Data Intake Report

Name: Bank Marketing Data Set
Report date: 25/07/2021
Internship Batch: LISUM01
Version:1.0
Data intake by: Samuel Alejandro Cueva Lozano
Data intake reviewer: Samuel Alejandro Cueva Lozano
Data storage location: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

Tabular data details:

Name of the file: Bank-additional.csv

Total number of observations	4119
Total number of files	1
Total number of features	21(target included)
Base format of the file	.csv
Size of the data	571 KB

Name of the file: Bank-additional-full.csv

Total number of observations	41188
Total number of files	1
Total number of features	21(target included)
Base format of the file	.csv
Size of the data	5.6 MB

Proposed Approach:

Duplicate values

To consider two o more records as duplicates these must have the same values in the most of the attributes (excluding unique identifier) or very similar values in the case that they are strings (e.g. ‘Samuel Cueva Lozano’ = ‘Samy Cueva Loz’ and the others attributes be the same).
A record in the data set will be considered duplicate only if there are records with the same value in each attribute.

Null values

The *Empty values, NaN, N/A values* and values like *“?”* will be treated as ***Null values***

Fields with ***too many null values*** would be removed or filled according to the following:

- Fields like *Age, Duration and Emp.var.rate* would be filled because they have missing values that depends on their hypothetical values or depend on other attributes.
- The rest of fields would be removed because they have missing values due to bad configuration, issues with data collection, or untraceable random reasons.

After cleaning at the field level, rows with null values would be removed.

Outlier detection

Outliers will be removed from numerical fields so that they don’t negatively affect the analysis.
A fixed threshold would be used for the *Age, Duration*, C*ampaign, Pdays* and *Previous* attribute to avoid inconsistent data, then the IQR Score will be used to filter out the outliers in all attributes.

Data Transformation

The fields *Education, Month and Day_of_the_week* could be transformed to a more readable format for better understanding and of course the others categorical attributes will be transformed to feed the model later.