Group Name: Data Pro 2023

Name: Ana Lilliam Recio Garcia

Email: anarecio.89@gmail.com

Country: Costa Rica

College/Company: Universidad Estatal a Distancia de Costa Rica

Specialization: Data Science

Problem description: Bank Marketing (Campaign). A bank wants to perform a commercial campaign in order to promote their term deposit product to clients, but they want it to be shortlisted to customers with a highest probability of buying their products. The idea is to use a ML model which allow this identification of customers to be utilized for their marketing channel.

Business understanding: The objective to develop this model is to identify the customers who are more likely to buy the product that the bank is offering (term deposit), based on the past interactions of these clients, this with the intention of saving costs and time to the bank's marketing department, so they can focus their marketing campaign in these clients.

Project lifecycle along with deadline

Tasks	19/11/2022	26/11/2022	3/12/2022	10/12/2022	17/12/2022	24/12/2022	31/12/2022
Business Understanding	Done						
Data understanding	Done						
Exploratory data Analysis							
Data Preparation							
Model Building							
Model Selection							
Performance reporting							
Deploy the model							
Converting ML metrics into Business metric							
Prepare presentation for non technical persons.							

Data Intake Report

Name: Bank Marketing (Campaign)
Report date: 14/11/2022
Internship Batch: LISUM14
Version: 1.0
Data intake by: Ana Lilliam Recio Garcia
Data intake reviewer: NA
Data storage location: https://github.com/AnaRecio/Bank-Marketing-Campaign

Tabular data details: bank-additional-full.csv

Total number of observations	864969
Total number of files	41188
Total number of features	21
Base format of the file	csv
Size of the data	372 KB

Tabular data details: bankfull.csv

Total number of observations	768604
Total number of files	45211
Total number of features	17
Base format of the file	csv
Size of the data	503 KB

What type of data you have got for analysis:

I’ve tried to analyze the numerical columns in both data sets to look for skewed data and outliers.

What are the problems in the data:

No NA values were found.

Although many of the data analyzed was skewed to the right (balance, age, duration)

We can see outliers in age where the mean is 40 and the max is 98 and in balance where the mean is 1362 and the max is 102127.

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier, etc., and why?

Still need to analyze the information in the data frames to decide if it’s important to keep the outliers or to delete them.