Logistic Regression & Support Vector Machines

PROJECT # 2

Deenadayalamuthu | DS 510 – Introduction to Data Science | 05-Dec-2017

Overview of the Dataset:

The data set heart.csv contains information about 303 patients from four different hospitals. The original data set had about 76 attributes of which only 14 are used available for research purpose. The data set contains information such as age, sex, chest pain type, cholesterol, fasting blood sugar level and other various attributes for each patient.

Exploratory Data Analysis & Data preparation:

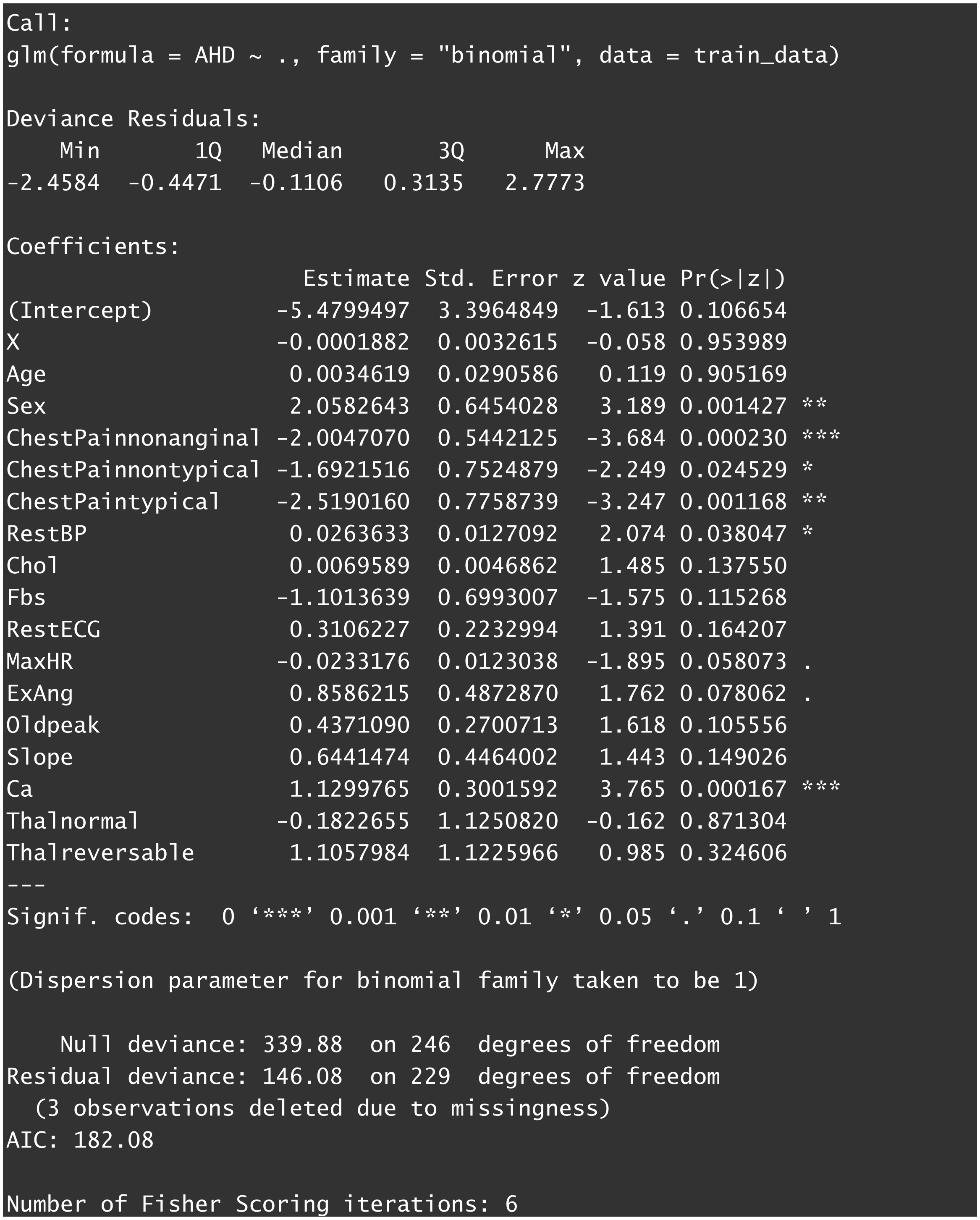
The data set contains lot of missing values which are distinguished by -9.0 as given in the heart.names file. Classifying whether a patient has heart disease is done by using logistic regression and support vector machine (SVM). Before creating the model the data is split into training data and test data. Model is created based on the training data and predictions are made using the test data.

```
# Loading the data set to R
proj_data <- read.csv("Heart.csv")

# Exploring the data set
head(proj_data)
summary(proj_data)
str(proj_data)


# Splitting up training data and test data
train_data <- proj_data[1:250, ]
test_data <- proj_data[251:303, ]
```

Building Logistic Regression model:

A logistic regression model is built using all the variables in the data set.

```
log_model <- glm(AHD ~ ., family = "binomial", data = train_data)
summary(log_model)
```



```
Call:
glm(formula = AHD ~ ., family = "binomial", data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4584  -0.4471  -0.1106   0.3135   2.7773

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.4799497  3.3964849  -1.613 0.106654
X                   -0.0001882  0.0032615  -0.058 0.953989
Age                  0.0034619  0.0290586   0.119 0.905169
Sex                  2.0582643  0.6454028   3.189 0.001427 **
ChestPainnonanginal -2.0047070  0.5442125  -3.684 0.000230 ***
ChestPainnontypical -1.6921516  0.7524879  -2.249 0.024529 *
ChestPaintypical    -2.5190160  0.7758739  -3.247 0.001168 **
RestBP               0.0263633  0.0127092   2.074 0.038047 *
Chol                 0.0069589  0.0046862   1.485 0.137550
Fbs                 -1.1013639  0.6993007  -1.575 0.115268
RestECG              0.3106227  0.2232994   1.391 0.164207
MaxHR               -0.0233176  0.0123038  -1.895 0.058073 .
ExAng                0.8586215  0.4872870   1.762 0.078062 .
Oldpeak              0.4371090  0.2700713   1.618 0.105556
Slope                0.6441474  0.4464002   1.443 0.149026
Ca                   1.1299765  0.3001592   3.765 0.000167 ***
Thalnormal          -0.1822655  1.1250820  -0.162 0.871304
Thalreversable       1.1057984  1.1225966   0.985 0.324606
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 339.88  on 246  degrees of freedom
Residual deviance: 146.08  on 229  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 182.08

Number of Fisher Scoring iterations: 6
```

Predictions based on the model:

Predictions are made based on the mode created, a confusion matrix is been created to calculate the accuracy of the model

```
mod_pred <- predict(log_model, newdata = test_data, type = "response")
cm <- table(test_data$AHD, mod_pred > 0.5)
accuracy <- sum(diag(cm))/sum(cm)
```

Accuracy of this model is calculated as 76%. From the model we can clearly see only the variables sex, chest pain, rest bp, ca, maximum heart rate, exang are significant.

Other Models:

I have also created various other models using different combinations of variables to choose the best possible model with higher accuracy.

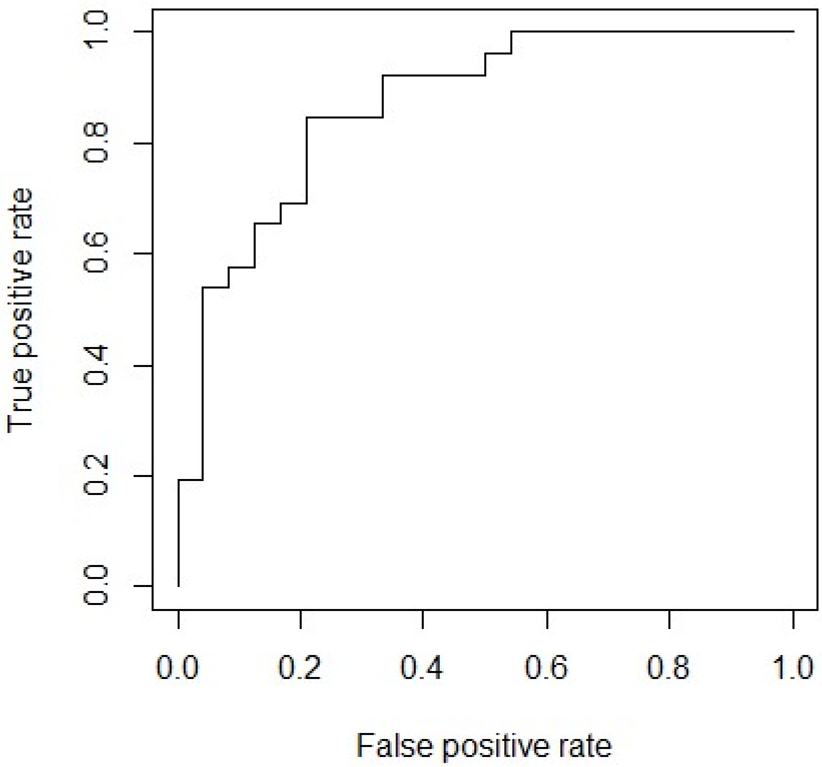
Model	Variables	Confusion Matrix	Accuracy
Model 1	All Variables	FALSE TRUE No 21 3 Yes 9 17	76%
Model 2	Age, Sex, Chest Pain, Rest BP, Ca	FALSE TRUE No 19 5 Yes 10 17	71%
Model 3	Chest Pain	FALSE TRUE No 19 7 Yes 8 19	72%
Model 4	Chest Pain + Ca	FALSE TRUE No 17 7 Yes 7 20	73%
Model 5	Age, Sex, Chest Pain, Rest BP, MaxHR, ExAng, Ca	FALSE TRUE No 20 4 Yes 10 17	73%
Model 6	Sex, Chest Pain, Rest BP, MaxHR, ExAng, Ca	FALSE TRUE No 20 4 Yes 10 17	73%

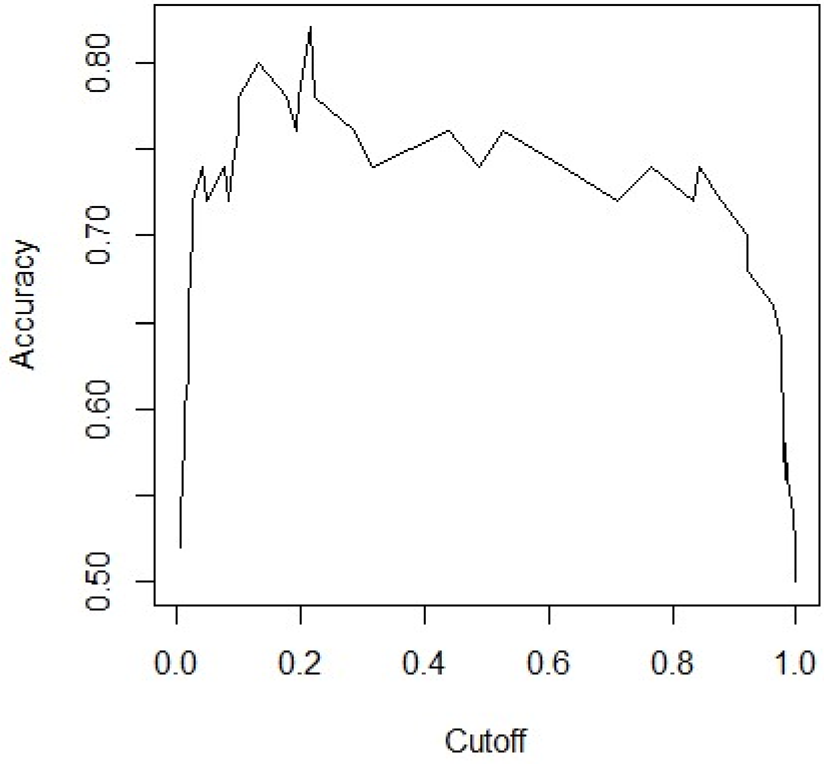
Of all the models the model using all the 14 variables for classifying whether the person has heart disease got highest accuracy.

Evaluation of the better model:

A ROCR charts are created to see the performance of the model and to see the accuracy of the model at various cut off points.

```
library(ROCR)
m_pred <- prediction(mod_pred, test_data$AHD)
eval <- performance(m_pred, "acc")
plot(eval)
roc <- performance(m_pred, "tpr", "fpr")
plot(roc)
```





From the above chart it is clear for the cut off probability of 0.5 the accuracy of the model will be 0.75

Support Vector Machine Model:

Similar to logistic regression model I have built various SVM models using different combinations of the variables to find a better model to classify whether the patient has got hear disease or not.

Preparing the data for SVM:

First the package to build a SVM model 'E1071' is loaded. Then the data set is split into two as training data set and test data set. A model is built using all the variables on the training data set. And predictions are made using the test data set.

```
library(e1071)

# Loading the data set to R
proj_data_svm <- read.csv("Heart.csv")

# Exploring the data set
head(proj_data_svm)
summary(proj_data_svm)
str(proj_data_svm)

# Splitting up training data and test data
train_data_svm <- proj_data_svm[1:250, ]
test_data_svm <- proj_data_svm[251:303, ]
test_data_svm <- na.omit(test_data_svm)
```

Building the model (SVM):

```
svm_proj_mod1 <- svm(AHD ~ ., data=train_data_svm)
summary(svm_proj_mod1)
```

Prediction based on the model:

```
svm_proj_pred1 <- predict(svm_proj_mod1, newdata=test_data_svm)
```

Confusion matrix and Accuracy of the model:

```
svm_cm1 <- table(svm_proj_pred1, test_data_svm$AHD)
svm_cm1
acc_svm1 <- sum(diag(svm_cm1))/sum(svm_cm1)
acc_svm1 # 78%
```

All SVM models:

Model	Variables	Confusion Matrix	Accuracy
Model 1	All Variables	No Yes No 20 7 Yes 4 19	78%
Model 3	Chest Pain	No Yes No 17 8 Yes 7 18	70%
Model 4	Chest Pain + Ca	No Yes No 17 8 Yes 7 18	70%
Model 5	Age, Sex, Chest Pain, Rest BP, MaxHR, ExAng, Ca	No Yes No 20 8 Yes 4 18	76%
Model 6	Sex, Chest Pain, Rest BP, MaxHR, ExAng, Ca	No Yes No 21 8 Yes 3 18	78%

Better Model in Support Vector Machine:

Model 5 seems to be a better SVM model with an accuracy of 78% using minimum attributes.

Summary of better SVM model:

```
Call:
svm(formula = AHD ~ Sex + ChestPain + RestBP + MaxHR + ExAng + Ca, data
= train_data_svm)
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.1111111
Number of Support Vectors:  125
 ( 62 63 )
Number of Classes:  2
Levels:
 No Yes
```

Codes for better SVM model:

```
# Building Model & Prediction
svm_proj_mod5 <- svm(AHD ~ Sex + ChestPain + RestBP +
                     MaxHR + ExAng + Ca, data=train_data_svm)
summary(svm_proj_mod5)
svm_proj_pred5 <- predict(svm_proj_mod5, newdata=test_data_svm)

svm_proj_pred5
# confusion matrix
svm_cm5 <- table(svm_proj_pred5, test_data_svm$AHD)
svm_cm5
acc_svm5 <- sum(diag(svm_cm5))/sum(svm_cm5)
acc_svm5 # 78%
```