The University of Hong Kong

STAT 7614

Advanced Statistical Learning

Heart Disease Dataset Analysis

Authors:
Chen Xiangyue 3035719052
Miao Peilin 3035535236
Xiong Jiaming
Wang Mingxun 3035712872

April 21, 2023

Contents

1 Introduction

Heart Disease is one of the leading cause of death in the world. One of the most prevalent type of heart disease is the coronary artery disease, which blocks the normal blood flow. The heart attack caused by the decreased blood flow is detrimental. In this report, we try to figure out whether there is a heart failure given the condition of a patient by implementing different kinds of statistical modelling techniques.

2 Data Preparation

? sponsored the dataset to Kaggle. The figure below includes the first three rows of the dataframe.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS
	<int>	<chr>	<chr>	<int>	<int>	<int>
1	40	M	ATA	140	289	0
2	49	F	NAP	160	180	0
3	37	M	ATA	130	283	0

RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
<chr>	<int>	<chr>	<dbl>	<chr>	<int>
Normal	172	N	0	Up	0
Normal	156	N	1	Flat	1
ST	98	N	0	Up	0

Figure 1: Demo of the data set

Most of the variables can be interpreted correctly according to the variable description. However, two of the variables need extra elaborations. For chest pain type, the four levels are typical angina(TA), atypical angina(ATA), non-anginal pain(NAP) and asymptomatic(ASY). For resting electrocardiogram results, the three levels are nomral results (Normal), having some wave abnormalities(ST) which means T wave inversions and ST elevation or depression greater than 0.05mV and showing probable or definite left ventricular hypertrophy by Estes' criteria (LVH) (Fedesoriano, 2021).

Variable	**Description**	**Unit/Level**
Age	patient's age	years
Sex	patient's gender	Male/Female
ChestPainType	chest pain type	TA/ATA/NAP/ASY
RestingBP	resting blood pressure	mm Hg
Cholesterol	serum cholesterol	mm/dl
FastingBS	fasting blood sugar	1: >120mg/dl, 0:else
RestingECG	resting electrocardiogram results	Normal/ST/LVH
MaxHR	maximum heart rate achieved	60-202
ExerciseAngina	exercise-induced angina	Y/N
Oldpeak	oldpeak	numeric value measured in depression
ST_Slope	slope of the peak exercise ST segment	Up/Flat/Down
HeartDisease	output class	1:heart disease, 0:normal

2.1 Data Pre-processing

Before we implement any statistical model to the dataset, we did some routine data processing works. There are no missing values in the dataframe, so we look at boxplots to identify potential outliers. We notice that for RestingBP, one data point deviates from the cluster significantly so we delete it. Furthermore, since all the categorical data are stored as numeric values, we convert all of them to factor levels.

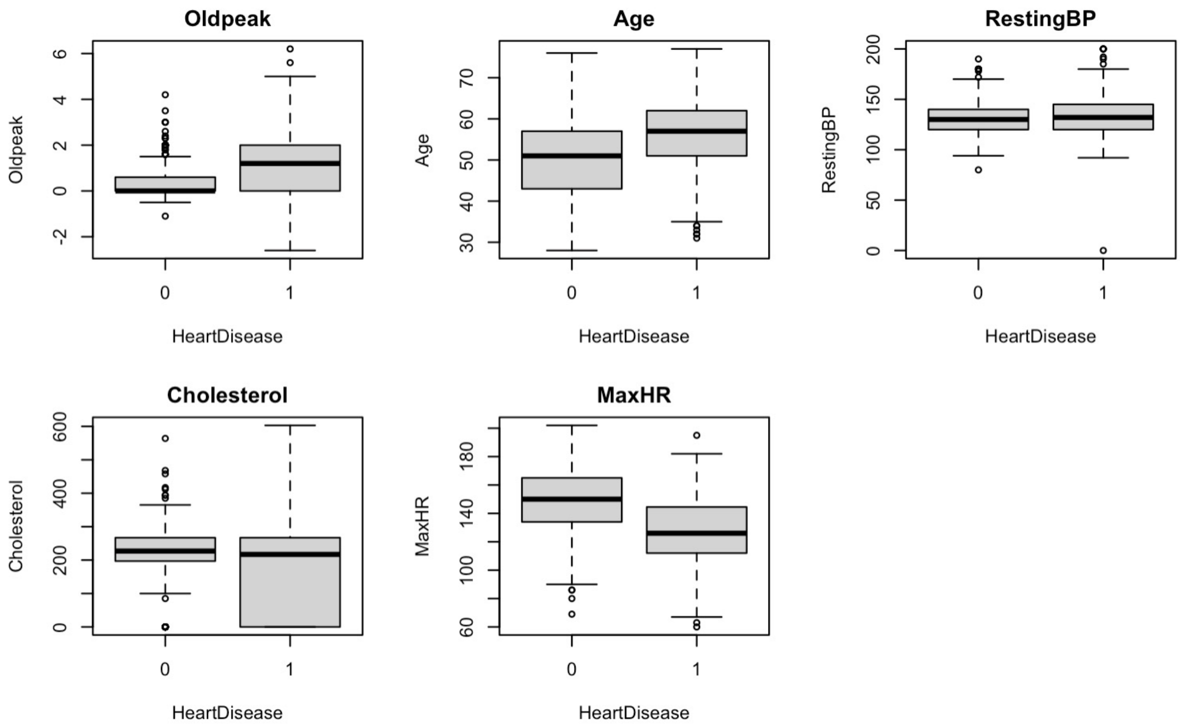

Figure 2: Boxplots of continuous variables

3 Parametric Methods

Firstly,we randomly select 20% of the full dataset as our test set, 80% of the full dataset as our training set. We perform model selection by AIC and BIC using stepwise greedy search. The optimal models using the two criteria are indeed very similar. The only difference is that AIC includes the variable Age. The phenomenon matches the fact that AIC tends to select more complex model for predictive purpose and BIC tends to select simpler model for inference purpose.

```
Call:  glm2(formula = HeartDisease ~ Age + Sex + ChestPainType + Cholesterol + 
    FastingBS + ExerciseAngina + Oldpeak + ST_Slope, family = binomial, 
    data = train)

Coefficients:
     (Intercept)               Age              SexM  ChestPainTypeATA  ChestPainTypeNAP   ChestPainTypeTA       Cholesterol        FastingBS1   ExerciseAnginaY  
       -2.345773          0.032929          1.573780         -1.880328         -1.914777         -1.265012         -0.003304          1.124661          0.846213  
         Oldpeak      ST_SlopeFlat        ST_SlopeUp  
        0.331429          1.447423         -1.055584  

Degrees of Freedom: 732 Total (i.e. Null);  721 Residual
Null Deviance:	    1008 
Residual Deviance: 482 	AIC: 506
```

Figure 3: Model selected by AIC

```
Call:  glm2(formula = HeartDisease ~ Sex + ChestPainType + Cholesterol + 
    FastingBS + ExerciseAngina + Oldpeak + ST_Slope, family = binomial, 
    data = train)

Coefficients:
     (Intercept)              SexM  ChestPainTypeATA  ChestPainTypeNAP   ChestPainTypeTA       Cholesterol        FastingBS1   ExerciseAnginaY           Oldpeak  
       -0.600940          1.571540         -1.893830         -1.893506         -1.223074         -0.003431          1.205071          0.925725          0.371055  
    ST_SlopeFlat        ST_SlopeUp  
        1.407633         -1.095983  

Degrees of Freedom: 732 Total (i.e. Null);  722 Residual
Null Deviance:	    1008 
Residual Deviance: 488.2 	AIC: 510.2
```

Figure 4: Model selected by BIC

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -2.345773   0.946304  -2.479 0.013180 *  
Age               0.032929   0.013328   2.471 0.013489 *  
SexM              1.573780   0.302183   5.208 1.91e-07 ***
ChestPainTypeATA -1.880328   0.363088  -5.179 2.23e-07 ***
ChestPainTypeNAP -1.914777   0.291904  -6.560 5.39e-11 ***
ChestPainTypeTA  -1.265012   0.478048  -2.646 0.008140 ** 
Cholesterol      -0.003304   0.001119  -2.951 0.003164 ** 
FastingBS1        1.124661   0.305880   3.677 0.000236 ***
ExerciseAnginaY   0.846213   0.259132   3.266 0.001092 ** 
Oldpeak           0.331429   0.126205   2.626 0.008636 ** 
ST_SlopeFlat      1.447423   0.483307   2.995 0.002746 ** 
ST_SlopeUp       -1.055584   0.501855  -2.103 0.035434 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1008.47  on 732  degrees of freedom
Residual deviance:  482.02  on 721  degrees of freedom
AIC: 506.02
```

Figure 5: Full regression summary

We are more interesting in the predictive purposes of the model, therefore we adopt the chosen model using AIC. The regression summary is presented on the previous page, all the variables are statistically significant under a 1% significance level. The residual deviance and degree of freedom ratio is 0.6685 which is roughly equal to one, so the model provides an acceptable fit.

We applied leave-one-out cross validation in our training set and obtain a MSE of 0.1063 which corresponds to an accuracy of 0.8937. In order to mimic the real life prediction scenario, we applied our model to the test dataset. We constructed a confusion matrix based on the model provided using the stepwise selection. The false positive rate is $\frac{12}{12+69} = 0.1481$ and the false negative rate is $\frac{9}{9+94} = 0.0874$. The overall prediction accuracy is 0.8859 which is very similar to our LOOCV result. Overall, it indicates that the logistic regression is a good fit to the dataset.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 69  9
         1 12 94

               Accuracy : 0.8859
                 95% CI : (0.8308, 0.9279)
```

4 Generalized Addictive Model

In this section, we constructed a Generalized Addictive Model (GAM) to predict the probability that an individual has Heart Disease based on the physical information collected.

The partial prediction plot and likelihood ratio test were used to fit a best model to the dataset.

Firstly, we fitted a logistic GAM model considering all numeric variables as non-parametric components. The summary of coefficients is shown below.

```
Family: binomial
Link function: logit

Formula:
HeartDisease ~ Sex + ChestPainType + RestingECG + ExerciseAngina +
    ST_Slope + FastingBS + s(Age) + s(RestingBP) + s(Cholesterol) +
    s(MaxHR) + s(Oldpeak)

Parametric coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.21567    0.58914  -2.063 0.039068 *
SexM              1.65732    0.31539   5.255 1.48e-07 ***
ChestPainTypeATA -1.91936    0.38129  -5.034 4.81e-07 ***
ChestPainTypeNAP -1.85303    0.30389  -6.098 1.08e-09 ***
ChestPainTypeTA  -1.08990    0.49978  -2.181 0.029200 *
RestingECGLVH    -0.07401    0.30948  -0.239 0.810994
RestingECGST     -0.23242    0.34545  -0.673 0.501058
ExerciseAnginaY   0.93565    0.27792   3.367 0.000761 ***
ST_SlopeFlat      1.75887    0.52598   3.344 0.000826 ***
ST_SlopeUp       -0.74146    0.54666  -1.356 0.174992
FastingBS1        1.02574    0.32432   3.163 0.001563 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                 edf Ref.df Chi.sq  p-value
s(Age)         1.000  1.000  4.481 0.034285 *
s(RestingBP)   1.000  1.000  1.237 0.266025
s(Cholesterol) 3.122  3.813 18.586 0.000873 ***
s(MaxHR)       3.567  4.477  3.906 0.414979
s(Oldpeak)     2.436  3.127 11.590 0.010484 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.609   Deviance explained = 55.5%
UBRE = -0.32682  Scale est. = 1         n = 733
```

Figure 6: GAM model1 for Heart Disease Data

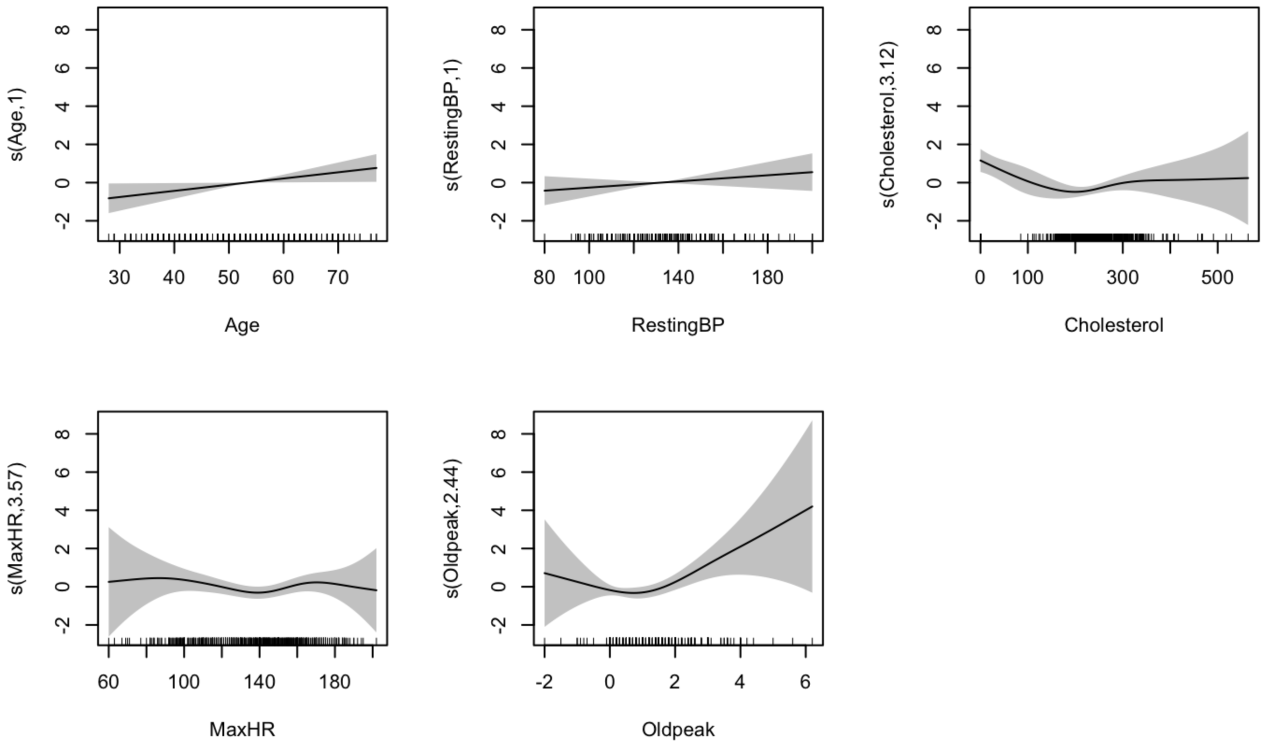

Figure 7: Partial prediction plots for GAM model1 for Heart Disease Data

We modified the model based on the summary of coefficients and the partial prediction plots for each nonparametric component in GAM model1. The parametric components of levels of RestingECG are not significant, since the p-value of coefficients of RestingECGLVH (0.810994) and RestingECGST (0.501058) are both greater than 0.05. The component plot for RestingBP and MaxHR suggest they have no influence to the response since the estimated fit lines are both close to the horizontal line on the x-axis together with large p-value for both smooth terms.

Based on this analysis, we removed RestingECG, Resting BP, MaxHR from the model and fitted the GAM model2. The summary of coefficients is shown below.

```
Family: binomial
Link function: logit

Formula:
HeartDisease ~ Sex + ChestPainType + ExerciseAngina + ST_Slope +
    FastingBS + Age + s(Cholesterol) + s(Oldpeak)

Parametric coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.08737    0.93677  -3.296 0.000982 ***
SexM              1.62257    0.31233   5.195 2.05e-07 ***
ChestPainTypeATA -1.84561    0.37076  -4.978 6.43e-07 ***
ChestPainTypeNAP -1.82407    0.29887  -6.103 1.04e-09 ***
ChestPainTypeTA  -1.10366    0.49818  -2.215 0.026735 *
ExerciseAnginaY   0.93544    0.26962   3.469 0.000522 ***
ST_SlopeFlat      1.73866    0.51738   3.361 0.000778 ***
ST_SlopeUp       -0.76785    0.53495  -1.435 0.151187
FastingBS1        1.00465    0.31993   3.140 0.001689 **
Age               0.03420    0.01356   2.521 0.011696 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                 edf Ref.df Chi.sq  p-value
s(Cholesterol) 3.172  3.874  20.13 0.000482 ***
s(Oldpeak)     2.556  3.281  12.35 0.008912 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.607   Deviance explained = 54.7%
UBRE = -0.33403  Scale est. = 1         n = 733
```

Figure 8: GAM model2 for Heart Disease Data

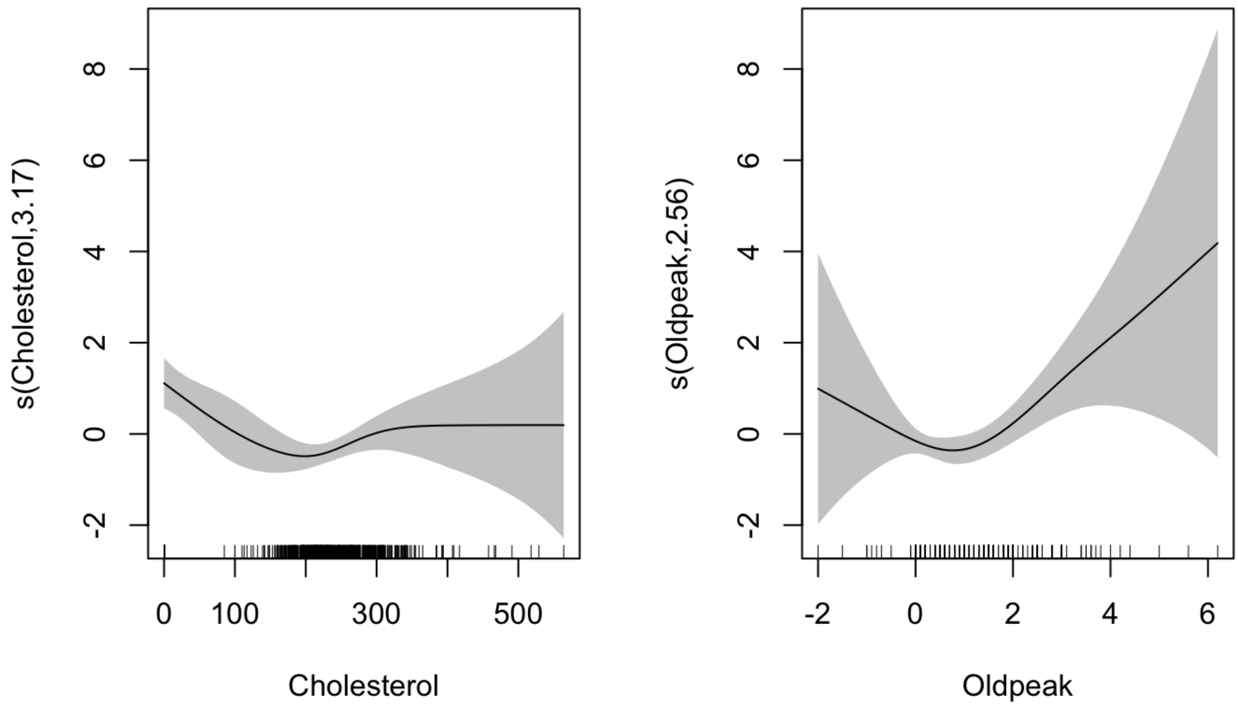

Figure 9: Partial prediction plots for GAM model2 for Heart Disease Data

Based on the summary of coefficients and the partial prediction plots, the variables in GAM model2 are all significant and the inclusion of nonparametric coefficients for Cholesterol and Oldpeak is reasonable.

```
Analysis of Deviance Table

Model 1: HeartDisease ~ Sex + ChestPainType + ExerciseAngina + ST_Slope +
    FastingBS + Age + s(Cholesterol) + s(Oldpeak)
Model 2: HeartDisease ~ Sex + ChestPainType + RestingECG + ExerciseAngina +
    ST_Slope + FastingBS + s(Age) + s(RestingBP) + s(Cholesterol) +
    s(MaxHR) + s(Oldpeak)
  Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
1    715.84     456.70
2    708.58     449.19 7.2627   7.5065   0.4057
```

Figure 10: Likelihood ratio test for GAM model1 and GAM model2

To compare GAM model1 and GAM model2, we conducted likelihood ratio test. The p-value for ANOVA

5 Bayesian Network

In this section, we constructed a Bayesian Network to predict the probability that an individual has heart disease conditional on various explanatory variables.

5.1 Structural Learning

As out data contains both categorical variables and continuous variables, hybrid structural learning is adopted. The summary of the learnt structure is as follows:

```
##   Bayesian network learned via Hybrid methods
##
##   model:
##     [Age][Sex][ChestPainType][FastingBS][RestingECG][RestingBP|Age]
##     [Cholesterol|Sex][MaxHR|Age][ExerciseAngina|ChestPainType]
##     [ST_Slope|ExerciseAngina][Oldpeak|ST_Slope][HeartDisease|ST_Slope]
```

```
## nodes:                                  12
## arcs:                                   7
##   undirected arcs:                      0
##   directed arcs:                        7
## average markov blanket size:            1.17
## average neighbourhood size:             1.17
## average branching factor:               0.58
##
## learning algorithm:                     Max-Min Hill-Climbing
## constraint-based method:                Max-Min Parent Children
## conditional independence test:          Mutual Information (cond. Gauss.)
## score-based method:                     Hill-Climbing
## score:                                  BIC (cond. Gauss.)
## alpha threshold:                        0.05
## penalization coefficient:               2.604743
## tests used in the learning procedure:   725
## optimized:                              TRUE
```

From the summary, the global joint distribution can be factorized as the following:

$$Pr(Age)Pr(Sex)Pr(ChestPainType)Pr(FastingBS)Pr(RestingECG)Pr(RestingBP|Age)$$
$$Pr(Cholesterol|Sex)Pr(MaxHR|Age)Pr(ExerciseAngina|ChestPainType)$$
$$Pr(ST\_Slope|ExerciseAngina)Pr(Oldpeak|ST\_Slope)Pr(HeartDisease|ST\_Slope)$$

The resulting directed acyclic graph (DAG) with continuous variables shaded in blue is as follows.

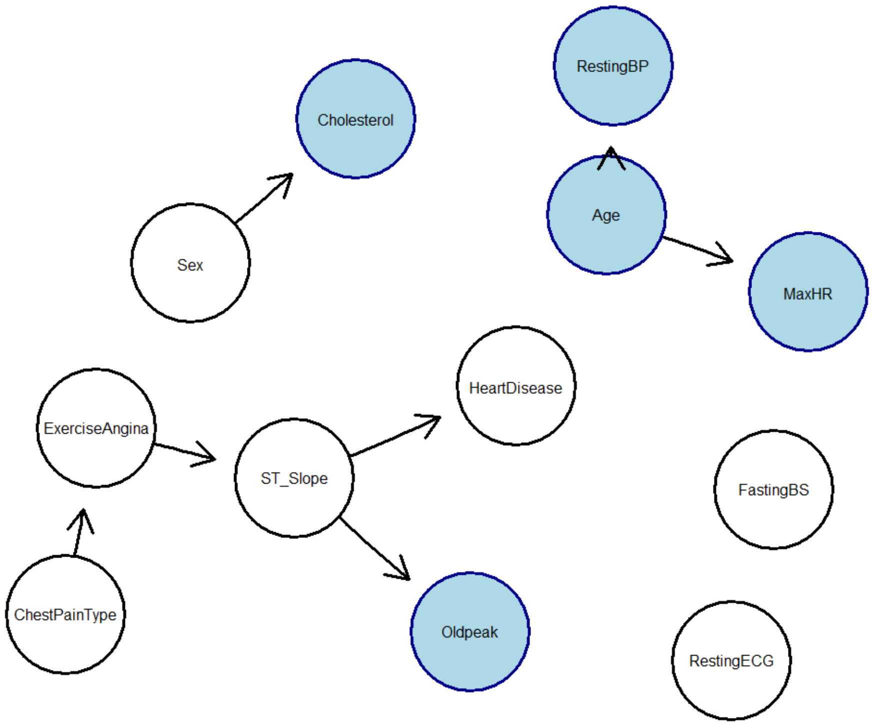

Figure 11: DAG for Heart Disease Data

From the DAG, one can see that the Markov Blanket for the target node $HeartDisease$ is $\{ST\_Slope\}$, which indicates that ST_Slope enough to perform inference on $HeartDisease$.

5.2 Parameter Learning

The resulting conditional probability table for node $HeartDisease$ is summarized as follow:

```
##   Bayesian network parameters
##
```

```
##   Parameters of node HeartDisease (multinomial distribution)
##
## Conditional probability table:
##
##             ST_Slope
## HeartDisease      Down      Flat        Up
##            0 0.1428571 0.1562500 0.8356164
##            1 0.8571429 0.8437500 0.1643836
```

From the fitted result, one can see that an individual is predicted to have heart failure with around 0.85 probability conditional on his ST_SLope is down or flat.

Using the learnt Bayesian network parameters, the resulting ROC curve is as follows and the AUC is 0.8037.

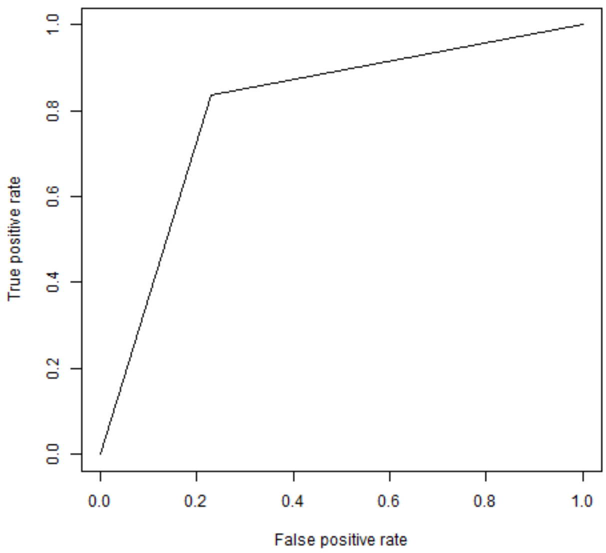

Figure 12: ROC for Bayesian Network