Heart Disease Prediction using Logistic Regression

Harshavardhan Ginkutla

4/27/2020

## Introduction

For a person, in order to stay healthy, its one of the important fact and good to know about is the working condition of heart. In th United States, around 647,000 people die because of cardiovascular diseases. That’s 1 in 4 people in the United States die due to heart disease every year. (www.cdc.gov)

Ofcourse, sometimes its not in our hand to always keep heart healthy and survive long, but its possible for us to predict the occurence of heart disease for the future with the help of indicators or symptoms that effect the disease.

In this research, the objective is to predict the occurrence of heart disease using a Machine Learning algorithm that helps to predict heart disease based on the past data. The past data is a set or collection of symptoms that affect the heart disease along with a dichotomous that is whether the heart disease is positve or negative belonging to the respective symptoms.

According to the observed scenario, a Shiny App (web-based interactive application) in R is developed in this study in order to assist a common person to predict the heart diease based on the past symptoms of the affected persons. In addition to the prediction of heart disease in R application, one more objective of this study is also to calculate the accuracy of the utilized model. In this study, a most popularly used machine learning algorithm called ‘Logistic Regression’ is used. There are several algorithms available in the market that can be utilized for prediction purpose, its upto us in order to choose which algorithm can tell us how much most accurate and each algorithm performance differs based on the data given.

For the system to learn about past symptoms and make predictions, a famous heart disease dataset from UCI (University of Calfornia Irvine) was used. This dataset contains the medical indicators that affect the heart disease. This dataset consists of heart disease history of mostly Switzerland and Hungarian origin. The dataset contains 76 attributes (indicators) originally but we are using 14 attributes which are more important in prediction of the heart disease in general.

The dataset is downloaded from “<https://www.kaggle.com/ronitf/heart-disease-uci>”

[Link](https://www.kaggle.com/ronitf/heart-disease-uci)

## Literature Survey

According to a journal published at Expert Systems with Applications in 2011, “A Predictive model for cerebrovascular disease using data mining Science, Vol. 8970- 8977, 2011.”, the research was found out that the Cerebrovascualar disase has been ranked the second or third of top 10 death causes in Taiwan and has caused about 13,00 death every year since 1986. In view of this, the future prediction of disease occurrence was convceived in the health and medi-care industry. They started to use machine learning algorithms to predict the occurrence of all kinds of health diseases like Cardiovacular (relating to heart and blood vessels), Cerebrovascular (disease that affect the blood vessels and blood supply to the brain). In the journal, they have researched about the utilization and performance of the supervised learning algorithms that helped in prediction. The classificaiton algorithms like Decision Tree, Bayesian Classifier, Back Propogation Neural Network were utilized and those models sensitivity and accuracy were 98.48% and 99.59% on average respectively. That’s an outstanding accuracy. In addition to these, they are trying to reduce the factors that affect the prediction keeping in mind that by reducing so can also predict the same with good accuracy. In view of this, prediction of heart disease by utilizing a supervised machine learning to predict and find out the accuracy of the model is being performed in this research.

## Theory

Hypothesis: A supervised learning algorithm called Logistic Regression can be used for prediction of the heart disease along with its probability.

## Data

In this systems, there has to be two datasets that should be uploaded, one is the training data set in which it contains the records of observations (means data about the heart disease, which we have got from Kaggle.com, and is a dataset of UCI) and the other is the testing dataset in which only one data record of same columns as in training dataset will be present.

Before going into uploading data, we will have a look over attributes and explain what those attributes are actually mean,

The dataset has 14 columns which means, 13 attributes are the independent variables and one column which is located at last is the dependent variable. We will discuss about what the indepedent variabe, and the dependent variable in Methodology section. But now we will know the meaning of the attributes here that are tagged to some outcome

1. **age** : patient’s age in year
2. **sex** : 1 = Male, 0 = Female
3. **cp** : type of chest pain

* 1 = Typical Angina (chest pain by physical efforts or emotional stress)  
   2 = Atypical Angina (when heart doesn't get enough oxygen, thus results in tightness in the central chest)  
   3 = Non-Anginal pain (when the chest pain duration is more than 30 minutes, it can affect the trunk or arm of the human body)  
   4 = Asymptotatic (No cause or chest pain)

1. **trestbps** : It is the resting blood pressure in mm Hg (120 mm-Hg is systolic is considered normal generally)
2. **chol** : Total cholesterol levels in the blood including High-Density Lipoprotein (good cholesterol) and Low-Density Lipoprotein (bad cholesterol) in mg/dL (milli-grams/deci-Litre).

* Desirable : <200mg/dL (for adults)  
   Borderlne : 200 - 239 mg/dL  
   High : >240mg/dL

1. **fbs** : If fasting blood sugar level > 120mg/dL, the 1 = true and 0 = false
2. **restecg** : Resting electorcardiographic measurement

* 0 = normal  
   1,2 = some abnormality

1. **thalach** : The maximum heart rate achieved by the patients in bpm (beats per minute)
2. **exang** : Exercise induced angina (means the discomforts acquired due to physical activities)

* 1 = Yes; 0 = No

1. **oldpeak** : ST depression (means less blood flow through out the body parts) during exercises
2. **slope** : The slope visualized in the abnormalities in the heart during ST depression at peak exercises

* 1 = Upsloping  
   2 = Flat (Normal)   
   3 = Downsloping

1. **ca** : Number of major blood vessels observed during fluroscopy (study of moving body structures similar to X-Ray such that body parts and motion is observed in the monitor)
2. **thal** : Thalassemina(on of the blood disorder due to decreased heamoglobin)

* Normal = upto 3,  
   Fixed defects = upto 6,  
   Reversible defects = 7

1. **target** : Presence of heart disease

* 1 = Yes,  
   0 = No

Once we got to know upto the data attributes, now we are ready to upload the data…  
*Training Data*

## age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal  
## 1 63 1 3 145 250 1 0 150 0 2.3 0 0 1  
## 2 37 1 2 130 250 0 1 187 0 3.5 0 0 2  
## 3 41 0 1 130 204 0 0 172 0 1.4 2 0 2  
## 4 56 1 1 120 236 0 1 178 0 0.8 2 0 2  
## 5 57 0 0 120 354 0 1 163 1 0.6 2 0 2  
## 6 57 1 0 140 192 0 1 148 0 0.4 1 0 1  
## target  
## 1 1  
## 2 1  
## 3 1  
## 4 1  
## 5 1  
## 6 1

## [1] "Number of rows: 303"

*Testing Data*

## age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal  
## 1 71 0 0 120 250 1 1 162 0 0.4 2 2 2  
## target  
## 1 NA

## [1] "Number of rows: 1"

*Cleaned Data (After Data Preprocessing)*

As we know before taking data into consideration, its important to clean it. The data is cleaned and the system will also omit records containing null values. Well the dataset is actually cleaned and in dignified manner and as it is UCI data and it might not need data cleansing but still the data is subjected to pre-process here.

## age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal  
## 2 37 1 2 130 250 0 1 187 0 3.5 0 0 2  
## 3 41 0 1 130 204 0 0 172 0 1.4 2 0 2  
## 4 56 1 1 120 236 0 1 178 0 0.8 2 0 2  
## 5 57 0 0 120 354 0 1 163 1 0.6 2 0 2  
## 6 57 1 0 140 192 0 1 148 0 0.4 1 0 1  
## 7 56 0 1 140 294 0 0 153 0 1.3 1 0 2  
## target  
## 2 1  
## 3 1  
## 4 1  
## 5 1  
## 6 1  
## 7 1

## Methodology

Once we have got the cleaned data, now the machine learning algorithm will be applied. One of the purpose of the learning algorithm is to get the proposed model well-trained. As it is a supervised learning algorithm, the model will be trained based on the given training dataset. In this case, the training dataset will be the so and so discussed heart dataset from UCI. The machine learning algorithm will always have some terms like ‘Independent Variable’ and ‘Dependent Variable’. The idependent variables are called the factors/indicators that affect the dependent variable. As we know, the training and testing data contains of 14 columns in which the first 13 columns are considered as independent variables here and the last column called ‘target’ will be considered as the dependent variable. In other words, first the system will learn or get trained by the records from the training dataset and then it will estimate the probability of the outcome based on the inference.

Before doing so, the data has to be divided into two parts called ‘Training’ and ‘Testing’. Although we have the testing data seperately in another CSV file, but still by dividing the training data, R will randomly choose the records required for training and divide them according to the splitting percentage given in the code such that whenever the code is run for different times, different trandom records will be selected at each time which can result in better training of the model.

Here the data is splitted on 0.75 : 0.25 ratio, meaning 75% of data is devoted to training and 25% is devoted to testing. We are not worried about the 25% splitted testing data because we already have the testing data in another CSV file. Now the prediction will happen based on that 75% of training data. We can choose how much ever percentage of partition to be done based on our choice. But on standards, 50% or 75% is quite good figure to split.

As target variable has binary value like ‘1’ or ‘0’, the values refers to ‘TRUE’ or ‘FALSE’. For example, the target variable after it is splitted will looklike, (just first few records),

## [1] TRUE FALSE TRUE FALSE FALSE TRUE

After splitting, the training dataset will be scrambled means parition is done randomly and the records will look like,

## age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal  
## 2 37 1 2 130 250 0 1 187 0 3.5 0 0 2  
## 4 56 1 1 120 236 0 1 178 0 0.8 2 0 2  
## 7 56 0 1 140 294 0 0 153 0 1.3 1 0 2  
## 8 44 1 1 120 263 0 1 173 0 0.0 2 0 3  
## 10 57 1 2 150 168 0 1 174 0 1.6 2 0 2  
## 11 54 1 0 140 239 0 1 160 0 1.2 2 0 2  
## target  
## 2 1  
## 4 1  
## 7 1  
## 8 1  
## 10 1  
## 11 1

## [1] "Number of rows: 227"

After splitting, now it is good to observe that the numbers of records became to **227**. This has happened as we have splitted the training data to 75%.

In addition to this, we can also see the descriptive statisitcs of training data as shown below,

## age sex cp trestbps   
## Min. :29.00 Min. :0.0000 Min. :0.0000 Min. : 94.0   
## 1st Qu.:47.25 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:120.0   
## Median :55.00 Median :1.0000 Median :1.0000 Median :130.0   
## Mean :54.34 Mean :0.6821 Mean :0.9603 Mean :131.6   
## 3rd Qu.:61.00 3rd Qu.:1.0000 3rd Qu.:2.0000 3rd Qu.:140.0   
## Max. :77.00 Max. :1.0000 Max. :3.0000 Max. :200.0   
## chol fbs restecg thalach   
## Min. :126.0 Min. :0.0000 Min. :0.0000 Min. : 71.0   
## 1st Qu.:211.0 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:133.2   
## Median :240.5 Median :0.0000 Median :1.0000 Median :153.0   
## Mean :246.3 Mean :0.1457 Mean :0.5298 Mean :149.6   
## 3rd Qu.:274.8 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:166.0   
## Max. :564.0 Max. :1.0000 Max. :2.0000 Max. :202.0   
## exang oldpeak slope ca   
## Min. :0.0000 Min. :0.000 Min. :0.000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:1.000 1st Qu.:0.0000   
## Median :0.0000 Median :0.800 Median :1.000 Median :0.0000   
## Mean :0.3278 Mean :1.035 Mean :1.404 Mean :0.7318   
## 3rd Qu.:1.0000 3rd Qu.:1.600 3rd Qu.:2.000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :6.200 Max. :2.000 Max. :4.0000   
## thal target   
## Min. :0.000 Min. :0.000   
## 1st Qu.:2.000 1st Qu.:0.000   
## Median :2.000 Median :1.000   
## Mean :2.318 Mean :0.543   
## 3rd Qu.:3.000 3rd Qu.:1.000   
## Max. :3.000 Max. :1.000

Now that we’ve got some information abut the data like mean, median, minimum, maximum, quartiles for each of the variable available in the dataset given. This is to know in depth statistics about each variable.

Once we had an overlook on the data, now we are ready to build the model and apply it to the data and do prediction. The dataset is in the form that we can say the last column is the target variable (Dependent Variable) and the rest of the columns before the target variable are called independent variables. In Data Analytics, the prediction will be made based on the training data i.e., what we have seen so far. All the independent variables acts as symptoms or indicators for the target variable.

In this research, we are using Logistic Regression for prediction. We can use this linear regression model for predicting a value which is in binary class. It means that the resultant outcome will have only two outcomes i.e., ‘1’ or ‘0’, some may read it as ‘True’ or ‘False’, or ‘Positive’ or ‘Negative’. According to this research study, as the data chosen is numerical dataset and the outcome variable which is target variable which has two outcomes and we can assume that variable as ‘1’ for ‘Postive’ and ‘0’ for ‘Negative’. It says that for a record if the outcome is ‘1’ (positive) means the person with that symptoms has heart disease and ‘0’ (negative) means the person doesn’t have the disease. And moreover, the target variable has the same variable name “target”.

In R, we have a linear regression model function available in “caret” package. The logistic regression model can be applied for target variable with all the available features or attributes. The independent variables (age, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal) will be into “X” variable and the dependent variable as “Y” (for prediction) in the model.

In this process, we have three main steps,

**Step 1:** Building the model and finding the probability about the occurence of disease,

**Step 2:** Plotting ROCR curver to know the threshold for calculating accuracy of the model, and

**Step 3:** Calculating the Accuracy (performance measure of the model).

First we will find out the intercepts (labelled constants, the mean expected value of “Y” when all the “X” = 0) in order to find the probability.

**Step 1:**

**Building the model and finding out intercepts for “Y”(target variable)**

Step 1a:

Here, we are breaking down the code to just applying Logistic Regression in order to know the intercepts. That is, as in the code “target~” means we are modelling the target variable using (~) for each and every available features by using (.) i.e., for all the independent variables and we are also setting “family = binomial()” in the code because we are predicting a binary outcome, 0 or 1.

Once we have done this, we can see the intercept values, significance levels of each variable identified with asterick(\*) symbol, and the detailed summary as,

## Loading required package: lattice

## Loading required package: ggplot2

##   
## Call:  
## glm(formula = target ~ age + trestbps + chol + fbs + restecg +   
## thalach + exang + oldpeak + slope + ca + thal, family = binomial,   
## data = qualityTrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3350 -0.5191 0.2887 0.6364 2.2570   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.197328 2.698440 -0.073 0.941705   
## age 0.017093 0.024007 0.712 0.476484   
## trestbps -0.009792 0.011208 -0.874 0.382306   
## chol -0.001374 0.003618 -0.380 0.704152   
## fbs 0.153229 0.546697 0.280 0.779261   
## restecg 0.553431 0.355523 1.557 0.119550   
## thalach 0.034043 0.010728 3.173 0.001508 \*\*   
## exang -1.441694 0.420077 -3.432 0.000599 \*\*\*  
## oldpeak -0.471561 0.222960 -2.115 0.034430 \*   
## slope 0.190446 0.373027 0.511 0.609672   
## ca -0.777652 0.195156 -3.985 6.75e-05 \*\*\*  
## thal -1.360049 0.332008 -4.096 4.20e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 313.10 on 226 degrees of freedom  
## Residual deviance: 184.13 on 215 degrees of freedom  
## AIC: 208.13  
##   
## Number of Fisher Scoring iterations: 5

Step 1b:

Now the testing data record, which is a one record containing same attributes (indepedent variables) as in the training dataset uploaded will be used for prediction. All the variables will be there except the target variable since we are predicting the target variable based on the trained model. By doing so, with the help of logistic regression summary, we are now calculating the probability as shown below..

## age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal  
## 1 71 0 0 120 250 1 1 162 0 0.4 2 2 2  
## target  
## 1 NA

## [1] NA

## [1] "Based on the Model, the probablity that given data will have chance to get the disease most likely 83.7%"

**Step 2:**

**Plotting ROCR curver to know the threshold for calculating accuracy of the model,**

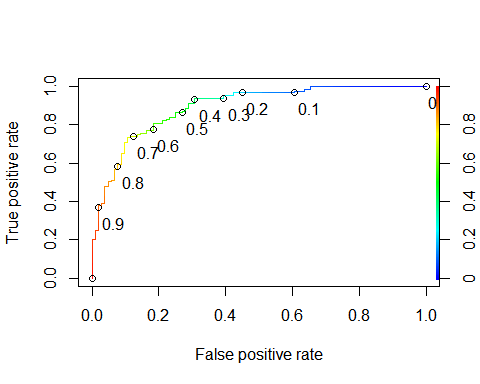
ROCR (Receiver Operating Characteristics Curve) helps in visualising the performance of the the classifer (logistic regression here) used. It’s like finding out the ability of binary classifer system based on the threshold value. Now it’s important for us to see how this model is predicting accurately.

From the following figure obtained, the ROC Curve says that the curve with threshold 0.7 is a good indication for a model to be good. And the as the true positve rate indicates that the maximum number of persons/patients with the disease are not identified as healthy.

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess



*Figure 1: ROCR curve showing the threshold*

**Step 3:**

**Calculating the Accuracy (performance measure of the model).**

Accuracy is very important in finding out the algorithm performance, because finding the outcome is quite easy for the algorithm to inference based on the training but how well the correspoding decision is made what is important here. The accuracy will take all the correct predictions done for the testing data (not one record, but the 25% testing data record taken from the training dataset by splitting. Remember, we have splitted the original dataset into two parts 75% for training and 25% for testing in order to calculate the accuracy for knowing the algorithm performance.) to the total observations, i.e., Accuracy = (Number of correct predictions) / (Number of total observations)

The accuracy of the model is as shown below,

## [1] "The Accuracy of this Model is 80%"

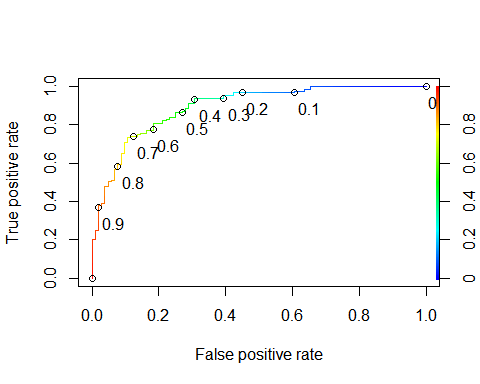
## Results

1. Upon uploading the testing dataset, the probability that the patient with that dataset record will have the disease is shown as below,

## [1] "Based on the Model, the probablity that given data will have chance to get the disease most likely 83.7%"

You can change the data values of all the attributes to see different predictions for different testing dataset.

1. ROCR graph for threshold,



1. Accuracy of the model,

## [1] "The Accuracy of this Model is 80%"

## Implications

In this research Logistic Regression was used. For classifying binary outcome, supervised machine learning algorihtms like Decision Trees, Support Vector Machine, K-Nearest Neighbor etc., can also be used. So in future, this prediction can be done using those algorithms as well. And different algorithms predicts different outcomes and performance of different algorithms are different. In this research, only 14 data columns were used from the dataset which is the most widely used dataset of UCI data repositories for applying data mining techniques. But there is also the original dataset of this which contains 86 columns originally in which it has lot of numerical and categorical data values. We can do predictions by seeing the association among all those variables in the dataset using some categorical data mining techniques also. But here, only 14 variables were considered according to the UCI data repository because there was a research that these attributes has major importance in predicting the target variable according to the study.

## Conclusion

In Data Analytics, its the process of understanding the data and making future predictions based on the current and the past data. Using this point, the chance of getting the heart disease to a patient was calculated using Logistic Regression by training upon the most famous dataset from UCI data repository for prediction based on the attributes that have some particular outcome i.e., either “Yes” or “No” tagged. This is helpful in estimating the new patient’s condition of getting the disease with different scenarios further.

## References

1. Apte S., & Dangare S., (2012). Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, *International Journal of Computer Applications (0975 – 888)*.
2. Jyoti S., Ujma A., Sharma S., & Soni S., (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, *International Journal of Computer Applications (0975 – 8887)*.
3. Kemphila A., & Boonjing V., Comparing Performances of Logistic Regression, Decision Trees, and Neural Networks for classifying Heart Disease Patients, *International Conference on Computer Information Systems and Industrial Management Applications (CISIM)*, 2011.
4. Duen Y., Ching-Hsue, C., & Yen-Wen C. (2011), A Predictive Model for Cerebrovascular Disease using Data Mining, *Expert Systems with Applications*, Vol. 8970-8977, 2011.
5. King R., (2019, September), *Heart Disease Prediction From Patient Data in R*, Riley King, “<https://rileyking.netlify.app/post/heart-disease-prediction-from-patient-data-in-r/>”
6. Mueller B., (2016, June), *Heart Disease Prediction*, www.rpubs.com, “<https://rpubs.com/mbbrigitte/heartdisease>”