Module 4 CT Option 1 (2)

Richard Klein

7/19/2020

# Import Data and Create DataFrame

bank.df <- read.csv(“Bankinfo.csv”)

# Data Exploration

str (bank.df) head (bank.df) summary (bank.df)

# Examine Target attribute treating as a categorical attribute, 1 strong, 0 weak

table(bank.df$Financial.Condition)

# Create Boxplot and Histogram using attribute Total Loans and Losses

boxplot(bank.df$TotLnsandL, xlab="TotLnsandL") hist(bank.df$TotLnsandL, xlab=“TotLnsandL”)

# Compute mean, standard dev., min, max, median, length, and missing values of TotLnsandL

mean(bank.dfTotLnsandL) min(bank.dfTotLnsandL) median(bank.dfTotLnsandL)

# Partition - no partition, use 100% of data

set.seed(1) # set seed for reproducing the partition train.index <- sample(c(1:20), 20)  
train.df <- bank.df[train.index, ] valid.df <- bank.df[train.index, ]

# Run Logistic Regression

logit.reg <-glm(Financial.Condition ~., data = train.df, family = “binomial”) options(scipen=999) summary(logit.reg)

# Logistic Fitting Function

Logit(Financial.Condition = 1) = -30.148 + Obs \* 6.508 + TotCap \* 2.375 - TotExp \* 260.470 - TotLnsandL \* 51.653

# Evaluating Model

# install.packages(“gains”)

library(gains) pred <- predict(logit.reg, valid.df) gain <- gains(valid.df$Financial.Condition, pred, groups=20)

plot(c(0,gainFinancial.Condition)) c(0,gain$cume.obs), xlab="# cases", ylab="Cumulative", main="", type="l") lines(c(0,sum(valid.df$Financial.Condition))c(0, dim(valid.df)[1]), lty=2)

# Install Packages for Confusion Matrix

# Install.packages (“caret”)

library(caret) # Install.packages(“e1071”) library(e1071) # Run confusion matrix confusionMatrix(as.factor(ifelse(pred > 0.5, 1, 0)), as.factor(valid.df$Financial.Condition))

**Critical Thinking 4: Option 1**

Critical Thinking 4, Option 1 assignment entitled, “Logistic Regression on Banks”, uses data from the Banks.csv file. The assignment included importing data, running exploration functions on the data set, modeling the Financial Condition attribute as the categorical attribute and function of the other attributes, running a logistical regression model, writing the logistic equation, evaluation the model, plotting and creating confusion matrix.

Importing the data cased several errors for the project early on. Result was that a new project was not initially created therefore not mapping to the correct folder. Once this was done the data imported without issue.

Data exploration included using the table function () to treat Financial Condition as a categorical attribute, the head function () to show the first 6 lines of every column, and the summary function () to provided statistical summary of the different attributes. Also used individual statistical functions to provide additional details on TotalLnsandL (column headers were edited in Excel for convenience).

Although no partition was need for assignment, created the train and valid data frames using all data.

Next, logistic regression was run using the attribute Financial Condition which provided the intercept and the coefficients of all the attributes. From this output the logistic fitting function calculation was written. The estimated logistic equation takes the coefficients calculated using the summary function () when running the logistic regression and returns coefficients that tell the positive or negative correlation between the predictor attributes and the Financial Condition of the 20 banks. Observation of course had the least correlation, as did TotCap (Total Capitalization). TotExp had the highest correlation as to whether the bank was considered to have a strong or weak financial condition and TotLnsandL (Total loans and loses) was the next highest positive correlation (Shmueli, Bruce, Yahav, Patel & Lichtendahl, 2018).

Finally, to plot the and evaluate the model, the packages “gains”, “caret” and “e1071” were installed. The attribute Financial Condition was plotted, and the confusion matrix created as shown in below figure. Accuracy of logistic regression model was 1 or 100% on Banks.csv dataset.

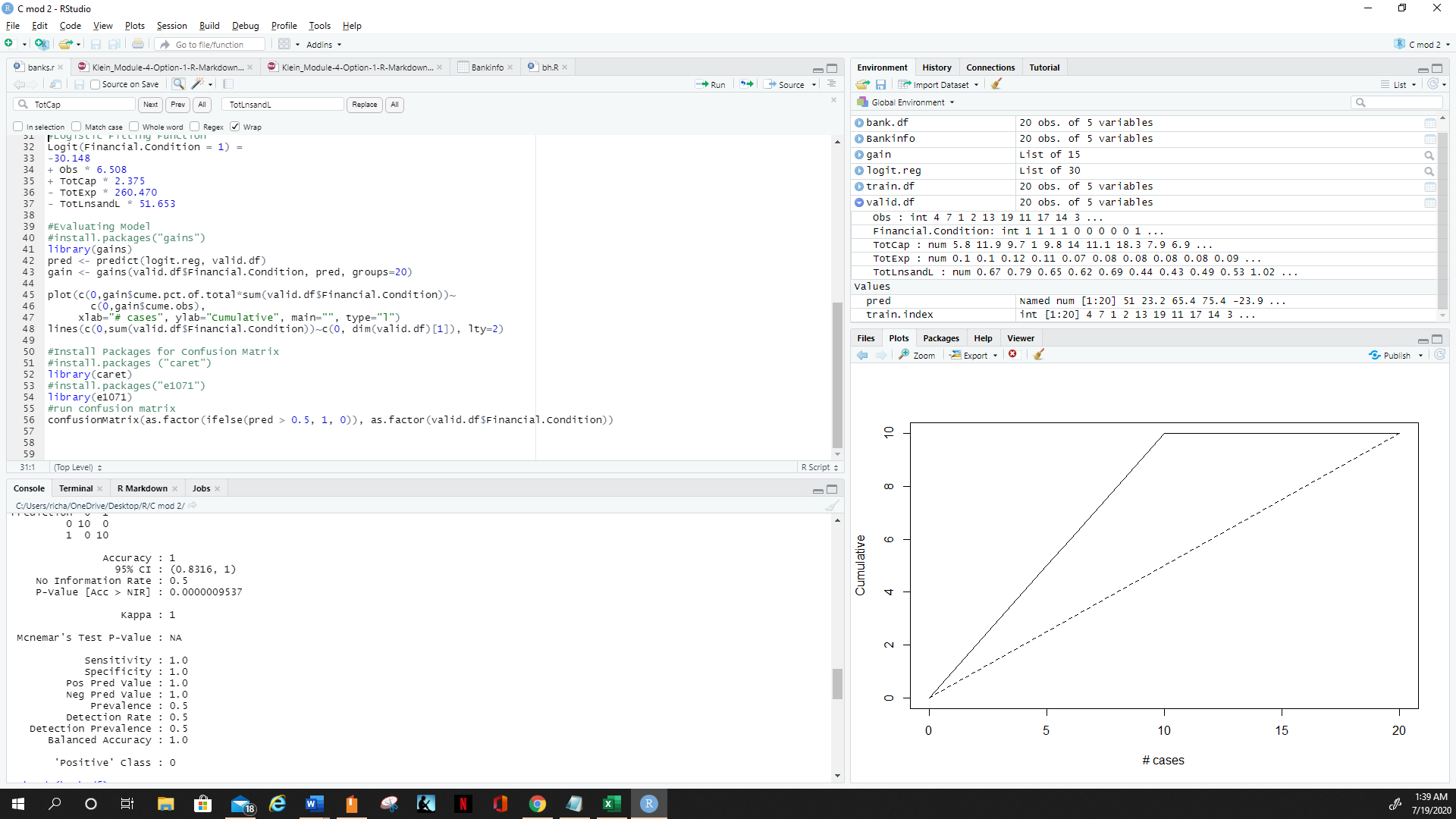


Figure 1. Shows plot and confusion matrix.

This week’s Critical Thinking exercise was challenging, but ultimately successful. Issues encountered included correctly creating the program and creating the train and valid data frames to produce the confusion matrix both of which were successfully corrected through trial and error.

**References**

Shmueli, G., Bruce, P., Yahav. I., Patel., N., & Lichtendahl, K. (2018).  Data Mining for Business Analytics: Concepts, Techniques, and Applications in R. Hoboken, NJ: John Wiley & Sons, Inc.