ISYE6420

Chong Zhang

Nov 25, 2019

Logistic Regression with Bayesian Approach

A Case Study of heart disease data

1.Introduction

Heart disease is very serious disease that threat human beings. It is a range pf conditions that adversely affect one's heart, which includes blood vessel diseases, such as Coronary heart disease, and heart defects. Symptoms of heart disease can include check pain, high or low blood pressure, racing heart rate, chest discomfort and short of breath, etc. Every year, millions of people lost their lives due to heart disease. In US along, 610000 people die of heart disease every year. It is the leading cause of death in the US. According to the CDC, more than half of the victims of heart disease are male. Thus, it is very critical for patients and doctors to diagnose the disease in the early stage. Also, it would be very helpful if doctors can use a combination of features such as age, sex and max heart rate to predict whether a patient is likely to develop heart conditions. If the chance of the patient to have heart condition is high, doctors can intervene in the early stage, and patient can respond according to reduce the chance by changing lifestyle or taking certain medications. Thus, it is very important to develop a model by using a combination of features to predict whether the patient will have heart disease or not.

Here I used the **Heart Disease UCI** dataset that I obtained from the Kaggle website(https://www.kaggle.com/ronitf/heart-disease-uci). The whole dataset was created by Andras Janosi, M.D, William Steinbrunn, M.D., Matthias Pfisterer, M.D and Robert Detrano, M.D., Ph.D. It composed of 303 samples with 13 features and 1 dependent variable.

The 13 features are:

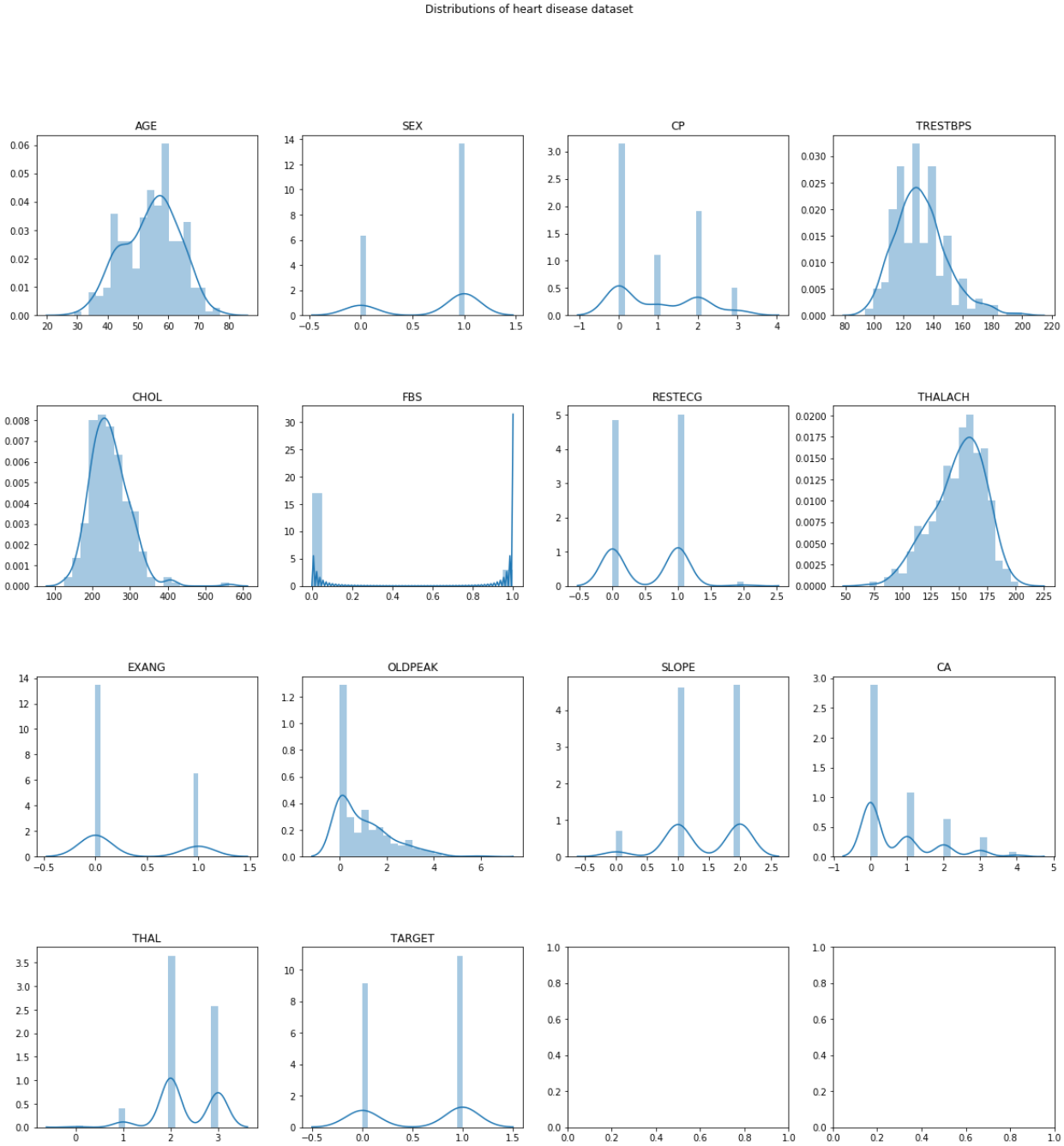
1. age
2. Sex, 1 for male, 0 for female
3. chest pain type (4 values)
4. resting blood pressure (in mm Hg on admission to the hospital)
5. serum cholestoral in mg/dl
6. Whether fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina (1 = yes; 0 = no)
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

All the 13 features are converted to numerical numbers for computation.

The one dependent variable: Target is whether the patient has heart disease or not, indicating by 1 or 0. 1 for yes, 0 for no.

Here I am going to use logistic regression to predict whether the patient has heart disease or not. I will use both Bayesian logistic regression model and classical frequentist logistic regression model. I have set aside 20% of the sample for testing purpose. In the end, I will AUC as a measure of accuracy to compare the performance of both models.

2. Data investigation

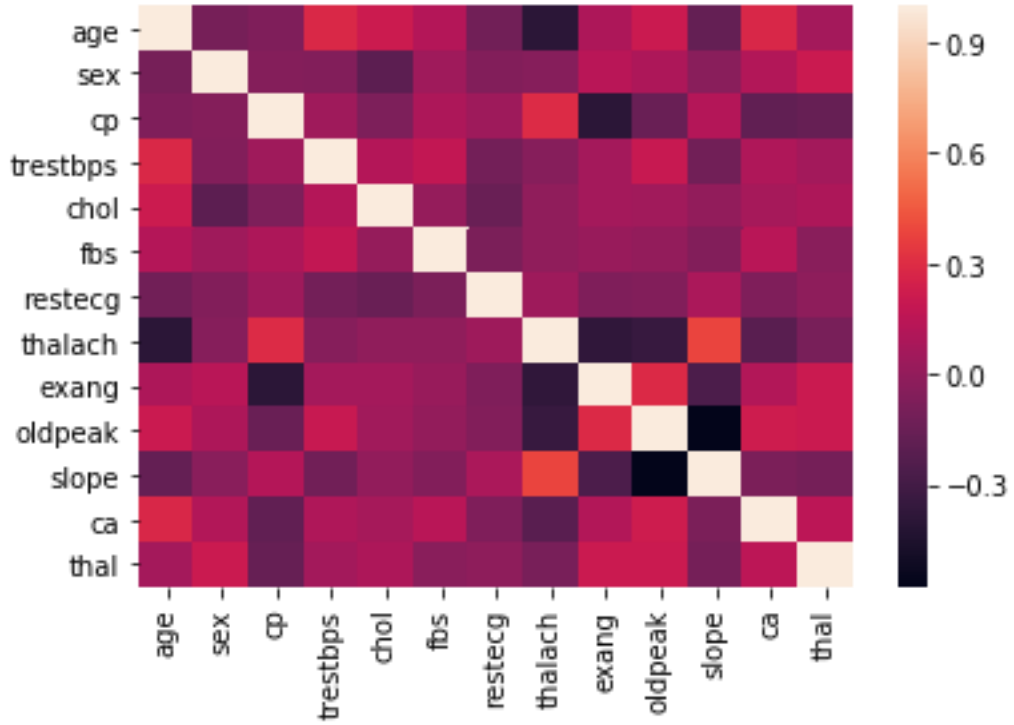


First, I looked at the distribution of each variable. The distribution of each variable is shown as below. As we can see that some of them have normal distribution (Age,

Trestbps, Chol, Thalach). Some of them have geometric distribution (CA, CP). Some have exponential distribution (OldPeak). Some of them have uniform distribution (SEX, Exang, Restecg, FBS). The dependent variable target has roughly 50% for 0 and 50% for 1.

For logistic regression, one of the assumptions is no or little multicollinearity. Thus, I check the correlation between each variable. It seems that there is no very high multicollinearity in the dataset.

Another assumption is that logistic regression typically requires a large sample size. For each independent variable, at least 10 samples with the least frequent outcome are needed. The whole dataset has 138 samples with the outcome as 0, which means the frequency is 0.45. Thus, we need as least (10*13)/0.45=289 samples. Since we have 303 samples, the sample number satisfy the assumption.



3. Model Identification & Results

Logistic regression model is presented as the formula below:

Logit(p) = beta1 + beta2 * age+beta3 * sex+ beta4 * cp+ beta5 * trestbps+
beta6 * chol + beta7 * fbs + beta8 * restecg + beta9 * thalach + beta10*
exang + beta11 * oldpeak+beta12*slope+beta13*ca+beta14*thal

Here p is the probability of the patient has heart disease. I set the threshold at 0.5, which means is p > 0.5, the patient is predicted to have heart disease. If p <0.5, then the patient is predicted to not have heart disease.

Beta1 to beta14 are coefficients. Since we have 13 features, there are 14 coefficients, 13 for each feature, 1 for the intercept.

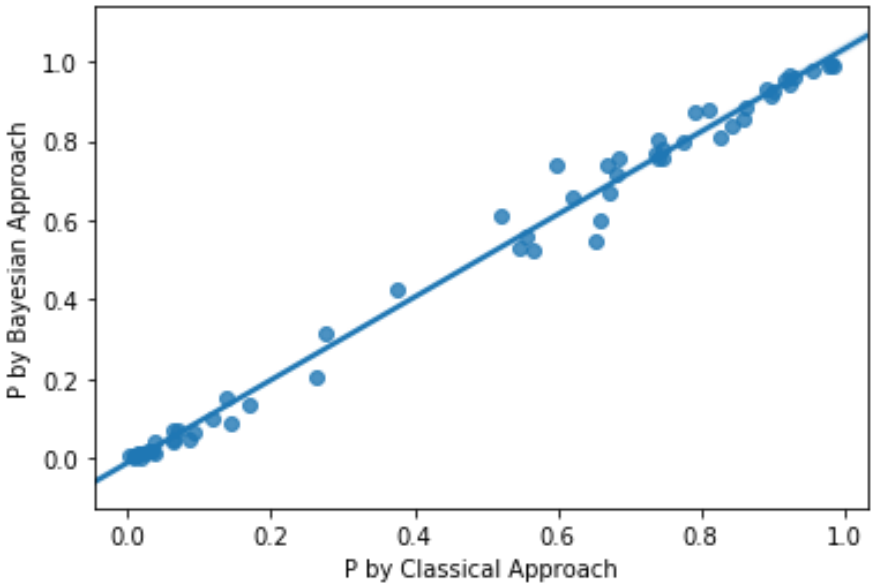
The whole 303 samples were split into two parts. One for training, which is consisted of 80% of the samples. The rest is for testing, which is consisted of 20% of the samples.

First, I implement logistic regression in OpenBugs. It is based on Bayesian Statistics. I used non-informative priors for all the 14 coefficients. Then I also implement logistic regression with Python Sklearns. It is based on the frequentist's approach. In the end, I compared the performance of both implementations. Here is what I got.

	OpenBugs	
	Predicted Yes	Predicted No
Actual Yes	25	8
Actual No	1	27

	Python	
	Predicted Yes	Predicted No
Actual Yes	25	8
Actual No	1	27

As we can see that the performance of both implementations is the same, with an accur acy at 85%. The ROC score of Bayesian implementation in Openbugs is 0.936. It is 0.9 37 for Python implementation. Thus, we can say that the performance of both implemen tations is almost the same. I plotted the p predicted by classical approach and Bayesian approach. We can see that they are comparable.



The final Bayesian model is:

Logit(p) = 2.839 + 0.02 * age -1.90* sex +0.79* cp-0.029* trestbps-0.007 * chol + 0.022 * fbs + 0.715* restecg + 0.036* thalach -0.80* exang -0.69* oldpeak+0.50*slope-0.80*ca-1.06 *thal

4. Conclusion

In my project, I showed that the performance of Bayesian Logistic Regression model in predicting whether the patient will have heart disease or not is almost the same as class ical frequentist's Logistic Regression approach.

For this analysis, I used non-informative priors. For future analysis, I will explore differen t priors for beta, such as informative priors. Also we can change the threshold of p to se e whether it can improve the prediction accuracy.