**Data Pre-processing:**

The bank marketing data set that we have used is almost a clean dataset and does not contain any missing values. However, there are some changes such as renaming of variables, identifying the outliers etc that have been made in few attributes as mentioned below:

* Firstly, we have selected a sample data from the original data which contains 22606 instances (50% of the original data). This has been done to get accurate models with minimum misclassification.
* Dataset observations – 22606, predictors – 16, outcome variable - 1
* There are no missing values in the dataset.
* The variable "default" has been renamed as "default\_credit" which means whether the client has any credit in default or not. The renaming has been done to make the variable more specific.
* The variable "y" (outcome variable) which indicates whether a client has subscribed a term deposit or not has been renamed as "term\_deposit\_susbscribed". The renaming variable gives a meaning to this variable which makes the outcome easy to understand.
* **Age:** The variable "age" had 180 outliers(Age >= 74 years) [min = 18, max = 95] , so the data was right-skewed. We couldn’t remove the outliers as they add meaning to the data, and they showed nature of the age of people(right-skewed). Just to have better understanding of the distribution of the age, it has been converted to categorical type. Three distinct age groups are created which are: Youth (age between 18 and 35), Workforce (between 36 and 59) and Retired (above 60).
* **Balance:** This variable had 2400 outliers with some outliers(balance < -1884 euros & balance > 3412 euros) under debt. The maximum account balance was recorded as 98417 euros and maximum debt was 8019 euros, which are plausible amounts therefore no imputation was carried out.
* **Duration:** This variable had 1600 outliers(duration >= 646 seconds) with minimum and maximum call duration of 2 and 3881 seconds. As, this was also reliable measure so here also no imputation was done.
* **Campaign:** This variable is right-skewed and has around 1500 outliers(campaign >= 7)[min = 1, max = 58]. In order to maximize the Clients for subscription of term deposit 58 is quite reliable count so no change was made in this variable as well.
* **Pdays:** This variable was also right skewed and had around 4000 outliers(pdays >= 1)[min = -1, max = 854]. Most values were -1 which means client was not previously contacted. Due to very high frequency of -1 the other values were treated as outliers. So, we didn’t remove them.
* **Previous:** This variable was also right skewed and had around 4000 outliers(previous >= 1)[min = 0, max = 275]. Most values were 0 which means client was not contacted before campaign. Due to very high frequency of 0 the other values were treated as outliers. So, we didn’t remove them.
* **Poutcome:** The observation "other" was merged with "unknown" as they belong to the same hierarchy according to the dataset and also there were total 900 records for ‘others’. And also, they both don’t convey any meaning.

Neither new variables nor any concept hierarchies were created. Neither any observations nor any variables were excluded from the analysis. Dataset after data preprocessing observation – 22606, predictors – 16, outcome variable – 1.