BANK MARKETING CAMPAIGN

NAME: CATHERINE SANDA
EMAIL: sandacate@gmail.com
COUNTRY: KENYA
SPECIALIZATION: DATA SCIENCE
GITHUB REPO LINK: https://github.com/cate6495/week9-bank

PROBLEM DESCRIPTION

We are given data related to direct marketing campaigns i.e. phone calls of a bank in Portugal.

The classification goal is to predict whether a client will subscribe or not(yes/no) to a term deposit (variable y).

DATA CLEANSING AND TRANSFORMATION

1.The attribute ‘duration’ highly affects our output target in that if duration= 0 then y=0, yet duration is not known before a call is made. At the end of the call then y is known. Hence duration should only be included for benchmark purposes. We will discard it as we intend to have a realistic predictive model.

2. To ensure that I have fewer columns after encoding I joined some attributes in job and education variables.

3.Copy the original data so that all the transformation is done on the duplicate data.

4. Convert the categorical variables to numeric using one hot encoding technique. In one-hot encoding, you create a new column for each unique value in that column. Then the value of the column is 1 if the sample has that unique value or 0 otherwise.

5. Scale the data.

6.Fix the imbalance part of the dataset.

7.Apply dimensionality reduction using PCA.

8.Save the preprocessed data for use in modelling