Bank Dataset Documentation

Project Documentation: Analysis of Customer Subscription Data

1. Introduction

This document provides a comprehensive overview of the data sources, methodologies, and actionable insights derived from the analysis of customer subscription data. The objective of this analysis is to understand the factors influencing subscription outcomes and to provide actionable recommendations for improving subscription rates.

2. Data Sources

Bank Dataset

- **Source:** The data for this project comes from the Data Analytics Elites association. It's based on a bank's marketing campaign, and the dataset includes several key details,
- **Description:** The dataset includes information on customer demographics, contact details, and subscription outcomes.

- **Key Variables:**
- • **age**: Client's age (numeric).
- • **job**: Type of job (categorical, e.g., "admin", "unknown", "unemployed", "management").
- • **marital**: Marital status (categorical, e.g., "married", "divorced", "single").
- • **education**: Education level (categorical, e.g., "unknown", "secondary", "primary", "tertiary").
- • **default**: Credit in default (binary, "yes" or "no").
- • **balance**: Average yearly balance in euros (numeric).
- • **housing**: Housing loan status (binary, "yes" or "no").
- • **loan**: Personal loan status (binary, "yes" or "no").
- • **contact**: Type of contact communication (categorical, e.g., "unknown", "telephone", "cellular").
- • **day**: Last contact day of the month (numeric).
- • **month**: Last contact month of the year (categorical, e.g., "jan", "feb", "mar").
- • **duration**: Duration of the last contact in seconds (numeric).
- • **campaign**: Number of contacts performed during this campaign (numeric).
- • **pdays**: Days since the client was last contacted from a previous campaign (numeric, -1 if never contacted).
- • **previous**: Number of contacts before this campaign (numeric).

- • **poutcome**: Outcome of the previous marketing campaign (categorical, e.g., "unknown", "failure", "success").
- • **y**: Subscription to a term deposit (binary, "yes" or "no").

3. Methodologies

Data Preprocessing

- **Data Collection**: The initial dataset is obtained in CSV format.

Preprocessing Steps in Excel:

- **Column Management**:
  - **Splitting Columns**: Use Excel to split combined columns into individual columns as needed.
  - **Dropping Columns**: Remove unnecessary columns that do not contribute to the analysis.
- **Data Assignment**:
  - **Assigning Values**: Utilize Excel to manually assign and adjust values in columns to ensure data consistency and correctness.
- **Data Formatting**:
  - **Standardization**: Modify string formats and adjust data types directly in Excel to standardize the dataset.
- **Cleaning**:
  - **Handling Missing Values**: Address null values in Excel by imputation or deletion.
  - **Removing Duplicates**: Identify and remove duplicate records in Excel to maintain data quality.

Transition to Python:

- **Importing Data**: After preprocessing in Excel, import the cleaned dataset into pandas for further processing.

Data Analysis

- **Analysis Objective**: To perform in-depth analysis of the cleaned dataset using Python libraries.
- **Data Processing in Pandas**:
  - **Data Exploration**: Explore the dataset using pandas to understand its structure and content.
  - **Further Cleaning**: Perform any additional cleaning or adjustments as needed in pandas.

Visualization and Summary:

  - **Chart Visualizations**: Create various charts using Matplotlib to visualize trends and patterns.

- **Summary Tables**: Generate summary statistics and tables to provide an overview of key metrics.

Tools and Libraries:

- Pandas.
- Matplotlib.

4. Insight

- The dataset consisted of 4,521 records, with 521 identified as active subscribers, indicating a modest subscription rate of approximately 11.5%.
- The average campaign duration was notably brief at 3.0 seconds, which could suggest potential data anomalies or interactions that were unusually short.
- Age analysis revealed that the 35-40 age group was the most represented demographic, suggesting a higher targeting or responsiveness within this range during the campaign.
- During the month of May, the contact type classified as 'unknown' had the highest engagement, with over 800 clients contacted, raising questions about the accuracy of contact classification or the presence of a significant segment with undefined contact methods.
- A comparison of subscription rates across job categories showed that individuals in blue-collar occupations were more likely to subscribe, reflecting either a focused demographic strategy or higher receptivity within this group.
- The age distribution highlighted that the 30-39 age group accounted for the largest proportion of subscribers at 36%, followed by the 40-49 age group at 24%, pinpointing key age demographics for future campaigns.
- Among job categories, management was the top-performing with 969 subscriptions, while the 'unknown' job category had the lowest at 38, providing insights into job-based engagement patterns

5. Summary

The analysis of the bank marketing campaign dataset uncovered significant insights into demographics, engagement trends, and the overall effectiveness of the campaign. The preprocessing steps, including the use of Excel for data cleaning and column management, were instrumental in transforming the raw data into a structured format suitable for in-depth analysis. Utilizing pandas for data manipulation and Matplotlib, for data visualization, the analysis identified critical trends, such as the moderate subscription rate, notably brief campaign interactions, and strong engagement among specific age groups and job categories.

These findings offer valuable guidance for enhancing future marketing strategies and addressing potential data quality concerns.