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 access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

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.

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 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")

Learning Outcomes:

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

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 (2.5 marks)
2. Read the data as a data frame (2.5 marks)
3. Perform basic EDA which should include the following and print out your insights at every step. (15 marks)
    a. Shape of the data (2 marks)
    b. Data type of each attribute (2 marks)
    c. Checking the presence of missing values (4 marks)
    d. 5 Point summary of numerical attributes (3 marks)
    e. Checking the presence of outliers (4 marks)
4. Prepare the data to train a model – check if data types are appropriate, get rid of the missing values etc. (15 marks)
5. Train a few standard classification algorithms, note and comment on their performances across different classification metrics. (15 marks)
6. Build the ensemble models and compare the results with the base models. Note: Random forest can be used only with Decision trees. (15 marks)
7. Compare performances of all the models (5 marks)

References:

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