Overview of outputs of R markdown – Session 15- Assignment

a. Predict the no of comments in next H hrs
b. Use regression technique
c. Report the training accuracy and test accuracy
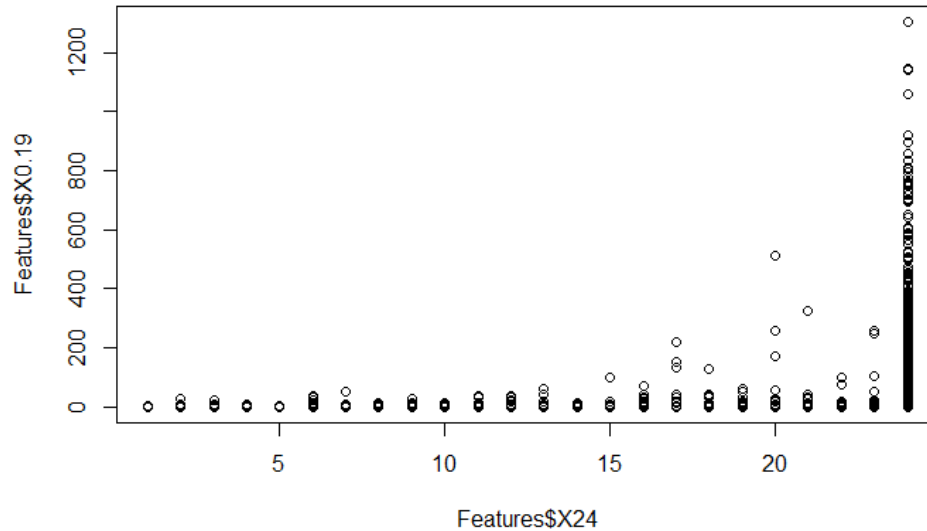
Feature-Training log regression

Over view

Attribute Information:

(39 - This describes the H hrs, for which we have the target variable/ comments received.

54 -Target Variable - Decimal Target The no of comments in next H hrs(H is given in Feature no 39).

plot(Features$X24,Features$X0.19)

The coding and the main output are given below

```
Features_Train<-Features_norm[1:28003,]
Features_Test<-Features_norm[28004:40948,]
library(neuralnet)
```

```
model<-glm(X0.19~X634995+X0+X463+X1+X0.0+X806.0+X11.291044776119403+X1.0+X70.
49513846124168+X0.0.1+X806.0.1+X7.574626865671642+X0.0.2+X69.435826365571+X0.
0.3+X76.0+X2.6044776119402986+X0.0.4+X8.50550186882253+X0.0.5+X806.0.2+X10.64
9253731343284+X1.0.1+X70.25478763764251+X.69.0+X806.0.3+X4.970149253731344,da
ta = Features_Train, family = binomial)

## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!

model

##
## Call:  glm(formula = X0.19 ~ X634995 + X0 + X463 + X1 + X0.0 + X806.0 +
##     X11.291044776119403 + X1.0 + X70.49513846124168 + X0.0.1 +
##     X806.0.1 + X7.574626865671642 + X0.0.2 + X69.435826365571 +
##     X0.0.3 + X76.0 + X2.6044776119402986 + X0.0.4 + X8.50550186882253 +
##     X0.0.5 + X806.0.2 + X10.649253731343284 + X1.0.1 + X70.25478763764251
+
##     X.69.0 + X806.0.3 + X4.970149253731344, family = binomial,
##     data = Features_Train)
##
## Coefficients:
##         (Intercept)              X634995                   X0
##             -8.1890              15.1657               0.6445
##                X463                   X1                 X0.0
##            -23.2171              -0.7102             -26.2338
##              X806.0  X11.291044776119403                 X1.0
##              2.0551             157.0660               4.4092
##  X70.49513846124168                X0.0.1             X806.0.1
##             -8.4151             -10.0358              -6.5078
##  X7.574626865671642                X0.0.2     X69.435826365571
##             10.7507              12.4957               3.7177
##              X0.0.3                X76.0  X2.6044776119402986
##            -11.5462               4.2625               1.7392
##              X0.0.4    X8.50550186882253               X0.0.5
##             -4.4118              -6.5585              18.3120
##            X806.0.2  X10.649253731343284               X1.0.1
##             -1.0349            -113.9644             -23.7730
##  X70.25478763764251                X.69.0             X806.0.3
##              8.9179               3.4798               5.6877
##  X4.970149253731344
##                  NA
##
## Degrees of Freedom: 28002 Total (i.e. Null);  27976 Residual
## Null Deviance:        734.5
## Residual Deviance: 505.6      AIC: 498.2

summary(model)

##
## Call:
```

```
## glm(formula = X0.19 ~ X634995 + X0 + X463 + X1 + X0.0 + X806.0 +
##     X11.291044776119403 + X1.0 + X70.49513846124168 + X0.0.1 +
##     X806.0.1 + X7.574626865671642 + X0.0.2 + X69.435826365571 +
##     X0.0.3 + X76.0 + X2.6044776119402986 + X0.0.4 + X8.50550186882253 +
##     X0.0.5 + X806.0.2 + X10.649253731343284 + X1.0.1 + X70.25478763764251
+
##     X.69.0 + X806.0.3 + X4.970149253731344, family = binomial,
##     data = Features_Train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59736  -0.07115  -0.05908  -0.02997   2.59350
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -8.1890     2.0025  -4.089 4.33e-05 ***
## X634995                   15.1657     7.1199   2.130  0.03317 *
## X0                         0.6445     0.7299   0.883  0.37721
## X463                     -23.2171     8.4336  -2.753  0.00591 **
## X1                        -0.7102     0.6881  -1.032  0.30200
## X0.0                     -26.2338    70.6481  -0.371  0.71039
## X806.0                     2.0551     4.4873   0.458  0.64696
## X11.291044776119403      157.0660   110.2234   1.425  0.15416
## X1.0                       4.4092    47.7658   0.092  0.92645
## X70.49513846124168        -8.4151    20.2766  -0.415  0.67813
## X0.0.1                   -10.0358    23.2964  -0.431  0.66662
## X806.0.1                  -6.5078     4.5938  -1.417  0.15658
## X7.574626865671642        10.7507    35.1815   0.306  0.75993
## X0.0.2                    12.4957    18.7866   0.665  0.50596
## X69.435826365571           3.7177     4.9712   0.748  0.45455
## X0.0.3                   -11.5462    19.7890  -0.583  0.55958
## X76.0                      4.2625     2.3504   1.813  0.06976 .
## X2.6044776119402986        1.7392     6.1798   0.281  0.77838
## X0.0.4                    -4.4118     5.1254  -0.861  0.38937
## X8.50550186882253         -6.5585     5.5094  -1.190  0.23388
## X0.0.5                    18.3120    64.8921   0.282  0.77780
## X806.0.2                  -1.0349     4.4386  -0.233  0.81564
## X10.649253731343284     -113.9644    86.4334  -1.319  0.18733
## X1.0.1                   -23.7730    34.6937  -0.685  0.49320
## X70.25478763764251         8.9179    19.8051   0.450  0.65250
## X.69.0                     3.4798     4.0256   0.864  0.38736
## X806.0.3                   5.6877     4.8682   1.168  0.24267
## X4.970149253731344             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 734.54  on 28002  degrees of freedom
## Residual deviance: 505.64  on 27976  degrees of freedom
```

```
## AIC: 498.16
##
## Number of Fisher Scoring iterations: 9

predict<-predict(model)
head(predict)

##         1         2         3         4         5         6
## -5.584117 -5.584117 -5.584117 -5.584117 -5.584117 -5.584117
```

```
fit <- glm(X0.19~X634995+X0+X463+X1+X0.0+X806.0+X11.291044776119403+X1.0+X70.
49513846124168+X0.0.1+X806.0.1+X7.574626865671642+X0.0.2+X69.435826365571+X0.
0.3+X76.0+X2.6044776119402986+X0.0.4+X8.50550186882253+X0.0.5+X806.0.2+X10.64
9253731343284+X1.0.1+X70.25478763764251+X.69.0+X806.0.3+X4.970149253731344,da
ta = Features_Train, family = binomial(link='logit'))

## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!

summary(fit)

##
## Call:
## glm(formula = X0.19 ~ X634995 + X0 + X463 + X1 + X0.0 + X806.0 +
##      X11.291044776119403 + X1.0 + X70.49513846124168 + X0.0.1 +
##      X806.0.1 + X7.574626865671642 + X0.0.2 + X69.435826365571 +
##      X0.0.3 + X76.0 + X2.6044776119402986 + X0.0.4 + X8.50550186882253 +
##      X0.0.5 + X806.0.2 + X10.649253731343284 + X1.0.1 + X70.25478763764251
+
##      X.69.0 + X806.0.3 + X4.970149253731344, family = binomial(link = "logi
t"),
##      data = Features_Train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.59736  -0.07115  -0.05908  -0.02997   2.59350
##
## Coefficients: (1 not defined because of singularities)
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -8.1890     2.0025   -4.089 4.33e-05 ***
## X634995              15.1657     7.1199    2.130  0.03317 *
## X0                    0.6445     0.7299    0.883  0.37721
## X463                -23.2171     8.4336   -2.753  0.00591 **
## X1                   -0.7102     0.6881   -1.032  0.30200
## X0.0                -26.2338    70.6481   -0.371  0.71039
```

```
## X806.0                      2.0551     4.4873    0.458   0.64696
## X11.291044776119403      157.0660   110.2234    1.425   0.15416
## X1.0                        4.4092    47.7658    0.092   0.92645
## X70.49513846124168         -8.4151    20.2766   -0.415   0.67813
## X0.0.1                     -10.0358    23.2964   -0.431   0.66662
## X806.0.1                    -6.5078     4.5938   -1.417   0.15658
## X7.574626865671642         10.7507    35.1815    0.306   0.75993
## X0.0.2                      12.4957    18.7866    0.665   0.50596
## X69.435826365571            3.7177     4.9712    0.748   0.45455
## X0.0.3                     -11.5462    19.7890   -0.583   0.55958
## X76.0                        4.2625     2.3504    1.813   0.06976 .
## X2.6044776119402986         1.7392     6.1798    0.281   0.77838
## X0.0.4                       -4.4118    5.1254   -0.861   0.38937
## X8.50550186882253           -6.5585     5.5094   -1.190   0.23388
## X0.0.5                       18.3120   64.8921    0.282   0.77780
## X806.0.2                     -1.0349    4.4386   -0.233   0.81564
## X10.649253731343284       -113.9644   86.4334   -1.319   0.18733
## X1.0.1                      -23.7730   34.6937   -0.685   0.49320
## X70.25478763764251           8.9179   19.8051    0.450   0.65250
## X.69.0                        3.4798    4.0256    0.864   0.38736
## X806.0.3                      5.6877    4.8682    1.168   0.24267
## X4.970149253731344              NA         NA       NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 734.54  on 28002  degrees of freedom
## Residual deviance: 505.64  on 27976  degrees of freedom
## AIC: 498.16
##
## Number of Fisher Scoring iterations: 9
```
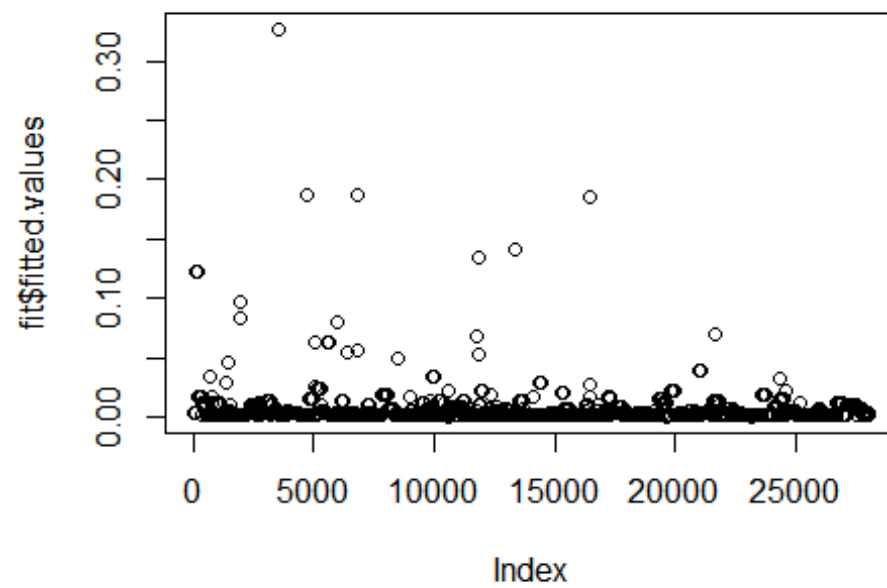
```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-2   2017-02-28
```

```
hoslem.test(Features_Train$X0.19, fitted(fit))
```
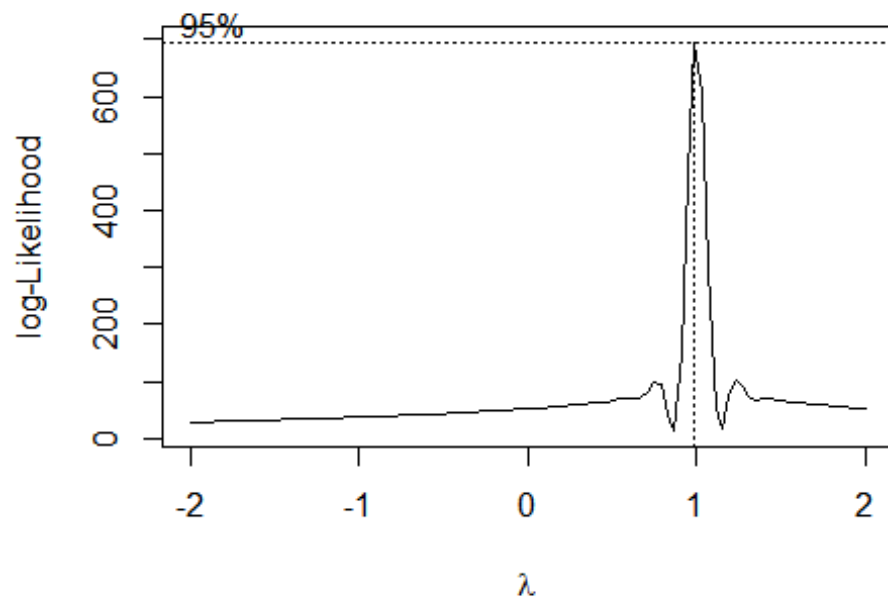
```
##
##   Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  Features_Train$X0.19, fitted(fit)
## X-squared = 14.466, df = 8, p-value = 0.07041
```
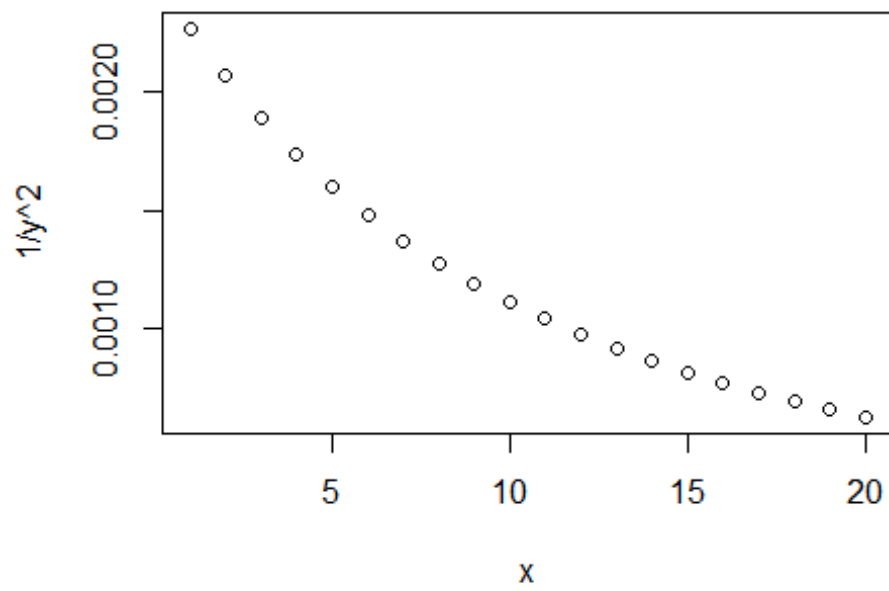
```
#plot the fitted model
plot(fit$fitted.values)
```

```
x <- 1:20
y <- 21:40
library(MASS)
boxcox(y~x)
```

```
plot(1/y^2~x)
```



```
g <- roc(X0.19~ pred, data = Features_Test)
```

```
g
```

```
plot(g)
```



```
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

#with default prob cut 0.50
Features_Test$pred_X0.19 <- ifelse(pred<0.7,'yes','no')

table(Features_Test$pred_X0.19,Features_Test$X0.19)

## 29788 0.0000000000        yes  -5.8735363
##  [ reached getOption("max.print") -- omitted 11160 rows ]
```

```
# summarize results
#confusionMatrix<- confusionMatrix(pred$predictions,pred$X0.19)
#confusionMatrix

barchart(Features_Test$pred_X0.19)
```
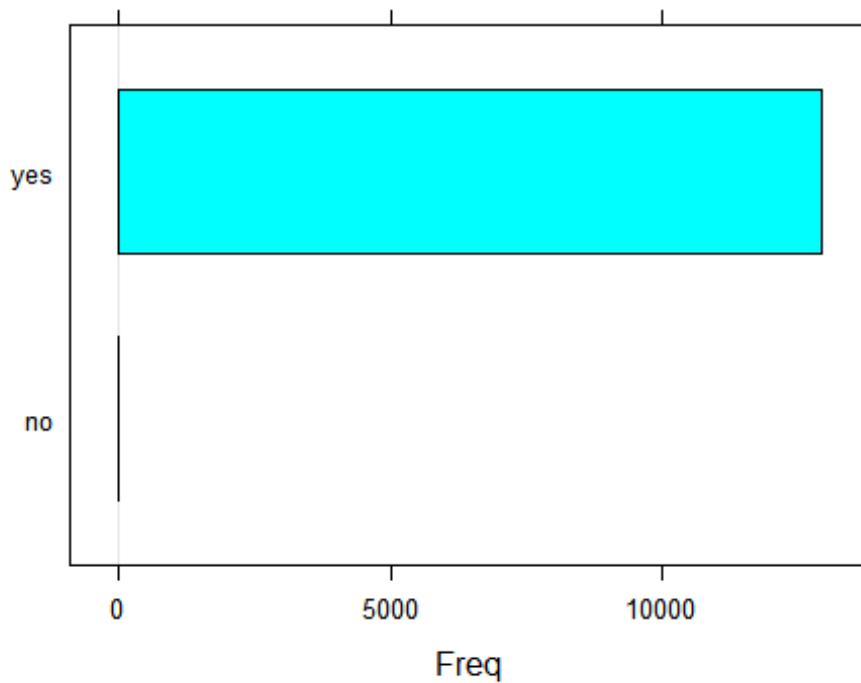

Neural Network model prediction

library(readr)

dataset <- read_delim("Dataset.zip", ";", escape_double = FALSE, trim_ws = TRUE)

View(Features_TestSet)

dim(Features_TestSet)

head(Features_TestSet)

str(Features_TestSet)

names(Features_TestSet)

```
dim(Features_Variant_1)
Features<-Features_Variant_1
names(Features)
head(Features)
head(Test_Case_1)
normalize<-function(x) {return((x-min(x))/(max(x)-min(x)))}
Features_norm<-as.data.frame(lapply(Features,normalize))
Features_norm
names(Features_norm)
Features_Train<-Features_norm[1:28003,]
Features_Test<-Features_norm[28004:40948,]
library(neuralnet)
set.seed(123)
Features_model<-neuralnet(X0_19~
X634995+X0+X463+X1+X0.0+X806.0+X11.291044776119403+X1.0+X70.49513846124168+X0.0_1+X806.
0_1+X7.574626865671642+X0.0_2+X69.435826365571+X0.0_3+X76.0+X2.6044776119402986+X0.0_4+
X8.50550186882253+X0.0_5+X806.0_2+X10.649253731343284+X1.0_1+X70.25478763764251+X.69.0+
X806.0_3+X4.970149253731344+X0.0_6+X69.85058043098057+X0_1+X0_2+X0_3+X0_4+X0_5+X65+X1
66+X2,data = Features_Train)

plot(Features_model)

model_results<-compute(Features_model,Features_Test[1:37])

predicted_X0_19<-model_results$net.result

nrow(predicted_X0_19)

head(predicted_X0_19)

round(cor( predicted_X0_19,Features_Test$X0_19),2)
```
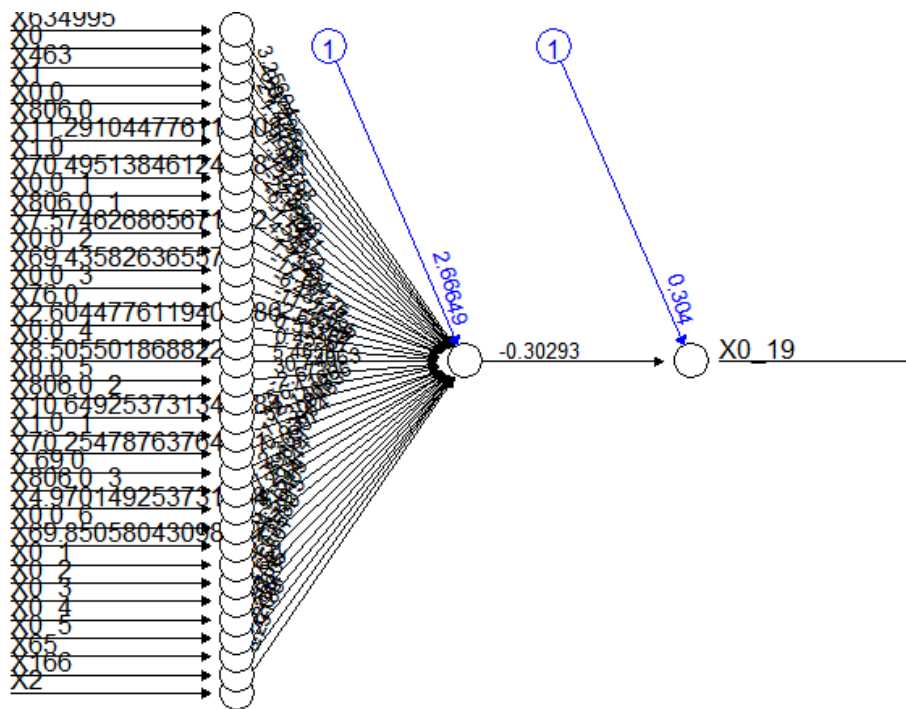
```
[1] 12945
                 [,1]
28004  0.001069433699
28005  0.001070214932
28006  0.001087235290
28007  0.001069451002
28008  0.001589321488
28009  0.001273782979
      [,1]
[1,]  0.65
```

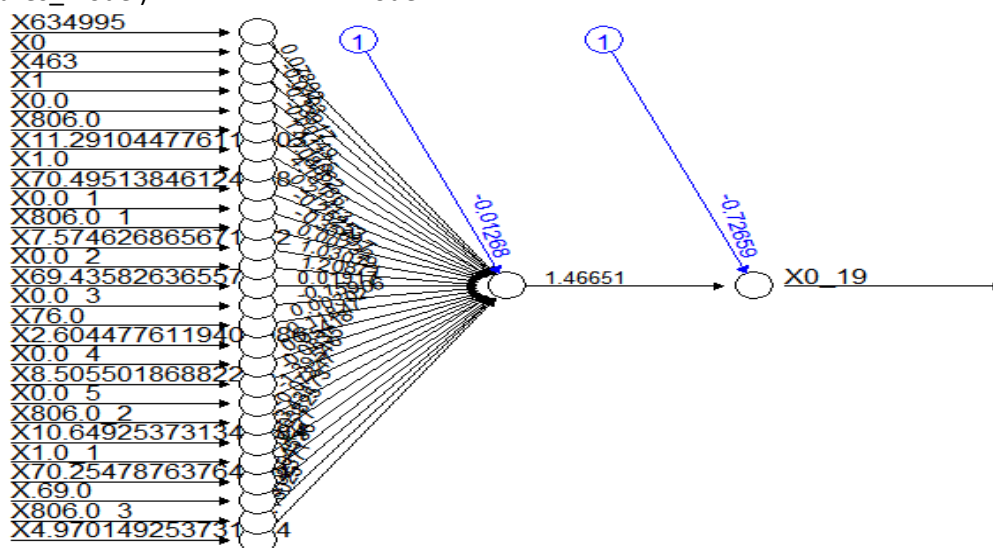**The correlation (0.65) is the highest than the others given below. Higher is better.**
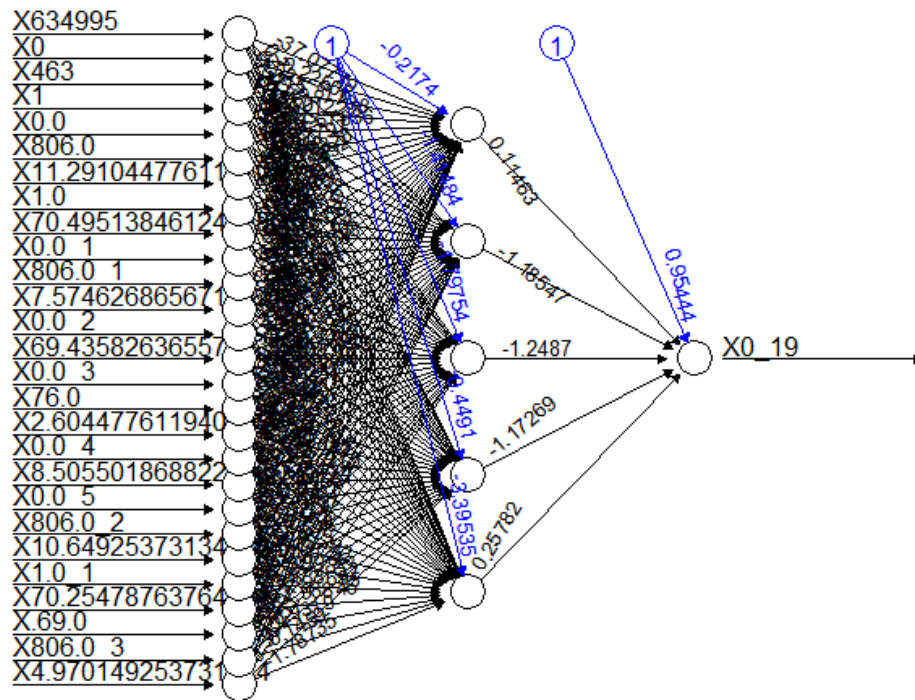
set.seed(123)

Features_model<-
neuralnet(X0_19~X634995+X0+X463+X1+X0.0+X806.0+X11.291044776119403+X1.0+X70.49513846124
168+X0.0_1+X806.0_1+X7.574626865671642+X0.0_2+X69.435826365571+X0.0_3+X76.0+X2.60447761
19402986+X0.0_4+X8.50550186882253+X0.0_5+X806.0_2+X10.649253731343284+X1.0_1+X70.254787
63764251+X.69.0+X806.0_3+X4.970149253731344,data = Features_Train)

plot(Features_model)   Model1

Model2

model_results<-compute(Features_model,Features_Test[1:27])

predicted_X0_19<-model_results$net.result

nrow(predicted_X0_19)

head(predicted_X0_19)

round(cor(predicted_X0_19,Features_Test$X0_19),2)

```
[1] 12945
                   [,1]
28004  0.002802219847
28005  0.002802219847
28006  0.002802219847
28007  0.002802219847
28008  0.002802219847
28009  0.002802219847
       [,1]
[1,]  0.34
```
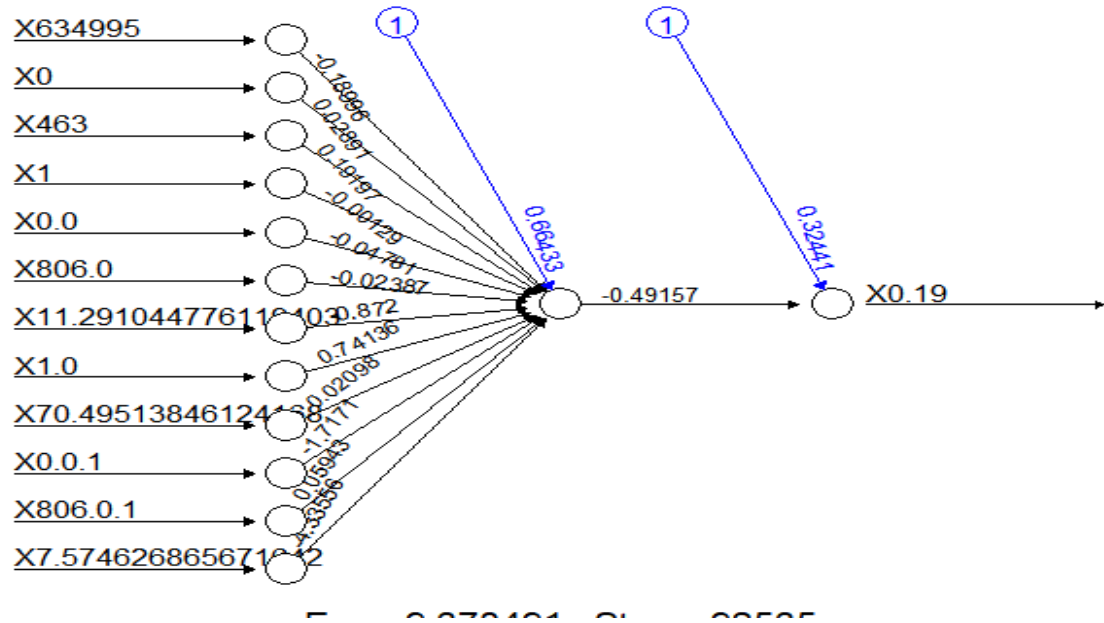
set.seed(123)

Features_model<-
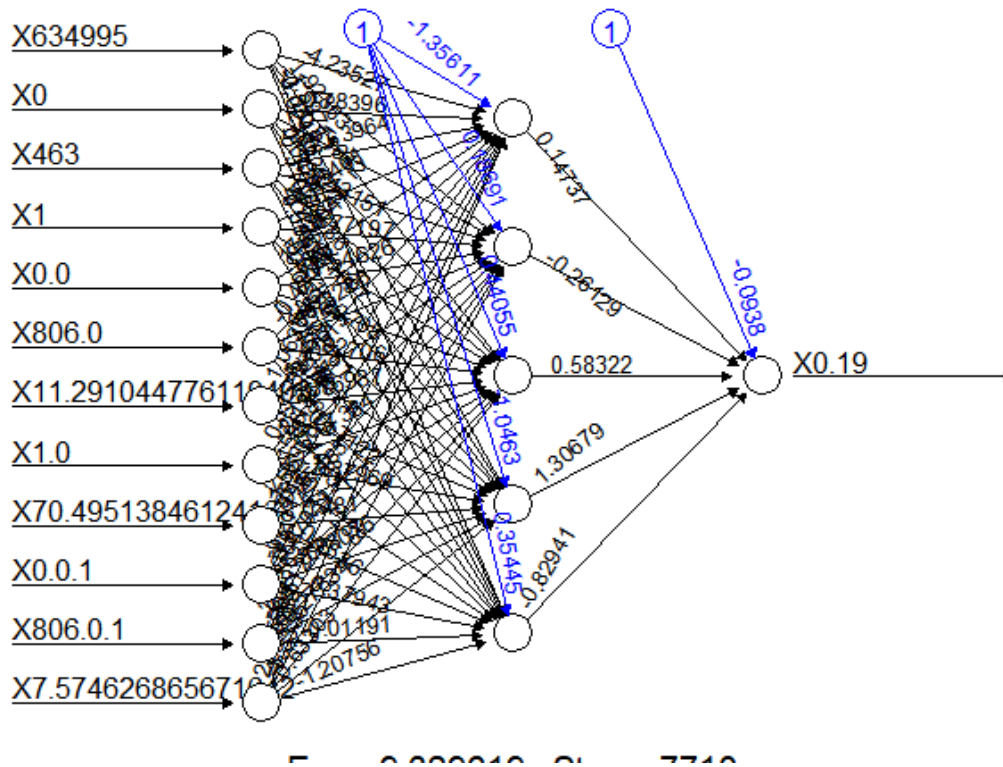neuralnet(X0.19~X634995+X0+X463+X1+X0.0+X806.0+X11.291044776119403+X1.0+X70.495138461241
68+X0.0.1+X806.0.1+X7.574626865671642,data = Features_Train)

plot(Features_model)

Model1

Model2

Model1

```
[1] 12945
                [,1]
28004  0.002660508374
28005  0.002660508374
28006  0.002660508374
28007  0.002660508374
28008  0.002660508374
28009  0.002660508374
       [,1]
[1,]  0.34
```

Model2

```
[1] 12945
                [,1]
28004  0.001997766546
28005  0.001997766546
28006  0.001997766546
28007  0.001997766546
28008  0.001997766546
28009  0.001997766546
       [,1]
[1,]  0.32
```

```
non-integer #successes in a binomial glm!
Call:  glm(formula = X0.19 ~ X634995 + X0 + X463 + X1 + X0.0 + X806.0 +
    X11.291044776119403 + X1.0 + X70.49513846124168 + X0.0.1 +
    X806.0.1 + X7.574626865671642 + X0.0.2 + X69.435826365571 +
    X0.0.3 + X76.0 + X2.6044776119402986 + X0.0.4 + X8.50550186882253 +
    X0.0.5 + X806.0.2 + X10.649253731343284 + X1.0.1 + X70.25478763764251 +
    X.69.0 + X806.0.3 + X4.970149253731344, family = binomial,
    data = Features_Train)

Coefficients:
        (Intercept)                 X634995                      X0
X463
         -8.1889873              15.1656626               0.6445092                    -
23.2171336
                 X1                    X0.0                  X806.0
X11.291044776119403
         -0.7102152             -26.2338417               2.0551480
157.0660451
               X1.0    X70.49513846124168                  X0.0.1
X806.0.1
          4.4092485              -8.4150733             -10.0357869                    -
6.5077822
 X7.574626865671642                  X0.0.2      X69.435826365571
X0.0.3
         10.7507051              12.4957414               3.7176835                    -
11.5461700
              X76.0  X2.6044776119402986                  X0.0.4
X8.50550186882253
          4.2624506               1.7391687              -4.4118131                    -
6.5584531
              X0.0.5                 X806.0.2  X10.649253731343284
X1.0.1
         18.3120490              -1.0348589            -113.9643962                    -
23.7729803
 X70.25478763764251                   X.69.0                X806.0.3
X4.970149253731344
          8.9179436               3.4798448               5.6876570
NA

Degrees of Freedom: 28002 Total (i.e. Null);  27976 Residual
Null Deviance:      734.5361
Residual Deviance: 505.6398     AIC: 498.1562
```