CMTH820 Big Data Analytic Program

Initial Results and the Code

[Peiran Liu]

[CMTH820]

####Abstract of the Project The theme I would like to work on is classification and regression with non-textural dataset. The reason I choose this theme is that I would like to become an data scientist in business area. Classification and regression are the most common tools to use to analyse data, so I want to practice the knowledge and techniques I learned in this program in the real world’s cases.

After reviewing all the datasets, I choose the Bank Marketing Dataset in UC Irvine Machine Learning Repository. The UC Irvine Machine Learning Repository dataset provide service to machine learning users and contain over 600 datasets. I choose this datasets because it come from real world and it contains variables which can be applied to machine learning techniques. The Portugal banking institution collect the data from their clints with direct marketing campaigns (phone calls). The dataset contains 41188 observers and 21 variables( from May 2008 to November 2010 ), it may contains missing values and replicate data which needed to be cleaned. The variables describe the clients’ socioeconomic status (e.g. martial, job, education, loan, housing), as well as their previous telemarketing campaigns (e.g. contact, campaign, previous, duration) . The independent variable is binary( “yes”, “no” ) data about if the client subscribed a term deposit. The bank is marketing a long-term deposit account such as bonds and saving account to their existing clients. They intend to improve their telemarketing strategies and to predict the clients who will subscribe more long-term deposit by analysing the dataset they provided.

My study question is help a bank to predict weather their clients will subscribe the long-term deposit. . By using the dataset which collected by the bank, my project can help the bank to determine such customers and finding an effective telemarketing strategy. This research can improve the efficiency of their marketing department and reduce their expenses. Our research can also help the bank to figure out which existing clients they should advertise a long-term deposit account such as bonds and saving account.

In order to achieve this goal, I will first have a clear explore of the dataset, clean the dataset which we can use for analysing it. Then I will choose those important attributes which I can use in the predictive model. Finally, I will develop two prediction models, including decision tree and Naïve Bayes and find the models’ accuracy and precision to judge if I successfully developed the prediction models.

My research questions are: • help this bank to predict whether their clients will subscribe the long-term deposit. • Compare decision tree and naïve bayes prediction methods • Find out if booster dataset will affect results • Find out if dataset need to be balanced and if balance the dataset will affect results ####

#### 1. Read the csv files in the folder.

library(class)  
library(gmodels)  
  
data\_bank<-read.csv(file = "bank.csv")  
data\_bank<-data.frame(data\_bank)

#### 2. Basic factors about the data

head(data\_bank)

## age job marital education default balance housing loan contact day  
## 1 30 unemployed married primary no 1787 no no cellular 19  
## 2 33 services married secondary no 4789 yes yes cellular 11  
## 3 35 management single tertiary no 1350 yes no cellular 16  
## 4 30 management married tertiary no 1476 yes yes unknown 3  
## 5 59 blue-collar married secondary no 0 yes no unknown 5  
## 6 35 management single tertiary no 747 no no cellular 23  
## month duration campaign pdays previous poutcome y  
## 1 oct 79 1 -1 0 unknown no  
## 2 may 220 1 339 4 failure no  
## 3 apr 185 1 330 1 failure no  
## 4 jun 199 4 -1 0 unknown no  
## 5 may 226 1 -1 0 unknown no  
## 6 feb 141 2 176 3 failure no

str(data\_bank)

## 'data.frame': 4521 obs. of 17 variables:  
## $ age : int 30 33 35 30 59 35 36 39 41 43 ...  
## $ job : chr "unemployed" "services" "management" "management" ...  
## $ marital : chr "married" "married" "single" "married" ...  
## $ education: chr "primary" "secondary" "tertiary" "tertiary" ...  
## $ default : chr "no" "no" "no" "no" ...  
## $ balance : int 1787 4789 1350 1476 0 747 307 147 221 -88 ...  
## $ housing : chr "no" "yes" "yes" "yes" ...  
## $ loan : chr "no" "yes" "no" "yes" ...  
## $ contact : chr "cellular" "cellular" "cellular" "unknown" ...  
## $ day : int 19 11 16 3 5 23 14 6 14 17 ...  
## $ month : chr "oct" "may" "apr" "jun" ...  
## $ duration : int 79 220 185 199 226 141 341 151 57 313 ...  
## $ campaign : int 1 1 1 4 1 2 1 2 2 1 ...  
## $ pdays : int -1 339 330 -1 -1 176 330 -1 -1 147 ...  
## $ previous : int 0 4 1 0 0 3 2 0 0 2 ...  
## $ poutcome : chr "unknown" "failure" "failure" "unknown" ...  
## $ y : chr "no" "no" "no" "no" ...

summary(data\_bank)

## age job marital education   
## Min. :19.00 Length:4521 Length:4521 Length:4521   
## 1st Qu.:33.00 Class :character Class :character Class :character   
## Median :39.00 Mode :character Mode :character Mode :character   
## Mean :41.17   
## 3rd Qu.:49.00   
## Max. :87.00   
## default balance housing loan   
## Length:4521 Min. :-3313 Length:4521 Length:4521   
## Class :character 1st Qu.: 69 Class :character Class :character   
## Mode :character Median : 444 Mode :character Mode :character   
## Mean : 1423   
## 3rd Qu.: 1480   
## Max. :71188   
## contact day month duration   
## Length:4521 Min. : 1.00 Length:4521 Min. : 4   
## Class :character 1st Qu.: 9.00 Class :character 1st Qu.: 104   
## Mode :character Median :16.00 Mode :character Median : 185   
## Mean :15.92 Mean : 264   
## 3rd Qu.:21.00 3rd Qu.: 329   
## Max. :31.00 Max. :3025   
## campaign pdays previous poutcome   
## Min. : 1.000 Min. : -1.00 Min. : 0.0000 Length:4521   
## 1st Qu.: 1.000 1st Qu.: -1.00 1st Qu.: 0.0000 Class :character   
## Median : 2.000 Median : -1.00 Median : 0.0000 Mode :character   
## Mean : 2.794 Mean : 39.77 Mean : 0.5426   
## 3rd Qu.: 3.000 3rd Qu.: -1.00 3rd Qu.: 0.0000   
## Max. :50.000 Max. :871.00 Max. :25.0000   
## y   
## Length:4521   
## Class :character   
## Mode :character   
##   
##   
##

#### 4.Check the missing value

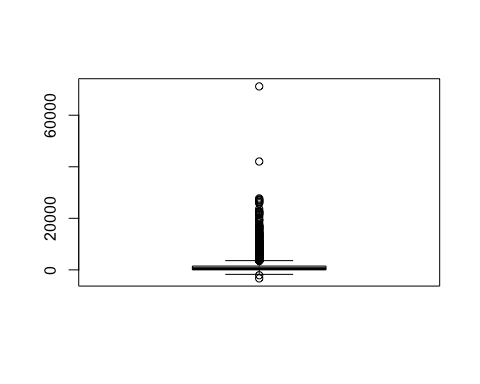
sum(is.na(data\_bank))

## [1] 0

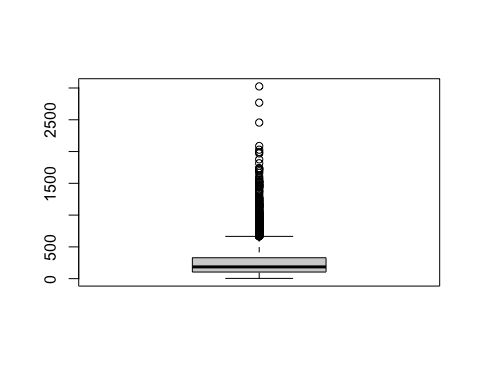
#### From here, we can see that there is no missing value.

#### 5. Graph of the dataset.

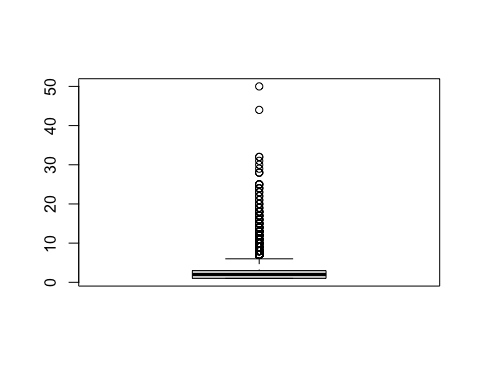
boxplot(data\_bank$balance)



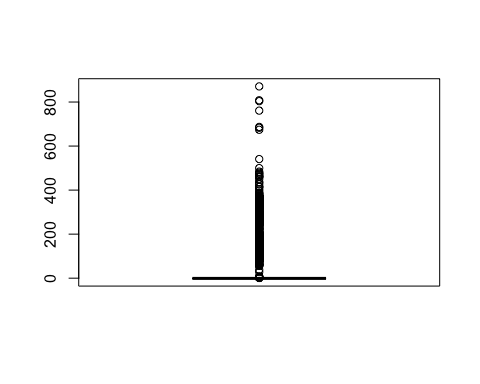
boxplot(data\_bank$duration)



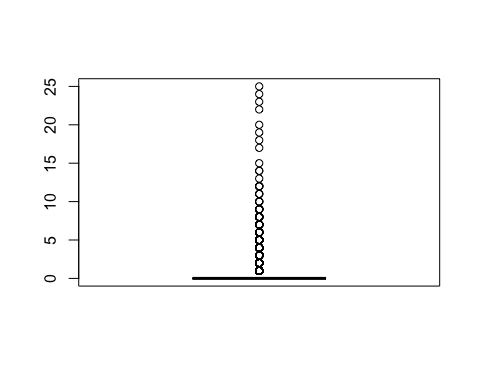
boxplot(data\_bank$campaign)



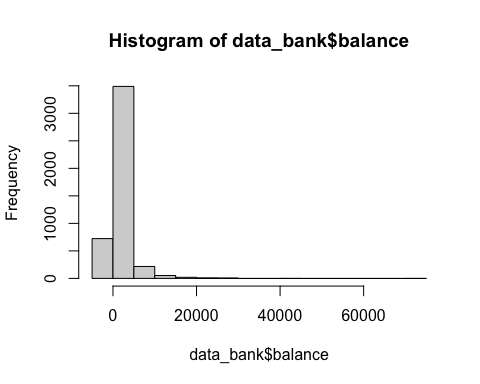
boxplot(data\_bank$pdays)



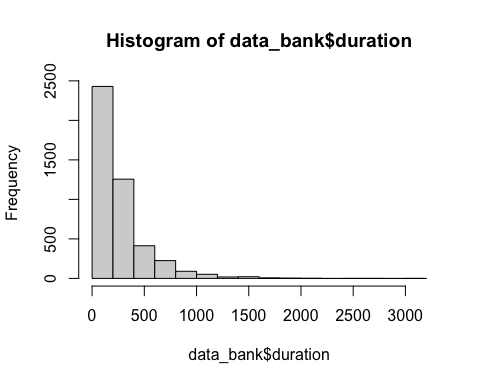
boxplot(data\_bank$previous)  
  
###histogram of all numeric attributs  
library(ggplot2)



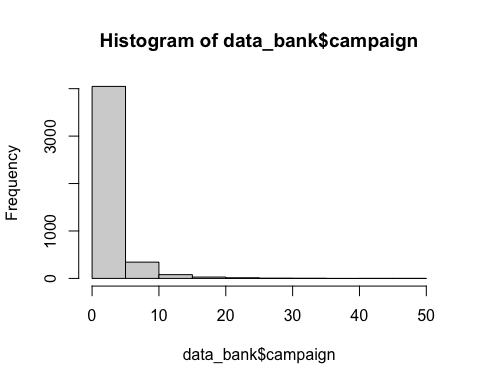
hist(data\_bank$balance)



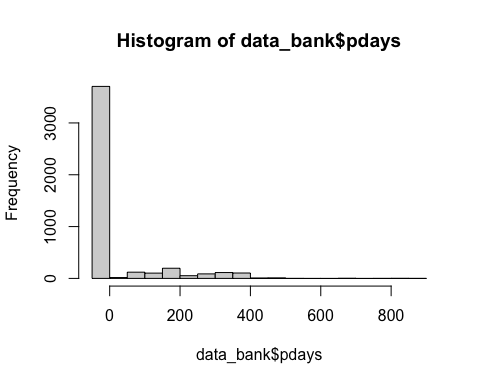
hist(data\_bank$duration)



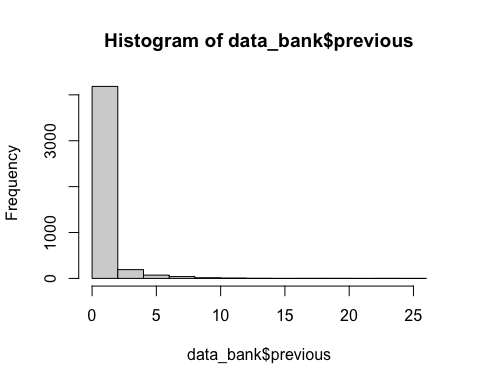
hist(data\_bank$campaign)



hist(data\_bank$pdays)



hist(data\_bank$previous)



#### 6.check the correlation of the dataset

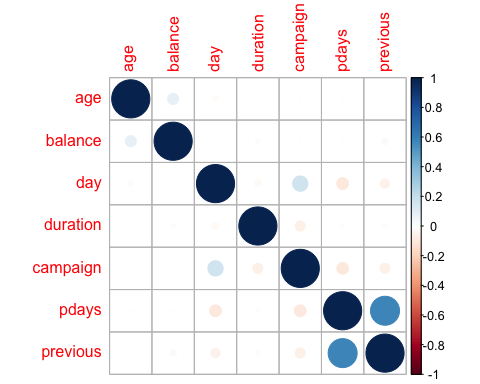
library(corrplot)

## corrplot 0.92 loaded

my\_num\_data <- data\_bank[, sapply(data\_bank, is.numeric)]  
cor(my\_num\_data)

## age balance day duration campaign  
## age 1.000000000 0.083820142 -0.017852632 -0.002366889 -0.005147905  
## balance 0.083820142 1.000000000 -0.008677052 -0.015949918 -0.009976166  
## day -0.017852632 -0.008677052 1.000000000 -0.024629306 0.160706069  
## duration -0.002366889 -0.015949918 -0.024629306 1.000000000 -0.068382000  
## campaign -0.005147905 -0.009976166 0.160706069 -0.068382000 1.000000000  
## pdays -0.008893530 0.009436676 -0.094351520 0.010380242 -0.093136818  
## previous -0.003510917 0.026196357 -0.059114394 0.018080317 -0.067832630  
## pdays previous  
## age -0.008893530 -0.003510917  
## balance 0.009436676 0.026196357  
## day -0.094351520 -0.059114394  
## duration 0.010380242 0.018080317  
## campaign -0.093136818 -0.067832630  
## pdays 1.000000000 0.577561827  
## previous 0.577561827 1.000000000

corrplot(cor(my\_num\_data))



#### 7. Apply normallization to the dataset

normdata <- prcomp(my\_num\_data, scale = TRUE, center= TRUE)

This is the end

Peiran Liu