Problem statement (Term Deposit Sale)

Goal

Using the 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)

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.

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. Presence of leverage points. Visualize the analysis using boxplots and pair plots, histograms or density curves. Select the most appropriate attributes.

3. Strategies to address the different data challenges such as data pollution, outliers and missing values.

Deliverable – 2 (Prepare the data for analytics)

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)

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 and compare the results with the base model. Note: Random forest can be used only with Decision trees.

Deliverable – 4 (Tuning the model)

1. Discuss some of the key hyper parameters available for the selected algorithm. What values did you initialize these parameters to?
2. Regularization techniques used for the model.
3. Range estimate at 95% confidence for the model performance in production.

Attribute information

Input variables:
# 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)

Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')