Multivariate analysis - D1 project delivery

Andreja Andrejic
David Candela
Nayara Costa
Julie Oppedal
Alexandros Tremopoulos

September 18, 2023

1. Members: Andreja Andrejic, David Candela, Nayara Costa, Julie Oppedal and Alexandros Tremopoulos.

2. Dataset: Bank Marketing

3. Download instructions: https://archive.ics.uci.edu/dataset/222/bank+marketing Going through this link there is a 'Download' button. When pressed, a zip folder called 'bank+marketing' is downloaded. When the contents are extracted, there is a zip folder called 'bank'. Inside this folder, there is the file 'bank.csv' that we are using. 'bank.csv' contains 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

4. Description: The Bank Marketing dataset compiles call information from a Portuguese banking institution with the objective of predicting whether a client will subscribe to a term deposit. This classification task relies on various categorical features, such as the client's education level, occupation, and marital status. Additionally, numerical features like the client's age and average account balance are considered in the analysis.

5. Contents:

- 4521 records in the sample (bank.csv)
- 17 total variables of which:
  - 7 are numerical ('age', 'balance', 'day', 'duration', 'campaign', 'pdays', 'previous')
  - 10 are categorical ('job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'poutcome', 'y')
  - 4 of the categorical are binary ('default', 'housing', 'loan', 'y')
  - 6 of the categorical are qualitative ('job', 'marital', 'education', 'month', 'contact', 'poutcome')
- 6.83% of cells with missing data distributed as follows:

- age 0%
- job 0.85% (38 entries)
- marital 0
- education 4.13% (187 entries)
- default 0%
- balance 0%
- housing 0%
- loan 0%
- contact 29.7% (1324 entries)
- day 0%
- month 0%
- duration 0%
- campaign 0%
- pdays 0%
- previous 0%
- poutcome 81.9% (3705 entries)
- y 0%