Bank Marketing (Campaign)

Group	Name	Email	Country	College	Specialization
Data Duo	Ayushi Malaviya	ayushimalaviya1999@gmail.com	USA	Stevens Institute of Technology	Data Science
Data Duo	Nilesh Rathi	nileshrathi99@gmail.com	USA	Indiana University Bloomington	Data Science

Problem Description:

ABC Bank is planning to launch a term deposit product and wants to determine the likelihood of customers subscribing to this product based on their past interactions with the bank or other financial institutions. To achieve this, ABC Bank aims to develop a predictive model that can assist in understanding whether a particular customer will buy their term deposit or not.

The available data for this project consists of information related to direct marketing campaigns conducted by a Portuguese banking institution. These campaigns were primarily conducted via phone calls, and multiple contacts were often made with the same client to assess their interest in subscribing to the bank's term deposit product. The outcome of each campaign was recorded as either a successful subscription ('yes') or a non-subscription ('no').

The primary objective of this project is to build a classification model that can accurately predict whether a client will subscribe to the term deposit product ('yes') or not ('no'). By leveraging the historical data on customer interactions, ABC Bank aims to identify patterns, trends, and factors indicative of a higher likelihood of subscription. This predictive model will enable the bank to target its marketing efforts more effectively, optimise resource allocation, and enhance the success rate of its future marketing campaigns.

Business Understanding:

Understanding the likelihood of customer subscription to a term deposit product is crucial for ABC Bank's marketing and sales strategies. By developing a predictive model, ABC Bank aims to gain insights into customer behaviour and preferences, allowing them to tailor their marketing efforts accordingly. This understanding will help the bank optimize its resources, minimize costs, and increase the efficiency of its marketing campaigns.

By analyzing the past interactions between the bank and customers, the model can identify key factors that influence the decision-making process. These factors may include demographic information, previous banking history, campaign-specific variables, and other relevant attributes. By accurately predicting customer behaviour, ABC Bank can focus its marketing efforts on individuals who are more likely to subscribe to the term deposit product, thereby improving the conversion rate and maximizing the return on investment (ROI) for its marketing campaigns.

Moreover, the predictive model can provide valuable insights into customer segmentation, allowing ABC Bank to differentiate its marketing strategies based on various customer profiles. By understanding which customer segments are more inclined towards subscribing to the term deposit product, the bank can tailor their messaging, offers, and communication channels to cater to each segment's specific preferences, thus increasing the overall effectiveness of its marketing campaigns.

Overall, developing an accurate predictive model will empower ABC Bank to make data-driven decisions, enhance customer targeting, optimize marketing efforts, and increase the success rate of their term deposit product.

Project Lifecycle:

Phase1	**Business Evaluation and EDA for further processes**	**By 6/26/23**
Phase2	**Data Modelling and Data Consistency Evaluation**	**By 6/03/23**
Phase3	**Web APP Development and Testing**	**By 7/10/23**
Phase4	**Deployment on Cloud**	**By 7/17/23**
Phase5	**Building Data Ingestion Pipeline**	**By 7/24/23**
Phase6	**Building Dashboard (Optional)**	**By 7/30/23**

Data Intake Report:

Name: Data Science:: Bank Marketing Campaign
Report date: 19-June-2023
Internship Batch: LISUM21
Version: 1.0
Data intake by: Data Duo (team)
Data intake reviewer:
Data storage location: https://archive.ics.uci.edu/dataset/222/bank+marketing

Tabular data details: bank-additional-full

Total number of observations	**41188**
Total number of files	**1**
Total number of features	**21**
Base format of the file	**.csv**
Size of the data	**5.8MB**

Proposed Approach:

- There are no missing values in the data set.
- Assumptions made:
  o Complete: since data has no missing values.
  o Consistent: data is uniform in its format, units of measurement and data types.
  o Relevant: data is totally related to the research question being studied.
  o Unbiased: data is not influenced by any personal or external factors.

Github Repo Link: https://github.com/Ayushimalaviya/Bank-Marketing-Campaign-Analysis

—----------xx—----------

Data Description:

The dataset contains 41,000 rows and 21 features.

- The 'y' feature represents the target variable, indicating whether a customer subscribed to the term deposit product or not.
- The distribution of the target variable is as follows:
  - No: 36,548 customers did not subscribe to the term deposit product.
  - Yes: 4,640 customers subscribed to the term deposit product.

The data is highly imbalanced, with a ratio of approximately 1:8 for 'no' to 'yes' subscriptions.

Feature Types: The features are categorized as categorical and numerical. There are 10 categorical features and 10 numerical features apart from the response variable ‘y’.

Numerical Features:

Out of the 10 numerical features:

- Three variables are discrete in nature, while the remaining seven are continuous.
- One of the features, 'pdays', has values ranging from 1 to 27 but also includes a value of 999, which appears to be an imputed value for missing data. The 999 value accounts for 96% of the data for this feature. The other two discrete variables seem to have valid values.

- Among the seven continuous variables

- 'age' and 'duration' exhibit outliers.
- Summary statistics for 'duration' when the response variable is 'yes':
    - Count: 4,640
    - Mean: 553.191164
    - Standard Deviation: 401.171871
    - Minimum: 37
    - 25th Percentile: 253
    - Median: 449
    - 75th Percentile: 741.250000
    - Maximum: 4199

- Summary statistics for 'duration' when the response variable is 'no':
    - Count: 36,548
    - Mean: 220.844807
    - Standard Deviation: 207.096293
    - Minimum: 0
    - 25th Percentile: 95
    - Median: 163.500000
    - 75th Percentile: 279
    - Maximum: 4918

Comparing the maximum values for 'duration' between the 'yes' and 'no' cases, we observe a difference. The maximum value for 'yes' is 4199, while for 'no' it is 4918. To address this, we are considering two options:

- Removing all rows with 'no' cases where the duration value exceeds 4199.
- Applying RobustScaler to scale the 'duration' variable.

We also examined the 'age' variable for differences, but we did not find any substantial variations. However, applying a transformation to the age variable would result in a loss of interpretability. Therefore, we will use RobustScaler to scale this variable.

Categorical Data:

1. Job:
    - The 'job' feature represents the occupation of the customers.
    - The most common job categories are:
        - Admin: 10,422
        - Blue-collar: 9,254
        - Technician: 6,743
        - Services: 3,969
        - Management: 2,924
    - There are also categories with lower frequencies such as retired, entrepreneur, self-employed, housemaid, unemployed, student, and **unknown**.

2. Marital:
    - The 'marital' feature describes the marital status of the customers.
    - The majority of customers fall into the following categories:
        - Married: 24,928
        - Single: 11,568
        - Divorced: 4,612
    - There is also a small number of customers with **unknown** marital status.

3. Education:
    - The 'education' feature represents the educational background of the customers.
    - The most common education levels are:
        - University degree: 12,168
        - High school: 9,515
        - Basic 9 years: 6,045
        - Professional course: 5,243
        - Basic 4 years: 4,176
    - There are also categories for basic 6 years, **unknown** education, and even a few customers labelled as illiterate.

4. Default:
    - The 'default' feature indicates whether the customers have credit in default.

- The majority of customers have no default status (32,588), while a smaller number have an **unknown** default status (8,597).
- Only three customers have a default status of 'yes'.

5. Housing:
    - The 'housing' feature represents whether the customers have a housing loan.
    - The data shows that:
    - 21,576 customers have a housing loan.
    - 18,622 customers do not have a housing loan.
    - There are 990 customers with an **unknown** housing loan status.

6. Loan:
    - The 'loan' feature indicates whether the customers have a personal loan.
    - The majority of customers do not have a personal loan (33,950), while a smaller number have a loan (6,248).
    - There are also 990 customers with **unknown** loan statuses.

7. Contact:
    - The 'contact' feature describes the communication method used to contact the customers.
    - The data reveals two main contact methods:
        - Cellular: 26,144 contacts
        - Telephone: 15,044 contacts

8. Month:
    - The 'month' feature represents the month in which the last contact was made with the customers.
    - The distribution of contacts by month is as follows:
        - May: 13,769 contacts
        - July: 7,174 contacts
        - August: 6,178 contacts
        - June: 5,318 contacts
        - November: 4,101 contacts
        - April: 2,632 contacts
        - October: 718 contacts
        - September: 570 contacts
        - March: 546 contacts
        - December: 182 contacts

9. Day of Week:

- The 'day_of_week' feature indicates the day of the week when the last contact was made.
- The distribution of contacts by day of the week is as follows:
    - Thursday: 8,623 contacts
    - Monday: 8,514 contacts
    - Wednesday: 8,134 contacts
    - Tuesday: 8,090 contacts
    - Friday: 7,827 contacts

10. Poutcome:
    - The 'poutcome' feature describes the outcome of the previous marketing campaign for each customer.
    - The distribution of outcomes is as follows:
        - **Nonexistent**: 35,563 customers had no previous campaign outcome.
        - Failure: 4,252 customers had a previous campaign that resulted in failure.
        - Success: 1,373 customers had a previous campaign that was successful.

When it comes to the missing values in the dataset, we have noticed that certain features have been replaced with values like 'unknown' or 'nonexistent'. Now, the question arises: how should we handle these missing values? Should we keep them as they are, or is there a better approach?

After careful consideration, here are some options for dealing with these missing values:

Firstly, one option is to leave the missing values as they are. In some cases, these 'unknown' or 'nonexistent' categories might carry important information and could be meaningful for our analysis. By keeping them in the dataset, we allow the model to consider them as distinct categories and potentially learn patterns associated with missing data.

Alternatively, we can treat the missing values as a separate category within each feature. This means creating a new category, such as 'Missing' or 'Unknown', to specifically capture the information of missing values. This approach allows us to acknowledge and account for the missing data in our analysis.

Another approach is to impute the missing values with estimated values. For categorical features, we can consider using imputation techniques such as filling the missing values with the mode, which represents the most frequent category. However, it's important to be cautious as imputing missing values may introduce biases or alter the original distribution of the data.

In cases where a feature has a high percentage of missing values and it's not feasible to reasonably impute them, we may need to consider dropping the entire feature from our analysis. It's crucial to carefully evaluate the relevance and importance of the feature before making the decision to remove it.

Ultimately, the choice of how to handle missing values depends on the specific context of our analysis and the goals we want to achieve. We need to assess the impact of missing values on our results and consider the potential consequences of each approach. It's important to take a thoughtful and informed approach to ensure the integrity and reliability of our analysis.

Numerical Data:

In the given data there are around 6 numerical data features:

i) age (discrete values)
ii) previous (discrete values)
iii) emp.var.rate (continuous values)
iv) cons.price.idx (continuous values)
v) cons.conf.idx (continuous values)
vi) nr.employed (discrete values) (could be scaled by applying Scaling approaches)

Skewness analysis is typically performed on continuous variables since it measures the distribution's departure from symmetry. In the dataset, there are three continuous features, and ideally, their skewness values should be close to zero. As you can see below:

Emp.var.rate(Employee variate rate) : -0.724096
Cons.price.idx (consumer price index) : -0.230888
cons.conf.idx (consumer confidence index) : 0.303180

The Employee variation feature in the dataset exhibits slight skewness, indicating a departure from symmetry. Common statistical techniques such as exponential or logarithmic transformations can be applied to address this skewness. In this case, due to the negative values and range in the tens place, applying a simple technique like exponential transformation significantly improves the skewness value as shown below:

Emp.var.rate(Employee variate rate) : -0.2439304620546845

Imbalance Dataset:

The dataset exhibits an imbalanced distribution between true positive and false negative values, with a ratio of 8:1. This imbalance can introduce bias when predicting true positive

values on unseen data. To address this issue, it is good practice to balance the data labels in the dataset. In this case, since the false negative values are significantly higher in ratio, undersampling can be applied to reduce their representation. Alternatively, oversampling the true positive values could also be considered, but it may result in more than half of the actual true positive records being duplicated, potentially affecting model performance.

Please note that these approaches are subject to change based on further analysis and experimentation.