Data Science Project

Menna Muhammad Ibrahim	Ahmed Samir Ali
Abdelrahman Mohamed Ahmed	Omnia Fathy Mwafy
Ebrahim Ahmed Mohamed	Eman Abdelrahman Attya
Hagar Reda Mohamed	Zahret Elislam Nashaat
Mohamed Ahmed Wahdan	Omar Mohamed Raafat

Abstract:

Direct marketing campaigns are a crucial aspect of the banking industry, serving as a primary avenue for engaging potential clients and promoting various financial products. Understanding the factors that influence the success of these campaigns is essential for optimizing resource allocation and improving overall effectiveness. In this project, we explore a dataset from a Portuguese banking institution that encompasses information on past marketing campaigns, including client demographics, contact details, and economic indicators. The primary objective is to develop a predictive model capable of determining whether a client will subscribe to a term deposit based on the provided attributes.

The data is related to a Portuguese banking institution's direct marketing campaigns (phone calls). The classification goal is to predict if the client will subscribe to a term deposit (variable y).

Data Set Information:

The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required, to assess if the product (bank term deposit)

would be ('yes') or not ('no') subscribed.

Attribute Information:

Bank client data:

- `Age` (numeric)
- `Job` : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- `Marital` : marital status (categorical: 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed)
- `Education` (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
- `Default`: has credit in default? (categorical: 'no', 'yes', 'unknown')
- `Housing`: has housing loan? (categorical: 'no', 'yes', 'unknown')
- `Loan`: has personal loan? (categorical: 'no', 'yes', 'unknown')
- Related with the last contact of the current campaign:
    - `Contact`: contact communication type (categorical: 'cellular','telephone')
    - `Month`: last contact month of year (categorical: 'jan', 'feb', 'mar',..., 'nov', 'dec')
    - `Day_of_week`: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
    - `Duration`: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known.Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

- `Campaign`: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- `Pdays`: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- `Previous`: number of contacts performed before this campaign and for this client (numeric)
- `Poutcome`: outcome of the previous marketing campaign (categorical:'failure','nonexistent','success')
- Social and economic context attributes:
    - `Emp.var.rate`: employment variation rate - quarterly indicator (numeric)
    - `Cons.price.idx`: consumer price index - monthly indicator (numeric)
    - `Cons.conf.idx:` consumer confidence index - monthly indicator (numeric)
    - `Euribor3m:` euribor 3 month rate - daily indicator (numeric)
    - `Nr.employed:` number of employees - quarterly indicator (numeric)

Output variable (desired target):

- `y`: has the client subscribed to a term deposit? (binary: 'yes', 'no')

Coming to work

- **Importing Libraries**

  This section is straightforward. We import all the necessary libraries needed for our data analysis and machine learning tasks. This typically includes libraries such as pandas, numpy, matplotlib, seaborn, and scikit-learn. We can write this as a code chunk.

1. **Data Exploration:**
   This section involves several steps:
   - Reading the data.
   - Displaying the head and tail of the dataset to get a glimpse of the data.
   - Checking data types.
   - Using the `describe()` function to get summary statistics.
   - Using the `info()` function to get information about the dataset, including data types and missing values.
   - Calculating correlations between numerical features.
   - Creating visualizations to explore the data further.

2. **Preprocessing Data**
   This section involves:
   - Handling missing values: replacing 'unknown' values with NaN.
   - Summarizing the number of NaN values in each column.
   - Removing rows or columns with NaN values.
3. **Dividing the Data**
   In this step, we split the dataset into input features (X) and the target variable (Y). The input features comprise various attributes such as age, job type, marital status, contact details, and socio-economic indicators, while the target variable (Y) indicates whether the client subscribed to a term deposit ('yes') or not ('no'). This division allows us to separate the independent variables from the dependent variable, facilitating the modeling process.
4. **Handling Duplicates**
   We checked for and removed any duplicate rows in the dataset to ensure the integrity of our data. Duplicate rows could potentially skew our analysis and model performance by inflating the importance of certain observations.

5. **Standardization and Scaling**
   To prepare the input features for modeling, we applied standardization and scaling techniques. Standardization involves transforming the features such that they have a mean of 0 and a standard deviation of 1, which helps algorithms converge faster and improves model performance. Scaling ensures that all features are on a similar scale, preventing features with larger magnitudes from dominating the model's learning process.

6. **Model Building**

   After preprocessing the data, we proceeded to build and evaluate several machine learning models to predict whether a client would subscribe to a term deposit. Here's a summary of the models we employed and the corresponding analysis:
   1. *Naive Bayes:*
      - We began by splitting the preprocessed data into training and testing sets to evaluate model performance.
      - Utilizing the Naive Bayes algorithm, we trained a classifier and evaluated its performance using confusion metrics and accuracy scores.
   2. *K-Nearest Neighbors (KNN):*
      - Next, we applied the KNN algorithm to the data.
      - We computed the error rate and examined confusion metrics to assess the model's effectiveness.
   3. *Decision Tree:*
      - Building on our analysis, we constructed a decision tree classifier and specified its parameters.
      - To optimize model performance, we performed a grid search to identify the best combination of hyperparameters.
   4. *Random Forest:*
      - We extended our analysis by implementing a random forest classifier.
      - Similar to the decision tree, we conducted a grid search to fine-tune the model's parameters for optimal results.
   5. *Support Vector Machine (SVM):*
      - We explored the SVM algorithm and specified its parameters, including the kernel type and regularization parameter.
      - To enhance model performance, we conducted a grid search to identify the optimal hyperparameters.

6. *Logistic Regression:*
    - In addition to tree-based and kernel-based algorithms, we employed logistic regression.
    - Visualizing the data, we gained insights into the relationships between input features and the target variable.
    - We evaluated model performance using confusion metrics and other relevant metrics.
7. K-Nearest Clustering:
    - Finally, we applied KNN clustering to identify patterns and groupings within the data.
    - Utilizing the Elbow method, we determined the optimal number of clusters for our dataset.
    - We visualized the clusters to gain insights into client segmentation and behavior.