**OBJECTIVE**

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 objectives include

* Cleaning the data and Impute missing values.
* Figure out which model to apply.
* Figuring out the training and testing set
* Removing bias and variance, if any.
* Optimizing the model
* The final question to be answered is if the client has subscribed to a term deposit or not?

The data set includes the following variable:

1 - age (numeric)

2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur",

"student","blue-collar","self-employed","retired","technician","services")

3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)

4 - education (categorical: "unknown","secondary","primary","tertiary")

5 - default: has credit in default? (binary: "yes","no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes","no")

8 - loan: has personal loan? (binary: "yes","no")

# related with the last contact of the current campaign:

9 - contact: contact communication type (categorical: "unknown","telephone","cellular")

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - duration: last contact duration, in seconds (numeric)

# other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")'

The dataset contains 45211 rows. The data set contains no missing values. Since many of these variables are categorical, we need to provide dummy variables in order to use them in the model.

Train Test Split

The dataset is unbalanced, as only 5289 (11.69%) records are related with successes. For evaluation purposes, a time ordered split was initially performed, where the records were divided into training (four years) and test data (one year).

Feature Selection

From the 16 features that have been given to us, we need to select the features that significantly contribute to our models. We could manually do this by analyzing the which features are relevant to our objective. This is time consuming and hence we will use inbuilt python libraries. We would want to check for co-relation among the features as well.

Model Selection

The output of our data set is a binary classification of yes or no. The models that we could use are

* Naïve Bayes