**Case Study: Building a model to predict who takes a term loan in a Campaign**

The goal of this task is to fit a statistical model to historical marketing data and then use the model to estimate which customers will take up a term loan offering from the bank. The task is broken up into seven steps which are explained in detail below.

This is a relatively straightforward task and should not take you too long to complete. We would like you to complete the task using a choice of SAS, R or Python notebook. We would like you to demonstrate that you have some familiarity with some core concepts in data science and follow a structured and logical thought process. Within these parameters there are still many ways to solve the task. If necessary you can make (and note) assumptions about the data and the interpretation of the task.

This analysis will follow the following steps:

1. Import the data
2. Draw some high-level visuals on the dataset
3. Prepare the data by creating dummy variables
4. Divide the data into a training set and a test set.
5. Run the model of choice on the training data set.
6. Predict the take up on the hold out sample.
7. Create a cumulative gains chart and Confusion Matrix
8. Summary of work

Overview of the data:

| **Feature Number** | **Column Name** | **Description** | **Type** |
| --- | --- | --- | --- |
| 1 | age | Age of the client | Numeric |
| 2 | job | Client's occupation | Categorial:   * admin * blue-collar * entrepreneur * housemaid * management * retired * self-employed * services * student * technician * unemployed * unknown |
| 3 | marital | Marital status | Categorial:   * divorced * married * single   **Note:** divorced means divorced or widowed |
| 4 | education | Client's education level | Categorial:   * primary * secondary * tertiary * unknown |
| 5 | default | Indicates whether the client has credit in default | Numerical:   * no = 0 * yes = 1 |
| 6 | Balance | average yearly balance, in euros | Numeric |
| 7 | housing | Indicates whether the client has a housing loan | Categorial:   * no = 0 * yes = 1 |
| 8 | loan | Indicates whether the client as a personal loan | Categorial:   * no = 0 * yes = 1 |
| 9 | contact | Type of contact communication | Categorial:   * cellular * telephone * unknown |
| 10 | day | Last contact day of the month | Numeric |
| 11 | month | Month that last contact was made | Categorial:   * jan * feb * ⋮ * dec |
| 12 | duration | Duration of last contact in seconds | Numeric |
| 13 | campaign | Number of contacts performed during this campaign for this client (including last contact) | Numeric |
| 14 | pdays | Number of days since the client was last contacted in a previous campaign | Numeric |
| 15 | previous | Number of contacts performed before this campaign for this client | Numeric |
| 16 | poutcome | Outcome of the previous marketing campaign | Categorial:   * failure * other * success * unknown |
| 17 | Take\_up | Indicates whether the client has subscribed for a term deposit | Numerical:   * no = 0 * yes = 1 |

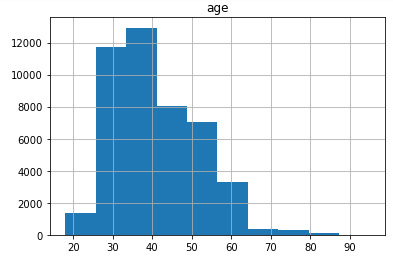
**Task 1: Import the data**

Load the supplied comma-separated data file (bank\_full.csv).

Check to ensure that your data file is completed. A correctly imported file has a Total number of records: 45,211 records with 17 columns

**Task 2 : Explore the relationship between some of the variables.**

An example plot could be a histogram of a particular field (such as the histogram of age plotted below) or a scatter plot to indicating the relationship between variables. Don’t spend too long on these plots, they needn’t be beautiful nor do they need to be exhaustive. The idea is simply to demonstrate that you can visualize relationships in the dataset.



**Task 3: Convert the categorical values to numerical values**

Convert the categorical values to dummy variables and replot some of the relationships **if needed**

**Task 4: Test and training data sets**

In order to evaluate this model (the next step) you will need to create a backtesting or hold-out set which you will not use to fit the model with. Use a holdout sample of 20%.

**Task 5: Run the model of choice.**

Using the newly created dataset on the above transformations. Run a model of your choice that will learn from the current dataset.

**Task 6: Use this model to now predict the take-up in the hold out sample**

**Task 7: Create a cumulative gains chart**

Calculate the area under the curve (AUC) for the test data and generate a confusion matrix to understand the impact of the predictions.

**Task 8: Summary**

Provide a short extract : 3-5 lines; summarising your outcomes.

* What have you learnt from the process?
* How effective is your model?
* What are things we can improve upon?