# Set up cartelle e creazione dataset

getwd()

## [1] "/Users/Americo/Documents/Education/Unitelma/tesi/data\_analysis/analysis"

setwd("/Users/Americo/Documents/Education/Unitelma/tesi/data\_analysis/dataset")  
bank0 <- read.csv(file = "bank\_full.csv", sep = ";")

# Informazioni sul dataset

## Input variables:

### bank client data

* 1 - age (numeric)
* 2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
* 3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
* 4 - education (categorical: "unknown","secondary","primary","tertiary")
* 5 - default: has credit in default? (binary: "yes","no")
* 6 - balance: average yearly balance, in euros (numeric)
* 7 - housing: has housing loan? (binary: "yes","no")
* 8 - loan: has personal loan? (binary: "yes","no")

# related with the last contact of the current campaign:

* 9 - contact: contact communication type (categorical: "unknown","telephone","cellular")
* 10 - day: last contact day of the month (numeric)
* 11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
* 12 - duration: last contact duration, in seconds (numeric)

### Other attributes:

* 13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
* 14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
* 15 - previous: number of contacts performed before this campaign and for this client (numeric)
* 16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

### Output variable (desired target)

* 17 - y - has the client subscribed a term deposit? (binary: "yes","no")

## Missing Attribute Values

None

getwd()

## [1] "/Users/Americo/Documents/Education/Unitelma/tesi/data\_analysis/analysis"

setwd("/Users/Americo/Documents/Education/Unitelma/tesi/data\_analysis/dataset")  
bank0 <- read.csv(file = "bank\_full.csv", sep = ";")  
str(bank0)

## 'data.frame': 45211 obs. of 17 variables:  
## $ age : int 58 44 33 47 33 35 28 42 58 43 ...  
## $ job : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...  
## $ marital : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...  
## $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...  
## $ default : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...  
## $ balance : int 2143 29 2 1506 1 231 447 2 121 593 ...  
## $ housing : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...  
## $ loan : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...  
## $ contact : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ day : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ month : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...  
## $ duration : int 261 151 76 92 198 139 217 380 50 55 ...  
## $ campaign : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ pdays : int -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...  
## $ previous : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...  
## $ y : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...

require(dplyr)

## Loading required package: dplyr  
##   
## Attaching package: 'dplyr'  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

bank0 %>%  
 group\_by (y) %>%   
 summarise(n= n()) %>%  
 mutate(freq = n / sum(n))

## Source: local data frame [2 x 3]  
##   
## y n freq  
## 1 no 39922 0.8830152  
## 2 yes 5289 0.1169848

bank0 %>%  
 group\_by (default, y) %>%   
 summarise(n= n()) %>%  
 mutate(freq = n / sum(n))

## Source: local data frame [4 x 4]  
## Groups: default  
##   
## default y n freq  
## 1 no no 39159 0.88203892  
## 2 no yes 5237 0.11796108  
## 3 yes no 763 0.93619632  
## 4 yes yes 52 0.06380368

fit <- glm(y ~ default, data = bank0, family = "binomial")  
summary(fit)

##   
## Call:  
## glm(formula = y ~ default, family = "binomial", data = bank0)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.501 -0.501 -0.501 -0.501 2.346   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.01188 0.01471 -136.737 < 2e-16 \*\*\*  
## defaultyes -0.67413 0.14408 -4.679 2.88e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 32631 on 45210 degrees of freedom  
## Residual deviance: 32604 on 45209 degrees of freedom  
## AIC: 32608  
##   
## Number of Fisher Scoring iterations: 5

ODDS = exp(-0.67413)