Module 4 CT Option 1

Ty Cook

4/11/2020

## Read File into Dataframe

banks.df <- read.csv("Banks.csv", header = TRUE) #Read Banks.csv into banks.df dataframe

## Data Exploration

### Dimension and Header Functions

dim(banks.df) #Find the Dimensions of the Data Frame Banks

## [1] 20 5

head(banks.df, n=20) #Show all 20 Rows and Columns from Data Frame

## 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  
## 11 11 0 11.1 0.08 0.43  
## 12 12 0 20.5 0.12 0.80  
## 13 13 0 9.8 0.07 0.69  
## 14 14 0 7.9 0.08 0.53  
## 15 15 0 9.6 0.09 0.73  
## 16 16 0 12.5 0.09 0.30  
## 17 17 0 18.3 0.08 0.49  
## 18 18 0 7.2 0.11 0.55  
## 19 19 0 14.0 0.08 0.44  
## 20 20 0 8.3 0.08 0.51

### Summary

knitr::kable( summary(banks.df) ) #Table of Summary Statistics for Banks.Csv

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Obs | Financial.Condition | TotCap.Assets | TotExp.Assets | TotLns.Lses.Assets |
|  | Min. : 1.00 | Min. :0.0 | Min. : 1.000 | Min. :0.0700 | Min. :0.3000 |
|  | 1st Qu.: 5.75 | 1st Qu.:0.0 | 1st Qu.: 7.125 | 1st Qu.:0.0800 | 1st Qu.:0.5250 |
|  | Median :10.50 | Median :0.5 | Median : 9.200 | Median :0.1000 | Median :0.6400 |
|  | Mean :10.50 | Mean :0.5 | Mean : 9.320 | Mean :0.1045 | Mean :0.6285 |
|  | 3rd Qu.:15.25 | 3rd Qu.:1.0 | 3rd Qu.:11.300 | 3rd Qu.:0.1200 | 3rd Qu.:0.7225 |
|  | Max. :20.00 | Max. :1.0 | Max. :20.500 | Max. :0.1600 | Max. :1.0200 |

### Data Exploration Functions

data.frame(mean=sapply(banks.df[-c(1:2)],mean,na.rm=TRUE), #Mean excluding Obs and Financial.Condition   
sd=sapply(banks.df[-c(1:2)],sd,na.rm=TRUE), #Standard Deviation excluding Obs and Financial.Condition   
min=sapply(banks.df[-c(1:2)],min,na.rm=TRUE), #Minmium excluding Obs and Financial.Condition   
max=sapply(banks.df[-c(1:2)],max,na.rm=TRUE), #Maxmium excluding Obs and Financial.Condition   
median=sapply(banks.df[-c(1:2)],median,na.rm=TRUE), #Median excluding Obs and Financial.Condition   
length=sapply(banks.df[-c(1:2)],length) #Legnth excluding Obs and Financial.Condition   
)

## mean sd min max median length  
## TotCap.Assets 9.3200 4.79721410 1.00 20.50 9.20 20  
## TotExp.Assets 0.1045 0.02605157 0.07 0.16 0.10 20  
## TotLns.Lses.Assets 0.6285 0.15977863 0.30 1.02 0.64 20

## Logistic Regression Model

banks.df <- banks.df[ , -c(3)] #Exclude 3 TotCap.Asset from Regression Model  
logit.reg <- glm(Financial.Condition ~ ., data = banks.df, family = "binomial") # Create Regression model based on financial Condition  
options(scipen=999) #Options command to prevent Scientfic Notation  
summary(logit.reg) #Summary Statistics for Logistics Regression Model

##   
## Call:  
## glm(formula = Financial.Condition ~ ., family = "binomial", data = banks.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.0000138316 -0.0000000211 0.0000000000 0.0000000211 0.0000150988   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 68.70 277487.32 0.000 1.000  
## Obs -11.12 17431.60 -0.001 0.999  
## TotExp.Assets 493.00 1541743.23 0.000 1.000  
## TotLns.Lses.Assets -21.84 251695.83 0.000 1.000  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 27.72588722239781 on 19 degrees of freedom  
## Residual deviance: 0.00000000061548 on 16 degrees of freedom  
## AIC: 8  
##   
## Number of Fisher Scoring iterations: 25

## Logistic Regression Equation

## Confusion Matrix

library(gains) # Loading gains library for gains funtion  
library(caret) # Loading caret library for confusionMatrix Function

## Warning: package 'caret' was built under R version 3.6.3

## Loading required package: lattice

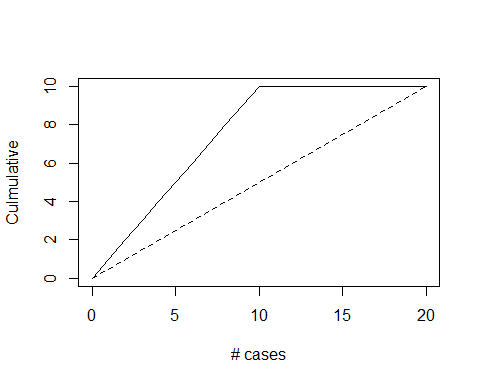
## Loading required package: ggplot2

pred <- predict(logit.reg, banks.df) # Creating Pred to gather prediction data from logistic regression model  
gain <- gains(banks.df$Financial.Condition, pred, groups=20) #Running gain function from true data and predictiondata  
pred\_factor <- ifelse(pred > 1, 1, 0) #creating binary array from prediction results  
pred\_factor <- factor(pred\_factor, levels = c(0,1), labels = c("S", "W")) #Turning pred\_factor into a factor, adding levels, and labels  
true\_factor <- factor(banks.df$Financial.Condition, levels = c(0,1), labels = c("S", "W")) #creating true\_factor array from financial.condition and turning into a factor, adding levels, and labels  
confusionMatrix(pred\_factor, true\_factor) #Creating a confusion Matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction S W  
## S 10 0  
## W 0 10  
##   
## Accuracy : 1   
## 95% CI : (0.8316, 1)   
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : 0.0000009537  
##   
## Kappa : 1   
##   
## Mcnemar's Test P-Value : NA   
##   
## Sensitivity : 1.0   
## Specificity : 1.0   
## Pos Pred Value : 1.0   
## Neg Pred Value : 1.0   
## Prevalence : 0.5   
## Detection Rate : 0.5   
## Detection Prevalence : 0.5   
## Balanced Accuracy : 1.0   
##   
## 'Positive' Class : S   
##

## Lift Chart

plot(c(0,gain$cume.pct.of.total\*sum(banks.df$Financial.Condition))~   
 c(0,gain$cume.obs),  
 xlab="# cases", ylab="Culmulative", main="", type="l") #Plot Function  
lines(c(0,sum(banks.df$Financial.Condition))~c(0, dim(banks.df)[1]), lty=2) #Logistic Regression Line



**Summary**

For this assignment I first examined the Banks.CSV in order to look over the dataset and confirm quantitative and qualitive categories of the data within the dataset, being Ops and Financial Condition being qualitive and TotCap/Assets, TotExp/Assets, and TotLns&Lses/Assets being quantitative. From there I proceeded to read the dataset into a data frame called Banks.df and created the general data exploratory functions that we have demonstrated over the past few weeks, DIM, HEAD, Summary, SD, Max, Min, Mean, Median, and length. I did experiment with creating a table in RMarkdown Knitr::kable command for the summary and did like the way it looked better as just showing the summary output had the results split on multiple lines do the long names of the variables, and number of the columns.

The next big part of the assignment was creating the logistic regression model within Rstudio, as with most things this in another new subject to me and was a challenge for me problem solve how I was going to get the logistic model commands from the reading to match what I wanted to model against. I first removed column 3 TotCap.Asset from the model due to the hint given and proceeded to follow to process the logistic regression model against Financial.Condition. After that I did further research on the syntax in order to understand how to show an equation within RMarkdown which was $FC = 68.70 + -11.12(Obs) + 493.00(TotExp.Assets) + -21.84(TotLns.Lses.Assets)$ Which is Finanical Condition is equal to 68.70 + -11.12 \* Obs + 493.00 \* TotExp.Assets + -21.84 TotaLns,Lses.Assets.

Last step was the confusion matrix and lift chart. Which the confusion matrix gave me the most trouble I understood that I needed to create factor variables and levels for it to work just the syntax in the book was causing me to receive several error codes so I opted to some further research and was able to find a way that made much more sense to me but was able to publish the same results I was looking for.