DOMAIN BACKGROUND:

The dataset I chosen is from UCI Machine Learning Repository. The data is about Direct Marketing campaigns of a Portuguese banking Institution.

Through phone calls The Marketing campaigns were held. The bank makes calls more than once to the client to ensure to clear the problem about whether he subscribes the bank term deposit or not.

https://danielabban.github.io/2017/04/predicting-the-success-of-bank-telemarketing/

https://rstudio-pubs-static.s3.amazonaws.com/115805_0a4a63f687984ebaae2655243389bbae.html#(1)

PROBLEM STATEMENT:

Generally Financial Institutions provides investments, loans and offers term deposits to its clients. These companies requires more customers to make more deposits in their banks to overcome the financial crisis. Companies using tele-marketing to improve their business inorder to meet the end client for a specific work for specific term deposit to develop their business.

DATASETS AND INPUTS:

The dataset used in the project had taken from the UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

The inputs are:
bank client data:
1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no', 'yes', 'unknown')
6 - housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
7 - loan: has personal loan? (categorical: 'no', 'yes', 'unknown')
related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular', 'telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - 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:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - 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)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical:

'failure','nonexistent','success')
social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric).

There are 45211 rows and 17 columns.

And There are 5289 – yes and "39922- No's are available in this dataset.

The main goal is to predict whether the client subscribes the term deposit or not.

SOLUTION STATEMENT:

Here I used Binary classification using supervised Learning algorithm to predict whether he subscribes the term deposit or not.

BENCHMARK MODEL:

Here I wish to use sklearn naive bayes GuassianNB algorithm as a Benchmark model.

This method can work on very large datasets effectively.

This GuassianNB algorithm has a simple approach, fast and works accurate for prediction.

EVALUATION METRICS:

Her the data is imbalanced so, we use to apply precision to the data using the below formula because, we have to know how much we have predicted from the actual data. And we have to know how much interest the people are wishing to make a term deposits.

Precision= True Positive/(True Positive+ False Positive)

PROJECT DESIGN:

Here we will perform some data cleaning and data pre-processing operations for the dataset.

And also check for the outliers, if any outliers exist, then we have to remove outliers. And we have to count the yes no values, and data type of the features

We use to split the data and train some data, and test the remaining data and wish to save the results.

After applying the GuassianNB as a Benchmark model, we have to consider the other algorithms to get the good score. And then after I perform Hyper-parameter optimisation to get the best results . And predict the results at ending using new data.