Week 8

1-Team members details :
Group name: Scientists
Members

Name	Email	Country	College/Company
Aly Ahmed Refaat	alyahmed1947@gmail.com	Egypt	Fresh computer engineering graduate, Cairo university
Mazen Hawwa	xotofloyt@gmail.com	United Arab Emirates	Self Employed - IT
Mohammed Wahba	mohammedwahba9@gmail.com	Saudi Arabia	computer science
Betty Wairegi	wairreb@gmail.com	Kenya	USIU student, IT

Specialization: Data science

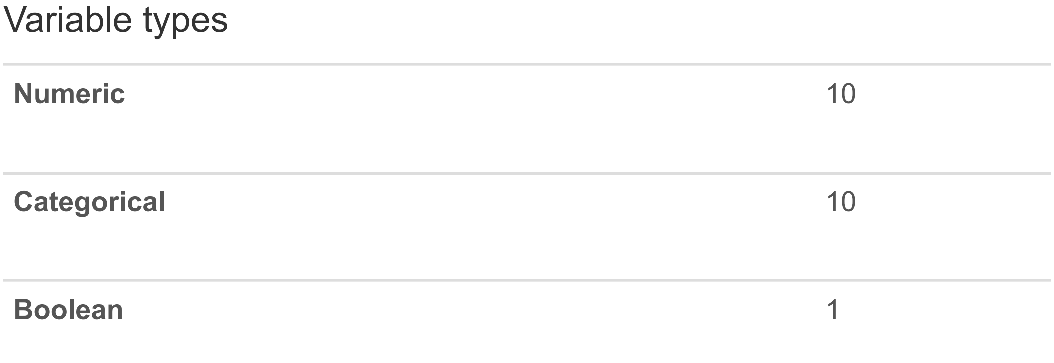
2-Problem description:

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 , the model will predict whether the customer will buy the product or not based on customer's past interaction with bank or other Financial Institution.

3. Data understanding
The list of data types for this dataset include:

```
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6  loan          41188 non-null  object
 7  contact        41188 non-null  object
 8  month          41188 non-null  object
 9  day_of_week    41188 non-null  object
 10  duration       41188 non-null  int64
 11  campaign       41188 non-null  int64
 12  pdays          41188 non-null  int64
 13  previous       41188 non-null  int64
 14  poutcome       41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m      41188 non-null  float64
 19  nr.employed    41188 non-null  float64
 20  y             41188 non-null  object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

- This dataset has no NA values.



Variable types

Numeric	10
Categorical	10
Boolean	1

- Duplicates=12 rows

Problems with the data

- Imbalanced class- target variable

Approaches to solve the problems

- Under-sampling: This method reduces the number of the majority class by randomly eliminating some of the training set observations; it could cause underfitting if the ratio of under-sampling was large.
- Over-sampling: This is the opposite of under-sampling where we duplicate the number of minority class observations up to a certain value to avoid overfitting.
- SMOTE (Synthetic Minority Over Sampling Technique): Here the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement. These introduced

synthetic examples are based along the line segments joining a defined number of k minority class nearest neighbours.