**Figure 1**  
*Load Data Set*

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.2 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(gains)  
library(carat)  
library(ISLR)  
library(InformationValue)   
mod4banks.df <- read.csv("MIS510banks.csv", header = TRUE)   
head(mod4banks.df)

## 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

**Figure 2**  
*Remove Extra Attributes*

mod4banks.df <- mod4banks.df[ , -c(1,5)]  
head(mod4banks.df)

## Financial.Condition TotCap.Assets TotExp.Assets  
## 1 1 9.7 0.12  
## 2 1 1.0 0.11  
## 3 1 6.9 0.09  
## 4 1 5.8 0.10  
## 5 1 4.3 0.11  
## 6 1 9.1 0.13

**Figure 3**  
*View and dim Functions*

View(mod4banks.df)  
dim(mod4banks.df)

## [1] 20 3

**Figure 4**  
*Mean and Median of Attributes*

data.frame(mean = sapply(mod4banks.df[,-1], mean),  
 median = sapply(mod4banks.df[,-1], median))

## mean median  
## TotCap.Assets 9.3200 9.2  
## TotExp.Assets 0.1045 0.1

**Figure 5**  
*Logistic Regression*

logit.mod4banks <- glm(Financial.Condition ~ ., family ="binomial", data = mod4banks.df)  
options(scipen=999)  
summary(logit.mod4banks)

##   
## Call:  
## glm(formula = Financial.Condition ~ ., family = "binomial", data = mod4banks.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.02818 -0.44160 -0.03606 0.27025 1.62010   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.3768 5.3683 -1.374 0.1694   
## TotCap.Assets -0.3873 0.2235 -1.732 0.0832 .  
## TotExp.Assets 109.8631 55.8316 1.968 0.0491 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 27.726 on 19 degrees of freedom  
## Residual deviance: 11.028 on 17 degrees of freedom  
## AIC: 17.028  
##   
## Number of Fisher Scoring iterations: 6

**Figure 6**  
*Predicted Probabilities*

logit.mod4banks.pred <- predict(logit.mod4banks, mod4banks.df, type = "response")  
summary(logit.mod4banks.pred)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.003421 0.101697 0.364492 0.500000 0.957639 0.999943

data.frame(actual = mod4banks.df$Financial.Condition[1:5],  
 predicted = logit.mod4banks.pred[1:5])

## actual predicted  
## 1 1 0.8859924  
## 2 1 0.9868868  
## 3 1 0.4598032  
## 4 1 0.7963242  
## 5 1 0.9544790

**Figure 7**  
*Confusion Matrix*

(optimal <- optimalCutoff(mod4banks.df$Financial.Condition, logit.mod4banks.pred))

## [1] 0.2399432

confusionMatrix(mod4banks.df$Financial.Condition, logit.mod4banks.pred, threshold = optimal)

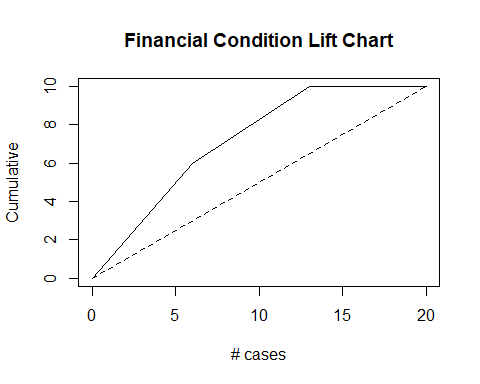
## 0 1  
## 0 9 0  
## 1 1 10

**Figure 8**  
*Lift Chart*

(mod4gain <- gains(mod4banks.df$Financial.Condition, logit.mod4banks.pred, groups = length(mod4banks.df)))

## Depth Cume Cume Pct Mean  
## of Cume Mean Mean of Total Lift Cume Model  
## File N N Resp Resp Resp Index Lift Score  
## -------------------------------------------------------------------------  
## 30 6 6 1.00 1.00 60.0% 200 200 0.98  
## 65 7 13 0.57 0.77 100.0% 114 154 0.53  
## 100 7 20 0.00 0.50 100.0% 0 100 0.06

plot(c(0,mod4gain$cume.pct.of.total\*sum(mod4banks.df$Financial.Condition))~c(0,mod4gain$cume.obs),  
 xlab = "# cases", ylab = "Cumulative", main = "Financial Condition Lift Chart", type = "l")  
lines(c(0, sum(mod4banks.df$Financial.Condition))~c(0, dim(mod4banks.df)[1]), lty = 2)



## Description and Lessons Learned

The MIS510Banks *csv* file was loaded to R Studio using the read.csv function and set as a data frame. Figure 1 displays this code as well as the libraries that were loaded for the assignment. Then since the data set is quite small, the head() function was used to view all attributes. The Obs attribute is just the numbered list of observations. It should be removed. Also, per the class instructor’s recommendations, one of the three remaining predictors needs to be removed to avoid warnings during the logistic regression model build. The predictor *TotLns.Lses.Assets* (column five) will be removed.

Figure 2 displays the code for removing columns one and five. The head() function was then used again to verify the attributes were removed. The View() and dim() data exploration functions are displayed in Figure 3. RMarkdown does not display the results of the View function. This makes sense as many data sets are too large for a standard page in a Word document. The data set dimensions dim() are three columns and 20 rows. To demonstrate another data exploration example, Figure 4 calculates the mean and median values for the two predictor attributes. Both attributes have similar statistics which does not provide much insight.

The logistic regression model is provided in Figure 5. As stated, the outcome or predicted attribute is *Financial.Condition* which represents the financial strength of the institution categorized into two bins, weak financial strength represented as a zero, and strong financial strength represented as one. The *family* option in the glm() function must be binomial since the outcome attribute is only one of two outcomes. The results of the regression can be seen using summary(). Using these results, a predictive equation can be written and used to predict future observation outcomes. The logistic equation for this data set is **Logit(Financial Condition = Strong) = (-7.3768) - 0.3873(TotCap.Assets) + 109.8631(TotExp.Assets)**.

Evaluation can now be performed on the logistic regression model. Predicted probabilities for the first 5 rows of observations will be calculated and then compared to the actual values. Figure 6 displays code for creating the predicted values model (function predict) and the creation of a new data frame that compares the predicted model values with the actual categorized results from the logistic regression. The summary function displays some extra information from the predicted model results. The comparison shows a possible discrepancy at the third observation with the actual value at one and the predicted value at 0.46. However, this result will be justified with the calculation of an optimal cutoff. The optimal cutoff is the optimized separation between classes and is used within a confusion matrix which displays how well the model predicted the correct classifications (See Figure 7). The optimal cutoff result is 0.2399. This would mean a predicted result of 0.46 would be classed as one or strong. Reading the matrix left to right, up to down, there were nine true weak financial condition results, no false weak results, one false strong result, and 10 true strong financial conditions. In the final evaluation, a lift chart will be created (See Figure 8). The lift chart curve represents the improvement the model provides in classifying results (the curved solid line) as opposed to random choice (the straight dashed line). If multiple models are made, a comparison of their lift charts provides a graphical method for determining which model performed the best (the chart that has the most area between the curve and the random guess line).

In this assignment, a financial data set was explored and a logistic regression model was created. Methods to evaluate the model performance were also explored including prediction probabilities, confusion matrix, and lift chart. A prediction equation was created from the model similar to how a prediction equation can be created from a linear regression. There are still many questions concerning R code and how it relates to the theories and math discussed in the textbook (Shmueli et al., 2018). The author of this paper’s goal is to understand each keystroke within the R code examples. The textbook does not always provide a thorough explanation between concepts and code.

## References

RStudio Team. (2021). *RStudio: Integrated development environment for R. RStudio*, PBC, Boston, MA. <http://www.rstudio.com/>.

Shmueli, G., Bruce, P.C., Yahav, I., Patel, N.R., and Lichtendahl, K.C. (2018). *Data mining for business analytics: Concepts, techniques, and application sin R.* Wiley Publishing. ISBN: 9781118879337.