Data Intake Report

Name: Bank Marketing (Campaign)

Report date: 06/19/2023

Internship Batch: LISUM21

Version: 1

Data intake by: Shiva Ramezani

Data intake reviewer:<intern who reviewed the report>

Data storage location: <https://github.com/ShivaRamezani/BankMarketing>

**Tabular data details: bank.csv**

|  |  |
| --- | --- |
| **Total number of observations** | 4522 |
| **Total number of files** | N/A |
| **Total number of features** | 17 |
| **Base format of the file** | .csv |
| **Size of the data** | 453 KB |

**Tabular data details: bank-full.csv**

|  |  |
| --- | --- |
| **Total number of observations** | 45212 |
| **Total number of files** | N/A |
| **Total number of features** | 17 |
| **Base format of the file** | .csv |
| **Size of the data** | 4.5 MB |

**Tabular data details: bank-names.txt**

|  |  |
| --- | --- |
| **Total number of observations** | 17 |
| **Total number of files** | N/A |
| **Total number of features** | N/A |
| **Base format of the file** | .csv |
| **Size of the data** | 4 KB |

**Tabular data details: bank-additional.csv**

|  |  |
| --- | --- |
| **Total number of observations** | 4120 |
| **Total number of files** | N/A |
| **Total number of features** | 21 |
| **Base format of the file** | .csv |
| **Size of the data** | 571 KB |

**Tabular data details: bank-additional-full.csv**

|  |  |
| --- | --- |
| **Total number of observations** | 41189 |
| **Total number of files** | N/A |
| **Total number of features** | 21 |
| **Base format of the file** | .csv |
| **Size of the data** | 5.6 MB |

**Tabular data details: bank-additional-names.txt**

|  |  |
| --- | --- |
| **Total number of observations** | N/A |
| **Total number of files** | N/A |
| **Total number of features** | N/A |
| **Base format of the file** | .csv |
| **Size of the data** | 6 KB |

**Proposed Approach:**

* **Mention approach of dedup validation (identification)**

Identify Key Fields: Determine the key fields that uniquely identify a customer. In this dataset, it could be a combination of attributes like name, contact number, or any other relevant information.

Data Sorting: Sort the dataset based on the key fields identified in the previous step. This step ensures that similar records are placed together.

Record Comparison: Compare consecutive records based on the key fields to identify potential duplicates. If the key fields match, it indicates a potential duplicate record.

Duplicate Identification: Flag or mark the potential duplicate records for further analysis or processing.

Data Quality Analysis: Perform an in-depth analysis of the potential duplicates to determine their validity. Some common techniques for deduplication include:

a. Data Sampling: Randomly select a subset of potential duplicates and manually review them to confirm if they are indeed duplicates.

b. Automated Methods: Utilize automated techniques like fuzzy matching, string similarity algorithms, or record linkage algorithms to compare and match potential duplicates.

c. Domain Knowledge: Leverage domain knowledge and business rules to identify duplicate records based on specific criteria.

Duplicate Handling: Decide on the appropriate action to handle the duplicates. Options include removing duplicates, merging duplicate records, or keeping only the most recent or most complete record.

* **Mention your assumptions (if you assume any other thing for data quality analysis**

Missing Values: Assume that missing values exist in the dataset and devise strategies to handle them appropriately. This can involve techniques like imputation, deletion, or treating missing values as a separate category.

Outliers: Assume the presence of outliers in numerical variables and determine their impact on the analysis. Decide whether to remove outliers or transform variables to mitigate their influence.

Data Consistency: Assume that inconsistencies might exist within the dataset, such as conflicting values or data format discrepancies. Address such inconsistencies through data cleansing and standardization techniques.

Data Integrity: Assume that the dataset is reliable and represents accurate information. If there are concerns about data integrity, explore methods to verify the data's accuracy and rectify any inconsistencies.

Data Balance: Consider the possibility of class imbalance in the target variable ('y') and evaluate techniques to handle it during model building, such as oversampling, undersampling, or generating synthetic samples.