



# Virtual Internship Data Science Data Intake Report

**Group Name: LISUM01: Data science Group 1**

**Members:**

1. William Ogweli Okomba, Kenya
2. Ece Kurnaz, Turkey
3. Collin Mburugu, Kenya
4. Udbhav Balaji, India

**Name:** Bank Marketing(Campaign)

**Report date:** 23/07/2021

**Internship Batch:** LISUM01

**Version:**1.0

**Data intake by:** William Ogweli Okomba

**Data intake reviewer:** Intern who viewed the report

**Data storage location:** [https://github.com/williamokomba/DataGlacier\\_Internship-Data-science/tree/main/Week\\_7\\_Group\\_Project\\_Data\\_Science](https://github.com/williamokomba/DataGlacier_Internship-Data-science/tree/main/Week_7_Group_Project_Data_Science)

## Problem statement

ABC bank (a Portuguese banking institution) has a term deposit product that is desired to be sold to clients. We will focus on customer's past interactions with the bank or other financial institutions to have a better understanding on whether these particular clients will buy this product or not. Developing a model with using machine learning for this aim is reasonable. With performing this project, our aim is to save resources and time for ABC bank.

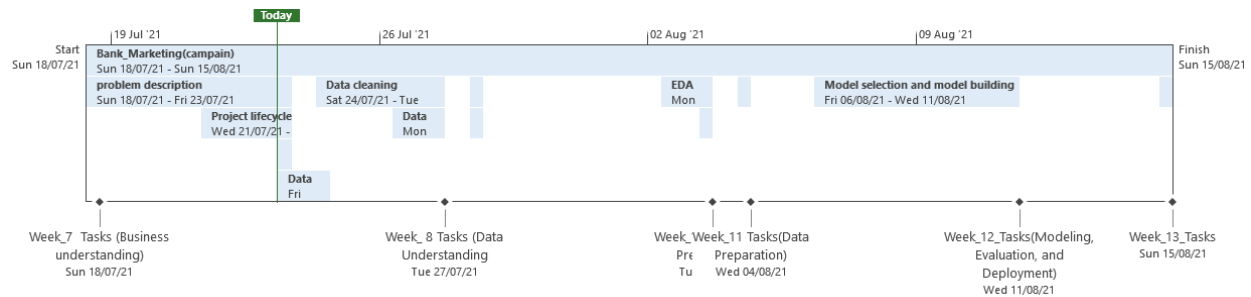
## Business Understanding

Bank term deposit is a deposit product by ABC Bank with is offered to their customers in Portugal. The potential customers are likely to buy the product when educated by marketing channel (tele marketing, SMS/email marketing etc) personnel.

The approval is based on a variety of information, from basic biographical data to the loan applications that come through daily.

We work with the product team as a data scientists to help create effective predictive model used to assess the customer chances of buying the product.

## Project lifecycle



### 1. File 1: Tabular data details: bank\_additional\_full.csv

<b>Total number of observations</b>	41188
<b>Total number of files</b>	2
<b>Total number of features</b>	21
<b>Base format of the file</b>	.csv
<b>Size of the data</b>	5834924 BYTES(5.56MB)

### 2. File 2: Tabular data details: bank\_additional.csv

<b>Total number of observations</b>	4119
<b>Total number of files</b>	2
<b>Total number of features</b>	21
<b>Base format of the file</b>	.csv
<b>Size of the data</b>	583898 BYTES(572KB)

## Proposed Approach of dedup validation (identification)

1. The data cover a period from May 2008 to November 2010.
2. There are 2 dataset, the second dataset is a sample of the first dataset.
3. There are 10 integers and 11 categorical variables.
4. The missing values in both datasets are presented by "unknown" string. We changed it to NaN.
5. There are missing values in six variables namely, job, marital status, education, default, housing, and loan. This will be imputed using various methods.
6. There are 12 duplicates in the first dataset and no duplicates in the sample dataset, this will be dropped since they are minimal and will not affect our analysis.

7. The target variable is unbalanced class, "no" class has more observation than "yes" class in both dataset.
8. Columns are not uniformed named for example "day\_of\_week", and "emp.var.rate". This need to be modified for make it easier to work with.
9. All variables in both datasets have the right datatypes.

### **Assumptions**

1. We assume the data provided is correct and up to date