Project Proposal

2024-04-16

### Using Logistic Regression to Predict Heart Disease

The main aim of this project is to develop a logistic regression model that can be able to predict the posibility of heart disease to an individual. According to the WHO, 12 million people die annually from Heart-related conditions. Furthermore, over half of the total deaths in America are due to cardiovascular diseases. An early identification of such conditions can help patients make important decisions on lifestyle changes to improve their health status. I intend to use this project to highlight the risk factors of heart disease and predict the overall risk using logistic regression.

## Data Explanation

I sourced this data from the Kaggle Website. The data was collected by the World Health Organization from the residents of Framingham, Massachusetts. It has over 4000 subjects and 16 variables.

# Data Dictionary

The data has both demographic, behavioral and medical risk factors.

**Demographic:** • Sex: male or female(Nominal) • Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

**Behavioral:** • Current Smoker: whether or not the patient is a current smoker (Nominal) • Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

**Medical( history):** • BP Meds: whether or not the patient was on blood pressure medication (Nominal) • Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal) • Prevalent Hyp: whether or not the patient was hypertensive (Nominal) • Diabetes: whether or not the patient had diabetes (Nominal)

**Medical(current):** • Tot Chol: total cholesterol level (Continuous) • Sys BP: systolic blood pressure (Continuous) • Dia BP: diastolic blood pressure (Continuous) • BMI: Body Mass Index (Continuous) • Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.) • Glucose: glucose level (Continuous)

**Predict variable (desired target):** • 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

First, let’s load some libraries to be used in the project

library(tidyverse)  
library(dplyr)  
library(cowplot)

## Warning: package 'cowplot' was built under R version 4.3.3

##   
## Attaching package: 'cowplot'

## The following object is masked from 'package:lubridate':  
##   
## stamp

library(pROC)

## Warning: package 'pROC' was built under R version 4.3.3

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

library(caret)

## Warning: package 'caret' was built under R version 4.3.3

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following objects are masked from 'package:yardstick':  
##   
## precision, recall, sensitivity, specificity

## The following object is masked from 'package:purrr':  
##   
## lift

library(janitor)

##   
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

Let me import the data to highlight some of the attributes of the data

framingham <- read.csv("C:/Users/karlm/OneDrive/Documents/STA 631/framingham.csv")

## Data Exploration

head(framingham)

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

cat('Number of Rows:',nrow(framingham),' ','Number of Fields:',ncol(framingham))

## Number of Rows: 4238 Number of Fields: 16

# View the structure of the dataset  
str(framingham)

## 'data.frame': 4238 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 : int 4 2 1 3 3 2 1 2 1 1 ...  
## $ currentSmoker : int 0 0 1 1 1 0 0 1 0 1 ...  
## $ cigsPerDay : int 0 0 20 30 23 0 0 20 0 30 ...  
## $ BPMeds : int 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 : 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 : int 0 0 0 1 0 0 1 0 0 0 ...

# Summary statistics  
summary(framingham)

## male age education currentSmoker   
## Min. :0.0000 Min. :32.00 Min. :1.000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:42.00 1st Qu.:1.000 1st Qu.:0.0000   
## Median :0.0000 Median :49.00 Median :2.000 Median :0.0000   
## Mean :0.4292 Mean :49.58 Mean :1.979 Mean :0.4941   
## 3rd Qu.:1.0000 3rd Qu.:56.00 3rd Qu.:3.000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :70.00 Max. :4.000 Max. :1.0000   
## NA's :105   
## cigsPerDay BPMeds prevalentStroke prevalentHyp   
## Min. : 0.000 Min. :0.00000 Min. :0.000000 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.:0.000000 1st Qu.:0.0000   
## Median : 0.000 Median :0.00000 Median :0.000000 Median :0.0000   
## Mean : 9.003 Mean :0.02963 Mean :0.005899 Mean :0.3105   
## 3rd Qu.:20.000 3rd Qu.:0.00000 3rd Qu.:0.000000 3rd Qu.:1.0000   
## Max. :70.000 Max. :1.00000 Max. :1.000000 Max. :1.0000   
## NA's :29 NA's :53   
## diabetes totChol sysBP diaBP   
## Min. :0.00000 Min. :107.0 Min. : 83.5 Min. : 48.00   
## 1st Qu.:0.00000 1st Qu.:206.0 1st Qu.:117.0 1st Qu.: 75.00   
## Median :0.00000 Median :234.0 Median :128.0 Median : 82.00   
## Mean :0.02572 Mean :236.7 Mean :132.4 Mean : 82.89   
## 3rd Qu.:0.00000 3rd Qu.:263.0 3rd Qu.:144.0 3rd Qu.: 89.88   
## Max. :1.00000 Max. :696.0 Max. :295.0 Max. :142.50   
## NA's :50   
## BMI heartRate glucose TenYearCHD   
## Min. :15.54 Min. : 44.00 Min. : 40.00 Min. :0.000   
## 1st Qu.:23.07 1st Qu.: 68.00 1st Qu.: 71.00 1st Qu.:0.000   
## Median :25.40 Median : 75.00 Median : 78.00 Median :0.000   
## Mean :25.80 Mean : 75.88 Mean : 81.97 Mean :0.152   
## 3rd Qu.:28.04 3rd Qu.: 83.00 3rd Qu.: 87.00 3rd Qu.:0.000   
## Max. :56.80 Max. :143.00 Max. :394.00 Max. :1.000   
## NA's :19 NA's :1 NA's :388

# Check for missing values  
colSums(is.na(framingham))

## male age education currentSmoker cigsPerDay   
## 0 0 105 0 29   
## BPMeds prevalentStroke prevalentHyp diabetes totChol   
## 53 0 0 0 50   
## sysBP diaBP BMI heartRate glucose   
## 0 0 19 1 388   
## TenYearCHD   
## 0

# Check the class distribution of the target variable  
table(framingham$TenYearCHD)

##   
## 0 1   
## 3594 644

janitor::tabyl(framingham$TenYearCHD)

## framingham$TenYearCHD n percent  
## 0 3594 0.8480415  
## 1 644 0.1519585

## Data Preprocessing

First, remove any duplicate observations

framingham <- framingham |>   
 distinct()

Then, Clean Null Observations

colSums(is.na(framingham))

## male age education currentSmoker cigsPerDay   
## 0 0 105 0 29   
## BPMeds prevalentStroke prevalentHyp diabetes totChol   
## 53 0 0 0 50   
## sysBP diaBP BMI heartRate glucose   
## 0 0 19 1 388   
## TenYearCHD   
## 0

# Handle missing values

# Remove rows with missing values  
framingham <- na.omit(framingham)

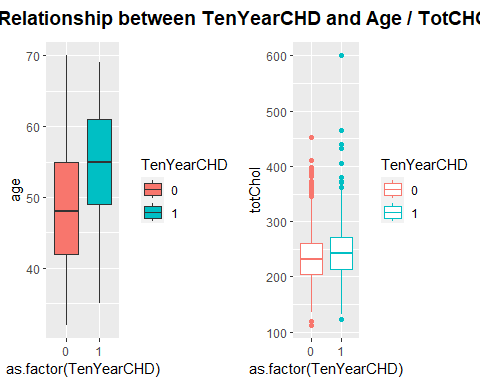
### EDA

# Convert binary variables to characters for better visualization

framingham <- framingham |>   
 mutate(male = as.character(male),  
 currentSmoker = as.character(currentSmoker),  
 prevalentHyp = as.character(prevalentHyp),  
 diabetes = as.character(diabetes),  
 TenYearCHD = as.character(TenYearCHD))

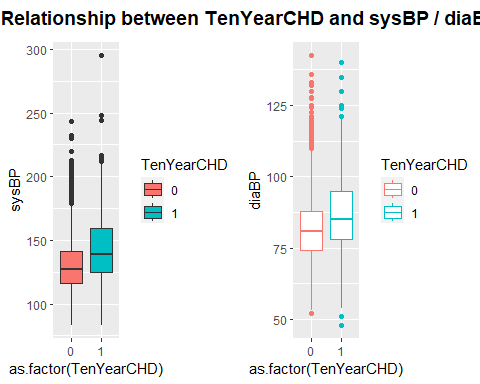
# Visualize the relationship between TenYearCHD and Age/TotCHOL

x <- ggplot(data = framingham, mapping = aes(x = as.factor(TenYearCHD), y = age, fill = TenYearCHD)) +  
 geom\_boxplot()  
y <- ggplot(data = framingham, mapping = aes(x = as.factor(TenYearCHD), y = totChol, color = TenYearCHD)) +  
 geom\_boxplot()  
p <- plot\_grid(x, y)   
title <- ggdraw() + draw\_label("1. Relationship between TenYearCHD and Age / TotCHOL", fontface='bold')  
plot\_grid(title, p, ncol=1, rel\_heights=c(0.1, 1))



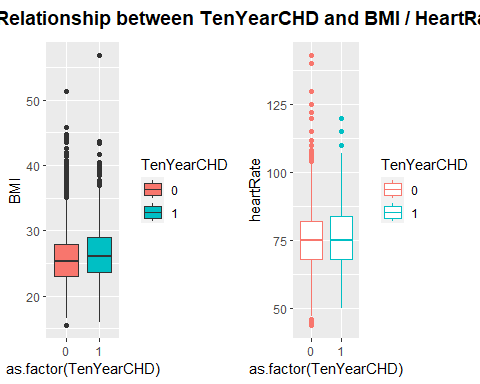
# Visualize the relationship between TenYearCHD and sysBP/diaBP

x <- ggplot(data = framingham, mapping = aes(x = as.factor(TenYearCHD), y = sysBP, fill = TenYearCHD)) +  
 geom\_boxplot()  
y <- ggplot(data = framingham, mapping = aes(x = as.factor(TenYearCHD), y = diaBP, color = TenYearCHD)) +  
 geom\_boxplot()  
p <- plot\_grid(x, y)   
title <- ggdraw() + draw\_label("2. Relationship between TenYearCHD and sysBP / diaBP", fontface='bold')  
plot\_grid(title, p, ncol=1, rel\_heights=c(0.1, 1))



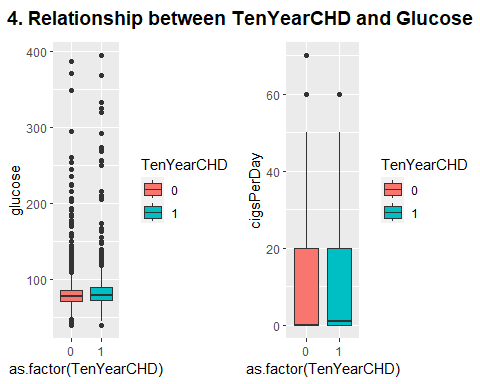
# Visualize the relationship between TenYearCHD and BMI/HeartRate

x <- ggplot(data = framingham, mapping = aes(x = as.factor(TenYearCHD), y = BMI, fill = TenYearCHD)) +  
 geom\_boxplot()  
y <- ggplot(data = framingham, mapping = aes(x = as.factor(TenYearCHD), y = heartRate, color = TenYearCHD)) +  
 geom\_boxplot()  
p <- plot\_grid(x, y)   
title <- ggdraw() + draw\_label("3. Relationship between TenYearCHD and BMI / HeartRate", fontface='bold')  
plot\_grid(title, p, ncol=1, rel\_heights=c(0.1, 1))



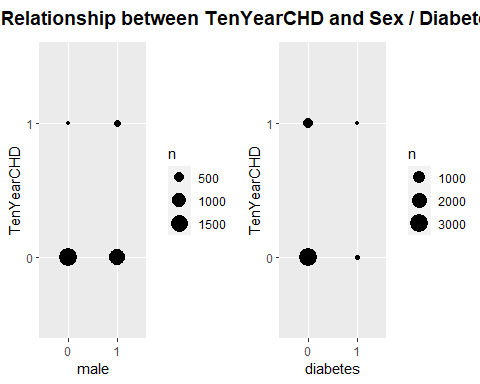
# Visualize the relationship between TenYearCHD and Glucose

x <- ggplot(data = framingham, mapping = aes(x = as.factor(TenYearCHD), y = glucose, fill = TenYearCHD)) +  
 geom\_boxplot()  
y <- ggplot(data = framingham, mapping = aes(x = as.factor(TenYearCHD), y = cigsPerDay, fill = TenYearCHD)) +  
 geom\_boxplot()  
p <- plot\_grid(x,y)   
title <- ggdraw() + draw\_label("4. Relationship between TenYearCHD and Glucose", fontface='bold')  
plot\_grid(title, p, ncol=1, rel\_heights=c(0.1, 1))



# Visualize the relationship between TenYearCHD and Sex/Diabetes

x <- ggplot(data = framingham) +  
 geom\_count(mapping = aes(x = male, y = TenYearCHD))  
y <- ggplot(data = framingham) +  
 geom\_count(mapping = aes(x = diabetes, y = TenYearCHD))  
p <- plot\_grid(x, y)   
title <- ggdraw() + draw\_label("5. Relationship between TenYearCHD and Sex / Diabetes", fontface='bold')  
plot\_grid(title, p, ncol=1, rel\_heights=c(0.1, 1))



## Observations

**Plot 1:** People with CHD have higher mean ages and choloesterol levels

**Plot 2:** People with CHD have higher mean systolic and diastolic blood pressures

**Plot 3:** People with CHD have a higher mean BMI, but very similar mean heart rates as people without CHD

**Plot 4:** People with CHD and without CHD have very similar mean glucose levels

**Plot 5:** More males than females have CHD and more non-diabetics have CHD compared to diabetics. This doesn’t take into account the population proportions.

### Build and Evaluate the Logistic Regression Model

**Build the logistic regression model**

framingham <- framingham |>   
 mutate(TenYearCHD = as.integer(TenYearCHD))

# Split the Data

n\_train <- round(0.7 \* nrow(framingham))  
train\_indices <- sample(1:nrow(framingham), n\_train)  
framingham\_train <- framingham[train\_indices, ]  
framingham\_test <- framingham[-train\_indices, ]

# Train the Model

dim(framingham\_train)

## [1] 2559 16

# Test the Model

dim(framingham\_test)

## [1] 1097 16

log.fit = glm(TenYearCHD ~ ., data=framingham\_train, family=binomial)  
round(summary(log.fit)$coefficients, 2)

## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -8.00 0.85 -9.38 0.00  
## male1 0.59 0.13 4.53 0.00  
## age 0.06 0.01 8.12 0.00  
## education -0.06 0.06 -1.07 0.29  
## currentSmoker1 0.14 0.19 0.73 0.46  
## cigsPerDay 0.01 0.01 1.46 0.14  
## BPMeds 0.31 0.27 1.15 0.25  
## prevalentStroke 0.24 0.63 0.38 0.70  
## prevalentHyp1 0.31 0.16 1.93 0.05  
## diabetes1 -0.05 0.39 -0.14 0.89  
## totChol 0.00 0.00 2.14 0.03  
## sysBP 0.01 0.00 2.74 0.01  
## diaBP -0.01 0.01 -0.89 0.37  
## BMI 0.01 0.02 0.60 0.55  
## heartRate 0.00 0.00 -0.23 0.82  
## glucose 0.01 0.00 2.54 0.01

*From the model, the variables that significantly effect the probability of a patient having CHD are: Sex, Age, Cigs per day, systolic blood pressure, and glucose levels.*

log.probs <- predict(log.fit, type="response")

log.preds = rep("No CHD", 2559)

log.preds[log.probs>0.5] = "CHD"

attach(framingham\_train)

table(log.preds,TenYearCHD)

## TenYearCHD  
## log.preds 0 1  
## CHD 13 21  
## No CHD 2154 371

cat("Training Error Rate:", round(((11+361)/2559)\*100,2),"%")

## Training Error Rate: 14.54 %

*Remove the insignificant variables and retrain the model*

log.fit = glm(TenYearCHD ~ male + age + cigsPerDay + sysBP + glucose, data=framingham\_train, family=binomial)  
round(summary(log.fit)$coefficients, 2)

## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -8.34 0.50 -16.77 0  
## male1 0.54 0.13 4.34 0  
## age 0.07 0.01 9.11 0  
## cigsPerDay 0.01 0.00 3.02 0  
## sysBP 0.02 0.00 6.18 0  
## glucose 0.01 0.00 3.42 0

## Observations

cat("Sex: Being a male increases the odds of having heart disease by", round((exp(0.51)-1)\*100,0), "%", "\n")

## Sex: Being a male increases the odds of having heart disease by 67 %

cat("Age: Every additional year in patient age increases the odds of heart disease by", round((exp(0.06)-1)\*100,0), "%", "\n")

## Age: Every additional year in patient age increases the odds of heart disease by 6 %

cat("Cigarrets Per Day: Every additional 5 daily ciggaret increases the odds of heart disease by", round(5\*(exp(0.02)-1)\*100,0), "%", "\n")

## Cigarrets Per Day: Every additional 5 daily ciggaret increases the odds of heart disease by 10 %

cat("Sys Blood Pressure: Every additional 10 units of Sys Blood Pressure increases the odds of heart disease by", round(10\*(exp(0.02)-1)\*100,0), "%", "\n")

## Sys Blood Pressure: Every additional 10 units of Sys Blood Pressure increases the odds of heart disease by 20 %

cat("Glucose: Every additional 10 units of glucose levels increases the odds of heart disease by", round(10\*(exp(0.01)-1)\*100,0), "%")

## Glucose: Every additional 10 units of glucose levels increases the odds of heart disease by 10 %

log.probs <- predict(log.fit, type="response")

log.preds = rep("No CHD", 2559)

log.preds[log.probs>0.5] = "CHD"

table(log.preds,TenYearCHD)

## TenYearCHD  
## log.preds 0 1  
## CHD 11 19  
## No CHD 2156 373

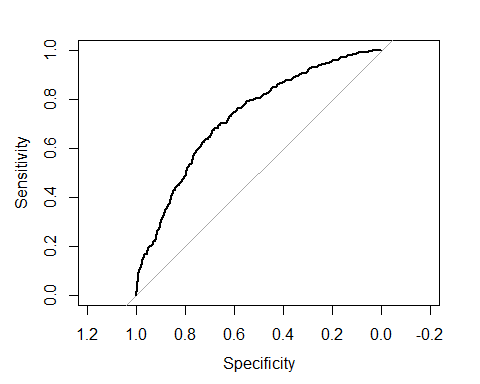
cat("Training Error Rate:", round(((17+362)/2559)\*100,2),"%")

## Training Error Rate: 14.81 %

roc(framingham\_train$TenYearCHD, log.fit$fitted.values, plot=TRUE)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases



##   
## Call:  
## roc.default(response = framingham\_train$TenYearCHD, predictor = log.fit$fitted.values, plot = TRUE)  
##   
## Data: log.fit$fitted.values in 2167 controls (framingham\_train$TenYearCHD 0) < 392 cases (framingham\_train$TenYearCHD 1).  
## Area under the curve: 0.7277

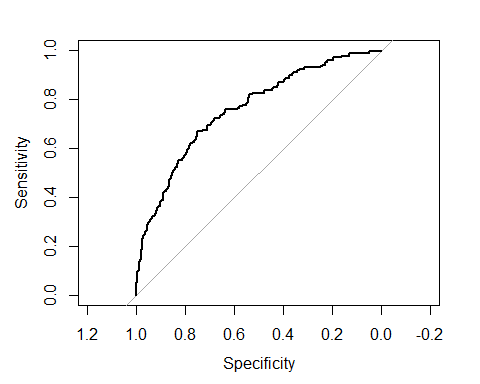
## Test the Model

log.fit = glm(TenYearCHD ~ male + age + cigsPerDay + sysBP + glucose, data=framingham\_test, family=binomial)

roc(framingham\_test$TenYearCHD, log.fit$fitted.values, plot=TRUE)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases



##   
## Call:  
## roc.default(response = framingham\_test$TenYearCHD, predictor = log.fit$fitted.values, plot = TRUE)  
##   
## Data: log.fit$fitted.values in 932 controls (framingham\_test$TenYearCHD 0) < 165 cases (framingham\_test$TenYearCHD 1).  
## Area under the curve: 0.7582

## Logistic regression equation

P=eβ0+β1X1/1+eβ0+β1X1P=eβ0+β1X1/1+eβ0+β1X1

When all variables are included:

logit(p)=log(p/(1−p))=β0+β1∗Sexmale+β2∗age+β3∗cigsPerDay+β4∗totChol+β5∗sysBP+β6∗glucoselogit(p)=log(p/(1−p))=β0+β1∗Sexmale+β2∗age+β3∗cigsPerDay+β4∗totChol+β5∗sysBP+β6∗glucose

**Interpreting the results: Odds Ratio, Confidence Intervals and P-values**

• This fitted model shows that, holding all other features constant, the odds of getting diagnosed with heart disease for males (sex\_male = 1)over that of females (sex\_male = 0) is exp(0.5815) = 1.788687. In terms of percent change, we can say that the odds for males are 78.8% higher than the odds for females.

• The coefficient for age says that, holding all others constant, we will see 7% increase in the odds of getting diagnosed with CDH for a one year increase in age since exp(0.0655) = 1.067644.

• Similarly , with every extra cigarette one smokes thers is a 2% increase in the odds of CDH.

• For Total cholesterol level and glucose level there is no significant change.

• There is a 1.7% increase in odds for every unit increase in systolic Blood Pressure.

**Model Evaluation - Statistics**

From the above statistics it is clear that the model is highly specific than sensitive. The negative values are predicted more accurately than the positives.

Predicted probabilities of 0 (No Coronary Heart Disease) and 1 ( Coronary Heart Disease: Yes) for the test data with a default classification threshold of 0.5

Lower the threshold

Since the model is predicting Heart disease too many type II errors is not advisable. A False Negative ( ignoring the probability of disease when there actually is one) is more dangerous than a False Positive in this case. Hence in order to increase the sensitivity, threshold can be lowered.

## Conclusions

• All attributes selected after the elimination process show P-values lower than 5% and thereby suggesting significant role in the Heart disease prediction.

• Men seem to be more susceptible to heart disease than women. Increase in age, number of cigarettes smoked per day and systolic Blood Pressure also show increasing odds of having heart disease

• Total cholesterol shows no significant change in the odds of CHD. This could be due to the presence of ’good cholesterol(HDL) in the total cholesterol reading. Glucose too causes a very negligible change in odds (0.2%)

• The model predicted with 0.88 accuracy. The model is more specific than sensitive. Overall model could be improved with more data

**Summary of the model**

summary(framingham)

## male age education currentSmoker   
## Length:3656 Min. :32.00 Min. :1.00 Length:3656   
## Class :character 1st Qu.:42.00 1st Qu.:1.00 Class :character   
## Mode :character Median :49.00 Median :2.00 Mode :character   
## Mean :49.56 Mean :1.98   
## 3rd Qu.:56.00 3rd Qu.:3.00   
## Max. :70.00 Max. :4.00   
## cigsPerDay BPMeds prevalentStroke prevalentHyp   
## Min. : 0.000 Min. :0.00000 Min. :0.000000 Length:3656   
## 1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.:0.000000 Class :character   
## Median : 0.000 Median :0.00000 Median :0.000000 Mode :character   
## Mean : 9.022 Mean :0.03036 Mean :0.005744   
## 3rd Qu.:20.000 3rd Qu.:0.00000 3rd Qu.:0.000000   
## Max. :70.000 Max. :1.00000 Max. :1.000000   
## diabetes totChol sysBP diaBP   
## Length:3656 Min. :113.0 Min. : 83.5 Min. : 48.00   
## Class :character 1st Qu.:206.0 1st Qu.:117.0 1st Qu.: 75.00   
## Mode :character Median :234.0 Median :128.0 Median : 82.00   
## Mean :236.9 Mean :132.4 Mean : 82.91   
## 3rd Qu.:263.2 3rd Qu.:144.0 3rd Qu.: 90.00   
## Max. :600.0 Max. :295.0 Max. :142.50   
## BMI heartRate glucose TenYearCHD   
## Min. :15.54 Min. : 44.00 Min. : 40.00 Min. :0.0000   
## 1st Qu.:23.08 1st Qu.: 68.00 1st Qu.: 71.00 1st Qu.:0.0000   
## Median :25.38 Median : 75.00 Median : 78.00 Median :0.0000   
## Mean :25.78 Mean : 75.73 Mean : 81.86 Mean :0.1524   
## 3rd Qu.:28.04 3rd Qu.: 82.00 3rd Qu.: 87.00 3rd Qu.:0.0000   
## Max. :56.80 Max. :143.00 Max. :394.00 Max. :1.0000

## Interpretation and Further Analysis

**The training AUC value of 0.73 and the test AUC of 0.74 suggest that the logistic regression model does a fair job of predicting whether a patient has CHD. The significant variables are the sex of the patient, age, cigaretts per day, systolic blood pressure, and glucose levels. The effects of these variables are as follows:**

**Sex:** The sex of the patient has the biggest effect. Being a male increases the odds of having heart disease by 67 % compared to being a female. This corresponds with the graph that shows a larger amount of people with heart disease are male.

**Age:** Every additional year in patient age increases the odds of heart disease by 6 %. This corresponds with the boxplots that show people with heart disease have much higher mean ages.

**Cigarettes Per Day:** Every additional 5 daily cigarette increases the odds of heart disease by 10 %.

**Systolic Blood Pressure:** Every additional 10 units of Systolic Blood Pressure increases the odds of heart disease by 20 %. This corresponds with the boxplots that show people with heart disease have much higher mean systolic blood pressure.

**Glucose:** Every additional 10 units of glucose levels increases the odds of heart disease by 10 %.