**Citation Request:**

**This dataset is public available for research. The details are described in [Moro et al., 2011].**

**Please include this citation if you plan to use this database: [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.**

**In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, GuimarÃ£es, Portugal, October, 2011. EUROSIS.**

**Available at: [pdf] http://hdl.handle.net/1822/14838 [bib] http://www3.dsi.uminho.pt/pcortez/bib/2011-esm-1.txt**

**1. Title: Bank Marketing**

**2. Sources Created by: Paulo Cortez (Univ. Minho) and SÃ©rgio Moro (ISCTE-IUL) @ 2012**

**3. Past Usage:**

**The full dataset was described and analyzed in:**

**S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.**

**In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, GuimarÃ£es,**

**Portugal, October, 2011. EUROSIS.**

**4. Relevant Information:**

**The data is related with direct marketing campaigns of a Portuguese banking institution.**

**The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.**

**There are two datasets:**

**1) bank-full.csv with all examples, ordered by date (from May 2008 to November 2010).**

**2) bank.csv with 10% of the examples (4521), randomly selected from bank-full.csv.**

**The smallest dataset is provided to test more computationally demanding machine learning algorithms (e.g. SVM).**

**The classification goal is to predict if the client will subscribe a term deposit (variable y).**

**5. Number of Instances: 45211 for bank-full.csv (4521 for bank.csv)**

**6. Number of Attributes: 16 + output attribute.**

**7. Attribute information:**

**For more information, read [Moro et al., 2011].**

**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")**

**8. Missing Attribute Values: None**

**1). Output variable -> y**

**y -> Whether the client has subscribed a term deposit or not**

**Binomial ("yes" or "no")**

**Ans:**

**library(readr)**

**> bank <- read\_csv("Assignment Questions/Assignment 6 Logistic Regression/bank.csv")**

Parsed with column specification:

cols(

age = col\_double(),

job = col\_character(),

marital = col\_character(),

education = col\_character(),

default = col\_character(),

balance = col\_double(),

housing = col\_character(),

loan = col\_character(),

contact = col\_character(),

day = col\_double(),

month = col\_character(),

duration = col\_double(),

campaign = col\_double(),

pdays = col\_double(),

previous = col\_double(),

poutcome = col\_character(),

y = col\_character()

)

**> View(bank)**

**> bank<-read.csv("Assignment Questions/Assignment 6 Logistic Regression/bank.csv")**

**> head(bank)**

age job marital education default balance housing

1 58 management married tertiary no 2143 yes

2 44 technician single secondary no 29 yes

3 33 entrepreneur married secondary no 2 yes

4 47 blue-collar married unknown no 1506 yes

5 33 unknown single unknown no 1 no

6 35 management married tertiary no 231 yes

loan contact day month duration campaign pdays previous

1 no unknown 5 may 261 1 -1 0

2 no unknown 5 may 151 1 -1 0

3 yes unknown 5 may 76 1 -1 0

4 no unknown 5 may 92 1 -1 0

5 no unknown 5 may 198 1 -1 0

6 no unknown 5 may 139 1 -1 0

poutcome y

1 unknown no

2 unknown no

3 unknown no

4 unknown no

5 unknown no

6 unknown no

> str(bank)

'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 ...

> summary(bank)

age job marital

Min. :18.00 blue-collar:9732 divorced: 5207

1st Qu.:33.00 management :9458 married :27214

Median :39.00 technician :7597 single :12790

Mean :40.94 admin. :5171

3rd Qu.:48.00 services :4154

Max. :95.00 retired :2264

(Other) :6835

education default balance housing

primary : 6851 no :44396 Min. : -8019 no :20081

secondary:23202 yes: 815 1st Qu.: 72 yes:25130

tertiary :13301 Median : 448

unknown : 1857 Mean : 1362

3rd Qu.: 1428

Max. :102127

loan contact day

no :37967 cellular :29285 Min. : 1.00

yes: 7244 telephone: 2906 1st Qu.: 8.00

unknown :13020 Median :16.00

Mean :15.81

3rd Qu.:21.00

Max. :31.00

month duration campaign

may :13766 Min. : 0.0 Min. : 1.000

jul : 6895 1st Qu.: 103.0 1st Qu.: 1.000

aug : 6247 Median : 180.0 Median : 2.000

jun : 5341 Mean : 258.2 Mean : 2.764

nov : 3970 3rd Qu.: 319.0 3rd Qu.: 3.000

apr : 2932 Max. :4918.0 Max. :63.000

(Other): 6060

pdays previous poutcome

Min. : -1.0 Min. : 0.0000 failure: 4901

1st Qu.: -1.0 1st Qu.: 0.0000 other : 1840

Median : -1.0 Median : 0.0000 success: 1511

Mean : 40.2 Mean : 0.5803 unknown:36959

3rd Qu.: -1.0 3rd Qu.: 0.0000

Max. :871.0 Max. :275.0000

y

no :39922

yes: 5289

**> sum(is.na(bank))**

**[1] 0**

**> #NA values are not present in the data,so no need of imputation**

**> # Preparing a linear regression**

**> lmmodel.bank<-lm(y~.,data = bank)**

Warning messages:

1: In model.response(mf, "numeric") :

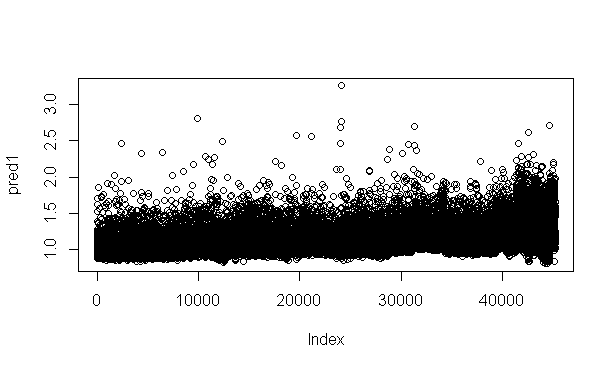
using type = "numeric" with a factor response will be ignored

2: In Ops.factor(y, z$residuals) : ‘-’ not meaningful for factors

> pred1<-predict(lmmodel.bank,bank)

**> # We can no way use the linear regression technique to classify the data**

**> plot(pred1)**



**> #From plot LR model cannot be used to classify the data.**

**> #So we will build the Glm model.**

**> # GLM function use sigmoid curve to produce desirable results**

**> # The output of sigmoid function lies in between 0-1**

**> glmmodel.bank<-glm(y~.,data = bank,family = binomial)**

**> summary(glmmodel.bank)**

Call:

glm(formula = y ~ ., family = binomial, data = bank)

Deviance Residuals:

Min 1Q Median 3Q Max

-5.7286 -0.3744 -0.2530 -0.1502 3.4288

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.536e+00 1.837e-01 -13.803 < 2e-16

age 1.127e-04 2.205e-03 0.051 0.959233

jobblue-collar -3.099e-01 7.267e-02 -4.264 2.01e-05

jobentrepreneur -3.571e-01 1.256e-01 -2.844 0.004455

jobhousemaid -5.040e-01 1.365e-01 -3.693 0.000221

jobmanagement -1.653e-01 7.329e-02 -2.255 0.024130

jobretired 2.524e-01 9.722e-02 2.596 0.009436

jobself-employed -2.983e-01 1.120e-01 -2.664 0.007726

jobservices -2.238e-01 8.406e-02 -2.662 0.007763

jobstudent 3.821e-01 1.090e-01 3.505 0.000457

jobtechnician -1.760e-01 6.893e-02 -2.554 0.010664

jobunemployed -1.767e-01 1.116e-01 -1.583 0.113456

jobunknown -3.133e-01 2.335e-01 -1.342 0.179656

maritalmarried -1.795e-01 5.891e-02 -3.046 0.002318

maritalsingle 9.250e-02 6.726e-02 1.375 0.169066

educationsecondary 1.835e-01 6.479e-02 2.833 0.004618

educationtertiary 3.789e-01 7.532e-02 5.031 4.88e-07

educationunknown 2.505e-01 1.039e-01 2.411 0.015915

defaultyes -1.668e-02 1.628e-01 -0.102 0.918407

balance 1.283e-05 5.148e-06 2.493 0.012651

housingyes -6.754e-01 4.387e-02 -15.395 < 2e-16

loanyes -4.254e-01 5.999e-02 -7.091 1.33e-12

contacttelephone -1.634e-01 7.519e-02 -2.173 0.029784

contactunknown -1.623e+00 7.317e-02 -22.184 < 2e-16

day 9.969e-03 2.497e-03 3.993 6.53e-05

monthaug -6.939e-01 7.847e-02 -8.842 < 2e-16

monthdec 6.911e-01 1.767e-01 3.912 9.17e-05

monthfeb -1.473e-01 8.941e-02 -1.648 0.099427

monthjan -1.262e+00 1.217e-01 -10.367 < 2e-16

monthjul -8.308e-01 7.740e-02 -10.733 < 2e-16

monthjun 4.536e-01 9.367e-02 4.843 1.28e-06

monthmar 1.590e+00 1.199e-01 13.265 < 2e-16

monthmay -3.991e-01 7.229e-02 -5.521 3.36e-08

monthnov -8.734e-01 8.441e-02 -10.347 < 2e-16

monthoct 8.814e-01 1.080e-01 8.159 3.37e-16

monthsep 8.741e-01 1.195e-01 7.314 2.58e-13

duration 4.194e-03 6.453e-05 64.986 < 2e-16

campaign -9.078e-02 1.014e-02 -8.955 < 2e-16

pdays -1.027e-04 3.061e-04 -0.335 0.737268

previous 1.015e-02 6.503e-03 1.561 0.118476

poutcomeother 2.035e-01 8.986e-02 2.265 0.023543

poutcomesuccess 2.291e+00 8.235e-02 27.821 < 2e-16

poutcomeunknown -9.179e-02 9.347e-02 -0.982 0.326093

(Intercept) \*\*\*

age

jobblue-collar \*\*\*

jobentrepreneur \*\*

jobhousemaid \*\*\*

jobmanagement \*

jobretired \*\*

jobself-employed \*\*

jobservices \*\*

jobstudent \*\*\*

jobtechnician \*

jobunemployed

jobunknown

maritalmarried \*\*

maritalsingle

educationsecondary \*\*

educationtertiary \*\*\*

educationunknown \*

defaultyes

balance \*

housingyes \*\*\*

loanyes \*\*\*

contacttelephone \*

contactunknown \*\*\*

day \*\*\*

monthaug \*\*\*

monthdec \*\*\*

monthfeb .

monthjan \*\*\*

monthjul \*\*\*

monthjun \*\*\*

monthmar \*\*\*

monthmay \*\*\*

monthnov \*\*\*

monthoct \*\*\*

monthsep \*\*\*

duration \*\*\*

campaign \*\*\*

pdays

previous

poutcomeother \*

poutcomesuccess \*\*\*

poutcomeunknown

---

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: 21562 on 45168 degrees of freedom

AIC: 21648

Number of Fisher Scoring iterations: 6

**> # Confusion matrix table**

**> prob<-predict(glmmodel.bank,bank,type = "response")**

**> # We are going to use NULL and Residual Deviance to compare the between different models**

**> # Confusion matrix and considering the threshold value as 0.5**

**> confusion<-table(prob>0.5,bank$y)**

**> confusion**

no yes

FALSE 38940 3456

TRUE 982 1833

**> # Model Accuracy**

**> Accuracy<-sum(diag(confusion)/sum(confusion))**

**> Accuracy**

[1] 0.901838