Data Intake Report

Name: Bank Marketing (Campaign)
Report date: 19 May 2024
Internship Batch: LISUM32: 01 Apr 24 - 01 July 24
Data intake by: Elif Nur Kemiksiz and Bisma Azeem
Data intake reviewer: Bisma Azeem
Data storage location: Github

Tabular data details:

Total number of observations	41189
Total number of files	4
Total number of features	21
Base format of the file	.csv
Size of the data	5.5 MB

Data Cleaning and Preparation

1. Duplicate Handling:
    - Sort the data based on any unique identifier present in the dataset.
    - Identify duplicates based on these unique identifiers and remove them.
2. Handling Missing Data:
    - Flag and analyze rows with missing data.
    - Depending on the nature and quantity of the missing data, apply imputation methods (such as mean, median, mode for numerical data, or most frequent category for categorical data) or remove the rows if imputation is not feasible.
3. Data Transformation:
    - Convert categorical variables into a format suitable for analysis using encoding techniques such as one-hot encoding or label encoding.
    - Normalize numerical data if required to ensure all features are on a comparable scale, typically using standardization (z-score normalization) or min-max scaling.

Assumptions:

1. Data Relevance:

- The data accurately reflects the customer base ABC Bank is targeting for the term deposit product.
- The data captures past interactions that are relevant to a customer's propensity to purchase term deposits (e.g., loan history, previous marketing campaign outcomes).

2. Data Accuracy:

- The data is free from significant errors in data entry or measurement.
- The recorded values for features like age, number of contacts, or loan status are accurate.

3. Data Completeness:

- The missing value rate for features is relatively low and doesn't significantly impact the ability to draw conclusions.
- There are no systematic biases in which data points are missing (e.g., missing data is not concentrated in a specific customer segment).

4. Data Consistency:

- The data uses consistent coding schemes for categorical features (e.g., "unemployed" is always spelled the same way).
- Dates are formatted consistently throughout the dataset.

5. Stationarity:

- The underlying relationships between features and the target variable (term deposit purchase) haven't changed significantly over time.
- This is important because the model is built on historical data and should generalize to future customer behavior.