
# *Data Science Final Project :*

# *Bank Marketing Campain prediction Using Machine Learning*


# *Data Science Final Project :Week8*

## *Plan:*

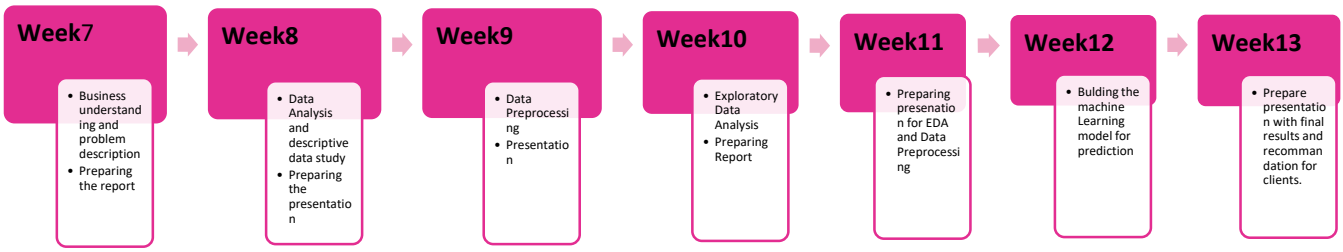
- *Project details*
- *Project Process*
- *Problem Description*
- *Data Analysis*
  - *1/ Dataset and Attribute Informations*
  - *2/Missing Values and Outliers*
- *Data Preprocessing recommandations*

## Project Informations:

| | |
|---|---|
| **Team Members Informations** | **Name of Group** : **Data Scientist Geeks** |
| | Member 1:***Refka Mejri*** - Tunisia/National Engineering School of Tunis |
| | Member 2 :***Tasnime Hamdeni*** - Tunisia/National Engineering School of Tunis |

### *Project Roadmap and Process:*

*This process is prepared according to the needs of the company and the submission of each week.*




## Problem Understanding and Business Need of the Company:

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 and other financial Institution)

In order to achieve their goal and need they demand to Data Glacier Company to help them with a AI model for prediction.

The Data Glacier company give us as a data scientist team this mission.
We will develop a Ml Model to shortlist customer whose chances of buying the product is more so that their marketing channel can focus only on those customers whose chances of buying the product is more.

# Data Analysis :

## Dataset and Attributes informations :

### Bank_additional_full_data details:

| | |
|---|---|
| **Total number of observations** | 41188 |
| **Total number of files** | 1 |
| **Total number of features** | 21 |
| **Base format of the file** | .csv |
| **Size of the data** | 6.6 MB |

## Attribute Information:

Input variables:

- **Age** (numeric)
- **Job** : type of job (categorical: 'admin.','blue collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
- **Marital** : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

- **Education** (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
- **Default**: has credit in default? (categorical: 'no','yes','unknown')
- **Housing**: has housing loan? (categorical: 'no','yes','unknown')
- **Loan**: has personal loan? (categorical: 'no','yes','unknown')

***Related with the last contact of the current campaign:***

- **Contact:** contact communication type (categorical: 'cellular','telephone')
- **Month**: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- **Day_of_week**: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- **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:***

- **Campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- **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)
- **Previous**: number of contacts performed before this campaign and for this client (numeric)
- **Poutcome**: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

***Social and economic context attributes***

- **Emp.var.rate**: employment variation rate - quarterly indicator (numeric)
- **Cons.price.idx**: consumer price index - monthly indicator (numeric)
- **Cons.conf.idx**: consumer confidence index - monthly indicator (numeric)
- **Euribor3m**: euribor 3 month rate - daily indicator (numeric)
- **Nr.employed:** number of employees - quarterly indicator (numeric)

***Output variable (desired target):***

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

We see from boxplot that we have outliers.Outliers are observations that a re far away from the other data points in a random sample of a population. Outliers may reveal unexpected knowledge about a population that is why we ought to handle in our EDA steps.
We dont have missing values in our dataset but,  generally in the real wor d, data has a lot of missing values.

The cause of missing values can be data corruption or failure to record da ta. Handling missing values is one of the data preprocessing essential ste p in EDA.

It's very important step as it help as clean our data and also due to many machine learning algorithms how don't support missing values.