Logistic Model for Heart Health Data

The Issue:

We have a dataset which deals with Heart Health Data, having 18 factors, of which delay days is one of the continuous variables given in fraction of days until the person sought medical treatment. We need to build a logistic model to predict whether a person seeks medical treatment in three different ways.

One is considering people seeking treatment in 2 days or less as "1" and longer than 2 days as "0". Second method is calculating cohort average delay and taking less than average value as "1" and longer than average value as "0". Third method is considering people seeking treatment in less than 1 day as "1" and longer than 1 day as "0".

Findings:

We have 18 variables, for which we are building logistic model such that delay days is depending on the remaining 17 variables. We can infer that these variables are having statistical significance or not, by observing their corresponding p-value and z-value.

From the model we built, we can infer that only cough(p-value=0.0153) is having statistical significance to some extent. Variable that has p-value less than 0.05 are generally considered statistically significant that is the association between predictor variable and dependent variable is not due to chance. So from our result we can say that there is a significant association between cough and delay days.

Coming to the remaining variables ID, age, gender, ethnicity, marital, livewith, education, palpitations, orthopnea, chestpain, nausea, fatigue, dyspnea, edema, PND, tightshoes, weightgain and DOE, if observed in our result, p-values are greater than 0.05, which suggests us that these variables have no significance.

We created a ROC curve which in general suggests the performance of our model. From our results we got ROC-AUC=0.635. AUC is area under the curve ROC, considering the value of AUC which is 0.635, suggests that the performance of our logistic model is just satisfactory.

Discussion:

In this logistic model with total 18 variables there are some general discussions that could be made:

The p-value corresponding to each factor plays an important role, if the p-value is greater than 0.05 it suggests that the variable has no significance and if the p-value is less than 0.05 it suggests that the variable is of some significance depending upon how less is the p-value compared to 0.05.

From the coefficient values we can infer that, if the value is positive then the dependent variable value increases as the value of predictor variable value increases and if the value is negative then the dependent variable value decreases as the predictor variable value increases.

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 poor.

Appendix A: Method

In order to construct the logistic model, we utilized the R programming language and implemented relevant packages such as pROC and caTools. Prior to use, we installed these packages and imported them into our code for execution. By incorporating these libraries into our analysis, we were able to effectively build the desired logistic model.

At the beginning of our analysis, we imported the .xls file into R studio and conducted an initial inspection to determine the number of rows and columns present in the dataset. Next, we separated the delay days from the rest of the variables and established it as the dependent variable. We then created a new dataset that contained all of the remaining variables, which we will use in our subsequent analyses.

To analyze the new dataset, we first partitioned it into two separate sets: training data and testing data. We then utilized the training data to construct the generalized logistic model, and applied the resulting model to generate predictions using the testing data. Our analysis of the model predictions allowed us to determine the variables that were significant.

Our next step is to create a ROC curve by utilizing the generalized logistic model and evaluating our predictions. We will calculate the AUC and utilize this metric to assess the effectiveness of our model.

We can also get the confusion matrix and from that we get accuracy and these values can be seen in console.

Appendix B: Results

From the given dataset containing 18 variables, we have build up a logistic model and below we will see the results.

The first result is when people seeking treatment in 2 days or less as "1" and longer than 2 days as "0", shown in figure 1 and figure 2.

```
Call:
glm(formula = Delayed ~ ., family = binomial(link = "logit"),
    data = train3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9960  -1.0619  -0.6225   1.0993   2.0429

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.994744   1.338181   0.743   0.4573
ID          -0.002169   0.001268  -1.710   0.0872 .
Age          0.006431   0.010737   0.599   0.5492
Gender       0.073563   0.250770   0.293   0.7693
Ethnicity   -0.153378   0.232864  -0.659   0.5101
Marital      0.063326   0.210591   0.301   0.7636
Livewith    -0.270408   0.303162  -0.892   0.3724
Education    0.089062   0.087122   1.022   0.3067
palpitations 0.195416   0.139267   1.403   0.1606
orthopnea   -0.075118   0.132894  -0.565   0.5719
chestpain    0.103470   0.144234   0.717   0.4731
nausea      -0.072799   0.151187  -0.482   0.6302
cough       -0.307403   0.126815  -2.424   0.0153 *
fatigue     -0.126532   0.157240  -0.805   0.4210
dyspnea     -0.070101   0.147143  -0.476   0.6338
edema       -0.259160   0.145488  -1.781   0.0749 .
PND         -0.139305   0.127027  -1.097   0.2728
tightshoes   0.159779   0.152087   1.051   0.2935
weightgain   0.227030   0.130464   1.740   0.0818 .
DOE         -0.111789   0.138493  -0.807   0.4196
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 442.15  on 318  degrees of freedom
Residual deviance: 412.91  on 299  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 452.91

Number of Fisher Scoring iterations: 4
```

Figure 1. Results of generalized logistic model for case 1

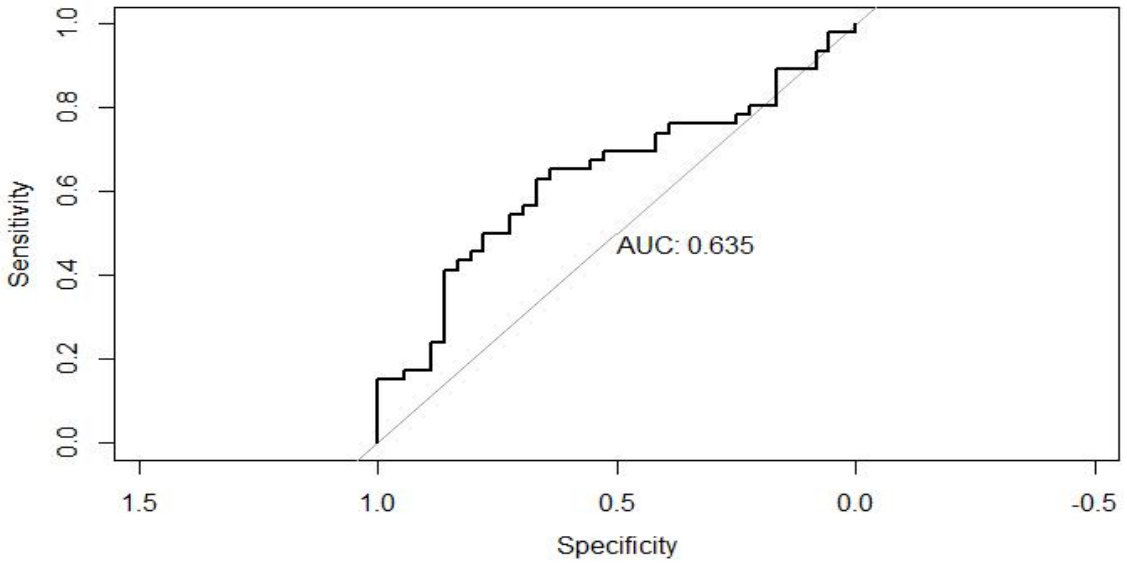

Figure 2. ROC curve and AUC for case 1.

Next case is calculating cohort average delay and taking less than average value as "1" and longer than average value as "0".

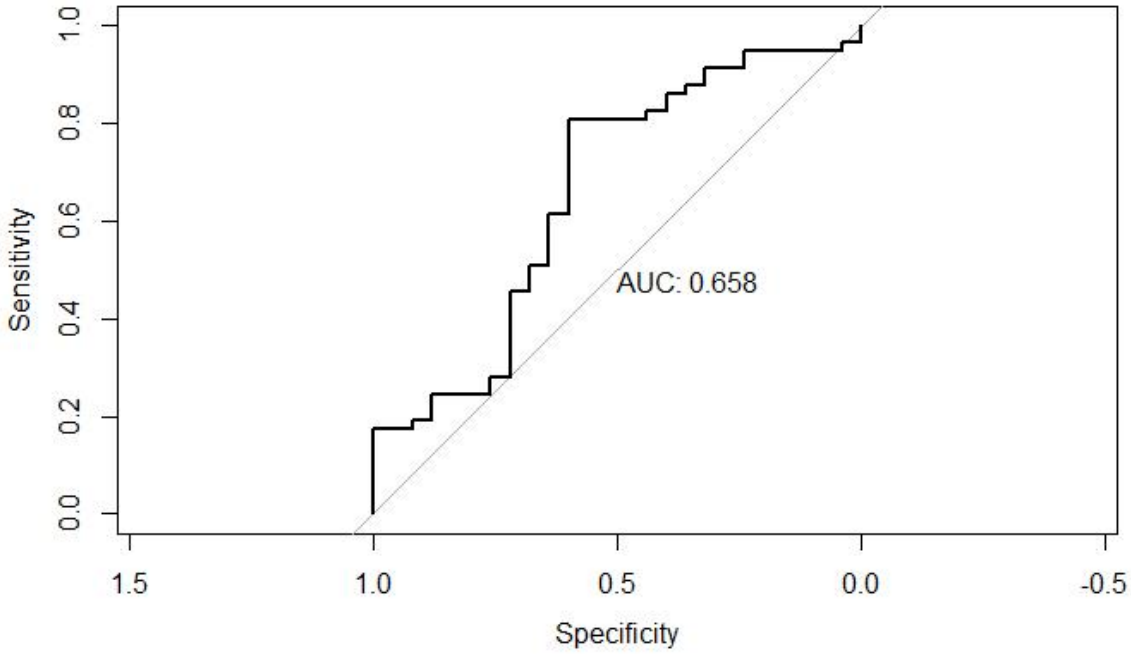

Figure 3. Roc curve and AUC for case 2.

```
Call:
glm(formula = Delayedx ~ ., family = binomial(link = "logit"),
    data = trainx)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9994  -1.2541   0.6839   0.8432   1.4234

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.605897   1.492061   1.076   0.2818
ID          -0.002352   0.001223  -1.923   0.0544 .
Age          0.003054   0.011248   0.272   0.7860
Gender      -0.022793   0.263059  -0.087   0.9310
Ethnicity   -0.071846   0.220266  -0.326   0.7443
Marital      0.102366   0.224616   0.456   0.6486
Livewith     0.020329   0.331227   0.061   0.9511
Education    0.025335   0.096869   0.262   0.7937
palpitations -0.097613   0.154716  -0.631   0.5281
orthopnea   -0.056734   0.144051  -0.394   0.6937
chestpain    0.081985   0.153196   0.535   0.5925
nausea      -0.272731   0.153860  -1.773   0.0763 .
cough       -0.072667   0.137965  -0.527   0.5984
fatigue      0.117493   0.173610   0.677   0.4986
dyspnea      0.066134   0.165906   0.399   0.6902
edema       -0.351611   0.152596  -2.304   0.0212 *
PND         -0.125158   0.131828  -0.949   0.3424
tightshoes   0.105856   0.156121   0.678   0.4977
weightgain   0.171132   0.142282   1.203   0.2291
DOE         -0.230727   0.155119  -1.487   0.1369
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 386.81  on 318  degrees of freedom
Residual deviance: 367.26  on 299  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 407.26

Number of Fisher Scoring iterations: 4
```

Figure 4. Results of generalized logistic model for case 2

Last case is considering people seeking treatment in less than 1 day as “1” and longer than 1 day as “0”.

```
Call:
glm(formula = Delayedy ~ ., family = binomial(link = "logit"),
    data = trainy)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5317  -0.8927  -0.6861   1.1884   2.1542

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.0109669  1.5146876  -0.007   0.9942
ID           -0.0003486  0.0013649  -0.255   0.7984
Age           0.0120281  0.0119470   1.007   0.3140
Gender       -0.0363213  0.2629431  -0.138   0.8901
Ethnicity    -0.3879258  0.3402509  -1.140   0.2542
Marital       0.0534642  0.2261312   0.236   0.8131
Livewith     -0.3719653  0.3205298  -1.160   0.2459
Education     0.0140195  0.0977992   0.143   0.8860
palpitations  0.0537209  0.1459577   0.368   0.7128
orthopnea    -0.3046084  0.1408632  -2.162   0.0306 *
chestpain    -0.1578701  0.1562519  -1.010   0.3123
nausea       -0.0026256  0.1683460  -0.016   0.9876
cough        -0.3405443  0.1383081  -2.462   0.0138 *
fatigue       0.1113219  0.1702127   0.654   0.5131
dyspnea       0.1613687  0.1575613   1.024   0.3058
edema        -0.2321483  0.1463388  -1.586   0.1127
PND           0.2271867  0.1348889   1.684   0.0921 .
tightshoes   -0.0316582  0.1633506  -0.194   0.8463
weightgain    0.0242004  0.1313445   0.184   0.8538
DOE          -0.1368777  0.1454559  -0.941   0.3467
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 407.83  on 319  degrees of freedom
Residual deviance: 379.27  on 300  degrees of freedom
  (4 observations deleted due to missingness)
AIC: 419.27

Number of Fisher Scoring iterations: 4
```

Figure 5. Results of generalized logistic model for case 3

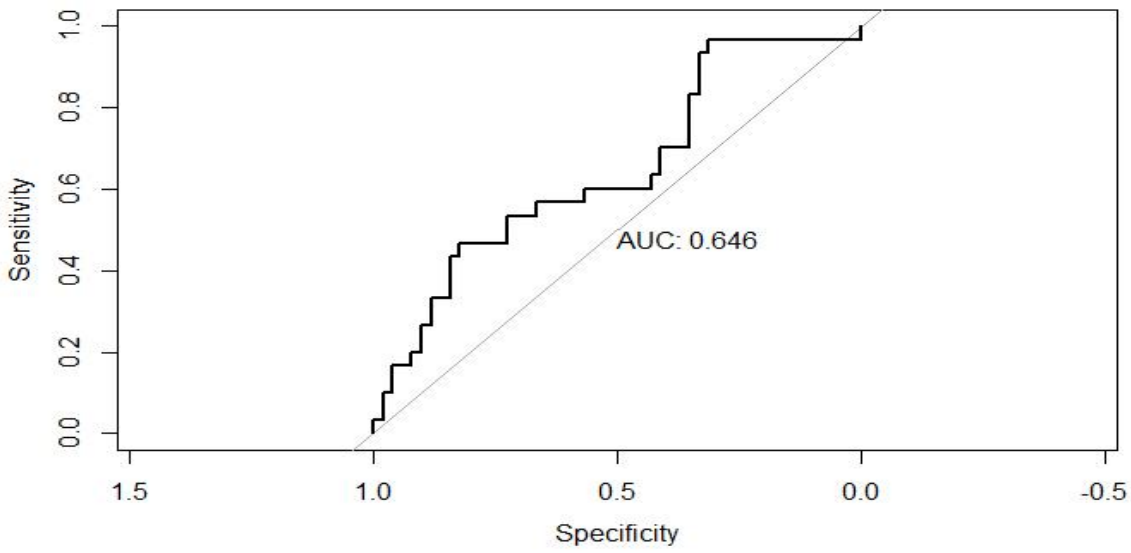

Figure 6. Roc curve and AUC for case 3.

Appendix C: Code

Case 1:

```
install.packages("caTools")
library(caTools)

#Reading input
install.packages("readxl")

file <- "C:/Users/DELL/Downloads/Heart Health Data.xls"
heartd <- readxl::read_xls(file)
View(heartd)
colnames(heartd)




heartd$Delayed <- ifelse(heartd$delaydays > 2, 0, 1)
data1 <- subset(heartd,select = -delaydays)
colnames(data1)
ncol(data1)
nrow(data1)

#splitting data
index <- sample(1:nrow(data1), size = 0.8*nrow(data1))
index
train3 <- data1[index,]
test3 <- data1[-index,]
```

```
#logistic model
model1 <- glm(Delayed ~ .,data = train3,family = binomial(link="logit"))
summary(model1)

pred <- predict(model1 , test3 , type="response")
pred

install.packages('pROC')
library(pROC)

#ROC curve & AUC value
test_roc <- roc(test3$Delayed,pred)
plot(test_roc , print.auc= TRUE)

#confusion matrix and accuracy
test3$predicted <- ifelse(pred>0.5 , 1, 0)
cm <- table(test3$Delayed , test3$predicted)
table(test3$Delayed , test3$predicted)
accuracy <- sum(diag(cm))/sum(cm)
accuracy
```

Case 2:

```
install.packages("caTools")
library(caTools)

#Reading input
```

```
install.packages("readxl")

file1 <- "C:/Users/DELL/Downloads/Heart Health Data.xls"
heartd <- readxl::read_xls(file1)
View(heartd)
colnames(heartd)

mean1 <- mean(heartd$delaydays, na.rm = TRUE)
mean1
heartd$Delayedx <- ifelse(heartd$delaydays >mean1, 0, 1)
datax <- subset(heartd,select = -delaydays)
colnames(datax)

#splitting data
index <- sample(1:nrow(datax), size = 0.8*nrow(datax))
index
trainx <- datax[index,]
testx <- datax[-index,]

#logistic model
model2 <- glm(Delayedx ~ .,data = trainx,family = binomial(link="logit"))
summary(model2)

predx <- predict(model2 , testx , type="response")
predx

install.packages('pROC')
```

```
library(pROC)

#ROC curve & AUC value
test_rocx <- roc(testx$Delayedx,predx)
plot(test_rocx , print.auc= TRUE)

#confusion matrix and accuracy
testx$predictedx <- ifelse(predx>0.5 , 1, 0)
cmx <- table(testx$Delayedx , testx$predictedx)
table(testx$Delayedx , testx$predictedx)

accuracyx <- sum(diag(cmx))/sum(cmx)
accuracy
```

Case 3:

```
install.packages("caTools")
library(caTools)

#Reading input
install.packages("readxl")

file2 <- "C:/Users/DELL/Downloads/Heart Health Data.xls"
heartd <- readxl::read_xls(file2)
View(heartd)
colnames(heartd)
```

```
heartd$Delayedy <- ifelse(heartd$delaydays >1, 0, 1)
datay <- subset(heartd,select = -delaydays)
colnames(datay)

#splitting data
indexy <- sample(1:nrow(datay), size = 0.8*nrow(datay))
indexy
trainy <- datay[indexy,]
testy <- datay[-indexy,]

#logistic model
model3 <- glm(Delayedy ~ .,data = trainy,family = binomial(link="logit"))
summary(model3)

predy <- predict(model3 , testy , type="response")
predy

install.packages('pROC')
library(pROC)

#ROC curve & AUC value
test_rocy <- roc(testy$Delayedy,predy)
plot(test_rocy , print.auc= TRUE)

#confusion matrix and accuracy
testy$predictedy <- ifelse(predy>0.5 , 1, 0)
cmy <- table(testy$Delayedy , testy$predictedy)
```

```
table(testy$Delayedy , testy$predictedy)

accuracyy <- sum(diag(cmy))/sum(cmy)
accuracyy
```