Medical Treatment-Heart Health Data

Issues

Data set of Heart Health Data consists of numerical characteristics of 18 factors in which variable “delay days” is a continuous variable given in fractions of days until the person sought medical treatment. We build the logistic model to predict whether a person seeks medical treatment in three different cases.

i. Built a logistic model to predict whether a person seeks medical treatment in 2 days or less (“1”) or takes longer than 2 days to seek medical treatment (“0”).
ii. Logistic model differ if it were to predict whether a person seeks medical treatment on or less than the cohort average delay days (“1”), or takes longer than the average number of days to seek medical treatment (“0”)
iii. Logistic model differ if it were to predict whether a person seeks medical treatment on or less than 1 day (“1”) or takes longer than 1 day to seek medical treatment (“0”)

Findings

Build the logistic model on Training data by splitting the given datasets into Training (70%) and Testing data (30%). And calculated the accuracy and error of the model by using the testing data. From the analysis-palpitations, cough and DOE are more significant when compared to the other factors (as p-value is less than 0.05). By using the Testing data, we find the accuracy (61.6%) and error (38.3%) of the model.

We created a ROC curve which in general suggests the performance of our model. From our results we got ROC-AUC=0.640. AUC is area under the curve ROC, considering the value of AUC, which is 0.640, suggests that the performance of our logistic model is just satisfactory.

Discussions

The dataset consists of 406 samples, we analyse the given dataset and establish the relationship between the variables. By looking through the model summary we determine the which variables are significant. From ROC (receiving operator characteristic) curve and the AUC (area under curve) value, we can infer the performance of our logistic model. Depending on the value we can infer whether the performance of our logistic model is satisfactory or not.

Appendix A: Method

We imported the .csv fie into the R- studio, imported the readxl, pROC packages. In order to do the analysis, we created the new factor named by considering as 1 if the delaydays are less than 1 or as 0. And created a new subset of original dataset by removing the delaydays factor and seen the number of rows and number of columns in the final data set.

Split the dataset into two parts-Training and Testing dataset (70% and 30%). Build the logistic model for delay variable and used the model to generate predictions using the testing data. Created a ROC curve by utilizing the generalized logistic model and evaluating our predictions. Calculated the AUC and utilize this metric to assess the effectiveness of our model. Build the confusion matrix and from that Accuracy and Misclassification error.

Appendix B: Results

We have built up a logistic model and below we will seen the result of three different case.

Case 1. Logistic model to predict whether a person seeks medical treatment in 2 days or less ("1") or takes longer than 2 days to seek medical treatment ("0").

```
> lm<-glm(delay~.,data=training,family='binomial')
> summary(lm)

Call:
glm(formula = delay ~ ., family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1784  -0.9932  -0.6458   1.0861   1.8313

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8196779  1.5105506   0.543  0.58738
ID            0.0004464  0.0012911   0.346  0.72952
Age           0.0124536  0.0119480   1.042  0.29726
```

```
Gender        0.1472851  0.2708836   0.544  0.58663
Ethnicity    -0.8489442  0.4379129  -1.939  0.05255 .
Marital       0.1269170  0.2203136   0.576  0.56456
Livewith     -0.4106740  0.3291038  -1.248  0.21208
Education     0.0032693  0.0920740   0.036  0.97167
palpitations  0.3757675  0.1615390   2.326  0.02001 *
orthopnea    -0.0830985  0.1457399  -0.570  0.56855
chestpain     0.0518232  0.1579488   0.328  0.74284
nausea       -0.0459340  0.1652839  -0.278  0.78108
cough        -0.3041879  0.1410510  -2.157  0.03104 *
fatigue      -0.0063682  0.1826256  -0.035  0.97218
dyspnea       0.1603186  0.1632900   0.982  0.32620
edema        -0.2697654  0.1582453  -1.705  0.08824 .
PND          -0.1211662  0.1425351  -0.850  0.39528
tightshoes    0.0552957  0.1682207   0.329  0.74238
weightgain    0.2498663  0.1407973   1.775  0.07596 .
DOE          -0.4561893  0.1604278  -2.844  0.00446 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 384.66  on 280  degrees of freedom
Residual deviance: 344.68  on 261  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 384.68

Number of Fisher Scoring iterations: 5
> #Confusion Matrix
> tab<-table(Prediction=pre1,Actual=testing$delay)
> tab
          Actual
Prediction  0  1
         0 50 30
         1 16 24
>
> #Accuracy,Missclassification error
> Accuracy<-sum(diag(tab))/sum(tab)
> Accuracy
[1] 0.6166667
> M_error<-1-Accuracy
> M_error
[1] 0.3833333
```

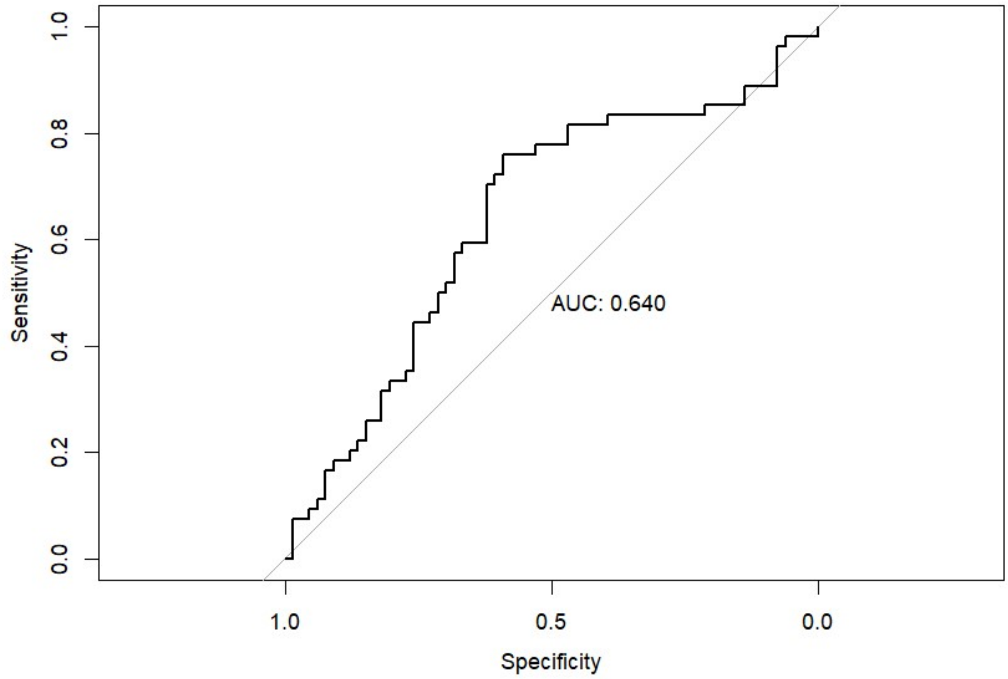

Figure 1 ROC curve and AUC for case 1.

Case 2. Logistic model differ if it were to predict whether a person seeks medical treatment on or less than the cohort average delay days (“1”), or takes longer than the average number of days to seek medical treatment (“0”)

```
> #logistic model
> lm<-glm(delay~.,data=training,family='binomial')
> summary(lm)

Call:
glm(formula = delay ~ ., family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0956  -1.1386   0.6427   0.8463   1.5632

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.771937   1.506608   1.840   0.0658 .
ID           -0.002276   0.001324  -1.719   0.0856 .
Age          -0.002013   0.012000  -0.168   0.8668
Gender       -0.081611   0.284535  -0.287   0.7742
Ethnicity    -0.264745   0.259163  -1.022   0.3070
Marital      -0.087946   0.225228  -0.390   0.6962
Livewith     -0.153416   0.345171  -0.444   0.6567
Education     0.058755   0.098862   0.594   0.5523
```

```
palpitations  0.001791   0.171427   0.010   0.9917
orthopnea    -0.091899   0.155108  -0.592   0.5535
chestpain    -0.023552   0.163488  -0.144   0.8855
nausea       -0.299769   0.163237  -1.836   0.0663 .
cough        -0.063337   0.145816  -0.434   0.6640
fatigue       0.137614   0.185417   0.742   0.4580
dyspnea       0.098166   0.174012   0.564   0.5727
edema        -0.384304   0.162690  -2.362   0.0182 *
PND          -0.155778   0.149571  -1.042   0.2976
tightshoes    0.091525   0.171874   0.533   0.5944
weightgain    0.280861   0.149532   1.878   0.0603 .
DOE          -0.262741   0.171960  -1.528   0.1265
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 346.14  on 280  degrees of freedom
Residual deviance: 319.42  on 261  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 359.42

Number of Fisher Scoring iterations: 4
> #Confusion Matrix
> tab<-table(Prediction=pre1,Actual=testing$delay)
> tab
          Actual
Prediction  0  1
         0  3  6
         1 30 81
>
> #Accuracy,Missclassification error
> Accuracy<-sum(diag(tab))/sum(tab)
> Accuracy
[1] 0.7
> M_error<-1-Accuracy
> M_error
[1] 0.3
```

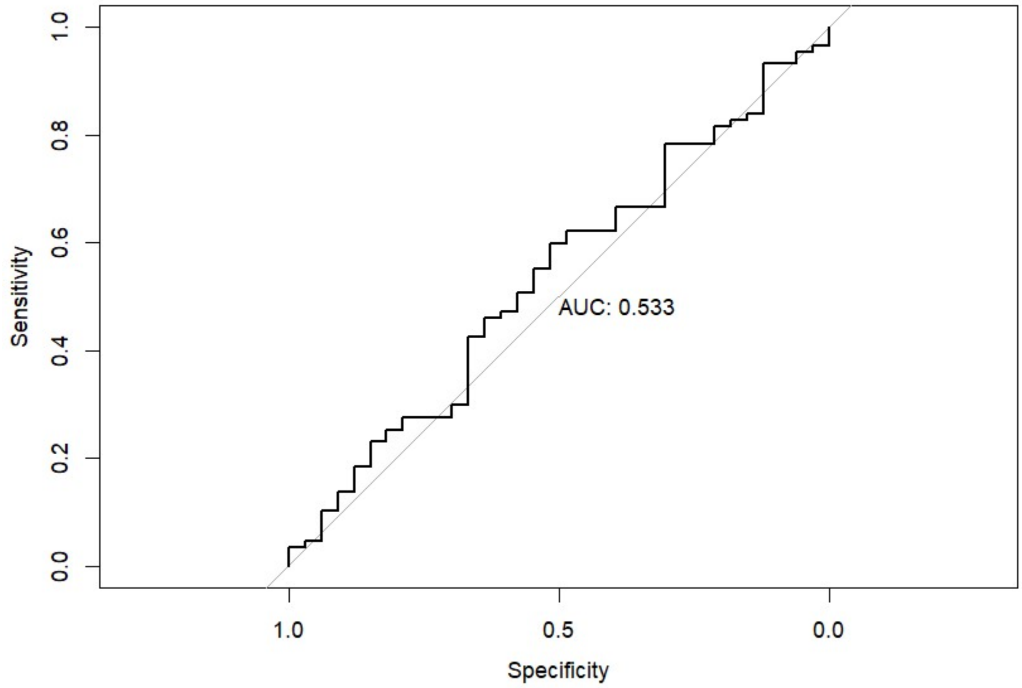

Figure 2 ROC curve and AUC for case 2.

Case 3. Logistic model differ if it were to predict whether a person seeks medical treatment on or less than 1 day (“1”) or takes longer than 1 day to seek medical treatment (“0”)

```
> #logistic model
> lm<-glm(delay~.,data=training,family='binomial')
> summary(lm)

Call:
glm(formula = delay ~ ., family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6362  -0.8833  -0.6131   1.1435   2.2006

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.015720   1.765013   1.142   0.2534
ID            0.001807   0.001360   1.329   0.1839
Age           0.003846   0.013230   0.291   0.7713
Gender        0.149796   0.289092   0.518   0.6043
Ethnicity    -1.236629   0.680362  -1.818   0.0691 .
Marital      -0.241117   0.244457  -0.986   0.3240
Livewith     -0.727954   0.357122  -2.038   0.0415 *
Education     0.053146   0.097705   0.544   0.5865
palpitations  0.225910   0.170614   1.324   0.1855
```

```
orthopnea    -0.264622   0.155942  -1.697   0.0897 .
chestpain    -0.336235   0.181914  -1.848   0.0646 .
nausea        0.191905   0.175223   1.095   0.2734
cough        -0.165668   0.153788  -1.077   0.2814
fatigue      -0.007183   0.189804  -0.038   0.9698
dyspnea       0.160929   0.175798   0.915   0.3600
edema        -0.228801   0.169325  -1.351   0.1766
PND           0.114553   0.152217   0.753   0.4517
tightshoes   -0.033053   0.184276  -0.179   0.8576
weightgain    0.127438   0.147877   0.862   0.3888
DOE          -0.345825   0.166482  -2.077   0.0378 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 344.48  on 280  degrees of freedom
Residual deviance: 310.45  on 261  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 350.45

Number of Fisher Scoring iterations: 6
> #Confusion Matrix
> tab<-table(Prediction=pre1,Actual=testing$delay)
> tab
          Actual
Prediction  0  1
         0 74 36
         1  9  1
>
> #Accuracy,Missclassification error
> Accuracy<-sum(diag(tab))/sum(tab)
> Accuracy
[1] 0.625
> M_error<-1-Accuracy
> M_error
[1] 0.375
```

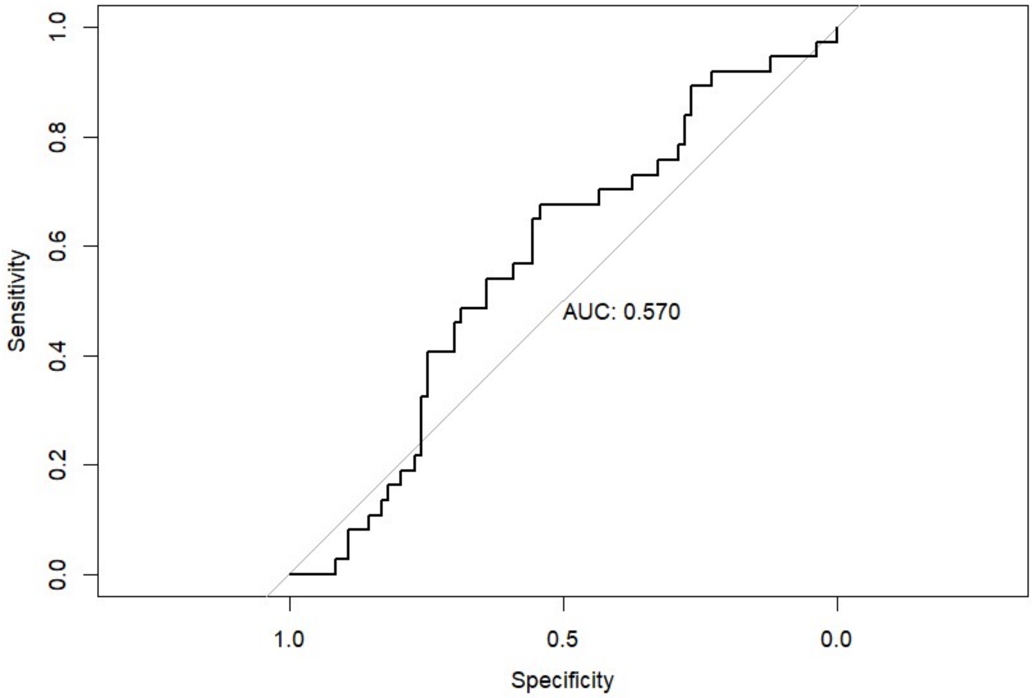

Figure 3 ROC curve and AUC for case 3.

Code

Case 1.

```
install.packages('readxl')

library(readxl)

install.packages('pROC')

library(pROC)

file <-"E:\\Assignments\\MTH 522\\Project 2\\heart-health-data.xls"

data <- read_excel(file, sheet = 1)

data

str(data)
```

```
summary(data)

colnames(data)

ncol(data)

nrow(data)


data$delay<-ifelse(data$delaydays<2,1,0)

colnames(data)

ncol(data)

nrow(data)

str(data)

ncol(data)

nrow(data)

summary(data)


#subset of original dataset by removing delaydays column

data1 <- subset(data,select = -delaydays)

str(data1)

colnames(data1)

ncol(data1)

nrow(data1)


#Spliting the data
```

```
set.seed(222)
div<-sample(2,nrow(data1),replace=T,prob=c(0.7,0.3))
training<-data1[div==1,]
testing<-data1[div==2,]
nrow(training)
nrow(testing)
training

#logistic model
lm<-glm(delay~.,data=training,family='binomial')
summary(lm)

#Prediction
pre<-predict(lm, testing, type='response')
pre

#ROC curve & AUC value
ROC <- roc(testing$delay,pre)
plot(ROC , print.auc= TRUE)

#Confusion Matrix
pre1<-ifelse(pre>0.5,1,0)
```

```
pre1
table(pre1)
tab<-table(Prediction=pre1,Actual=testing$delay)
tab
#Accuracy,Missclassification error
Accuracy<-sum(diag(tab))/sum(tab)
Accuracy
M_error<-1-Accuracy
M_error
```

Case 2.

```
install.packages('readxl')
library(readxl)
install.packages('pROC')
library(pROC)
file <-"E:\\Assignments\\MTH 522\\Project 2\\heart-health-data.xls"
data <- read_excel(file, sheet = 1)
data
str(data)
colnames(data)
ncol(data)
nrow(data)
```

```
#mean for delaydays
mean_d<-mean(data$delaydays,na.rm=TRUE)
mean_d
ncol(data)
data$delay<-ifelse(data$delaydays<mean_d,1,0)
ncol(data)
colnames(data)
data1<-subset(data,select = -delaydays)
colnames(data1)
ncol(data1)

#Spliting the data
set.seed(222)
div<-sample(2,nrow(data1),replace=T,prob=c(0.7,0.3))
training<-data1[div==1,]
testing<-data1[div==2,]
nrow(training)
nrow(testing)
training

#logistic model
```

```
lm<-glm(delay~.,data=training,family='binomial')
summary(lm)

#Prediction
pre<-predict(lm, testing, type='response')
pre

#ROC curve & AUC value
ROC <- roc(testing$delay,pre)
plot(ROC , print.auc= TRUE)

#Confusion Matrix
pre1<-ifelse(pre>0.5,1,0)
pre1
table(pre1)
tab<-table(Prediction=pre1,Actual=testing$delay)
tab
#Accuracy,Missclassification error
Accuracy<-sum(diag(tab))/sum(tab)
Accuracy
M_error<-1-Accuracy
M_error
```

Case 3.

```
install.packages('readxl')

library(readxl)

install.packages('pROC')

library(pROC)

file <-"E:\\Assignments\\MTH 522\\Project 2\\heart-health-data.xls"

data <- read_excel(file, sheet = 1)

data

str(data)

summary(data)

colnames(data)

ncol(data)

nrow(data)


data$delay<-ifelse(data$delaydays<1,1,0)

colnames(data)

ncol(data)

nrow(data)

str(data)

ncol(data)

nrow(data)

summary(data)
```

```
#subset of original dataset by removing delaydays column
data1 <- subset(data,select = -delaydays)
str(data1)
colnames(data1)
ncol(data1)
nrow(data1)

#Spliting the data
set.seed(222)
div<-sample(2,nrow(data1),replace=T,prob=c(0.7,0.3))
training<-data1[div==1,]
testing<-data1[div==2,]
nrow(training)
nrow(testing)
training

#logistic model
lm<-glm(delay~.,data=training,family='binomial')
summary(lm)

#Prediction
```

```
pre<-predict(lm, testing, type='response')
pre

#ROC curve & AUC value
ROC <- roc(testing$delay,pre)
plot(ROC , print.auc= TRUE)

#Confusion Matrix
pre1<-ifelse(pre>0.5,1,0)
pre1
table(pre1)
tab<-table(Prediction=pre1,Actual=testing$delay)
tab

#Accuracy,Missclassification error
Accuracy<-sum(diag(tab))/sum(tab)
Accuracy
M_error<-1-Accuracy
M_error
```