



Heart Disease Prediction by Logistic Regression

AUTHORS:

HET PATEL - 8680821

NEEL PATEL - 8682068

Objective

Every year millions of deaths occur all over world due to heart related conditions. If we can predict beforehand whether any given person will have a heart related problem in the near future, it can be easily overcome by changing their lifestyle or by taking other precautionary measures. The goal of this project is to predict all the important factors which leads to diagnosis of heart diseases and how likely the person in question is to suffer from it in the next 10 years.

For data, we have used the Framingham Data set. Framingham Heart study Is a long-term and ongoing cardiovascular study of the residents of Framingham city which is located in Massachusetts. Our goal of this project is to develop a working model which will take all the variables given in this dataset and try to predict whether the given person will have a heart disease or not based on the different factors used in the model.

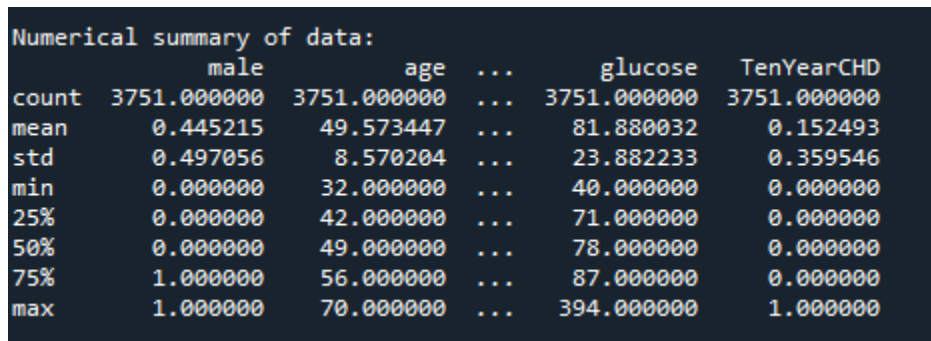
Logistic Regression will be used for prediction as both, the outcome and the structure of data in the dataset, are in binary format.

Logistic Regression Predictive Modeling will have the following steps:

1. Descriptive Data Analysis
2. Logistic Regression Model Creation and Interpretation
3. Model Evaluation
4. Final Prediction

Step 1: Descriptive Data Analysis

- Data Preparation:
In this step, first we prepare data by checking data types, duplicate values and rows with missing values if any. Missing values were found which constituted 12% of the dataset and hence were removed for better results.
- Numerical Summary:

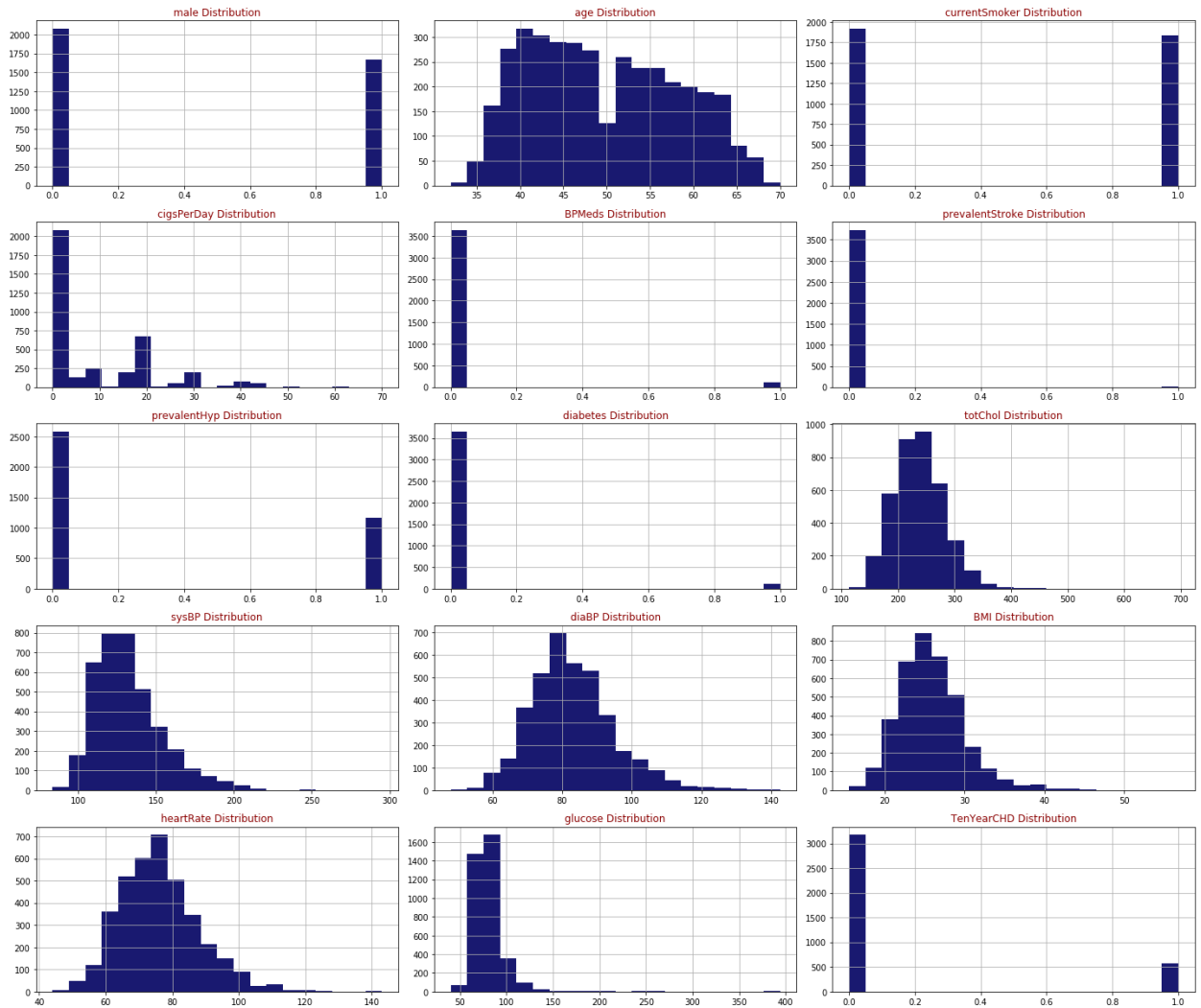


```
Numerical summary of data:
```

	male	age	...	glucose	TenYearCHD
count	3751.000000	3751.000000	...	3751.000000	3751.000000
mean	0.445215	49.573447	...	81.880032	0.152493
std	0.497056	8.570204	...	23.882233	0.359546
min	0.000000	32.000000	...	40.000000	0.000000
25%	0.000000	42.000000	...	71.000000	0.000000
50%	0.000000	49.000000	...	78.000000	0.000000
75%	1.000000	56.000000	...	87.000000	0.000000
max	1.000000	70.000000	...	394.000000	1.000000

In Numerical Summary, we check for 50% (mean) values whether they are symmetrical which can suggest that the data is normal and there is no unusuality.

- Graphical Summary:



The curve of each variable resembles to that of a normalized data, hence we can go ahead with performing Logistic Regression on the data.

Step 2: Logistic Regression Model Creation and Interpretation

- Baseline Model:

First, we create a model which considers all the variables, regardless of how strongly or weakly they are correlated to developing CHD (Coronary Heart Disease).

Baseline Model:						
Logit Regression Results						
=====						
Dep. Variable:	TenYearCHD	No. Observations:		3751		
Model:	Logit	Df Residuals:		3736		
Method:	MLE	Df Model:		14		
Date:	Sun, 19 Apr 2020	Pseudo R-squ.:		0.1170		
Time:	12:41:30	Log-Likelihood:		-1414.3		
converged:	True	LL-Null:		-1601.7		
Covariance Type:	nonrobust	LLR p-value:		2.439e-71		
=====						
	coef	std err	z	P> z	[0.025	0.975]

const	-8.6532	0.687	-12.589	0.000	-10.000	-7.306
male	0.5742	0.107	5.345	0.000	0.364	0.785
age	0.0641	0.007	9.799	0.000	0.051	0.077
currentSmoker	0.0739	0.155	0.478	0.633	-0.229	0.377
cigsPerDay	0.0184	0.006	3.000	0.003	0.006	0.030
BPMeds	0.1448	0.232	0.623	0.533	-0.310	0.600
prevalentStroke	0.7193	0.489	1.471	0.141	-0.239	1.678
prevalentHyp	0.2142	0.136	1.571	0.116	-0.053	0.481
diabetes	0.0022	0.312	0.007	0.994	-0.610	0.614
totChol	0.0023	0.001	2.081	0.037	0.000	0.004
sysBP	0.0154	0.004	4.082	0.000	0.008	0.023
diaBP	-0.0040	0.006	-0.623	0.533	-0.016	0.009
BMI	0.0103	0.013	0.827	0.408	-0.014	0.035
heartRate	-0.0023	0.004	-0.549	0.583	-0.010	0.006
glucose	0.0076	0.002	3.409	0.001	0.003	0.012
=====						

- Backward Selection Model:

In this model, we only consider variables which contributes to developing the heart disease and eliminate all other nonsignificant variables to make the model more efficient. (Baseline models are only used to create better models and not actually predict something for most of the cases).

Backward Selection Model:						
Logit Regression Results						
Dep. Variable:	TenYearCHD	No. Observations:	3751			
Model:	Logit	Df Residuals:	3744			
Method:	MLE	Df Model:	6			
Date:	Sun, 19 Apr 2020	Pseudo R-squ.:	0.1149			
Time:	12:41:30	Log-Likelihood:	-1417.7			
converged:	True	LL-Null:	-1601.7			
Covariance Type:	nonrobust	LLR p-value:	2.127e-76			
	coef	std err	z	P> z	[0.025	0.975]
const	-9.1264	0.468	-19.504	0.000	-10.043	-8.209
male	0.5815	0.105	5.524	0.000	0.375	0.788
age	0.0655	0.006	10.343	0.000	0.053	0.078
cigsPerDay	0.0197	0.004	4.805	0.000	0.012	0.028
totChol	0.0023	0.001	2.106	0.035	0.000	0.004
sysBP	0.0174	0.002	8.162	0.000	0.013	0.022
glucose	0.0076	0.002	4.574	0.000	0.004	0.011

- Interpretation of the model:

### Interpreting the results: Odds Ratio, Confidence Intervals and Pvalues ###				
	CI 95%(2.5%)	CI 95%(97.5%)	Odds Ratio	pvalue
const	0.000043	0.000272	0.000109	0.000
male	1.455242	2.198536	1.788687	0.000
age	1.054483	1.080969	1.067644	0.000
cigsPerDay	1.011733	1.028128	1.019897	0.000
totChol	1.000158	1.004394	1.002273	0.035
sysBP	1.013292	1.021784	1.017529	0.000
glucose	1.004346	1.010898	1.007617	0.000

Based on the above output, we can interpret that:

- All the changes in the variable noticed are based on the keeping all the other features or variable constant.
- From the Odds Ratio of the Male variable it is seen that Males are 78.8% more likely to get a heart disease compared to that of female.
- Similarly, every year a person grows older, his/her chances of getting a heart disease rise by 7% (1.067644)
- Also, for every extra cigarette a person smokes, his/her odds of getting CHD increase by 2%.
- There is minimal increase in the chances of getting a CDH for increase in the variable values of cholesterol, systolic BP and glucose.

- v. Also, as we can see after splitting the data into training and test data the accuracy of the model, we created is 0.8696 which means that the model predicts the outcome of having a CHD accurately 88% of the time.

Step 3: Model Evaluation

```
##### Computing Accuracy #####
Accuracy of Model: 86.96808510638297 %

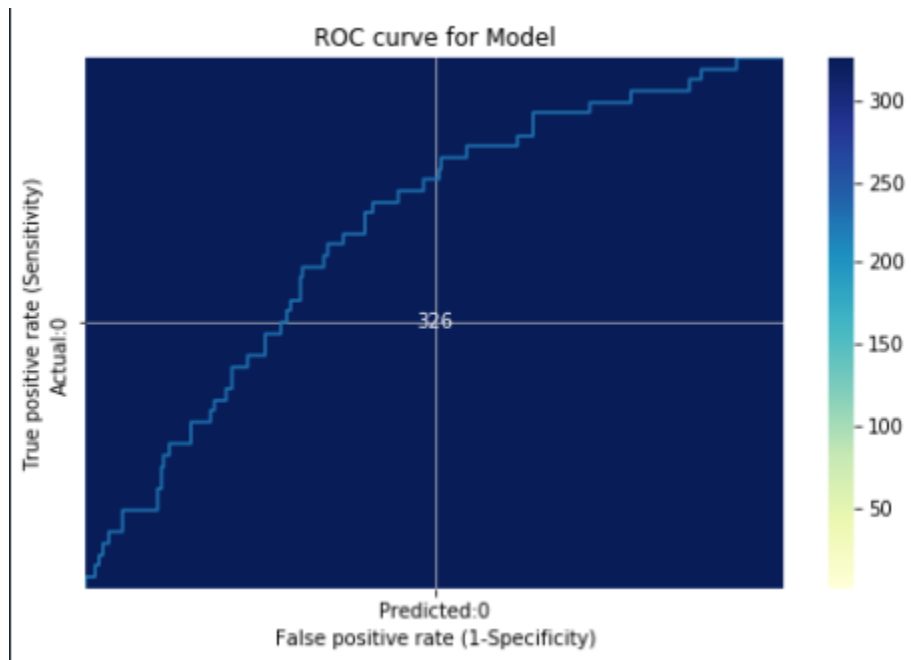
##### Model Evaluation by Confusion Matrix #####

The accuracy of the model =  $TP+TN/(TP+TN+FP+FN)$  = 0.8696808510638298
The Missclassification =  $1-Accuracy$  = 0.13031914893617025
Sensitivity or True Positive Rate =  $TP/(TP+FN)$  = 0.02083333333333332
Specificity or True Negative Rate =  $TN/(TN+FP)$  = 0.9939024390243902
Positive Predictive value =  $TP/(TP+FP)$  = 0.3333333333333333
Negative predictive Value =  $TN/(TN+FN)$  = 0.8739946380697051
Positive Likelihood Ratio =  $Sensitivity/(1-Specificity)$  = 3.4166666666666634
Negative likelihood Ratio =  $(1-Sensitivity)/Specificity$  = 0.9851738241308793
```

The accuracy of the model is approximately 87% which can be considered a good prediction model. A confusion matrix is also created in order to the evaluate the model more thoroughly. Some terms to keep in mind as we evaluate are:

- Accuracy: Accuracy is the proportion of cases correctly classified by the model. It ranges from 0 to 1 (the higher the better).
- Specificity (True Negative rate): Correct negative predictions divided total negatives.
- Sensitivity (True Positive Rate) : Correct positive predictions divided by total positives.
- Precision: Percent of 1's predicted as 1 by the model.

ROC and AUC:



```
##### ROC Curve and AUC #####  
AUC: 0.6768927845528455
```

ROC (Receiver operating characteristic) Curve is just a plot which shows how much more true positives are than false positives, which is a way to interpret that the model is good. In terms of evaluation by numbers, we use AUC (Area under curve) which just shows the area under the ROC Curve (the higher the better).

The AUC Curve is 0.67 in a 0 to 1 scale which indicates a good model.

Step 4: Final Prediction

```
##### FINAL PREDICTION #####
```

	Prob of no heart disease (0)	Prob of Heart Disease (1)
0	0.939611	0.060389
1	0.779009	0.220991
2	0.953600	0.046400
3	0.361612	0.638388
4	0.845033	0.154967
..
371	0.366422	0.633578
372	0.886308	0.113692
373	0.726589	0.273411
374	0.959471	0.040529
375	0.940064	0.059936

This is the final prediction where 0 indicates the probability of not having a heart disease and 1 is the probability of having the heart disease for a particular person in the next 10 years.

It is important to note that the dataset was initially split into Train(90%), Validate(5%) and Test(5%) dataset. The Final Prediction consists of predictions made from the test dataset.

Conclusion

The Backward Selection Model is selected as it produces more accurate output than the Baseline Model. The model can be used on any new data and would be extremely helpful to predict the disease beforehand and act on it in a timely manner to reduce the effects or avoid suffering from it altogether.

APPENDIX ONE: FRAMINGHAM DATA DICTIONARY

Male	Gender of the person (1- Male, 0- Female)
Age	Age of the person at the time getting the data
Education	1 to 4 (From High School to College)
CurrentSmoker	Whether the person is a smoker or not(1 -Smoker , 0 – Not Smoker)
cigsperDay	No of cigarettes per day if a person is a smoker
BPMeds	Whether the person is on blood pressure medications or not (1- Is on Medications, 0 – Not on medication)
prevalentStroke	Whether the person had a Stroke in the past
prevalentHyp	Whether the person has Hypertension
diabetes	Whether the person has diabetes or not
totCol	the cholesterol of the patient
sysBP	The Systolic Pressure of the patient
diaBP	The Diastolic Pressure of the patient
BMI	The body mass Index of the person
heartrate	The heartrate of the patient
glucose	The glucose of the person
TenyearCHD	The chances of whether the patient will have coronary heart disease(CHD) in 10 Year