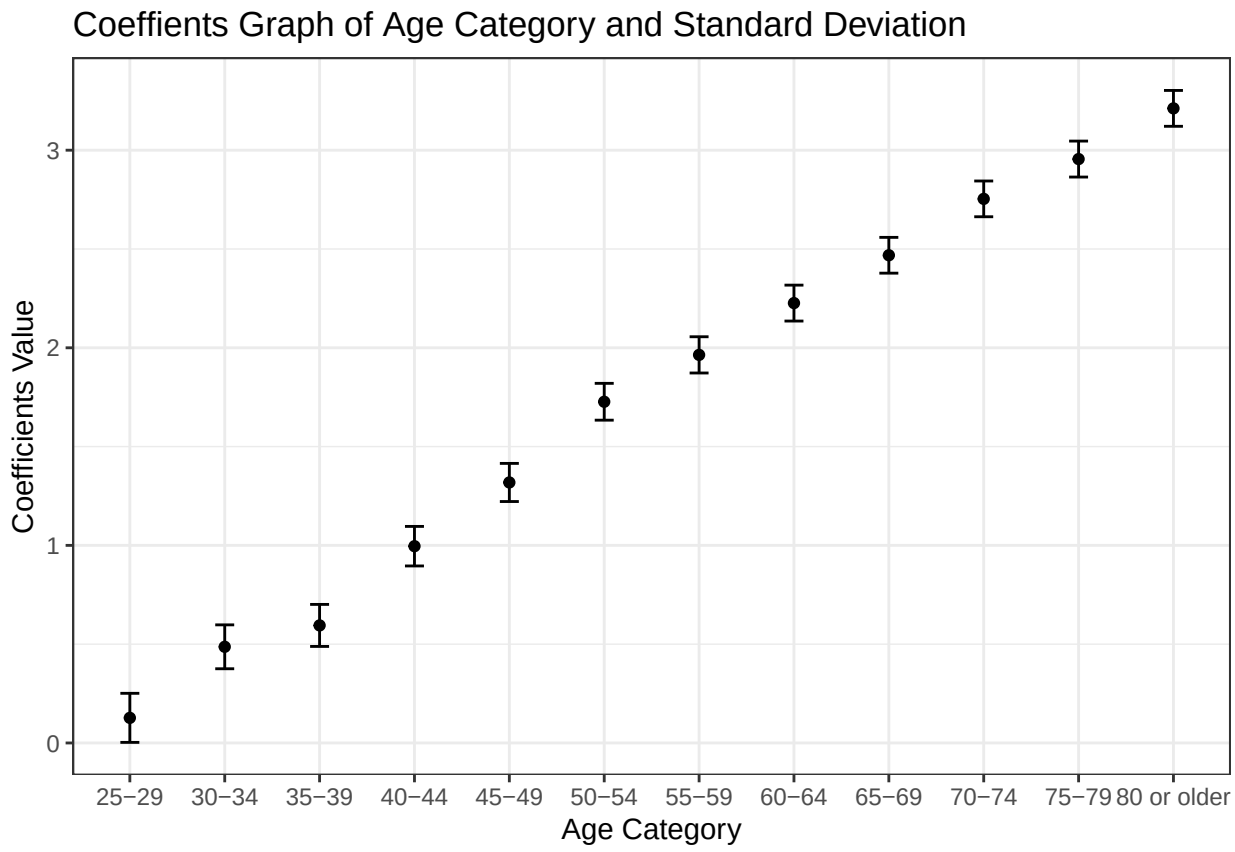
```
## glm(formula = HeartDisease ~ BMI + Smoking + AlcoholDrinking +
##     Stroke + PhysicalHealth + MentalHealth + DiffWalking + Sex +
##     AgeCategory + Race + Diabetic + GenHealth + SleepTime + Asthma +
##     KidneyDisease + SkinCancer, family = "binomial", data = heart_2020)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1305  -0.4110  -0.2440  -0.1295   3.6092
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -6.2716984  0.1133720 -55.320  < 2e-16 ***
## BMI                              0.0083590  0.0011397   7.335 2.22e-13 ***
## SmokingYes                       0.3555262  0.0143578  24.762  < 2e-16 ***
## AlcoholDrinkingYes              -0.2407292  0.0335307  -7.179 7.00e-13 ***
## StrokeYes                        1.0468845  0.0226319  46.257  < 2e-16 ***
## PhysicalHealth                   0.0031114  0.0008615   3.612 0.000304 ***
## MentalHealth                     0.0046882  0.0008823   5.313 1.08e-07 ***
## DiffWalkingYes                   0.2096299  0.0179435  11.683  < 2e-16 ***
## SexMale                          0.7088911  0.0145490  48.724  < 2e-16 ***
## AgeCategory25-29                 0.1270495  0.1241807   1.023 0.306260
## AgeCategory30-34                 0.4866544  0.1110910   4.381 1.18e-05 ***
## AgeCategory35-39                 0.5948967  0.1063721   5.593 2.24e-08 ***
## AgeCategory40-44                 0.9958578  0.1000683   9.952  < 2e-16 ***
## AgeCategory45-49                 1.3184067  0.0964936  13.663  < 2e-16 ***
## AgeCategory50-54                 1.7268540  0.0931507  18.538  < 2e-16 ***
## AgeCategory55-59                 1.9641295  0.0916889  21.422  < 2e-16 ***
## AgeCategory60-64                 2.2261981  0.0908493  24.504  < 2e-16 ***
## AgeCategory65-69                 2.4682742  0.0905748  27.251  < 2e-16 ***
## AgeCategory70-74                 2.7536005  0.0905089  30.424  < 2e-16 ***
## AgeCategory75-79                 2.9551423  0.0910431  32.459  < 2e-16 ***
## AgeCategory80 or older           3.2117057  0.0907736  35.381  < 2e-16 ***
## RaceAsian                       -0.5362281  0.0841224  -6.374 1.84e-10 ***
## RaceBlack                       -0.3459589  0.0577519  -5.990 2.09e-09 ***
## RaceHispanic                    -0.2557943  0.0588256  -4.348 1.37e-05 ***
## RaceOther                       -0.0606252  0.0639967  -0.947 0.343478
## RaceWhite                       -0.0778680  0.0515924  -1.509 0.131224
## DiabeticNo, borderline diabetes  0.1291314  0.0418238   3.088 0.002018 **
## DiabeticYes                      0.4764399  0.0167149  28.504  < 2e-16 ***
## DiabeticYes (during pregnancy)   0.1240304  0.1050694   1.180 0.237816
## GenHealthFair                    1.5162827  0.0327643  46.278  < 2e-16 ***
## GenHealthGood                    1.0439427  0.0295570  35.320  < 2e-16 ***
## GenHealthPoor                    1.8948865  0.0407730  46.474  < 2e-16 ***
## GenHealthVery good               0.4709907  0.0303698  15.509  < 2e-16 ***
## SleepTime                       -0.0251464  0.0043376  -5.797 6.74e-09 ***
## AsthmaYes                        0.2777669  0.0192190  14.453  < 2e-16 ***
## KidneyDiseaseYes                 0.5681036  0.0244022  23.281  < 2e-16 ***
## SkinCancerYes                    0.1153223  0.0194905   5.917 3.28e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 186906  on 319794  degrees of freedom
```

```
## Residual deviance: 145151  on 319758  degrees of freedom
## AIC: 145225
##
## Number of Fisher Scoring iterations: 7
```

- 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

```
library(ggplot2)
coef <- c(0.1270495,
          0.4866544,
          0.5948967,
          0.9958578,
          1.3184067,
          1.7268540,
          1.9641295,
          2.2261981,
          2.4682742,
          2.7536005,
          2.9551423,
          3.2117057)



coef_sd <- c(
  0.1241807,
  0.1110910,
  0.1063721,
  0.1000683,
  0.0964936,
  0.0931507,
  0.0916889,
  0.0908493,
  0.0905748,
  0.0905089,
  0.0910431,
  0.0907736
)


ageCategroy <- c( '25-29',
'30-34',
'35-39',
'40-44',
'45-49',
'50-54',
'55-59',
'60-64',
'65-69',
'70-74',
'75-79',
'80 or older')
```

```r
coefframe <- data.frame(ageCategroy,coef,coef_sd)

ggplot(coefframe, aes(ageCategroy, coef)) +
  geom_errorbar(aes(ymin = coef - coef_sd, ymax = coef + coef_sd), width = 0.2) +
  geom_point() +
  labs(title = "Coeffients Graph of Age Category and Standard Deviation") +
  xlab("Age Category") +
  ylab("Coefficients Value")+
  theme_bw()
```



Model Interpreatation