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 AvailableThe 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) 1. Univariate analysis   
 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 Jupiter notebook regarding the steps you take and  
 insights drawn from the plots.

2. Multivariate analysis   
 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.  
 Visualize the analysis using boxplots and pair plots, histograms or density curves.  
 Select the most appropriate attributes.  
 b. Please provide comments in Jupiter notebook regarding the steps you take and  
 insights drawn from the plots

Deliverable – 2 (Prepare the data for analytics) 1. Ensure the attribute types are correct. If not, take appropriate actions.  
2. Get the data model ready.  
3. Transform the data i.e. scale / normalize if required  
4. Create the training set and test set in ratio of 70:30

Deliverable – 3 (create the ensemble model)1. First create models using Logistic Regression and Decision Tree algorithm. Note the  
model performance by using different matrices. Use confusion matrix to evaluate  
class level metrics i.e. Precision/Recall. Also reflect the accuracy and F1 score of the  
model.   
2. Build the ensemble models (Bagging and Boosting) and note the model performance  
by using different matrices. Use same metrics as in above model. (at least 3  
algorithms)   
Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited  
3. Make a DataFrame to compare models and their metrics. Give conclusion regarding  
the best algorithm and your reason behind it.

Attribute informationInput 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 month  
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 times the client has been contacted before for the last  
campaign to subscribe term deposit  
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)