Group Detail

Name: Jose Vicente Solorzano

Email: solorzano.vco@gmail.com

Country: Argentina

College: Montpellier Business School

Specialization: Data Science

Problem Description

ABC bank is about to launch its new product, a term deposit. Before the launching, they want to develop a model which help 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).

We were provided with a dataset of clients and features in csv format ("bank-additional-full.csv").

The dataset has the following information:

Data related with clients:

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')

Data related with the last contact:

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.

Data related with the campaing:

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')

Data related with 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)

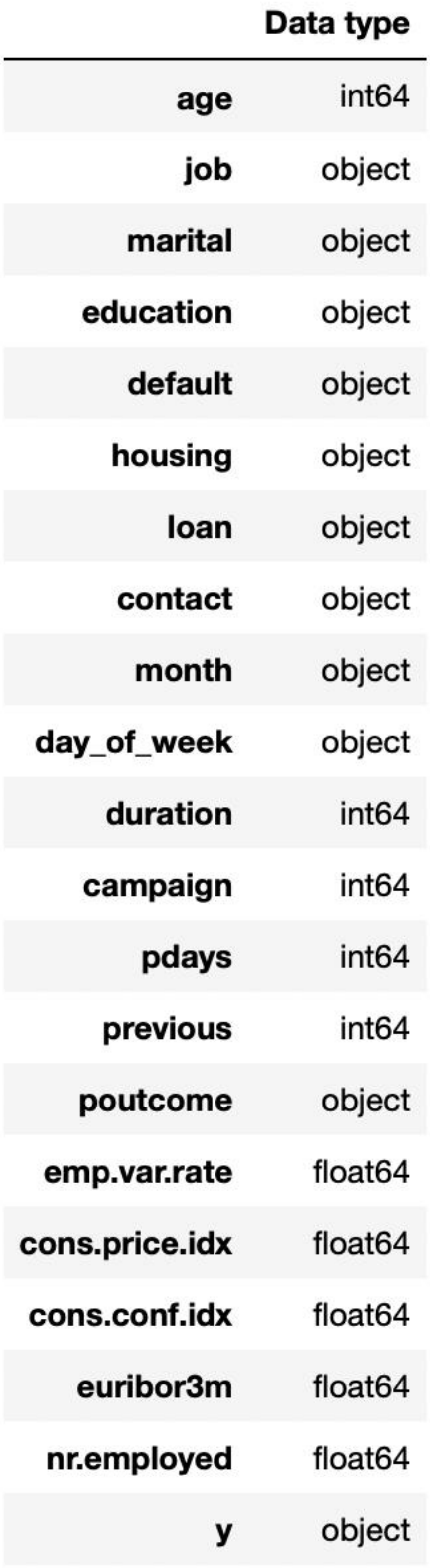
19) euribor3m: euribor 3 month rate - daily indicator (numeric)

20) nr.employed: number of employees - quarterly indicator (numeric)

Target:

21) y - has the client subscribed a term deposit? (binary: 'yes','no')

Features types:



	Data type
age	int64
job	object
marital	object
education	object
default	object
housing	object
loan	object
contact	object
month	object
day_of_week	object
duration	int64
campaign	int64
pdays	int64
previous	int64
poutcome	object
emp.var.rate	float64
cons.price.idx	float64
cons.conf.idx	float64
euribor3m	float64
nr.employed	float64
y	object

What are the problems found in data?

- There are instances with an “unknown” string in some features. This is going to be treated as null values. (Features with “unknown”: job 0.8%, marital 0.2%, education 4.2%, default 20.9%, housing 2.4% and loan 2.4%).
- Target “y” is imbalanced. No = 89% and yes = 11%.
- There is a lot of outliers in the features.
- “Duration” is left skewed. Equally, it is a feature to use with careful because after the end of the call y is obviously known. It could not be useful, so for now, the skewed is not a problem.

Problems Approach

For nulls, we will analyze if it is more convenient imputation or deletion. We have to be careful not to cause bias while imputing. We will experiment with both approaches in order to choose the best for the model.

Outliers going to be treated with IQR and detection by clustering (dbscan). The idea is to use a statistic option and contrast its results with one more datum based. We will keep the one better for the model.

Regarding the unbalance of the target “y”, we will consider two approaches. In one hand, a more passive approach, we will not generate nothing new. We will just consider the right metric for the model, in order to avoid the impact of the unbalance (recall, F1 score, ROC, etc).

In the other hand, we will apply SMOTE (Synthetic minority over-sampling technique), a common solution for this kind of problems in machine learning classification models.

Gitub Repo Link

https://github.com/jvsolorzano96/bank_marketing_campaign/tree/main/Week%208%20-%20Problems%20approach