ML PROJECT

BANK MARKETING

Student:

Valerio Botta

PROJECT PRESENTATION

Motivation

- **Problem**: Predict if the client will subscribe (yes/no) a term deposit
- **Dataset**: The data is related with direct marketing campaigns of a Portuguese banking institution

The dataset was updated in 2014 but the issue is still relevan: the sell of bank/financial product via telephone. It is a common way to sell products. The analysis of this dataset has the goal to create a program that could predict a sale and validate the results.

Method

Data Set 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, to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Dataset url: https://archive-beta.ics.uci.edu/ml/datasets/bank+marketing

My dataset has named “bank-additional-full.csv” with 41188 examples and 20 features (plus the target one).

Dataset structure

Input variables:
 # bank client data:
1 - **age** (numeric)
2 - **job** : type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown")
3 - **marital** : marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)
4 - **education** (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown")
5 - **default**: has credit in default? (categorical: "no","yes","unknown")
6 - **housing**: has housing loan? (categorical: "no","yes","unknown")
7 - **loan**: has personal loan? (categorical: "no","yes","unknown")
 # related with the last contact of the current campaign:
8 - **contact**: contact communication type (categorical: "cellular","telephone")
9 - **month**: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
10 - **day_of_week**: last contact day of the week (categorical: "mon","tue","wed","thu","fri")
11 - **duration**: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

 # other attributes:
12 - **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - **previous**: number of contacts performed before this campaign and for this client (numeric)
15 - **poutcome**: outcome of the previous marketing campaign (categorical: "failure","nonexistent","success")
social and economic context attributes
16 - **emp.var.rate**: employment variation rate - quarterly indicator (numeric)
17 - **cons.price.idx**: consumer price index - monthly indicator (numeric)
18 - **cons.conf.idx**: consumer confidence index - monthly indicator (numeric)
19 - **euribor3m**: euribor 3 month rate - daily indicator (numeric)
20 - **nr.employed**: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21 - **y** - has the client subscribed a term deposit? (binary: "yes","no")

Columns that I could use and relevat to the analysis are: age, job, marital, default, housing, loan, campaign, previous, poutcome, emp-var.rate, cons.price.idx, cons.conf.idx.

The last column ("y") is the target to predict. It's a simple binary choice: yes or no based on possibile subscribtion of the term deposit.

Cleaning Dataset

The dataset need an alteration to obtain a "cvs" file without column separation, with comma delimiter between data and without blank space. It'll be checked to delete all null row or any problem.

ML Algorithms to perform

Algorithms to perform:

- Logistic regression
- Neural network for classification
- Decision tree classifier

Intended experiments

The goal is to create a program that can predict the sale of a banking product. I'll divide into train set, validation set and test the dataset and perform the k-folds cross validation.

For the classification algorithm I'll compute the metrics of accuracy, confusion matrix and ROC curve to validate the results.

The last step consists in comparing the results to choose the better algorithm for this dataset.