Ensemble Techniques

Domain: Banking

Context:

Leveraging customer information is paramount for most businesses. In the case of a bank, attributes of customers like the ones mentioned below can be crucial in strategizing a marketing campaign when launching a new product.

Data Description:

The data is related with 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

Attribute Information:

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. balance: average yearly balance, in euros (numeric)
7. housing: has a housing loan? (categorical: 'no','yes','unknown')
8. loan: has personal loan? (categorical: 'no','yes','unknown')
9. contact: contact communication type (categorical: 'cellular','telephone')
10. day: last contact day of the month (numeric 1 -31)
11. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
12. 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.
13. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14. 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)
15. previous: number of contacts performed before this campaign and for this client (numeric)
16. poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
17. target: has the client subscribed a term deposit? (binary: "yes","no")

Objective:

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Steps and tasks:

1. Import the necessary libraries and read the data as a data frame (2 marks)
2. Perform basic EDA which should include the following and print out your insights at every step. (20 marks)
    a. Shape and data type of the data (1 marks)
    b. Check info of the dataset (1 marks)
    c. Report statistical summary of the dataset. (3 marks)
    d. Check the presence of missing values and impute if there is any (3 marks)
    e. Checking the presence of outliers and impute if there is any (3 marks)
    f. Report the distribution of independent variables. (2 marks)
    g. Check frequency distribution of target feature and comment on your findings. (2 marks)
    h. Perform bivariate analysis using pairplot and mention your findings. (3 marks)
    i. Check correlation among independent features and mention if there is any collinearity. (2 marks)
3. Prepare the data to train a model – check if data types are appropriate, get rid of the missing values etc. (3 marks)
4. Train a decision tree model, note and comment on their performances across different classification metrics. (5 marks)
5. Build the ensemble models (random forest, bagging classifier, Adaboosting, and gradient boosting, and stacking classifier) and compare the results. (15 marks)
6. Compare performances of all the models and comment on your findings. (5 marks)

Learning Outcomes:

- Exploratory Data Analysis
- Preparing the data to train a model
- Training and making predictions using an Ensemble Model
- Tuning an Ensemble model

References:

- Data analytics use cases in Banking
- Machine Learning for Financial Marketing