10.1 Final Project: Cardiovascular Disease Predictor

Roni Kaakaty

8/6/2020

Cardiovascular disease is the leading cause of death in the World. According to the CDC, 80% of these deaths are preventable. I wanted to examine what factors would be able to accurately predict if an individual is at risk of being diagnosed with cardiovascular disease based on certain factors. By knowing which factors contribute the most to a positive diagnosis, patients would be able to get an early indication of what conditions need to improve to improve their chances of avoiding a tragic event. Utilizing the Framingham dataset located on Kaggle, I was able to test various variables to see which would or wouldn't have a significant impact in predicting a possible cardiovascular even within the next 10 years. I first wanted to identify if there would be a drastic difference between males and females and if the variables impacted one more greatly than the other. I found that men were more likely to experience a cardiac event by a factor of 0.535621. Since the other variables seemed to impact the individual almost the same (not dependent on gender), I decided to incorporate both into my logistic regression model instead of producing two separate models. With the use of the logistic regression model, I was able to eliminate variables that had no significance on the output. These included education, current smoker, BP meds, diabetes, diastolic bp, BMI, and heart rate. BMI made sense to remove since that isn't the best reflection of an individual's body composition. The interesting thing to me was the model implied that glucose was a significant variable, but diabetes was not. I assumed they were correlated, but the logistic model stated otherwise based on the data. The variables that showed the most significance were gender, age, cigarettes per day, prevalent hypertension, total cholesterol, systolic bp, and glucose. Individuals with prevalent hypertension had increased odds of a cardiovascular event by a factor of 0.234235. This aligns with what I assumed would be the case when I first began this project. It makes sense that individuals who struggle with blood pressure are the most at risk, especially if they aren't currently using blood pressure medication to try to limit their risk as much as possible. I was able to create a prediction model that would be able to predict whether or not the individual would experience a cardiovascular event with an accuracy of 85%. If someone were to replicate my study, I would try to locate a dataset with even more applicable variables such as family history, time spent exercising on a weekly basis, and profession (office job/active job). Individuals with high BMIs don't automatically make them "overweight" as that also depends on their body composition. I also wonder if demographic played a role at all since this was limited to one area in the nation.

Load the libraries that will be used

```
library(ggplot2)
library(tidyr)
library(dplyr)
```

```
##
## 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
```

```
library(foreign)
library(class)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

Set directory

```
setwd("/Users/Roni Kaakaty/Documents/Github/dsc520")
```

Open up dataset

```
cardiovascular_df <- read.csv("data/cardio-dataset.csv")
```

Convert variables to factors

```
cardiovascular_df$education <- as.factor(cardiovascular_df$education)
cardiovascular_df$currentSmoker <- as.factor(cardiovascular_df$currentSmoker)
cardiovascular_df$BPMeds <- as.factor(cardiovascular_df$BPMeds)
cardiovascular_df$prevalentHyp <- as.factor(cardiovascular_df$prevalentHyp)
cardiovascular_df$diabetes <- as.factor(cardiovascular_df$diabetes)
```

Remove NA values from dataset

```
cardiovascular_df2 <- na.omit(cardiovascular_df)
```

Clean up the data so that 0 reflects F and 1 reflects M. Replace 0 with No and 1 with Yes.

```
cardiovascular_df2[cardiovascular_df2$gender == 0,]$gender <- "F"
cardiovascular_df2[cardiovascular_df2$gender == 1,]$gender <- "M"

cardiovascular_df2[cardiovascular_df2$TenYearCHD == 0,]$TenYearCHD <- "No"
cardiovascular_df2[cardiovascular_df2$TenYearCHD == 1,]$TenYearCHD <- "Yes"
```

```r
cardiovascular_df2$gender <- as.factor(cardiovascular_df2$gender)
cardiovascular_df2$TenYearCHD <- as.factor(cardiovascular_df2$TenYearCHD)
```

Verify all variables are appropriate.

```r
str(cardiovascular_df2)
```

```
## 'data.frame':    3656 obs. of  15 variables:
##  $ gender       : Factor w/ 2 levels "F","M": 2 1 2 1 1 1 1 1 2 2 ...
##  $ age          : int  39 46 48 61 46 43 63 45 52 43 ...
##  $ education    : Factor w/ 4 levels "1","2","3","4": 4 2 1 3 3 2 1 2 1 1 ...
##  $ currentSmoker: Factor w/ 2 levels "0","1": 1 1 2 2 2 1 1 2 1 2 ...
##  $ cigsPerDay   : int  0 0 20 30 23 0 0 20 0 30 ...
##  $ BPMeds       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ prevalentHyp : Factor w/ 2 levels "0","1": 1 1 1 2 1 2 1 1 2 2 ...
##  $ diabetes     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ totChol      : int  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    : int  80 95 75 65 85 77 60 79 76 93 ...
##  $ glucose      : int  77 76 70 103 85 99 85 78 79 88 ...
##  $ TenYearCHD   : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 2 1 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:582] 15 22 27 34 37 43 50 55 71 73 ...
##   ..- attr(*, "names")= chr [1:582] "15" "22" "27" "34" ...
```

Confirm that there is enough data to run Male/Female in the same model.

```r
xtabs(~TenYearCHD + gender, data = cardiovascular_df2)
```

```
##           gender
## TenYearCHD    F    M
##        No  1784 1315
##        Yes  250  307
```

Create Logistic Regression Model with just gender

```r
cardio_gender <- glm(TenYearCHD ~ gender, data = cardiovascular_df2, family = "binomial")
summary(cardio_gender)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ gender, family = "binomial", data = cardiovascular_df2)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -0.6478  -0.6478  -0.5121  -0.5121   2.0476
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.96515    0.06753 -29.100  < 2e-16 ***
## genderM      0.51041    0.09262   5.511 3.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3120.5  on 3655  degrees of freedom
## Residual deviance: 3090.0  on 3654  degrees of freedom
## AIC: 3094
##
## Number of Fisher Scoring iterations: 4
```

Create a Logistic Regression Model with all variables.

```
cardio_log <- glm(TenYearCHD ~., data = cardiovascular_df2, family = "binomial")
summary(cardio_log)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = "binomial", data = cardiovascular_df2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9344  -0.5945  -0.4244  -0.2828   2.8634
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.260514   0.709994 -11.635  < 2e-16 ***
## genderM         0.535621   0.109865   4.875 1.09e-06 ***
## age             0.062328   0.006750   9.233  < 2e-16 ***
## education2     -0.193875   0.123391  -1.571  0.11613
## education3     -0.193894   0.150092  -1.292  0.19642
## education4     -0.067496   0.164514  -0.410  0.68160
## currentSmoker1  0.070758   0.156667   0.452  0.65152
## cigsPerDay      0.017917   0.006232   2.875  0.00404 **
## BPMeds1         0.201740   0.232434   0.868  0.38543
## prevalentHyp1   0.239374   0.138052   1.734  0.08293 .
## diabetes1       0.023640   0.315945   0.075  0.94035
## totChol         0.002361   0.001129   2.092  0.03643 *
## sysBP           0.015346   0.003808   4.030 5.58e-05 ***
## diaBP          -0.003927   0.006440  -0.610  0.54202
## BMI             0.005545   0.012773   0.434  0.66418
## heartRate      -0.003160   0.004209  -0.751  0.45281
## glucose         0.007237   0.002236   3.237  0.00121 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3120.5  on 3655  degrees of freedom
## Residual deviance: 2753.9  on 3639  degrees of freedom
## AIC: 2787.9
##
## Number of Fisher Scoring iterations: 5
```

Remove variables without significance from function.

```
cardio_log2 <- glm(TenYearCHD~ gender + age + cigsPerDay + prevalentHyp + totChol + sysBP + glucose, da
summary(cardio_log2)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ gender + age + cigsPerDay + prevalentHyp +
##     totChol + sysBP + glucose, family = "binomial", data = cardiovascular_df2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0153  -0.5990  -0.4292  -0.2841   2.8626
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -8.749106   0.522375 -16.749  < 2e-16 ***
## genderM        0.554155   0.106964   5.181 2.21e-07 ***
## age            0.065531   0.006436  10.182  < 2e-16 ***
## cigsPerDay     0.019375   0.004178   4.637 3.53e-06 ***
## prevalentHyp1  0.234235   0.134858   1.737   0.0824 .
## totChol        0.002239   0.001122   1.996   0.0459 *
## sysBP          0.014257   0.002855   4.993 5.95e-07 ***
## glucose        0.007327   0.001674   4.378 1.20e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3120.5  on 3655  degrees of freedom
## Residual deviance: 2759.5  on 3648  degrees of freedom
## AIC: 2775.5
##
## Number of Fisher Scoring iterations: 5
```

Removed variables without significance from dataset.

```
cardiovascular_df3 <-cardiovascular_df2[, -c(3,4,6,8,11,12,13)]
```

Train and Test a model

```
##install.packages("caTools")
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 4.0.2
```

```
set.seed(9850)
cardio_split <- sample.split(cardiovascular_df3$TenYearCHD, SplitRatio = 0.7)
cardio_train <- subset(cardiovascular_df3, cardio_split == TRUE)
cardio_test <- subset(cardiovascular_df3, cardio_split == FALSE )
```

Create prediction model

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.2
```

```
## Loading required package: lattice
```

```
cardio_predict <- predict(cardio_log2, type = "response", newdata = cardio_test)
```

Determine accuracy of prediction model

```
cardio_acc <- table(cardio_test$TenYearCHD, cardio_predict >= 0.5)
accuracy <- sum(diag(cardio_acc))/sum(cardio_acc) * 100
accuracy
```

```
## [1] 85.32361
```