Module 4 CT Option 1

Didem Aykurt

2022-12-08

Logistic Regression with Banks dataset in R

LOAD the BANKS' DATA

```
banks.df <-read.csv("C:/Users/didem/OneDrive/Documents/CSUG Master DA/MIS510-
1 Data Mining_4 term/Module 4 Naive model/MIS510banks.csv", header=TRUE)
```

Show all the data in a new tab

```
View(banks.df)
```

Check to null object(result False means there isn't null)

```
is.null(banks.df)

## [1] FALSE
```

Find total observations and variables in the dataset

```
nrow(banks.df)

## [1] 20

ncol(banks.df)

## [1] 5
```

Display the first ten rows of each column

```
banks.df[1:10, ]

##    Obs Financial.Condition TotCap.Assets TotExp.Assets TotLns.Lses.Assets
## 1    1                   1           9.7          0.12               0.65
## 2    2                   1           1.0          0.11               0.62
## 3    3                   1           6.9          0.09               1.02
## 4    4                   1           5.8          0.10               0.67
## 5    5                   1           4.3          0.11               0.69
## 6    6                   1           9.1          0.13               0.74
## 7    7                   1          11.9          0.10               0.79
## 8    8                   1           8.1          0.13               0.63
## 9    9                   1           9.3          0.16               0.72
## 10  10                   1           1.1          0.16               0.57
```

Print the list in a useful column format

```
t(t(names(banks.df)))
```

```
##      [,1]
## [1,] "Obs"
## [2,] "Financial.Condition"
## [3,] "TotCap.Assets"
## [4,] "TotExp.Assets"
## [5,] "TotLns.Lses.Assets"
```

Give column name

```
colnames(banks.df)<- c(
  "obs",
  "Fcondition",
  "TotCapAsset",
  "TotExpAsset",
  "TotLnsLsesAsset"
)
```

View summary statistics of the dataset

```
summary(banks.df)
```

```
##       obs          Fcondition   TotCapAsset      TotExpAsset    
##  Min.   : 1.00   Min.   :0.0   Min.   : 1.000   Min.   :0.0700  
##  1st Qu.: 5.75   1st Qu.:0.0   1st Qu.: 7.125   1st Qu.:0.0800  
##  Median :10.50   Median :0.5   Median : 9.200   Median :0.1000  
##  Mean   :10.50   Mean   :0.5   Mean   : 9.320   Mean   :0.1045  
##  3rd Qu.:15.25   3rd Qu.:1.0   3rd Qu.:11.300   3rd Qu.:0.1200  
##  Max.   :20.00   Max.   :1.0   Max.   :20.500   Max.   :0.1600  
##  TotLnsLsesAsset 
##  Min.   :0.3000  
##  1st Qu.:0.5250  
##  Median :0.6400  
##  Mean   :0.6285  
##  3rd Qu.:0.7225  
##  Max.   :1.0200
```

Drop obs column

```
banks.df<- banks.df[,-c(1)]
```

Let’s check the frequency of bank condition

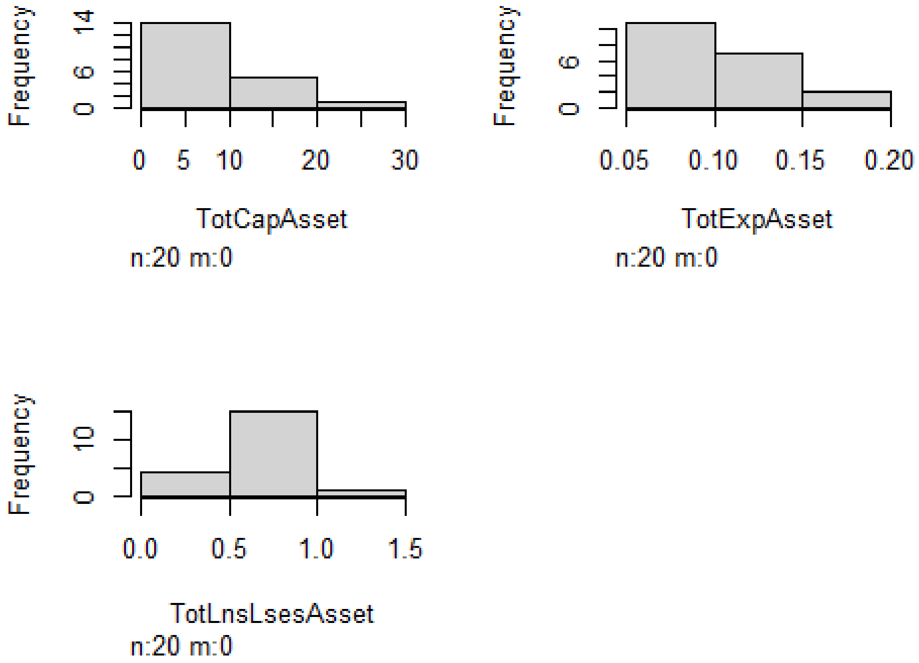
```
xtabs(~Fcondition, data = banks.df )
```

```
## Fcondition
##  0  1 
## 10 10
```

Calculate frequencies of total capital, expenses, and loan and lease assets with a histogram

```
library(Hmisc)
```

```
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

hist.data.frame(banks.df)
```



Create training and test sample dataset

```
#make this example duplicatable
set.seed(2)
#use  60% of the dataset as a training set and the remaining 40% as a testing
set
samplebanks<- sample(c(TRUE, FALSE), nrow(banks.df), replace = TRUE, prob = c
(0.6,0.4))
train<- banks.df[samplebanks, ]
test<- banks.df[!samplebanks, ]
dim(train)

## [1] 13  4

dim(test)

## [1] 7 4
```

Fit the logistic regression model

```
installed.packages('glmnet')

model<- glm(Fcondition~ TotCapAsset+TotExpAsset+TotLnsLsesAsset, family = "bi
nomial", data=train)
summary(model)


## Call:
## glm(formula = Fcondition ~ TotCapAsset + TotExpAsset + TotLnsLsesAsset,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.511e-05  -2.110e-08  -2.110e-08   2.110e-08   2.677e-05
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -702.21  706799.82  -0.001    0.999
## TotCapAsset           -14.11   16799.47  -0.001    0.999
## TotExpAsset          3468.93 5426410.22   0.001    0.999
## TotLnsLsesAsset       689.83  862598.77   0.001    0.999
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.7945e+01  on 12  degrees of freedom
## Residual deviance: 1.5390e-09  on  9  degrees of freedom
## AIC: 8
##
## Number of Fisher Scoring iterations: 25
```

Compute McFadden's R square for model

```
pscl::pR2(model)["McFadden"]

## fitting null model for pseudo-r2

## McFadden
##        1
```

Calculate the significance of predictor variable

```
caret::varImp(model)

##                      Overall
## TotCapAsset     0.0008398753
## TotExpAsset     0.0006392687
## TotLnsLsesAsset 0.0007997121
```

Compute the Variance Inflation Factor(VIF)

```
car::vif(model)
```

```
##     TotCapAsset     TotExpAsset TotLnsLsesAsset
##        5.908282        4.930755        1.862145
```

Code for using logistic regression to generate predicted probabilities

```
#calculate the probability of default for each individual in the test dataset
fcondition.prob<- predict(model, test, type="response")

#first five actual and predicted records
data.frame(actual= test$Fcondition[1:5], predict=fcondition.prob[1:5])
```

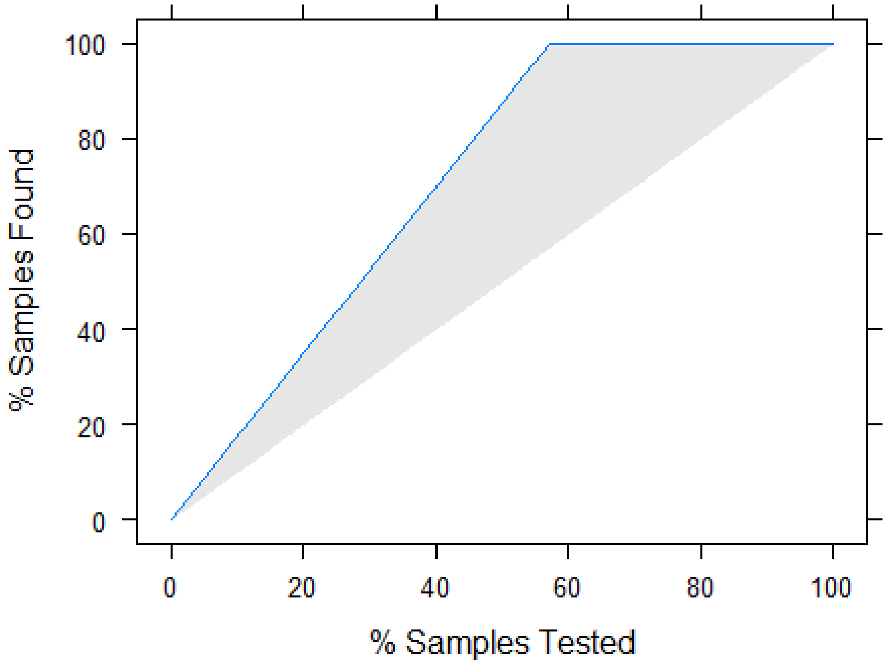
```
##    actual      predict
## 2       1 1.000000e+00
## 5       1 1.000000e+00
## 6       1 1.000000e+00
## 8       1 1.000000e+00
## 13      0 2.220446e-16
```

Analyze how well our model performs on the test dataset

```
library(caret)


## Attaching package: 'caret'

## The following object is masked from 'package:survival':
##
##     cluster

pred= round(fcondition.prob)
confusionMatrix(factor(test$Fcondition),factor(pred))

## Confusion Matrix and Statistics
##
##            Reference
## Prediction 0 1
##          0 3 0
##          1 0 4
##
##                Accuracy : 1
##                  95% CI : (0.5904, 1)
##     No Information Rate : 0.5714
##     P-Value [Acc > NIR] : 0.01989
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.0000
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 1.0000
##                Prevalence : 0.4286
##            Detection Rate : 0.4286
##      Detection Prevalence : 0.4286
##         Balanced Accuracy : 1.0000
##
##          'Positive' Class : 0
##
```

Calculate misclassification error rate

```
library(caret)
lift.test <- lift(relevel(as.factor(test$Fcondition), ref="1")~ fcondition.pr
ob, data = test)
xyplot(lift.test, plot="gain")
```



Logistic Regression

Logistic Regression approximates the possibility of the event, such as categorical variables (yes/no, male/female, online sale/ store sale, low/intermediate/high, surgery/medication/radiation..etc.). Logistic regression divides each categorical variable record into classes. The model has used variables that may include ordinal or nominal, as ordinal variables have essential order like 1=small, 2=medium, 3=high. The nominal variable has no essential order as variables are labeled in different categories like 1=furniture, 2=Appliances, and 3= Electronic.

The logistic regression models the maximum chance evaluation to find an equation $\log[p(X) / (1-p(X))] = \beta0 + \beta1X1 + \beta2X2 + \ldots + \beta pXp$ as Xj is the jth predictor variable and βj is the accent impression for the jth predictor variable. The right side of the equation forecast the log odds of winning to the response variable or log odds of losing. ODDS = probability of winning/Probability of losing Logit Function= log(P/1-p).

OBSERVATION

In this case, I worked on logistic regression with the Banks dataset. The dataset contains 20 observations, and five variables are Obs, financial.condition (the bank's financial condition) , Totcap.Assets (the ratio of total capital to total assets), TotExp.Assets (the ratio of total expenses to total assets) and TotLns.Lses.Assets (the ratio of total loans and leases to total assets). I used the outcome categorical variable is financial.contition shows each off-balance sheet value as poor or strong with three predictor variables. As a regression result, I ignore all variable P-value because all is higher than 0.05. The estimated logistic equation follows;

Logit(FinancialCondition=1)=-702.21-14.11*TotCapAsset+3468.93*TotExpAsset+689.83*TotLnsLsesAsset.

The positive coefficients for dummy variables TotExpAsset and TotLnsLsesAsset highest probabilities of the effect of the positive, strong condition of finance. The TotCapAsset has negative coefficients demonstrating a high value to affect the intense extreme condition.

Usually, linear regression has an R square to decide how the model fits the dataset. I use McFadden's R square to show if values over the 0.4 means model fit the dataset as our case result 1 model fits the data very well. The variable importance result of logistic regression with varImp() function. Higher values are the more critical predictor variable, followed by TotCapAsset, TotLnsLsesAsset, and TotExpAsset. The VIF values' results show how correlated to each other. If the correlation is high between two variables, that can cause a problem. Thus, TotCapAsset has a high value than 5. that multicollinearity can issue in the model.

At that point, the model fit the dataset and then applied the test dataset's first five actuals to make predictions. Predictions records result in the first four records' probability assumption being higher than 0.5. Thus, they are genuinely an acceptor(actual=1). The fifth one has a lower than 0.5 as nonacceptors(actual=0).

Lastly, analyze how well the logistic regression model fits the test dataset with the confusion matrix as a statistical result of the Accuracy of 59% which is lower than 70%, and P-value 0.01 is the dataset statistically significant. Positive and negative prediction values are 1. The plot of the ROC(Receiver Operating Characteristics) curve shows division of actual positive probability is minor from 1 to 0. The area under the curve shows how models accurately predict outcomes. The test dataset of the lift chart shows a sharp increase above the source line at 60%, then leveling as the test dataset fits the model.

Reference

Data Mining for Business Analytics… concepts, Techniques, and Applications in R by Galit Shmueli: Peter C. Bruce: Inbal Yahav: Nitin R. Patel: Kenneth C. Lichtendahl, Jr. Page 56, chapter 3, Figures 3.2 and 3.3. Page 94, chapter 4, Figures 4.3 and 4.4, chapter 6, Figure 6.3, page 157, Chapter 10 from 237 pages to 265.