CORONARY HEART DISEASE PREDICTION USING LOGISTIC REGRESSION

COURSE NAME: STATISTICAL METHODS

COURSE CODE: QMST 5334

YEAR: FALL 2021

TEAM MEMBERS:

ANJALI GOEL

AKSHITHA KARTHICK KUMAR

HARSHAL Y SHELARE

INDU DHULIPALA

NIHARIKA REDDY KOTA

PROBLEM STATEMENT:

It has been estimated by World Health Organization that there are at least 12 million deaths occurring worldwide every year due to heart diseases. Half the deaths in the United States and other developed countries are due to cardiovascular diseases. The primary objective of the research is to find out the most relevant/risk factors of heart disease as well as predict the overall risk of having heart disease in the next 10 years.

The dataset available on the Kaggle website is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD), which is the dependent variable in this study. From the dataset, variables like age, sex, number of cigarettes smoked per day, and systolic Blood Pressure are some of the key independent variables of interest as they are found to be potential risk factors of heart disease from WHO's research study.
https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity, and harmful use of alcohol. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. We are using R to perform Logistic Regression in our case study.

Contents

1 Executive summary

To predict whether the patient has a 10-year risk of future coronary heart disease (CHD), we have taken the data with 4238 data rows in csv file. The dataset contains different information about a patient's sex, age, education, current smoker, cigarettes per day, BP medication, general stroke, prevalent hypertension, BMI, heart rate, glucose, diabetes, and 10-year risk of coronary heart diseases. Here we have considered 10-year risk of future coronary heart disease (CHD) as the dependent variable and variables like age, sex, number of cigarettes smoked per day, and systolic Blood Pressure are some of the key independent variables of interest as they are found to be potential risk factors of heart disease from WHO's research study. After building logistic regression we have performed Hypothesis test 1 to calculate p-value and Hypothesis test 2 to evaluate the simple and complicated models. Finally, with a confusion matrix we have tried understanding the impact of each risk factor on the variable to be predicted in each patient.

2 Context and overview

The objective of the research is to find out the most relevant/risk factors of coronary heart disease (CHD) as well as predict the overall risk of having heart disease in the next 10 years with a dataset involving a subset of population from the town of Framingham, Massachusetts. This is achieved using a Logistic Regression model explained in the paper.

Logistic regression is a type of regression analysis in statistics used for the prediction of the outcome of a categorical dependent variable from a set of predictor or independent variables. In logistic regression, the dependent variable is always binary. Logistic regression is primarily used for prediction and calculating the probability of success. The backward elimination approach or the P-value approach will be used to eliminate attributes out of the 15 independent variables in the given dataset and that have the least impact on the regression test. Thereby evaluating the statistics of only the potential risk factors causing heart disease. These findings are then interpreted using Confusion Matrix to understand the impact of each risk factor on the variable to be predicted i.e.., the 10-year risk of coronary heart disease in each patient.

3 Analysis of descriptive statistics and outliers

The data relating to heart disease is acquired from the Kaggle. The dataset contains information about a patient's sex, age, education, current smoker, cigarettes per day, BP medication, general stroke, prevalent hypertension, BMI, heart rate, glucose, diabetes, and 10-year risk of coronary heart diseases.

According to National Heart, Lung, and Blood Institute, coronary heart disease, also known as CHD, is a leading cause of death in the United States. The risk of coronary heart disease increases with the number of risk factors one has and how serious they are. Some risk factors are high blood pressure and high cholesterol, and other risk factors such as sex, age, medical history cannot be changed. We can use these risk factors from our dataset and predict if the patient has a higher risk of coronary heart disease.

The plot of the columns with outliers is shown in the figure 9.1.6. If we look at the BMI column, the data interquartile ranges from 20 - 30 with many outliers with BMI 35 and above. This means that there are outlier patients who have higher BMI, which affects determining coronary heart disease.

For column heartRate, the interquartile range is 40 - 105. We also observe that there are lower and upper outliers. Here, the outliers show that the patients have a higher heart rate, which means the heart is under much pressure, affecting coronary heart disease. The next column is total cholesterol; the interquartile range is between 100 - 350 with a median of 250. We have many outliers, which reach 700. The glucose column has the most outliers in the data, and the median lies at 75, and the outlier's range to 400. The column ‘sysBP’ has an interquartile range 50 - 275, and outliers' range to 300. The column ‘diaBP’ has an interquartile range of 50 - 110, and outliers' range to 140.

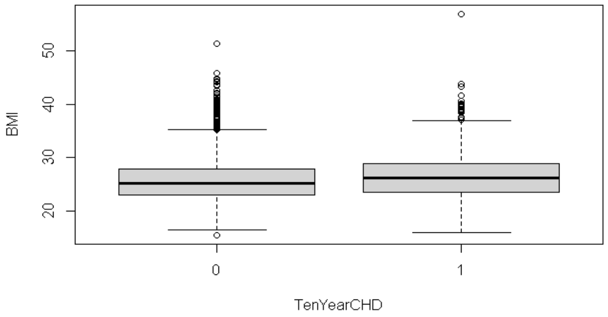
With the understanding of columns with outliers, we will visualize different columns to understand their relation. With heartRate and currentSmoker columns in the figure 9.1.1, patients who do not smoke have more outliers and have less interquartile range. Both smokers and not smokers have the equivalent median. With the male and age columns in figure 9.1.5, the female median is higher than males', and the range of ages of females is broader than that of a male. With age and currentSmoker columns in figure 9.1.2, the median of non-smokers concerning age is more significant than smokers, and the non-smoker has a broader interquartile range. For diabetes and totChol columns in figure 9.1.4, the interquartile range of patients with diabetes concerning cholesterol is low. However, patients without diabetes have more outliers in the column. For columns - totChol and currentSmoker in figure 9.1.3, both smokers and non-smokers have similar plots containing outliers. Lastly, for columns male and totChol in figure 9.1.7, females have a broader interquartile range than males with the same outliers.

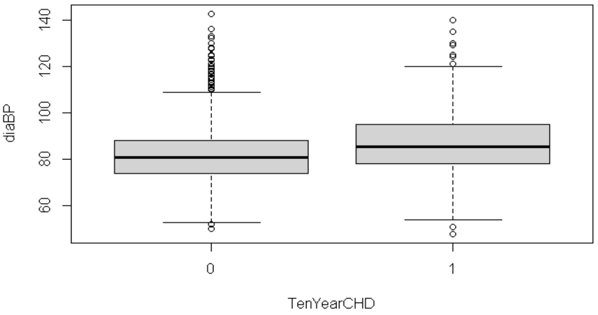
Variable	Type
Sex (male = 1 = Male & male = 0 = Female)	Nominal
Age	Continuous
currentSmoker	Nominal
cigsPerDay	Continuous
BPMeds	Nominal
prevalentStroke	Nominal
prevalentHyp	Nominal
diabetes	Nominal
totChol	Continuous
sysBP	Continuous
diaBP	Continuous
BMI	Continuous

heartRate	Continuous
glucose	Continuous

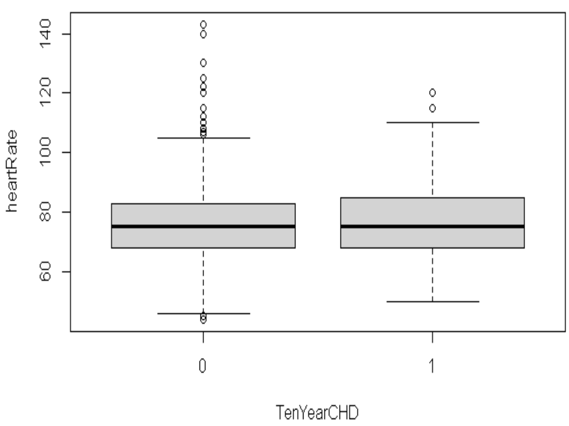
4 Bivariate Analysis

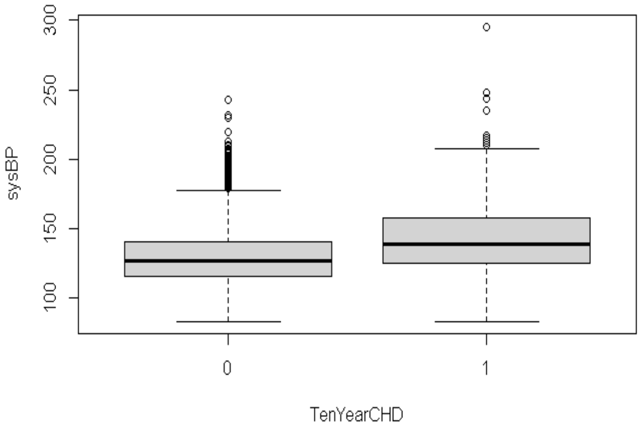
The selection of independent variables for building the regression model is based on the bivariate analysis shown below, which shows how these variables affect the variability of the dependent variable ‘TenYearCHD’



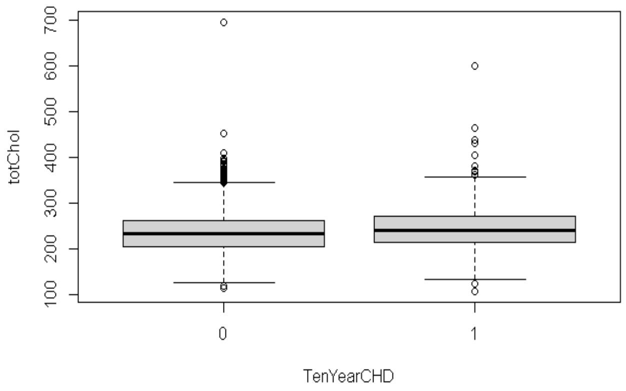


- The median BMI is a little higher in people with ten-year CHD risk than those who do not have ten-year CHD risk.
- The diastrophic BP is higher in people with ten-year than those without the risk of ten-year CHD





- The number of people ranging from bottom to up in category 1 of TenYearCHD have higher average heart rate than people in category 0.
- Most people with CHD risk have higher systrophic BP than people without the risk.



- Total cholesterol in people with risk of CHD is higher than those without the risk of CHD.

5 Model building and variable transformations

Now we have followed multiple steps to build logistic regression model to predict whether the patient has a 10-year risk of future coronary heart disease (CHD), which is the dependent variable in this study.

Assumptions

- We have dropped the education column as it was not playing any key role as a risk factor to predict the outcome of disease in our model.
- Threshold value (alpha) is set to be 0.5 in our model as it is morally neutral
- We have assumed 10-year risk of future coronary heart disease (CHD) as the dependent variable and variables like age, sex, number of cigarettes smoked per day, and systolic Blood Pressure as the key independent variables of interest as they are found to be potential risk factors of heart disease from WHO's research study.

Step 1: Data clean-up and transformation –

- **To check and remove missing values and insignificant variables:** The dataset has 3656 complete rows, and 582 incomplete rows. For data clean up, the column 'education' has been dropped and the rows with NULL values have been removed. After cleaning up the dataset, we get 3749 complete rows.
- **Converting data type of categorical variables into factors:** As categorical variables are binary values, to use them in Logistic regression modelling, we are converting 0s and 1s in int/string format into Factor data type.

Step 2: Building Logistic Regression model:

- **Data Sampling for Regression Modelling**: The cleaned dataset into training dataset and testing dataset to build and test Regression model. The ratio of Training: Testing data is 70:30 where 2624 is the number of training dataset rows obtained and 1125 is the testing dataset.

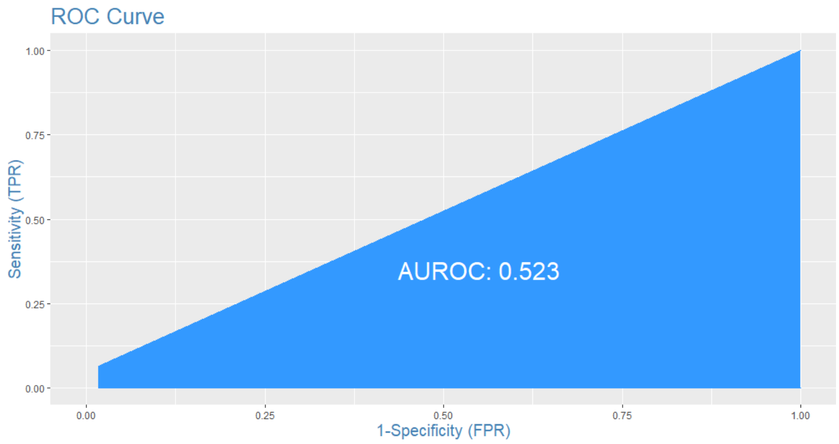
- **Modelling of Logistic Regression Model**: We have used the function 'glm' in R on the training dataset to build the model. As logistic regression works on binomially distributed variables, we have used appropriate class in the function.
- **Hypothesis Test 1**: In hypothesis test 1,
    1. Null hypothesis H_0 is stated as the fit of the model with independent variables is as good with the fit of model without independent variable.
    2. Alternate hypothesis H_a is stated as the fit of the independent variable of model is better than the model without independent variable.
    3. From the summary of the logistic regression model, we have calculated p-value using the expression 1-pchisq () and it has been compared with the threshold value (0.5).
    4. Here in this model, we found that Residual deviance is less than the Null deviance. So, the model fits better p -values are less than the threshold value so we are keeping the H_0 and rejecting the Ha.
- **Hypothesis Test 2**: In the Hypothesis Test 2,
    1. Null hypothesis H0 is stated as simple and complicated has the same fit to the data.
    2. Alternate hypothesis Ha is stated as the complicated model has better fit than the simple model. Simple model as 'model2' and Complicated model as 'model3'
    3. Here we have used independent variables totChol, sysBP, diabetes, age, currentSmoker, male, prevalentStroke, and diaBP in the complicated model to estimate the simple and complicated model performances.
    4. We have used anova() function to compare and analyze variance levels in models 2 and 3.
    5. As result from anova(), the p-value is very low (less than the threshold value). SO, we are rejecting the H0 and accepting the Ha. And hence, determining that the more complicated model is the better fit.

6 Final Model

The best model from the above hypothesis tests is model3, which will be used further for predicting the risk of having CHD in the patients in the test dataset. We have followed the steps below for estimating predictions:

Step 1: Using predict method on the test dataset:

- We have used the test dataset on the model built using training dataset to estimate predicted values of the dependent variable.
- An ROC (Receiver Operating Characteristic) curve is plotted to measure the performance of the model and an Area Under Curve (AUC) is obtained for the same.
- AUC ranging from 0 to 1 where if AUC is 0, it means that the predictions are 100% wrong and if AUC is 1, it means that the predictions are 100% true.
- In the designed model for test_heart_data Vs TenYearCHD, the deduced AUC is 0.523 which represents that the model fairly predicts the positive class.



Step 2: Estimating accuracy for the predicted model:

- The predicted model obtained from the research is tested for accuracy by calculating its sensitivity, sensitivity, and concordance.
- We then calculate the optimum cut-off from the formula (sensitivity + specificity) - 1 and use the value obtained as the threshold value of the model.
- The optimum cut-off value which is 0.04 in this case is then used to construct a Confusion Matrix, where the columns represent actuals, while rows represent predicted values.
- Finally, we obtain the percentage accuracy of the overall model as 84.27% and accuracy for the predicted values as 93.60%.

7 Inferences

- From Hypothesis Test 1, we are accepting the Null hypothesis H_0, which states that the fit of the model with independent variables is as good as the fit of model without independent variable. Hence, the model with independent variables is being selected for further analysis.
- From Hypothesis Test 2, we are accepting the alternate hypothesis H_a, which determines that the more complicated model is the better fit model for this regression. This indicates that the dependent variable TenYearCHD is highly dependent on many independent variables or factors.
- We have an AUROC = 0.523 which moderately greater than the initial threshold 0.5. This AUC value is classified as positive and indicates that the model will generate predictions that has more than 52.3% correct results.
- There is a high specificity of 0.983 in the research which indicates that there are only a few false positive results in the prediction, this will in turn avoid unnecessary screening of people who do not have the risk of heart disease.
- However, the sensitivity is low at about 0.06 which indicates that there can be more false negative results. This can happen in most clinical studies and is usually adjusted with some trade-offs.
- There is an overall accuracy of 84.27% in the model, indicating that the independent variables chosen from the patient's medical history i.e., totChol, sysBP, diabetes, age, currentSmoker, sex, prevalentStroke, and diaBP are substantial risk factors for heart disease.

- Also, an accuracy of 93.60% has been achieved in the predictions, which indicates that the model has rightly predicted the number of people having the risk of the heart disease in the next ten years.
- Moreover, the performance and accuracy of predictions gradually reduce below 85% when the independent variables age, sex, or currentSmoker is not included in the regression model. This means that these independent variables have a significant impact on the dependent variable thereby affecting the risk of having heart disease in the next ten years.

8 Conclusion

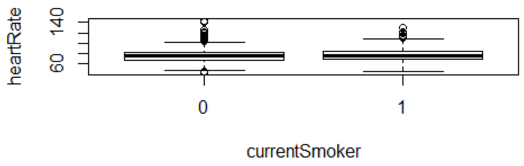
To summarize the above experiment, we have found the potential risk factors of coronary heart disease and have predicted the number of people who are at risk of getting the disease in the next ten years using Logistic Regression. After analyzing independent factors in the dataset, several regression models were built, and the best fit model was then identified from the hypothesis tests. This model was used on the training dataset to generate predictions. From the test results, we have an accuracy of 95.25% on predictions for those who have 10-year risk of coronary heart disease with an optimal threshold 0.05, found using the amount of sensitivity and specificity of the prediction model.

Similarly, there are so many usages of logistic regression modelling in other medical areas like in “Medical Appointment No Shows” to improve the online medical appointment process which helps a lot of actual patients to get appointment.

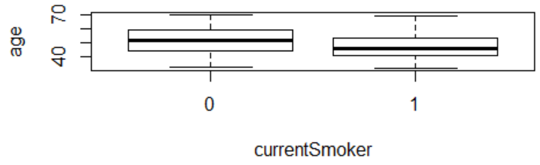
9 Appendix

9.1 Plots

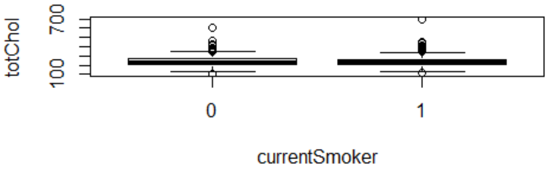
9.1.1 Fig 9.1.1:



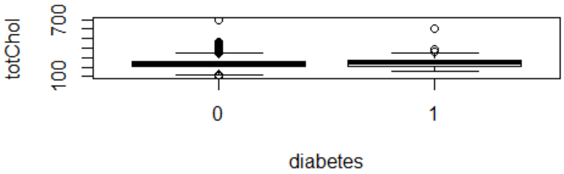
9.1.2 Fig 9.1.2:



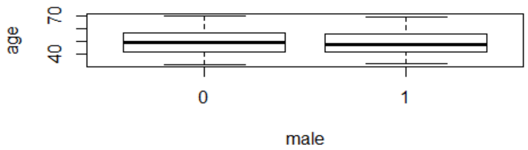
9.1.3 Fig 9.1.3:



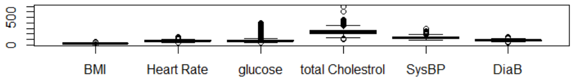
9.1.4 Fig 9.1.4:



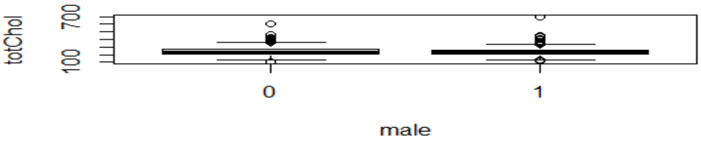
9.1.5 Fig 9.1.5:



9.1.6 Fig 9.1.6:



9.1.7 Fig 9.1.7:



9.2 R Code

```
#Loading Libraries for this project

install.packages("caTools")

library(tidyverse) # metapackage with lots of helpful functions

library(ggplot2) # Data visualization

library(plotly) # Interactive data visualizations

library(psych) # Will be used for correlation visualizations

library(rattle) # Graphing decision trees
```

CORONARY HEART DISEASE PREDICTION USING LOGISTIC REGRESSION

```
library(caret) # Machine learning
library(caTools)
library(InformationValue)

#Loading the dataset into R environment
heart_rel_data <-
read.csv("/Users/apoor/Desktop/Anjali/MSDAIS/Stats/Project/framingham.csv",header=TRUE,sep=",")
#Attaching the data with headers
attach(heart_rel_data)

#Data Exploration
str(heart_rel_data)

paste("There are",nrow(heart_rel_data),"data rows")
paste("There are",sum(complete.cases(heart_rel_data)),"complete data rows")
paste("There are",sum(!complete.cases(heart_rel_data)),"incomplete data rows")

#Visualization with Outliers
boxplot(totChol)
boxplot(BMI)
boxplot(heartRate)
boxplot(glucose)

#Correlation----------Explanation needed
boxplot(heartRate~currentSmoker)
boxplot(age~male)
boxplot(age ~ currentSmoker)
boxplot(age ~ diabetes)

#Summary of entire data
summary(heart_rel_data)

#Checking NA values in each column
colSums(is.na(heart_rel_data))

#Removing education column
heart_rel_data_clean <- subset(heart_rel_data, select = -c(education))
heart_rel_data_clean
```

```
#Removing NA values
heart_data<-(na.omit(heart_rel_data_clean))
heart_data
#Checking NA values in each column
colSums(is.na(heart_data))

#Converting categorical variables(data types) into factors
heart_data$male <- as.factor(heart_data$male)
heart_data$currentSmoker <- as.factor(heart_data$currentSmoker)
heart_data$BPMeds <- as.factor(heart_data$BPMeds)
heart_data$prevalentStroke <- as.factor(heart_data$prevalentStroke)
heart_data$prevalentHyp <- as.factor(heart_data$prevalentHyp)
heart_data$diabetes <- as.factor(heart_data$diabetes)
heart_data$TenYearCHD <- as.factor(heart_data$TenYearCHD)

#Checking data types after conversion to factors
str(heart_data)

#Taking Training dataset and Testing dataset
set.seed(30)
heart_datasplit <- sample.split(heart_data$TenYearCHD, SplitRatio = 0.7)
training_heart_data <- heart_data[heart_datasplit==T,]
testing_heart_data <- heart_data[heart_datasplit==F,]
nrow(training_heart_data)
nrow(testing_heart_data)

 #Creation of Logistic regression model
?glm()
logistic_model <- glm(heart_data$TenYearCHD~., data = training_heart_data, family = "binomial")
summary(logistic_model)

#Threshold value(alpha) is set to be 0.5 in our model as it is morally neutral
#Hypothesis Test 1
#H0: The fit of the model with independent variables is as good with the fit of model without independent variable
#Ha: The fit of the independent variable of model is better than the model without independent variable
#p-value calculation
1-pchisq(2240.5 - 1958.9,2623-2609 )
```

```
#Here in this model, we can see that Residual deviance is less than the Null deviance. So, the model fits better
#p values is less than the threshold value so we are keeping the H0 and rejecting the Ha

#Hypothesis Test 2
#Identifying Simple and Complicated model
model2<-glm(training_heart_data$TenYearCHD~training_heart_data$totChol+training_heart_data$sysBP,family=binomial(link="logit"))
summary(model2)
#Null deviance: 3202.8  on 3748  degrees of freedom
#Residual deviance: 3031.2  on 3746  degrees of freedom

model3<-glm(training_heart_data$TenYearCHD~training_heart_data$totChol+training_heart_data$sysBP+
training_heart_data$diabetes+training_heart_data$age+training_heart_data$currentSmoker+training_heart_data$male+training_heart_data$prevalentStroke+ training_heart_data$diaBP,family=binomial(link="logit"))
summary(model3)
#Null deviance: 3202.8  on 3748  degrees of freedom
#Residual deviance: 2852.6  on 3740  degrees of freedom

#Now Null deviance's DOF is matching with model2 null deviance's DOF. So, we'll compare residual deviance
anova(model2,model3,test="Chisq")

#Since the p-value is very low(less than the threshold value). SO, we are rejecting the H0 and accepting the Ha.
#And hence, determing that the more complicated model is better fit.
pred_model <- predict(model3,newdata = testing_heart_data,type = "response")
summary(pred_model)
pred_test_data <- ifelse(pred_model>0.5,1,0)
summary(pred_test_data)

# The columns are actuals, while rows are predicteds.
#predicted=predict(model3,type="response")
plotROC(testing_heart_data$TenYearCHD, pred_test_data)
#AUROC: 0.523
 Concordance(testing_heart_data$TenYearCHD, pred_test_data)
# The columns are actuals, while rows are predicteds.
sensitivity(testing_heart_data$TenYearCHD, pred_test_data, threshold = 0.5)
#[1] 0.06395349
specificity(testing_heart_data$TenYearCHD, pred_test_data, threshold = 0.5)
```

```
# [1] 0.9832109

confusionMatrix(testing_heart_data$TenYearCHD, pred_test_data, threshold = 0.5)

#0   1

#0 937 161

#1  16  11

#Though we have accuracy as 84.82262 % but this model is not able to predict people who have the risk of heart disease correctly

(937+11)/(161+16+937+11)

#Overall Accuracy with threshold as 0.5 as 84.27%

#So, we ll take optimum cut off value using sensitivity and specificity

opt_cut_off =  (  0.06+0.9832109)-1

opt_cut_off

#[1] 0.0432109

confusionMatrix(testing_heart_data$TenYearCHD, pred_test_data, threshold = opt_cut_off)

(161/(161+11))

#Accuracy for predicted values with optimum cut off as 0.04 is 93.60%
```