Data Intake Report

Name: Data Science:: Bank Marketing (Campaign)
Report date: 19 January 2024
Internship Batch: LISUM28
Version:1.0
Data intake by: Trofymova Anastasiia
Data intake reviewer: Data Glacier
Data storage location:
https://archive.ics.uci.edu/dataset/222/bank+marketing

Tabular data details: bank_full

Total number of observations	45211
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	5.9 MB

Tabular data details: bank

Total number of observations	4521
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	600.6 KB

Tabular data details: bank_additional_full

Total number of observations	41188
Total number of files	1
Total number of features	21
Base format of the file	.csv
Size of the data	6.6 MB

Tabular data details: bank_additional

Total number of observations	4119
Total number of files	1
Total number of features	21
Base format of the file	.csv
Size of the data	675.9 KB

Proposed Approach:

There are four datasets:

1. bank-additional-full.csv - all examples (41188) and 21 inputs
2. bank-additional.csv - 10% of the examples (4119), randomly selected, and 21 inputs.

Missing Attribute Values: There are several missing values in some categorical attributes, all coded with the "unknown" label. These missing values can be treated as a possible class label or using deletion or imputation techniques.

3. bank-full.csv - all examples (45211) and 17 inputs.
4. bank.csv - 10% of the examples (4521) and 17 inputs, randomly selected.

Missing Attribute Values: None

> The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

Because there are no specific markers like names or customer IDs, we assume the data doesn't have many or any repeated entries. The data doesn't have any missing or blank spots.

However, since some people in the customer survey chose not to share details for privacy reasons, we're putting those cases in a separate group, like "Unknown" or "None." We'll decide during the data cleaning stage whether to replace these cases with something else (like an average or common value) or just remove them.