Abstract :

Companies build many marketing campaigns to increase subscribed persons. Many companies work on analyzing the market before starting campaigns, and the prediction of subscriptions size and classify of the audience are considered as the basic points in their work. Machine Learning algorithms have shown remarkable performance on several classify and prediction tasks. In this project, we cleaned and processed the dataset then studied one of supervised machine learning algorithm : Logistic Regression, Random Forest and Naive Bayes classifiers to build the models to predict and anticipate the client's cases in terms of whether he will subscribe to a deposit or not based on his /her information, and finally, will test the model accuracy.

Design :

The data is related to 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 assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The data set available on the UCI site.

Data set :

The data set includes 18 columns, 17 columns containing independent features and one dependent variable which will be predicted by a machine learning model. And 45211 rows. some features about age, job, marital, education, default, balance, housing, loan, contact, duration, campaign and "y" which is the output variable to know if the client subscribed a term deposit? (binary: 'yes','no').

Algorithms:

Feature Engineering

i. Applied a scaling method in data to make the learning process more easy.
ii. Applied Principal component analysis (PCA) is a technique for reducing the dimensionality of datasets.
iii. Split Data set : Training and Testing
iv. Checking Missing Values

Models:

Logistic Regression Classifier, Random Forest and Naive Bayes classifiers used to classify the clients.

Model Evaluation and Selection:

The entire training dataset of records was split into 75/25 train vs. testing and all scores reported below were calculated.

The official metric was classification rate (accuracy); however, class weights were included to improve performance against F1 score

	precision	recall	f1-score	support
0	0.90	0.99	0.94	1010
1	0.48	0.08	0.14	121
accuracy			0.89	1131
macro avg	0.69	0.54	0.54	1131
weighted avg	0.85	0.89	0.86	1131

Accuracy : 89.2%

Naive Bayes Model:

	precision	recall	f1-score	support
0	0.91	0.97	0.94	1010
1	0.44	0.22	0.30	121
accuracy			0.89	1131
macro avg	0.67	0.59	0.62	1131
weighted avg	0.86	0.89	0.87	1131

Accuracy score is: 0.8859416445623343

Random Forest Classifier:

	precision	recall	f1-score	support
0	0.91	0.96	0.93	1010
1	0.37	0.20	0.26	121
accuracy			0.88	1131
macro avg	0.64	0.58	0.60	1131
weighted avg	0.85	0.88	0.86	1131

Accuracy score is: 0.8779840848806366

Tools:

i. Scikit-learn for modeling
ii. Matplotlib and Seaborn for plotting and visualizations
iii. Jupyter Notebook to write code
iv. Numpy and Pandas for data manipulation