

Mariama SOUMAHORO

Heart Disease

ABSTRACT

Heart Diseases are mainly describing a range of conditions that can affect the heart. It is most of the time triggered by the blockage of blood vessels that carry blood to the heart. According to the CDC, around 610,000 deaths caused by a heart disease are registered in the United States every year. The data presented to us has for purpose to understand the frequency of heart disease among 303 subjects between the ages of 29 to 77. It is our interest to figure out how each variable participates in being a risk factor for the heart disease by finding and developing a model in order to predict whether or not a person has a heart disease. Logistic regression model performed the best with an accuracy rate of 85.9%. Sex, cp, exang, oldpeak, slope, ca and thal were found to be the important factors that contribute into determining if a patient has a heart problem.

1 DATA CHARACTERISTICS

Data Collection

The goal is to be able to identify the involvement of a heart disease within the patient. The data has been categorized into 2 different groups: 0 meaning there is no presence of a heart disease and 1, meaning we have presence of a heart disease. The dataset used for this analysis contains 303 observations with 14 variables. The covariates are:

- age: The patient's age in year
- sex: The patient's sex (1 = male, 0 = female)
- cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, value 3: non-anginal pain, Value 4: asymptomatic)
- trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
- chol: The person's cholesterol measurement in mg/dl
- fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
- restecg : Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
- thalach: The person's maximum heart rate achieved
- exang: Exercise induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest
- slope: The slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: down sloping)
- ca: The number of major vessels (0-3)
- thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7= reversable defect)
- target: Heart disease (0 = no, 1 = yes); dependent variable

2.DATA VISUALISATION AND EXPLORATION

We take a closer look at the information given in order to have a general idea. We

have no missing data, however all the categorical variables appeared as non-

numerical. We then proceeded by recoding some of the variables into categorical (sex, fbs, exang, target, restecg, slope, cp, thal, ca) variables giving us a total of 9 categorical and 5 continuous variables.

We then plot the data for the continuous covariates to see if there exists a relation between them and the dependent variable target.

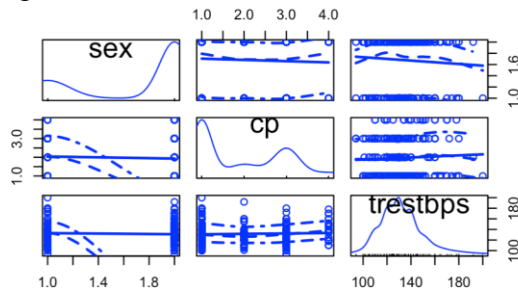
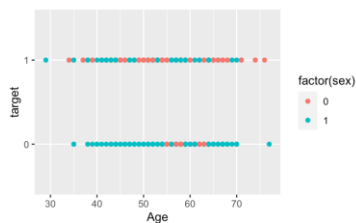


Figure 1 represents the scatterplots of some of the numerical variables which are drawn on the left part of the figure. Variable distribution is available on the diagonal. It is important to note that all variables seem to have a relationship with the dependent variable which is not linear.



Also, looking at figure 2, we notice a lot more men between the age of 40 to 70 who do not have heart disease compare to the opposite sex which only have 5 women who did not have a heart problem. For those who are sick, we observe that it is a mix of women and men.

After viewing the data, the next step into the analysis is to fit the data to select the best model.

3 MODEL SELECTION AND INTERPRETATION

Section 2 established that there were some patterns between dependent variable “target” and the different predictors. This section aims to summarize those patterns. For our analysis, we will be taking in consideration three different methods to put in place a best model in order to predict a heart disease: Logistic regression, LDA and QDA methods.

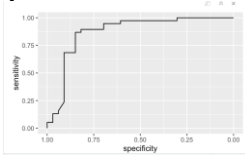
We first proceed by splitting the data into two parts. The training set is the first part representing 80% of the dataset. The second part is then the validation set which contains 20% of the data and is used to test the model chosen.

The first employed method is Logistic Regression. Indeed, Logistic Regression Analysis is a statistical method for analyzing a dataset and predict a binary outcome given a set of independent variables. With this method, we fit the training data with target being the response variable. The remaining of the covariates are the input variables. Using best selection model, we find the model. Looking at the results: sex, exang, pain type (cp), oldpeak, slope, and number of major vessels (ca), and thal were found to be significant. The model is run using the test data and then used to make a confusion matrix. From the confusion matrix we were able to calculate the accuracy of the model in hand which is in this case: 0.8592 and the misclassification rate is 0.1408.

With Logistic Regression, we were able to find a model that works. We next look at LDA and QDA analysis methods in order to see if we can find a model that works even better. Linear discriminant analysis (LDA) is a method used in statistics to find a linear combination of features which characterizes or separates two or more classes of events. Before performing the test, we check the distribution of each variable and normalize the data. We then proceed with the analysis. The LDA outputs contains different elements. Indeed, the prior probabilities of

groups represent the proportion of training observation in each group. In our case, 45% are in the group 0 (no heart problem) and 54.74% are in group 1 (presence of heart disease). The Group means is another component and shows the mean of each variable in each group. Regarding the coefficients of linear discriminants, this part gives us the linear combination of the predictors used to form the LDA decision rule. To see how effective the model is we make a confusion matrix and calculate the accuracy of the model which is equal to 0.86 with misclassification rate=0.14.

Following the LDA method, we proceed with the QDA method: Quadratic discriminant analysis. QDA is little bit more flexible than LDA, since it does not assume the equality of variance/covariance at each level. It is also recommended if the training set is very large, so that the variance of the classifier is not a major issue. After fitting the training data, we run the model on the test data. Calculating the accuracy of the model gives us that about 80% of the data has been correctly categorized. The misclassification rate is equal to 0.197. To summarize, each model had respectively an accuracy of 0.86;0.86;0.80. That being said, both logistic regression and least discriminant analysis (LDA) are doing a good job at fitting the data. We therefore choose the model obtained with the Logistic Regression since we get an area under the curve equal to 0.872 as the following figure

shows.  Therefore, the

final model is: $\text{Logit}(\hat{\pi}) = 1.31 - 1.92(\text{sex}=\text{male}) + 0.63(\text{cp}=\text{typical angina}) + 2.49(\text{cp}=\text{atypical angina}) + 3.09(\text{cp}=\text{non anginal pain}) - 0.98(\text{exang}=\text{yes}) - 0.80(\text{oldpeak}) - 1.65(\text{slope}=\text{upsloping}) + 0.23(\text{slope}=\text{flat}) - 2.68(\text{ca}=0 \text{ vessel}) - 3.88(\text{ca}=1 \text{ vessel}) - 2.28(\text{ca}=2 \text{ vessels}) + 10.67(\text{ca}=3 \text{ vessels}) + 3.57(\text{thal}=\text{normal}) + 2.78(\text{thal}=\text{fixed defect}) + 0.87(\text{thal}=\text{reversal defect})$

4. LIMITATION

At the end of the analysis, the statistical methods used to fit the data produced good results although logistic method gave us better results. Despite the results, more can be done to better the model. Other risk factors such as smoking, excessive drinking, stress levels among others could have been included in the attributes as they are part of the leading risk factors of heart diseases. One other factor is to check for outliers as LDA method is very sensitive. Other methods not as sensitive can be used as well to analyze the data in order to have better results.