

Appendimento Statistico

Nenna Giulio

Data Analysis applied to a Marketing Dataset

1 Introduction

The main goal of this report is to apply different machine learning techniques to a [dataset](#) that can be found on the [UCI Machine Learning Repository](#). The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. Each entry represents a phone call to a client of whom several features are provided.

The response variable is binary (yes or no) and indicates whether a phone call successfully convinced the client to purchase a financial product the bank is trying to sell (bank term deposit). Among the 20 provided features, 11 are categorical and the remaining are numerical. An extensive description of each feature will now follow.

1.1 Attribute information

Table [1](#) provides an extensive overview about each attribute featured in the dataset.

Important notes about some of the attributes are the following:

- Attributes from **Contact** to **duration** are related to the last contact of the current campaign
- The attribute **duration** refers to the call duration and it's not known until the call is completed. It highly affects the prediction (e.g. if **duration**=0 then the outcome will always be *not successful*), hence it will be deleted since the model is meant to be deployed with real world data.

| Name | Description | Type | Possible values |
|--------------------------------|--|-------------|---|
| Age | Age of each client | numeric | - |
| Job | Type of job | categorical | Admn, blue-collar, entrepreneur... |
| Marital | Marital Status | categorical | Divorced, married, single, unknown |
| Education | level of education | categorical | basic.4y, basic.6y, basic.9y, high-school,... |
| Default | whether a client has credit in default | Categorical | Yes, No, Unknown |
| Housing | whether a client has an housing loan | Categorical | Yes, No, Unknown |
| Loan | whether a client has a personal loan | Categorical | Yes, No, Unknown |
| Contact | contact communication type | Categorical | cellular, telephone |
| Month | last contact month of the yeah | Categorical | jan, feb, mar... |
| Day of the week | last contact day of the week | Categorical | mon, tue, ... |
| Duration | Last contact duration in seconds (see notes) | Numeric | - |
| Campaign | number of contacts performed during this campaign for that client | Numeric | - |
| pdays | number of days that passed by after the client was last contacted from a previous campaign | numeric | - |
| previous | number of contacts performed before this campaign for this client | numeric | - |
| poutcome | outcome of the previous marketing campaign | Categorical | Failure, nonexistent, success |
| emp.var.rate | employment variation rate - quarterly indicator | Numeric | - |
| cons.price.idx | consumer price index - monthly indicator | Numeric | - |
| cons.conf.idx | consumer conficence index - monthly indicator | Numeric | - |
| euribor3m | euribor 3 month rate - daily indicator | Numeric | - |
| nr.employed | Number of employees - quarterly indicator | Numeric | - |
| Responce variable: y | Has the client subscribed to a term deposit | Binary | Yes, No |

Table 1: Extensive description for each attribute in the dataframe

2 Data Exploration and Preprocessing

We first import the dataset and encode each categorical feature using *one-hot encoding*. This means that numerical features will be left unchanged while for each categorical feature the process is the following:

1. Determine all the distinct values of that feature (categories)
2. For each category generate a new binary column
3. Assign values to the binary columns according to categories featured in each line.

| feature | category 1 | category 2 | category 3 |
|------------|------------|------------|------------|
| category 1 | 1 | 0 | 0 |
| category 2 | 0 | 1 | 0 |
| category 3 | 0 | 0 | 1 |
| category 2 | 0 | 1 | 0 |

```

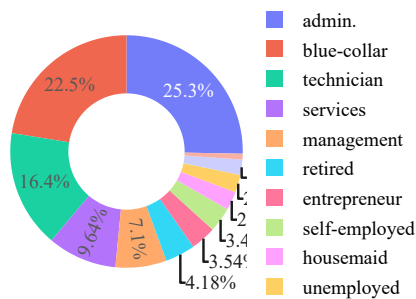
1 import pandas as pd
2
3 df = pd.read_csv('Data/bank/bank-additional-full.csv', sep=';')
4 display(df.head())
5 cat_col = df.dtypes=='O'
6 df_enc = pd.get_dummies(df.loc[:, cat_col], prefix=df.columns[cat_col])
7 df_enc = df_enc.join(df.loc[:, np.logical_not(cat_col)])
8
9 df_enc = df_enc.drop('y_no', axis=1) #deleting column since attribute "y" is
10 df_enc = df_enc.drop('duration', axis = 1) #drop the duration column (see
    attribute information)

```

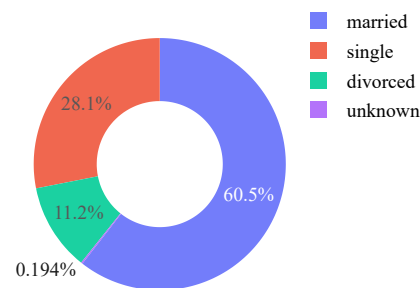
Listing 1: Data encoding

2.1 Data Exploration

Independently from encoding we can also perform some data exploration, in order to gain useful insights about some of the attributes featured in the dataset. There are 11 categorical features and the remaining are numerical. Categorical data will be explored by means of *pie charts* while numerical data will be explored by means of a *scatterplot matrix*



(a) caption1



(b) caption2