Mini Project 1 – Banking Campaign Output Prediction

Data Description:

The data is related to direct marketing campaigns of a banking institution. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required, to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

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 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 to a term deposit? (binary: 'yes','no')

Task Description:

The task will be to use **two** different Machine Learning Algorithms to predict whether a client would respond positively or negatively to the campaign.
You need to:
1- Perform data preprocessing
a. What features are you using for your models and why?
b. How do you process the data to be used in your models?
2- Train the models with at least 80% accuracy.
3- Visualize the performance of your models.
4- Compare the results from the different models, why is that difference?
You need to hand in your Python code (preferably Jupyter Notebook) alongside a written report.
Write a scientific report that includes
• Introduction (what is the problem you are solving?)
• Data processing (what are the choices you made in data processing and how you performed it?)
• Modelling(What are the algorithms that you chose, and why? How did you perform them? How did you improve the performance of your models? Were you able to reach the required accuracy in the first training round?)
• Conclusion (what were the “scientific” bottlenecks? How did you overcome them? Which algorithm was better?)