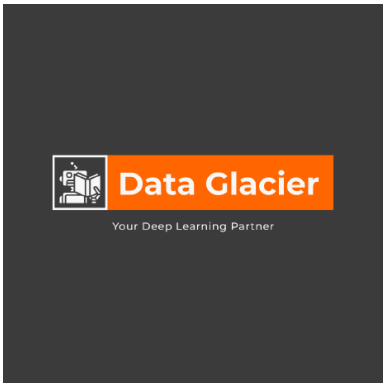
DATA SCIENCE INTERNSHIP - DATA GLACIER

Project: Bank Marketing (Campaign) -- Group Project



Group Name: **Datazoids**

Name: **Efe KARASIL - Sefa Sözer**

E-mail: **ekarasil@sabanciniv.edu - sefasozer9@gmail.com**

Country: **Turkey - Turkey**

College: **Sabancı University - Trakya University**

Specialization : **Data Science**

Problem Description:

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which helps them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Business Understanding:

Bank wants to use the ML model to shortlist customers whose chances of buying the product is more so that their marketing channel (telemarketing, SMS/email marketing etc) can focus only on those customers whose chances of buying the product is more.

This will save resources and their time (which is directly involved in the cost (resource billing)).

The data is related to 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 ('yes') or not ('no') subscribed.

The classification goal is to predict if the client will subscribe (yes/no) to a term deposit (variable y).

Citation Request:
This dataset is publicly available for research. The details are described in [Moro et al., 2014].
Please include this citation if you plan to use this database:

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001

Available at: [pdf] http://dx.doi.org/10.1016/j.dss.2014.03.001
[bib] http://www3.dsi.uminho.pt/pcortez/bib/2014-dss.txt

1. Title: Bank Marketing (with social/economic context)

2. Sources
Created by: Sérgio Moro (ISCTE-IUL), Paulo Cortez (Univ. Minho) and Paulo Rita (ISCTE-IUL) @ 2014

3. Past Usage:

The full dataset (bank-additional-full.csv) was described and analyzed in:

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems (2014), doi:10.1016/j.dss.2014.03.001.

4. Relevant Information:

This dataset is based on the "Bank Marketing" UCI dataset (please check the description at: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing).

The data is enriched by the addition of five new social and economic features/attributes (nationwide indicators from a ~10M population country), published by the Banco de Portugal and publicly available at: https://www.bportugal.pt/estatisticasweb.

This dataset is almost identical to the one used in [Moro et al., 2014] (it does not include all attributes due to privacy concerns).

Using the rminer package and R tool (http://cran.r-project.org/web/packages/rminer/), we found that the addition of the five new social and economic attributes (made available here) lead to substantial improvement in the prediction of a success, even when the duration of the call is not included. Note: the file can be read in R using:
d=read.table("bank-additional-full.csv",header=TRUE,sep=";")

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

5. Number of Instances: 41188 for bank-additional-full.csv

6. Number of Attributes: 20 + output attribute.

7. Attribute information:

For more information, read [Moro et al., 2014].

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")

8. Missing Attribute Values: There are several missing values in some categorical attributes, all coded with the "unknown" label. These missing values can be treated as a possible class label or using deletion or imputation techniques.

EDA- Exploratory Data Analysis:

**Most of the explanations are done in the python notebook. Please refer to that file.

Recommendations:

- In correlation heat map, “euribo3rn”, “cons.priceidx”, “nr.employed” and “emp.var.rate” features have very high correlation between them. 2 or 3 of these features can be dropped from the data, since their existence will not be extra useful-providing good, new information- for machine learning model which will be deployed in next steps.
- Call duration can be increased to convince customers

- Older people can be called more, since they have higher rates to accept subscription
- Students and retired people can be reached more as well, again their rate of acceptance are higher according to different jobs' subscriptions
- Calls made in the December, March, October and September months can be increased, because they have the least amount of calls while they have the highest rate of acceptance
- There is so little knowledge and data about illiterate people, but increasing the amount of calls to those people can be beneficial