**Group Name:** Project on my own.
**Name:** Aditi Dadariya
**Email:** aditi.dadariya@gmail.com
**Country:** United Kingdom
**College/Company:** MSc (Data Science and Advanced Computing) from University of Reading
**Specialization:** Data Science
**Project Name:** Bank Marketing (Campaign)

**Problem description:** ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

**Data understanding:** Bank Marketing Campaign dataset is focused on a supervised classification problem, aiming to predict whether a customer will purchase the bank's term deposit product.

The dataset consists of 41188 records and 21 variables. The variables are 'age', 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays', 'previous', 'poutcome', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed', 'y'. The variable 'y' is the target variable, while the rest of the variables are considered as features.

The variables 'age', ''duration', 'campaign', 'pdays' and 'previous' are of int64 data type. The variables 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m' and 'nr.employed' are of float64 data type. The variables 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'poutcome' and ‘y’ are of object data type. Please refer Fig 1 below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
None
```

Fig 1: Data info.

Age varies between 17 to 98. Duration varies between 0 to 4918. Campaign varies between 1 to 56. Pdays varies between 0 to 999. Previous varies between 1 to 7. Emp.var.ra te varies between -3.4 to 1.4. Cons.price.idx varies between 92.2 to 94.77. Cons.conf.idx var ies to -50.8 to -26.9. Euribor3m varies between 0.63 to 5.04. Nr.employed varies between 49 63.6 to 5228.1. Please refer Fig 2 below for these details.

| | age | duration | campaign | pdays | previous |
|---|---|---|---|---|---|
| count | 41188.00000 | 41188.000000 | 41188.000000 | 41188.000000 | 41188.000000 |
| mean | 40.02406 | 258.285010 | 2.567593 | 962.475454 | 0.172963 |
| std | 10.42125 | 259.279249 | 2.770014 | 186.910907 | 0.494901 |
| min | 17.00000 | 0.000000 | 1.000000 | 0.000000 | 0.000000 |
| 25% | 32.00000 | 102.000000 | 1.000000 | 999.000000 | 0.000000 |
| 50% | 38.00000 | 180.000000 | 2.000000 | 999.000000 | 0.000000 |
| 75% | 47.00000 | 319.000000 | 3.000000 | 999.000000 | 0.000000 |
| max | 98.00000 | 4918.000000 | 56.000000 | 999.000000 | 7.000000 |

| | emp.var.rate | cons.price.idx | cons.conf.idx | euribor3m | nr.employed |
|---|---|---|---|---|---|
| count | 41188.000000 | 41188.000000 | 41188.000000 | 41188.000000 | 41188.000000 |
| mean | 0.081886 | 93.575664 | -40.502600 | 3.621291 | 5167.035911 |
| std | 1.570960 | 0.578840 | 4.628198 | 1.734447 | 72.251528 |
| min | -3.400000 | 92.201000 | -50.800000 | 0.634000 | 4963.600000 |
| 25% | -1.800000 | 93.075000 | -42.700000 | 1.344000 | 5099.100000 |
| 50% | 1.100000 | 93.749000 | -41.800000 | 4.857000 | 5191.000000 |
| 75% | 1.400000 | 93.994000 | -36.400000 | 4.961000 | 5228.100000 |
| max | 1.400000 | 94.767000 | -26.900000 | 5.045000 | 5228.100000 |

Fig 2: Data describe.

**Challenges Faced in the Dataset and Implemented Solutions:**

1. Missing Values: The dataset does not contain any null (NaN) values. However, there are 'unknown' values present in 'job', 'marital', 'education', 'default', 'housing', 'loan' columns, which can be treated as missing values. The variable 'job' has 330 "unknown" values. Th e variable 'marital' has 80 "unknown" values. The variable 'education' has 1731 "unknow n" values. The variable 'default' has 8597 "unknown" values. The variable 'housing' has 9 90 "unknown" values. The variable 'loan' has 990 "unknown" values. Please refer Fig 3 b elow. The "unknown" values in the dataset have been replaced with NaN values. Subseq uently, the NaN values have been replaced with their corresponding indices.

```
# Variables consisting "unknown" values
columns_with_unknown = df_data.columns[df_data.isin(['unknown']).any()]
columns_with_unknown
```

```
Index(['job', 'marital', 'education', 'default', 'housing', 'loan'], dtype='object')
```

```
# Find the number the "unknown" in each variable
unknown_counts = df_data[df_data == 'unknown'].count()
unknown_counts
```

```
age                  0
job                330
marital             80
education         1731
default           8597
housing            990
loan               990
contact              0
month                0
day_of_week          0
duration             0
campaign             0
pdays                0
previous             0
poutcome             0
emp.var.rate         0
cons.price.idx       0
cons.conf.idx        0
euribor3m            0
nr.employed          0
y                    0
dtype: int64
```

Fig 3: "unknown" values

2. <u>Inconsistent datatypes:</u> After converting the “unknown” values to NaN values, the categorical variables ‘job’, ‘marital’, ‘education’, ‘default’, ‘housing’, ‘loan’ now has multiple datatypes such as str and int. Please refer Fig 4 below. The datatype of all these categorical variables has been converted to string.

```
Column 'job' has multiple data types.
Data types in 'job': [<class 'str'> <class 'int'>]

Column 'marital' has multiple data types.
Data types in 'marital': [<class 'str'> <class 'int'>]

Column 'education' has multiple data types.
Data types in 'education': [<class 'str'> <class 'int'>]

Column 'default' has multiple data types.
Data types in 'default': [<class 'str'> <class 'int'>]

Column 'housing' has multiple data types.
Data types in 'housing': [<class 'str'> <class 'int'>]

Column 'loan' has multiple data types.
Data types in 'loan': [<class 'str'> <class 'int'>]
```

Fig 4: Multiple datatypes

3. <u>Encoding the categorical value:</u> The categorical variables ‘job’, ‘marital’, ‘education’, ‘default’, ‘housing’, ‘loan’, ‘contact’, ‘month’, ‘day_of_week’, ‘poutcome’ and ‘y’ are converted to numeric using Label Encoder to maintain the consistency of data.

4. <u>Outlier Detection:</u> The Interquantile Range (IQR) method was employed for detecting outliers in the dataset. The dataset exhibits numerous outliers in the 'age', 'duration', 'campaign', 'pdays', 'previous', and 'cons.conf.idx' columns. Please refer Fig 5 to see the number of outliers in each column.

```
469 number of row are Outliers in column age
2963 number of row are Outliers in column duration
2406 number of row are Outliers in column campaign
1515 number of row are Outliers in column pdays
5625 number of row are Outliers in column previous
447 number of row are Outliers in column cons.conf.idx
```

Fig 5: Number of outliers

5. <u>Rescaling data:</u> Standardization (StandardScaler library) is a technique used to rescale data to have zero mean and unit variance.

6. <u>Imbalance data:</u> The target variable 'y' in the dataset consists of 36,548 values for the category 'no' and 4,640 values for the category 'yes'. This reveals an imbalance in the dataset. To address this issue, the SMOTE (Synthetic Minority Over-sampling Technique) method was employed to balance the dataset, ensuring a more equitable distribution for subsequent analysis.

**Github Repo link:** https://github.com/aditidadariya/BankMarketingCampaign