Problem statement (Term Deposit Sale)

Goal

Using the data collected from existing customers, build a model that will help the marketing team identify potential customers who are relatively more likely to subscribe term deposit and thus increase their hit ratio.

Resources Available

The historical data for this project is available in file

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Deliverable – 1 (Exploratory data quality report reflecting the following) – (20)

1. Univariate analysis (**12 marks**)
   a. Univariate analysis – data types and description of the independent attributes which should include (name, meaning, range of values observed, central values (mean and median), standard deviation and quartiles, analysis of the body of distributions / tails, missing values, outliers.
   b. Strategies to address the different data challenges such as data pollution, outlier's treatment and missing values treatment.
   c. Please provide comments in jupyter notebook regarding the steps you take and insights drawn from the plots.

2. Multivariate analysis (**8 marks**)
   a. Bi-variate analysis between the predictor variables and target column. Comment on your findings in terms of their relationship and degree of relation if any. Presence of leverage points. Visualize the analysis using boxplots and pair plots, histograms or density curves. Select the most appropriate attributes.
   b. Please provide comments in jupyter notebook regarding the steps you take and insights drawn from the plots

Deliverable – 2 (Prepare the data for analytics) – (10)

1. Load the data into a data-frame. The data-frame should have data and column description.
2. Ensure the attribute types are correct. If not, take appropriate actions.
3. Transform the data i.e. scale / normalize if required
4. Create the training set and test set in ration of 70:30

Deliverable – 3 (create the ensemble model) – (30)

1. Write python code using scikitlearn, pandas, numpy and others in Jupyter notebook to train and test the ensemble model.
2. First create a model using standard classification algorithm. Note the model performance.
3. Use appropriate algorithms and explain why that algorithm in the comment lines.
4. Evaluate the model. Use confusion matrix to evaluate class level metrics i.e..Precision and recall. Also reflect the overall score of the model.

5. Advantages and disadvantages of the algorithm.
6. Build the ensemble models (Bagging and Boosting) and compare the results with the base model. Note: Random forest can be used only with Decision trees.
7. Give conclusion regarding the best algorithm and your reason behind it.

Attribute information

Input variables:

Bank client data:

1. age: Continuous feature
2. job: Type of job (management, technician, entrepreneur, blue-collar, etc.)
3. marital: marital status (married, single, divorced)
4. education: education level (primary, secondary, tertiary)
5. default: has credit in default?
6. housing: has housing loan?
7. loan: has personal loan?
8. balance in account

Related to previous contact:

9. contact: contact communication type
10. month: last contact month of year
11. day: last contact day of the week
12. duration: last contact duration, in seconds*

Other attributes:

13. campaign: number of contacts performed during this campaign and for this client
14. pdays: number of days that passed by after the client was last contacted from a previous campaign (-1 tells us the person has not been contacted or contact period is beyond 900 days)
15. previous: number of contacts performed before this campaign and for this client
16. poutcome: outcome of the previous marketing campaign

Output variable (desired target):

17. Target: Tell us has the client subscribed a term deposit. (Yes, No)