Deliverables – Week 8

BANK MARKETING (CAMPAIGN) – GROUP PROJECT

➢ TEAM MEMBER'S DETAILS

- **Group Name:** Evolve Data
- **Name:** Giovanna Carolina Magalhães Vieira
- **E-mail:** giihvieira1703@gmail.com
- **Country:** Brazil
- **College/Company:** Hashtag Treinamentos
- **Specialization:** Data Science
- **GitHub Repo Link:** https://github.com/vgiih/EvolveData.git

➢ PROBLEM'S DESCRIPTION

Problem Statement:

ABC Bank wants to sell it's term deposit product to customers and before launching the product 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).

Problem Description:

The Portuguese Bank's strategy is to use a prediction model to shortlist customer whose chances of buying the product is more so that their marketing channel can focus only to those customers whose chances of buying the product is more.

Knowing who is more likely to buy their product, they can save resource and time. The more targeted campaign, the more successful they should be!

➢ DATA UNDERSTANDING

The data is related with 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) a term deposit (variable y).

- **WHAT TYPE OF DATA YOU GOT FOR ANALYSIS?**

Attribute Information:

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')

- **WHAT ARE THE PROBLEMS WITH THE DATA?**

The data seems to be cleaned and a little bit skewed, however you can see that the variables have outliers that need to be cleaned with data cleaning process.

There are no null values, but there are some unknown values.

We've got 41188 rows, some of the columns has 85% of the values repeated.

The value types seem to be correct but as said there are some "unknown" values.

- **WHAT APPROACHES ARE YOU TRYING TO APPLY ON YOUR DATASET TO OVERCOME THESE PROBLEMS AND WHY?**

1. As outliers can change the meaning of data, we will observe them and there's a high probability of dropping them all.

2. We will count values like "unknown" and decide, we have 3 possibilities:
   - Drop the row, when it's a small amount of this value.
   - Drop the column if almost all of them have this "unknown" value.
   - Replace the value in the most fitting way (for example, mode, mean and median).

3. About columns with more than 85% of the values repeated we can simply drop the column if it will not be important for our prediction model. But not the 'y' column, that has the answer for "Has the client subscribed a term deposit?"

- **GITHUB REPO LINK**

https://github.com/vgiih/EvolveData.git