**DATA 630**

**Assignment 4: Neural Network**

**Introduction**

**Objective**

The objective of this analysis is to use a neural network that can predict the main factors the determine if a person will survive or not survive post heart attack after the first follow up. The type of analysis that is being conducted in an explanatory analysis and the specific modeling type that will be used is a Neural Network. With the explanatory analysis and neural network modeling the type of questions that could be answered are “Does age play a role in people not surviving post heart attack?” and “What makes a person more or less likely to surviving post heart attack after the first follow up”.

**Problem Domain**

The background information for problem domain of this dataset is “to study factors and time trends associated with long-term survival following acute myocardial infarction (MI) amongst the residents of the Worcester, Massachusetts.” (Applied Survival Analysis) according to the dataset. In terms of context, the idea is seeing what main factors lead to a person surviving post heart attack. Some interesting statistics are in the article below as follows:

“If treatment is delivered within three or four hours, much of the permanent muscle damage can be avoided. But if treatment is delayed beyond five or six hours, the amount of heart muscle that can be saved drops off significantly. After about 12 hours, the damage is often irreversible. Cardiac arrests can occur within the first few hours of a heart attack or during recovery. If a cardiac arrest occurs in the hospital, there is an excellent chance it can be treated. Unfortunately, the risk of sudden cardiac arrest is heightened after a heart attack, especially within the first year.”(How Many People Survive a Heart Attack?)

**Method Rationale**

The method of rationale that will be used for understanding this data will be by using a neural network. By using a neural network, every single node is also linked to the output nodes which also have weights attached to them. It is applicable for the problem at hand, because it can easily identify the variables that will affect one leaving the hospital and dying or not dying after having a heart attack. Once completed one easily view which variables are the main connectors to one surviving after a heart attack in the WHAS dataset.

**Analysis**

**Data**

The nature of this data is heart attack survival data. It was collected in Worcester, Massachusetts and named Worecester Heart Attack Study(WHAS). The official name of the dataset is Heart Disease Survival (WHAS1.DAT), size is 481 observations, 14 variables, and the source comes from Hosmer D.W. and Lemeshow, S. (1998) Applied Survival Analysis:Regression Modeling of Time to Event Data, John Wiley and Sons Inc., New York, NY. A detailed description of the variables is below:

LIST OF VARIABLES:

Variable Description Codes / Units

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ID Identification Code 1 - 481

AGE Age (per chart) years

SEX Gender 0 = Male

1 = Female

CPK Peak Cardiac Enzyme International Units (iu)

SHO Cardiogenic Shock 0 = No

Complications 1 = Yes

CHF Left Heart Failure 0 = No

Complications 1 = Yes

MIORD MI Order 0 = First

1 = Recurrent

MITYPE MI Type 1 = Q-wave

2 = Not Q-wave

3 = Indeterminate

YEAR Cohort Year 1 = 1975

2 = 1978

3 = 1981

4 = 1984

5 = 1986

6 = 1988

YRGRP Grouped Cohort Year 1 = 1975 & 1978

2 = 1981 & 1984

3 = 1986 & 1988

LENSTAY Length of Hospital Stay Days

Days in Hospital

DSTAT Discharge Status 0 = Alive

from Hospital 1 = Dead

LENFOL Total Length of Follow-up Days

from Hospital Admission

FSTAT Status as of Last 0 = Alive

Follow-up 1 = Dead

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(*SOURCE: Hosmer D.W. and Lemeshow, S. (1998) Applied Survival Analysis: Regression Modeling of Time to Event Data,*

*John Wiley and Sons Inc., New York, NY)*

**Exploratory Analysis**

**Str/ Summary/ Target Variable**

This is the str which is the data frame for the WHAS dataset. The str discusses the number of observations which is 481 and the number of variables that are in the dataset which is 14. In addition, all the items in the dataset are integers. Below is the output of the str command and the summary command. The summary only displays the min, max, 1st Qu , Median , Mean , and the 3rd Qu. All the variables are integers. Likewise, the target variable chosen for this dataset was FSTAT this variable displays whether a person leaves after their first visit dead or alive. This variable was chosen as the target variable, because I will be able to better analyze which independent variables lead to a person staying alive or being dead after a heart attack.

#view data frame

> str(whas1)

'data.frame': 481 obs. of 14 variables:

$ ID : int 1 2 3 4 5 6 7 8 9 10 ...

$ AGE : int 62 78 81 79 60 72 60 83 78 72 ...

$ SEX : int 1 1 1 1 1 0 1 1 0 1 ...

$ CPK : int 485 910 320 3290 2500 99 1200 160 66 99 ...

$ SHO : int 1 0 1 1 1 0 0 0 0 0 ...

$ CHF : int 1 1 1 1 1 0 0 0 1 0 ...

$ MIORD : int 0 1 0 1 1 0 0 0 1 0 ...

$ MITYPE : int 1 1 1 1 1 1 1 1 1 1 ...

$ YEAR : int 1 1 1 1 1 1 1 1 1 1 ...

$ YRGRP : int 1 1 1 1 1 1 1 1 1 1 ...

$ LENSTAY: int 1 1 1 1 2 2 2 3 3 3 ...

$ DSTAT : int 1 1 1 1 1 1 0 1 1 0 ...

$ LENFOL : int 1 1 1 1 2 2 2 3 3 5586 ...

$ FSTAT : int 1 1 1 1 1 1 1 1 1 0 ...

Summary

Below is the output of the summary:

#viewing data in summary

> summary(whas1)

ID AGE SEX

Min. : 1 Min. :24.00 Min. :0.0000

1st Qu.:121 1st Qu.:59.00 1st Qu.:0.0000

Median :241 Median :68.00 Median :0.0000

Mean :241 Mean :67.48 Mean :0.4033

3rd Qu.:361 3rd Qu.:77.00 3rd Qu.:1.0000

Max. :481 Max. :98.00 Max. :1.0000

CPK SHO CHF

Min. : 10.0 Min. :0.000 Min. :0.0000

1st Qu.: 270.0 1st Qu.:0.000 1st Qu.:0.0000

Median : 587.0 Median :0.000 Median :0.0000

Mean : 941.5 Mean :0.079 Mean :0.4075

3rd Qu.:1146.0 3rd Qu.:0.000 3rd Qu.:1.0000

Max. :9000.0 Max. :1.000 Max. :1.0000

MIORD MITYPE YEAR

Min. :0.0000 Min. :1.00 Min. :1.000

1st Qu.:0.0000 1st Qu.:1.00 1st Qu.:2.000

Median :0.0000 Median :1.00 Median :3.000

Mean :0.3597 Mean :1.43 Mean :3.422

3rd Qu.:1.0000 3rd Qu.:2.00 3rd Qu.:5.000

Max. :1.0000 Max. :3.00 Max. :6.000

YRGRP LENSTAY DSTAT

Min. :1.000 Min. : 1.00 Min. :0.0000

1st Qu.:1.000 1st Qu.: 8.00 1st Qu.:0.0000

Median :2.000 Median :12.00 Median :0.0000

Mean :1.971 Mean :13.86 Mean :0.1705

3rd Qu.:3.000 3rd Qu.:17.00 3rd Qu.:0.0000

Max. :3.000 Max. :71.00 Max. :1.0000

LENFOL FSTAT

Min. : 1 Min. :0.0000

1st Qu.: 150 1st Qu.:0.0000

Median :1420 Median :1.0000

Mean :1735 Mean :0.5177

3rd Qu.:2551 3rd Qu.:1.0000

Max. :5843 Max. :1.0000

Stacked Bar Chart

The stacked bar chart in the Appendix as Figure 1. displays the ages people have died or survived post heart attack. From this chart it can be said that between the ages of 70-74 many people are likely to survive post heart attack visit. Over 20% of people in that age group are likely to survive. In contrast people in the age group of 54-65 are more likely to die from a heart attack post visit. People younger than 54 but older than 86 are less likely to even experience a heart attack according to the histogram.

Histogram

The histogram in the Appendix Figure 2. shows the total percentage of people dead or alive after their first visit post heart attack. The 0 represent people who have died and the 1 represents people who are still alive. From this histogram is can be said that between 46-48 percentage of people after their first heart attack have passed away. However over 50 percent of people have still survived after their first visit post heart attack.

**Preprocessing**

For pre-processing the WHAS data set, there were no missing values, and all the values

were already integers, so the histogram and bar chart were easily made. However I did a scale command in order to scale in the 14 variables in the dataset. I used the set the seed

data command so that way I could split the data in the train and test data. For splitting the

data into a train and test set I was able to split 70% of the data into train and 30% into test. I did not remove any of the variables because my algorithm method of choice was able to handle the needed variables for analysis. This is not a linear model, so no outliers were found.

**Algorithm Intuition**:

The intuition for the Neural network is to be able to visualize all the links and relationships in the WHAS dataset as it relates to the variable FSTAT. The neural network will be able to provide the key parameters of the different types of nodes input, hidden layer, and output nodes. The logic for the input nodes is that they are linked to every single hidden layer node, and each relationship boundary has a weight labeled on it. Every single hidden layer node is also linked to the output nodes which also have weights attached to them. With the labels on the nodes in the visual one will be able to see the ‘intercepts for the activation functions.

**Model Fitting**:

For model fitting I used a neural network and I modify a code so that the WHAS data would fit into as needed. As you can see below the only variable, I decided to not use was ID because it does not benefit the analysis of the data or the performance of the model. However, the rest of the 13 variables were all used and fitted into the code below with FSTAT being the dependent variables and others being independent.

#Building neural network changed ce to sse due to only working with binary reponse

nn<-neuralnet(formula =FSTAT~AGE+SEX+CPK+SHO+CHF+MIORD+MITYPE+YEAR+YRGRP+LENSTAY+DSTAT+LENFOL,data = train.data, hidden=2, err.fct="sse", linear.output = FALSE)

#displays the shown neural network properties for evaluation

names(nn)

**Result**s

**Output**:

Output of the neural network is shown as Figure 3. in the Appendix. To interpret the model there are three main layers. The primary which has FSTAT the beginning, two concealed layer nodes, and output nodes following behind. All the individual input nodes are connected to all the individual hidden layer nodes. The lines that connect them all have a certain weight to define them. The blue circle is for the intercept activation functions and to better explain the model property in the next section describes how strong the intercepts are to the dependent variable FSTAT which will explain how the overall objective has been met.

**Model Properties**

Below are the commands and outputs used for evaluating the model properties. For the WHAS dataset I decided modify a code to display the first 10 predicted values which were predicted at a strength of 0.8868. That was the result for the neural net predicted values. Likewise, I used a results matrix as well to determine the main intercepts of the neural network. The Intercept.to.1layhid1 is-0.630000926 for the first hidden layer node. The second section display the Intercept.to.1layhid2 at 0.326335870 for the second hidden layer node. In addition, in the last section it shows the output nodes at Intercept.to.FSTAT -0.727489764, 1layhid1.to.FSTAT 0.191414413, and 1layhid2.to.FSTAT 2.978288408.

Furthermore, the error is shown at an exceptionally low 26.1428 and the steps taking by the neural network were 73. For intercept 1layhid1 the strongest intercept is DSTAT which is the discharge status from hospital and for 1layhid2 is MIROD which measure if the heart attack was a recurrence. These two intercepts are the strongest amongst the datasets and are the main predictors for if one died after by their first follow up. Also just to mention DSTAT could have been a dependent variable, however I FSTAT shows a person’s status even after they were discharged from the hospital. The objective from this neural net has been met because we now know that if a person was discharged from the hospital and or their heart was a recurrence, they are more likely to dying from the heart attack by their first follow up. The output is below.

.

nn$net.result[[1]][1:10] # first 10 predicted probabilities

[1] 0.8868924 0.8868924 0.8868924 0.8868924 0.8868924

[6] 0.8868924 0.8868924 0.2851812 0.8868924 0.8868924

> nn$result.matrix # This command shows the amount of trainings steps, the error, and the weights

[,1]

error 26.142808598

reached.threshold 0.006571881

steps 73.000000000

Intercept.to.1layhid1 -0.630000926

AGE.to.1layhid1 0.084881135

SEX.to.1layhid1 0.369549320

CPK.to.1layhid1 1.108909511

SHO.to.1layhid1 1.278840571

CHF.to.1layhid1 0.898523546

MIORD.to.1layhid1 0.227029776

MITYPE.to.1layhid1 -0.437187945

YEAR.to.1layhid1 0.697896763

YRGRP.to.1layhid1 1.132377960

LENSTAY.to.1layhid1 0.726514319

DSTAT.to.1layhid1 1.507246776

LENFOL.to.1layhid1 1.075651818

Intercept.to.1layhid2 0.326335870

AGE.to.1layhid2 0.738305896

SEX.to.1layhid2 2.226415751

CPK.to.1layhid2 0.339488374

SHO.to.1layhid2 6.336484094

CHF.to.1layhid2 2.972352437

MIORD.to.1layhid2 6.647773742

MITYPE.to.1layhid2 -0.904884699

YEAR.to.1layhid2 0.214401268

YRGRP.to.1layhid2 -0.554413435

LENSTAY.to.1layhid2 0.262330181

DSTAT.to.1layhid2 0.876237155

LENFOL.to.1layhid2 -0.458938475

Intercept.to.FSTAT -0.727489764

1layhid1.to.FSTAT -0.191414413

1layhid2.to.FSTAT 2.978288408

>

**Evaluation**:

For evaluation I used a confusion matrix and the confusion matrix as seen below has the accuracies for the train and test set. For the train set 143 individuals’ predictions are true negatives. The matrix displays those 102 predictions are true positive. Equation for correct predictions are 143+102= 245. The prediction accuracy for training set is 245/315=0.7777% or 0.78% to round. The test data set prediction was 143 individuals’ predictions are true negatives. The matrix displays those 64 predictions are true positive. Equation for correct predictions are 64+59= 123. The prediction accuracy for training set is 123/166= 0.7409% or 0.74%. The train set’s prediction is around 3% stronger than the test set. The overall accuracy for the confusion matrix is around 74 to 77 percent.

Confusion matrix

**#Model evaluation; Round the predicted probabilities**

**> mypredict<-compute(nn, nn$covariate)$net.result**

**> mypredict<-apply(mypredict, c(1), round)**

**> mypredict [1:9]**

**5 6 7 8 11 12 14 15 16**

**1 1 1 1 1 1 1 0 1**

**> #Confusion matrix for the training set**

**> table(mypredict, train.data$FSTAT, dnn =c("Predicted", "Actual"))**

**Actual**

**Predicted 0 1**

**0 143 57**

**1 13 102**

**> mean(mypredict==train.data$FSTAT)**

**[1] 0.7777778**

**> #Confusion matrix for the test set**

**> testPred <- compute(nn, test.data[, 0:14])$net.result**

**> testPred<-apply(testPred, c(1), round)**

**> table(testPred, test.data$FSTAT, dnn =c("Predicted", "Actual"))**

**Actual**

**Predicted 0 1**

**0 64 31**

**1 12 59**

**> mean(testPred==test.data$FSTAT)**

**[1] 0.7409639**

**>**

**Conclusion**:

**Summary**:

The stated objective of this analysis was to find out what the leading variables are to one dying after their first follow up after having a heart attack. The key findings were that age does play a role in one being more likely to die from a heart attack. If you are middle aged it is likely that you could die before your first follow up after having a heart attack. The other key finding was that DSTAT and MIROD are what makes a person more likely to die or survive a heart attack after the first follow up. DSTAT is your discharge status from the hospital and MIROD is if you had a heart recurrence.

**Limitations**:

The limitation of the algorithm is that is was a little harder to view the weighted lines in the neural network visual. I wanted to view all of the variables besides one, but when viewing them all it is rather crowded together their variables there are.

**Improvement Areas**:

I can improve on my neural network visualization so I can explain better from looking at the visual rather than using the results matrix to interpret. It should be easy to interpret just from looking at the visual alone. So, next time I would shorten the number of variables I use.

References

*Applied Survival Analysis Data* . Sign in to your account. (n.d.). <https://learn.umgc.edu/content/enforced/583899-027339-01-2215-GO1-9040/Dataset%20Descriptions/whas1.txt>.

Lemeshow, H. D. W. (n.d.). Applied Survival Analysis: Regression Modeling of Time to Event Data, 2nd Edition. Wiley-Interscience.

UMUC. (n.d.). Neural Network Analysis Using R. Reading.

**Appendix**

**Figure 1. Age of People that were Released after Heart Attack that are Alive vs Dead**

Chart, bar chart

Description automatically generated

**Figure 2. Percentage of People Dead or Alive after First Visit**

**Chart

Description automatically generated**

**Figure 3. Neural Network**

**Chart, diagram

Description automatically generated**