Bank Loan Term Deposit Sales

The data is from direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. The classification goal is to predict if the customer will subscribe to a term deposit (variable y). Often, more than one contact to the same customer was required, in order to access if the product (bank term deposit) would be 'yes' or 'no' for subscription.

The data is provided with a spreadsheet called bank-full.csv with 17 input and 1 target variables and the attributes are:

Input Variables

1. **age**: Age of the client
2. **job** : Type of job diversities the clients belong to
3. **marital**: Marital status of the clients
4. **education**: Clients educational background
5. **default**: Has credit in default?
6. **balance**: Average yearly balance, in euros
7. **housing**: Has housing loan?
8. **loan**: Has personal loan?
   related to the last contact person of the current campaign:
9. **contact**: Contact communication type
10. **day**: Last contact day of the month
11. **month**: Last contact month of year
12. **duration**: Last contact duration, in seconds
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
15. **previous**: Number of contacts performed before this campaign and for this client
16. **poutcome**: Outcome of the previous marketing campaign for this client

Target Variable

1. **Target**: Has the client subscribed to a term deposit?

Task:

Deliverable -1 (Exploratory Data Analysis):

1. 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. Bivariate analysis between the predictor variables and 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 Ensure the attribute types are correct. If not, take appropriate actions
2. Transform the data ie. scale/normalize if required
3. Create the training set and test set in ratio of 70:30

Deliverable - 3 (Model Creation)

1. Write python code using scikit learn, pandas, numpy and others in Jupyter notebook to train and test the ensemble model
2. First create a model using a standard classification algorithm. Note the model performance
3. Use appropriate algorithm/s 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

Deliverable - 5 Final Conclusion

1. Final Conclusion about the best model and the key features.

Hint: Ensemble Technique to design a classification model which can predict the outcome of a potential client but compare the results with non - ensemble modedel.