Explain the Data using any Exploratory Data Analysis (EDA) you have used

The dataset includes client data and marketing interactions, aiming to predict if the client subscribed to a term deposit.

The provided train dataset contains 39,211 records and 16 columns, here is a summary of the key findings from the exploratory data analysis:

1. **Shape of the dataset:**
   * The dataset has 39,211 rows and 16 columns.
2. **Data Types:**
   * There are 6 numerical columns (age, balance, duration, campaign, pdays, previous) and 10 categorical columns (last contact date, job, marital, education, default, housing, loan, contact, poutcome, and target).
3. **Missing Values:**
   * job: 229 missing values.
   * education: 1,467 missing values.
   * contact: 10,336 missing values.
   * poutcome: 29,451 missing values (many instances labeled as "unknown"). This can be ignored while fitting a model.
4. **Numerical Data Summary:**
   * **Age:** The average age is 42 years, with a minimum of 18 and a maximum of 95 years.
   * **Balance:** The average yearly balance is 5,442 euros, with a wide range from -8,019 euros to 102,127 euros.
   * **Duration:** The mean duration of the last contact is 439 seconds, with a maximum of 4,918 seconds. The duration is critical as it significantly affects the target variable.
   * **Campaign:** The average number of contacts during the campaign is around 5, with a maximum of 63 contacts.
   * **Pdays:** This feature has a minimum of -1 (indicating clients not contacted before) and a mean of 72 days for those previously contacted.
   * **Previous:** The average number of previous contacts is about 12, with a maximum of 275 contacts.
5. **Categorical Data Distribution:**
   * **Job:** The most common job categories are blue-collar (7,776), management (7,462), and technician (6,078). The least common categories are student (1,393) and housemaid (1,656).
   * **Marital Status:** Most clients are married (22,691), followed by single (11,250), and divorced (5,270).
   * **Education:** The majority have secondary education (19,584), followed by tertiary (11,615) and primary (6,545).
   * **Default:** Most clients do not have credit in default (36,954), while a smaller portion does (2,257).
   * **Housing Loan:** About 21,657 clients have a housing loan, while 17,554 do not.
   * **Personal Loan:** 7,391 clients have a personal loan, while 31,820 do not.
   * **Contact Type:** The most common contact type is via cellular (25,030), followed by telephone (3,845).
   * **Poutcome (Outcome of Previous Campaign):** The outcomes are mostly failure (4,949) or success (2,251), while 25,451 records are labeled as "unknown."
   * **Target (Subscription to Term Deposit):** The dataset is imbalanced, with 33,384 clients not subscribing to the term deposit and 5,827 clients subscribing to it.
6. **Imbalance in Target Variable:**
   * The target variable is significantly imbalanced, with about 85% of clients not subscribing (no) and 15% subscribing (yes). This imbalance should be addressed during model training, possibly using techniques such as oversampling, undersampling, or other class-balancing methods.

**Key Insights:**

* **Campaign Effectiveness:** Most clients have been contacted only a few times (median campaign value of 2). Previous campaigns' outcomes indicate more failures than successes.
* **Loan and Default Information:** Most clients do not have personal loans or housing loans and are not in default.
* **Potential for Model Improvement:** Features like contact duration, previous, and balance could be highly predictive of whether a client subscribes to a term deposit, especially considering their significant variability across clients.
* **Correlation from Heat Map:**
  + **balance**, **previous**, **campaign**, and **duration** are more tightly connected, suggesting they could have a stronger influence on the outcome of interest (like subscribing to a term deposit).
  + **Age**, on the other hand, shows weaker correlations, indicating it may play a less direct role in influencing outcomes compared to the other factors.

References:

ChatGPT