2020

Project By:

Ritesh Natekar

2/5/2020

A picture containing drawing

Description automatically generated

**Predicting Heart Disease**

-PROJECT REPORT-

**Table of Contents**

1. Introduction 3

2. Approach 4

3. Data Preparation

* + Data Cleaning
  + Explotaroy Data analysis
  + Feature Selection

4. Logistic Regression Model 6

5. Project Results and it's Interpretion

* + Model Evaluation
  + ROC
  + AUC

7. Disease decisions after model analysis 10

Abstract

World Health Organization has estimated 12 million deaths occur worldwide; every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression.

This project is intended to find out the patient who could be having risk of heart disease in next ten years. So that we can plan to make some changes in lifestyle or habits to prevent this life-threatening disease.

Dataset contains information of over 40000 patients with all details like gender, age, current smoker, cigs per day, BP medication, heart rate, BMI, glucose etc. Which would be heling us to predict whether patient can have heart disease or not.

We have used Logistic Regression classification method here. To classify the patients in two categories i.e. With Heart disease and without heart disease. Logistic Regression is supervised machine learning technique to classify qualitative dependent variable.

Again, used statistical terms like Odds ratio, confidence interval and P-value. We have checked ROC Curve and AUC value in order to make it sure that the model is performing well.

Approach

* Getting Dataset

Dataset has been collected in a “.csv” format. Examined the dataset with all variables included. Understand the importance of each variable in the dataset. Checking with data type of variables and how much relevant information they are providing for model building.

* Choosing Regression Model

In the given dataset, having variables with categorical and continuous type of data. Dependent variable i.e. Ten-year Coronary Heart Disease is of type categorical and other independent variable which correlate the dependent variable more accurately are of type categorical. We have explanatory variable as qualitative and quantitative both and Response variable as qualitative, so **Logistic** **Regression** model is best fit for predicting the response variable i.e. Ten-Year CHD. In Logistic Regression dependent variable is always binary. Logistic Regression is mainly used for prediction and also calculating the probability of success.

* Cleaning dataset

Once regression model is confirmed, to start the working on dataset, it is very important to clean it. Finding missing values/null values and replacing it with relevant values in order to avoid discrepancy in data.

* Development Model and Testing Model

Divide dataset into two different model:

1. Development Model
2. Testing Model

Development Model, It is used to develop regression model. Considering different explanatory variables to predict response variable and examined accuracy of model.

* In the predictive modeling, the data need to be partitioned into train and test sets. Here 80% of the data will be partitioned for training purpose and 20% of the data will be partitioned for testing purpose.
* Developing Regression Model

Logistic Regression model, considering all relevant explanatory variables to predict response variable. Check the P-value, if variable having least p-values then it more accurate in predicting response variable.

Check AUC Curve value for every logistic regression model. The Area Under the Curve (**AUC**) is a metric for binary classification. It concludes that classifier is good or bad. How well it performing for Given a collection of models for the data.

* Testing Model

Once model developed fully then perform testing on it. And validate it.

* Inferring data insights and predict outcome for new data record.

Now model is ready to analyses the thing and make predictions.

## **Data Preparation**

### Source:

The dataset is publicly available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the Massachusetts, USA. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients’ information. It includes over 4,000 records and 15 attributes.

Variables :

Each attribute is a potential risk factor. There are both demographic, behavioural 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)

Behavioural

currentSmoker: Whether or not the patient is a current smoker (Nominal)

cigsPerDay: The number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarretts, even half a cigarette.)

Medical( history):

BPMeds: whether or not the patient was on Blood Pressure medication (Nominal)

prevalentStroke: whether or not the patient had previously had a stroke (Nominal)

prevalentHyp: whether or not the patient was hypertensive (Nominal)

diabetes: whether or not the patient had diabetes (Nominal)

Medical(current):

totChol: total cholesterol level (Continuous)

sysBP: systolic blood pressure (Continuous)

diaBP: diastolic blood pressure (Continuous)

BMI: Body Mass Index (Continuous)

heartRate: 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”)

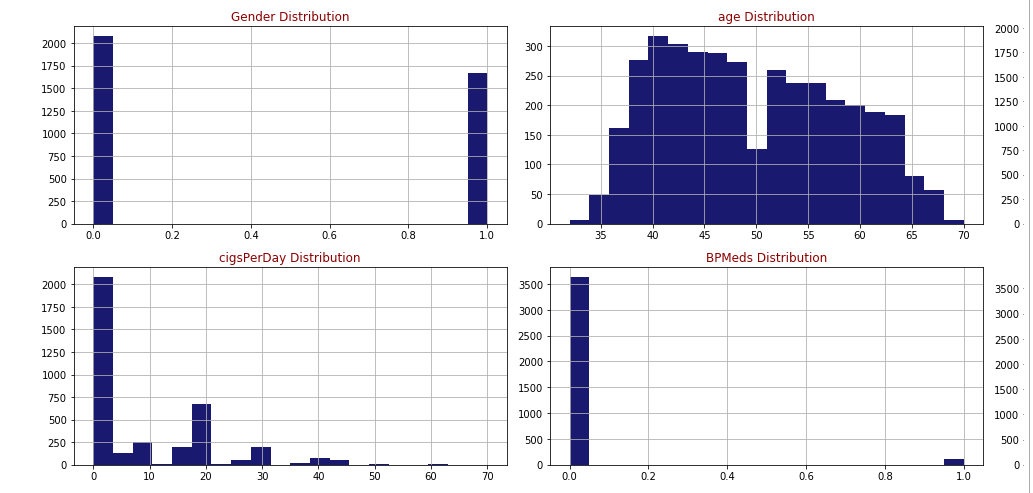
* **Data cleaning:**
  + 1. Removing not required columns.

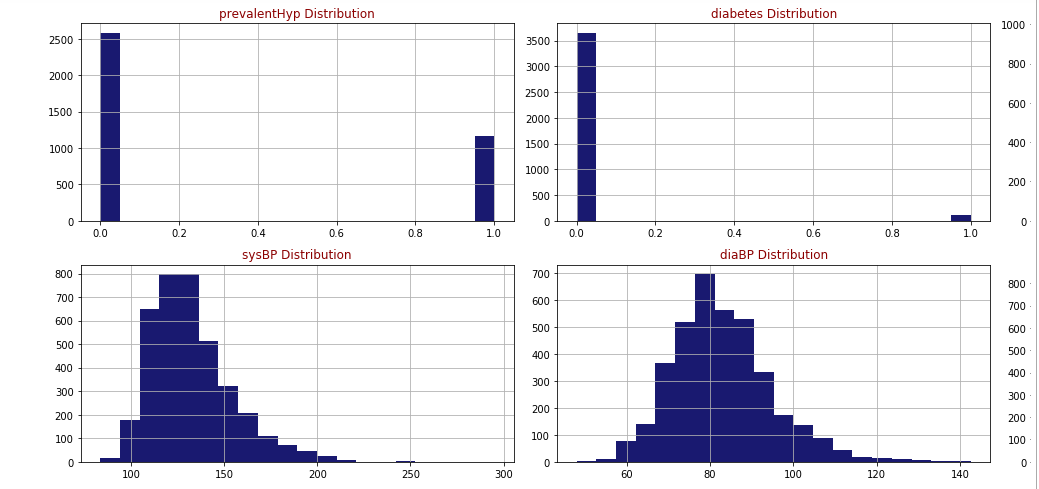
Here we can “education” does not make any sense in predicting whether person will have risk of heart disease in next Ten Year. So better to remove such columns in order to make model simpler.

* + 1. Removing Null values.

Null values should be removed. If there are only 10-20% of rows of the entire dataset contains null value, then its feasible to directly remove those records from the dataset. Here in our dataset, we prefer to remove those null values.

* **Exploratory Data Analysis**

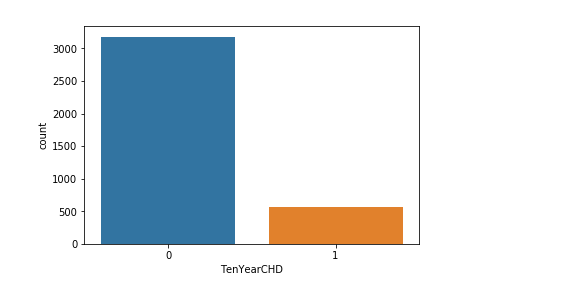




Studying Response Variable:

Counting number of patients with heart disease and patients with no heart disease.

There are **3179** patents with no heart disease and **572** patients with risk of heart disease.

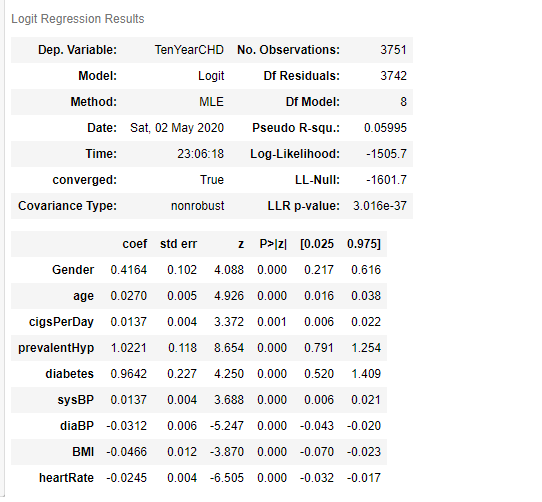


### Feature Selection: Backward elimination (P-value approach)

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.

Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

If there are the attributes with P value higher than the preferred alpha(5%) and thereby showing low statistically significant relationship with the probability of heart disease. Backward elemination approach is used here to remove those attributes with highest Pvalue one at a time follwed by running the regression repeatedly until all attributes have P Values less than 0.05.



We have gotten the variables with p-value less that.

Logistic Regression classification

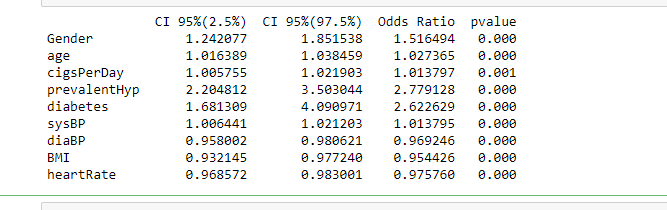
Logistic regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable from a set of predictor or independent variables. In logistic regression the dependent variable is always binary. Logistic regression is mainly used to for prediction and also calculating the probability of success.

P=eβ0+β1X1/1+eβ0+β1X1

When all features plugged in:

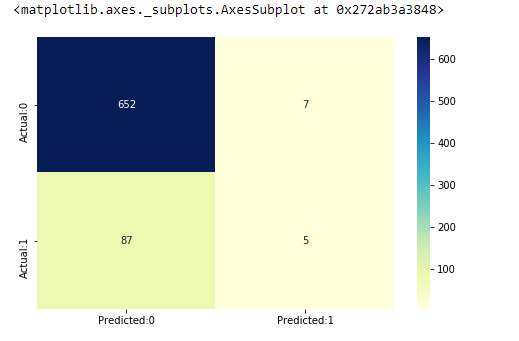
logit(p)=log(p/(1−p))=β0+β1∗Sexmale+β2∗age+β3∗cigsPerDay+β4∗totChol+β5∗sysBP+β6∗glucose

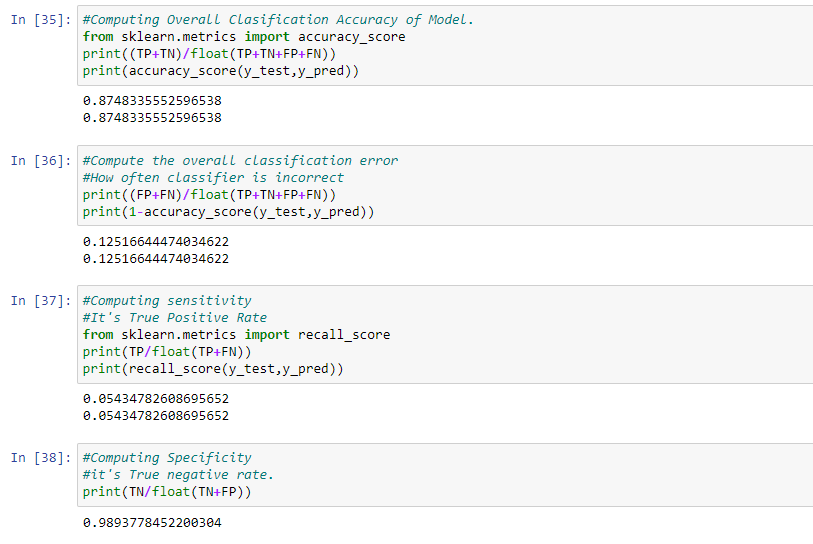
Interpreting the result: Odds ratio, Confidence Interval and P-value.



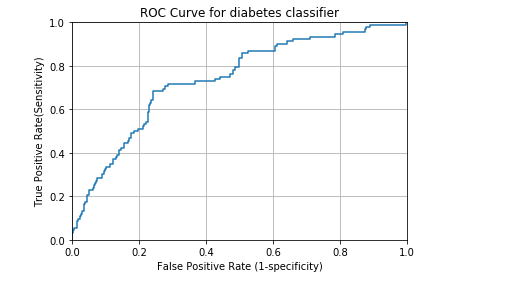
* This fitted model shows that, holding all other features constant, the odds of getting diagnosed with heart disease for males (sex\_male = 1) over that of females (sex\_female = 0) is 1.516494. In terms of percent change, we can say that the odds for males are 51.64% higher than the odds for females.
* The coefficient for age says that, holding all others constant, we will see 2% increase in the odds of getting diagnosed with CDH for a one year increase in age since exp(0.0270) = 1.067644.
* Similarly, with every extra cigarette one smokes there is a 2% increase in the odds of CDH.
* For Total cholesterol level and glucose level there is no significant change.
* There is a 1.2% increase in odds for every unit increase in systolic Blood Pressure.
* **Model Evaluation:**

Confusion Matrix:

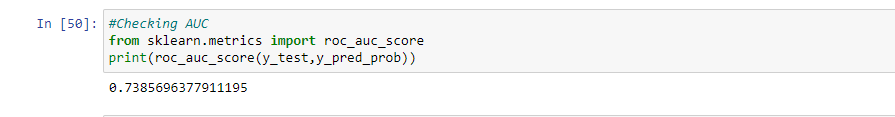




* From the above statistics it is clear that the model is highly specific than sensitive. The negative values are predicted more accurately than the positives.
* Predicted probabilities of 0 (No Coronary Heart Disease) and 1 ( Coronary Heart Disease: Yes) for the test data with a default classification threshold of 0.5.
* **ROC Curve:**



* A common way to visualize the trade-offs of different thresholds is by using an ROC curve, a plot of the true positive rate (# true positives/ total # positives) versus the false positive rate (# false positives / total # negatives) for all possible choices of thresholds. A model with good classification accuracy should have significantly more true positives than false positives at all thresholds.
* The optimum position for roc curve is towards the top left corner where the specificity and sensitivity are at optimum levels
* **AUC**



The area under the ROC curve quantifies model classification accuracy; the higher the area, the greater the disparity between true and false positives, and the stronger the model in classifying members of the training dataset. An area of 0.5 corresponds to a model that performs no better than random classification and a good classifier stays as far away from that as possible. An area of 1 is ideal. The closer the AUC to 1 the better.

Disease Decisions after model analysis

* 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.
* The Area under the ROC curve is 73.5 which is somewhat satisfactory.
* Overall model could be improved with more data.