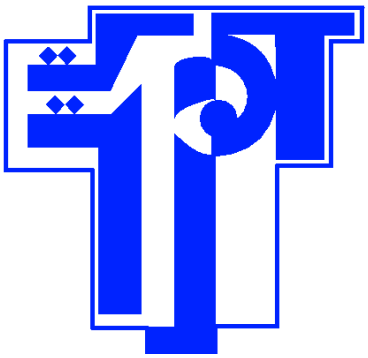
Université de Carthage



Ecole Polytechnique de Tunisie

Data Analysis Porject Report

Data Analysis of Bank Marketing Campaign

Work done by : Myriam El Amri
Rahma Kharrat
Chawki Hjaiji

Supervised by : Mr. Amor Messaoud

Academic Year : 2023-2024

Abstract

This report provides insights into the bank marketing campaign through a detailed exploration of a Python-based notebook. Our analysis encompasses key phases, including data cleaning, exploratory data analysis, and feature engineering, all executed within a Python programming environment. We employed various classification models to analyze customer behavior effectively. Furthermore, we explained the methods and statistical approaches used throughout the analysis.We also presented the results obtained from the employed algorithms along with their appropriate interpretations. Overall, this work serves the purpose of understanding customer behavior and facilitating classification tasks within the context of the bank marketing campaign.

Contents

Table of Figures

Introduction

The contemporary business landscape emphasizes data-driven decision-making, where insights derived from extensive datasets significantly influence strategies and outcomes. In this context, our objective is to analyze and extract insights from a bank's marketing campaign dataset. This dataset encompasses various attributes related to customer demographics, financial indicators, and interaction history with the bank. By exploring and analyzing these data points, we aim to identify key factors that influence the success of marketing campaigns and predict customer response to future initiatives. The report is structured to cover essential aspects of the data analysis process, including data cleaning, exploratory data analysis (EDA), feature engineering, and the application of classification models. Each stage of the analysis is meticulously documented, offering insights into the methodologies employed and the rationale behind key decisions.

I Problem Statement and Data

I.1 Problem Statement

The Portuguese bank is facing a decline in revenue and seeks guidance on the necessary actions to address this issue. Upon investigation, it was revealed that the primary cause of this decline is a decrease in client deposits. Term deposits, which allow banks to retain funds for a fixed period, are crucial for banks to invest in higher-yield financial products and generate profits. Furthermore, term deposit clients present opportunities for banks to promote additional products such as funds or insurance, thus bolstering revenue streams. Consequently, the Portuguese bank aims to identify existing clients with a higher likelihood of subscribing to term deposits and prioritize its focus accordingly.

I.2 Data

The data originates from (https ://archive.ics.uci.edu/dataset/222/bank+marketing) and details the outcomes of marketing campaigns conducted by a bank in Portugal. These campaigns primarily involved direct phone calls, where clients were offered the opportunity to subscribe to a term deposit. The target variable, labeled 'yes' if the client agreed to deposit and 'no' if not, reflects the success of these marketing efforts. The data we used, named bank-additional-full.csv, contains all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010).

I.2.1 Attribute Information

Bank client data :

- **age** : (numeric)
- **job** : type of job (categorical)
- **marital** : Marital status (categorical)
- **education** : Education level (categorical)
- **default** : Indicates if he has credit in default (categorical)
- **housing** : Indicates if he has housing loan (categorical)
- **loan** : Indicates if he has personal loan (categorical)

Related with the last contact of the current campaign :

- **contact** : Contact communication type (categorical)
- **month** : Last contact month of the year (categorical)
- **day_of_week** : Last contact day of the week (categorical)
- **duration** : Last contact duration, in seconds (numeric).

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)

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 :

y : Indicates if the client subscribed to a term deposit (binary : 'yes', 'no')

II Data Visualization and Missing Values

After examining the initial rows of our data, we notice that there are missing values denoted by terms such as 'Unknown', 'nonexistent', and '999'. Addressing missing values is crucial as they can introduce bias and inaccuracies in data analysis, potentially distorting statistical measures such as means, medians, and correlations, thereby yielding erroneous insights. Hence, it is essential to handle these missing values appropriately. Also, the presence of duplicates in the dataset necessitates their removal to ensure data integrity and prevent redundant information from influencing the analysis.

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	campaign	pdays	previous	poutcome
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	...	1	999	0	nonexistent
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	...	1	999	0	nonexistent
2	37	services	married	high.school	no	yes	no	telephone	may	mon	...	1	999	0	nonexistent
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	...	1	999	0	nonexistent
4	56	services	married	high.school	no	no	yes	telephone	may	mon	...	1	999	0	nonexistent

FIGURE 1 – Head of data

The categorical features with the 'unknown' category include 'job', 'marital', 'education', 'default', 'housing', and 'loan'. To investigate potential patterns or correlations between unknown values across different features for the same individual, we generated a correlation matrix, giving the following results : The correlation coefficient serves as an indicator of the relationship between features with regard to the

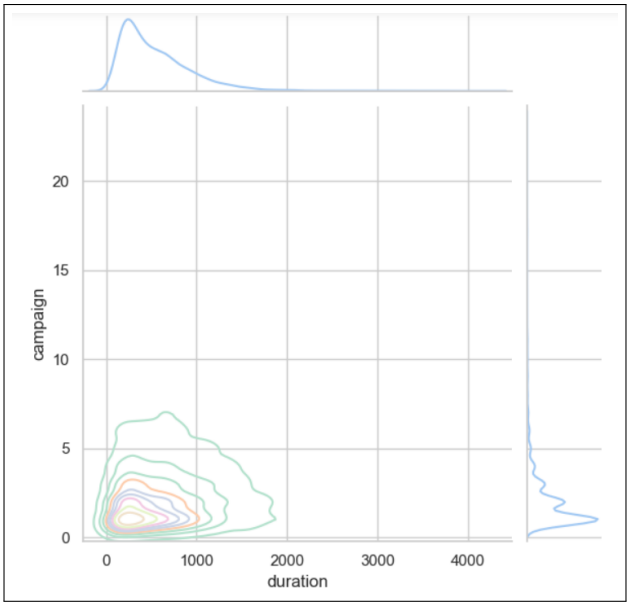


FIGURE 2 – Correlation Between Unknown Values in Categorical Features

presence or absence of "unknown" values. A coefficient of 1 indicates a perfect positive association, implying that whenever one feature has an "unknown" value, the other feature also has an "unknown" value, and vice versa. Conversely, a coefficient of -1 suggests a perfect negative association, meaning that when one feature has an "unknown" value, the other feature does not. Analyzing our correlation matrix, we observe a positive association between 'loan' and 'housing', indicating they share the same

percentage of missing values. Additionally, a moderate negative association is noted between 'education' and 'default'. While the 'job' distribution has a low percentage of 'Unknown' values (0.8%), facilitating consideration for dropping or replacing them with the 'admin' category, 'default' presents a challenge with a high percentage of missing values (20%).

For further exploration, let's examine the information gain for the 'default' feature. To do so, we need to define mutual information (MI), a measure of the mutual dependence between two random variables in probability theory and information theory. Specifically, MI quantifies the "amount of information" obtained about one random variable by observing the other. This concept is closely tied to the entropy of a random variable, which measures the expected "amount of information" held in a random variable.

The mutual information between variables X and Y is given by :

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x) \cdot p(y)} \right)$$

Here, $p(x,y)$ represents the joint probability mass (or density) function of variables X and Y, while $p(x)$ and $p(y)$ are their marginal probability mass (or density) functions, respectively.

The obtained information gain for the 'default' feature is 0.0058, indicating that it provides only a small amount of information about the target variable 'y'. Consequently, we removed the 'default' column as it contributes insignificantly to the predictive power of the model.

The relationship between 'housing' and 'loan' categories when 'loan' is 'no' is as follows :

$$\text{yes} : 17885$$

$$\text{no} : 16065$$

From this, it's evident that 'housing' and 'loan' are associated. When considering replacing 'unknown' values in 'housing' with the most frequent value, it's notable that the corresponding values for 'no' in 'housing' can be either 'yes' or 'no' with almost equal probability. Hence, we choose to replace the missing values in 'housing' with the most frequent category as well.

Now, let's examine the relationship between 'education' and 'job'. We can utilize the chi-square test, a statistical method for assessing the association between categorical variables in a contingency table. The chi-square statistic (χ^2) is computed based on the disparities between observed and expected frequencies in each cell of the contingency table.

Education categories and their respective counts are as follows :

Education	Count
basic.4y	4176
basic.6y	2291
basic.9y	6045
high.school	9512
illiterate	18
professional.course	5240
university.degree	12164
unknown	1730

The chi-square test gives a p-value of 0.0, indicating a significant association between education levels and the chosen categorical variable 'job'. This suggests that education levels may influence or be related to the distribution of categories in the 'job' feature. Consequently, understanding the job of individuals with 'unknown' education levels may aid in filling in the missing values. So let's further investigate the relationship between job and the unknown values for education to determine if we can replace some missing values. We obtained the following figure. Blue-collar workers typically engage in manual labor

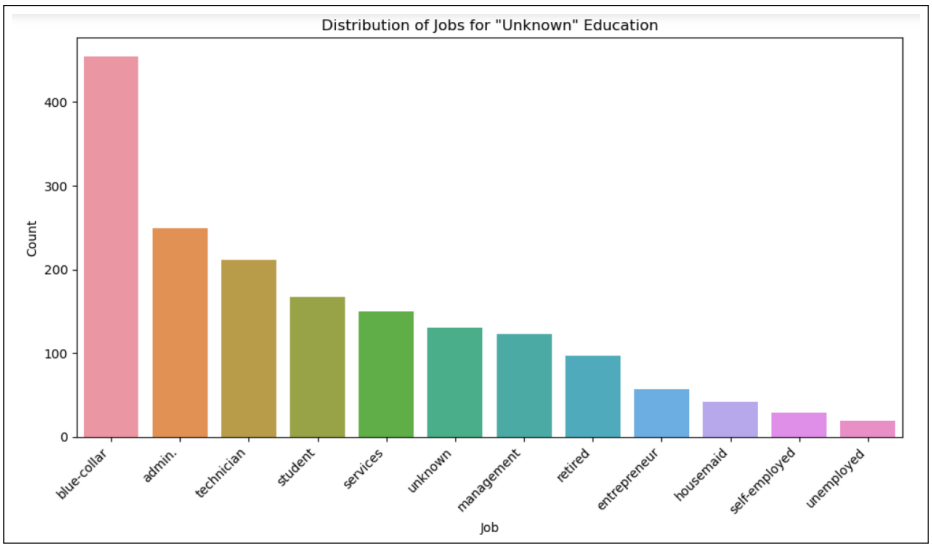


FIGURE 3 – Distribution of jobs for 'Unknown' Education

or factory jobs and don't typically work in office environments. These jobs often require less formal education, usually equivalent to a high school or junior high school level. Therefore, we found it appropriate to replace missing education values with 'basic'.

Conclusion : Handling Missing Values in Categorical Features

For the education feature : We will group individuals with only basic education into one category called 'basic'. We will replace the missing values in education with 'basic'.

For the default column : We will drop it as it doesn't provide significant added value. For other categorical features : We will replace missing values with the most frequent category.

The remaining categorical data also contain missing values, but they are not labeled as 'Unknown'. Through further exploration, we found that a substantial portion of the data (marked as '999') indicates that clients were not previously contacted. In the subset of data where clients were contacted previously, comprising a small fraction of the overall dataset, the distribution between 'No' and 'Yes' outcomes is nearly equal. Consequently, this column does not significantly contribute to the prediction task and can be dropped from further analysis to streamline the dataset without sacrificing predictive accuracy. This observation is supported by the following figure :

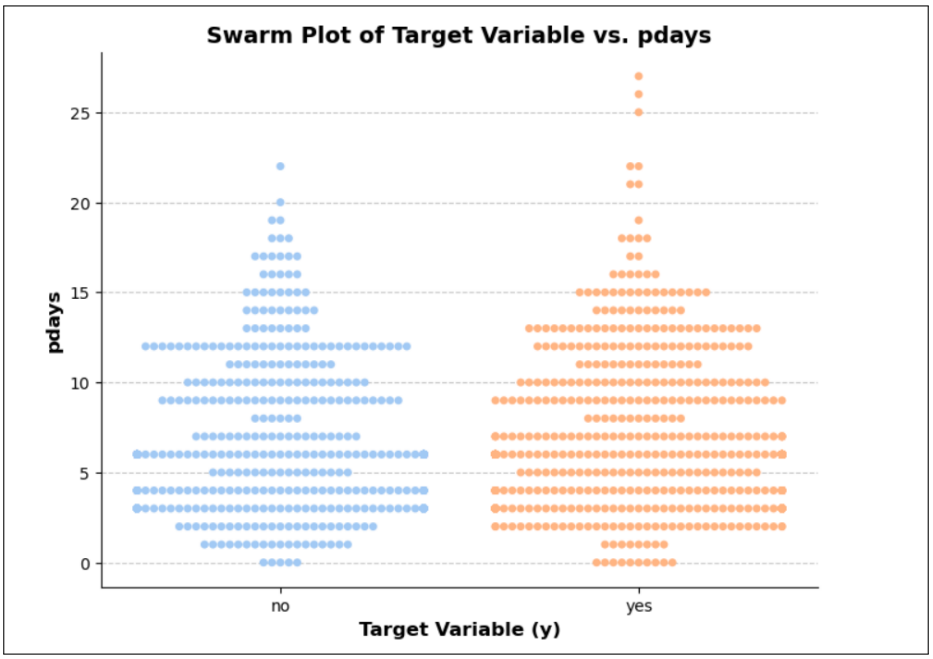


FIGURE 4 – Swarm Plot of Target Variable vs. pdays

Analyzing the distribution of call duration concerning the target variable reveals a notable trend : longer call durations are associated with a higher likelihood of the target variable ('y') being 'yes'. Conversely, a significant proportion of customers who experienced shorter call durations declined to subscribe to the deposit ('y' equal to 'no'). Hence, it is evident that this feature significantly influences the target variable.

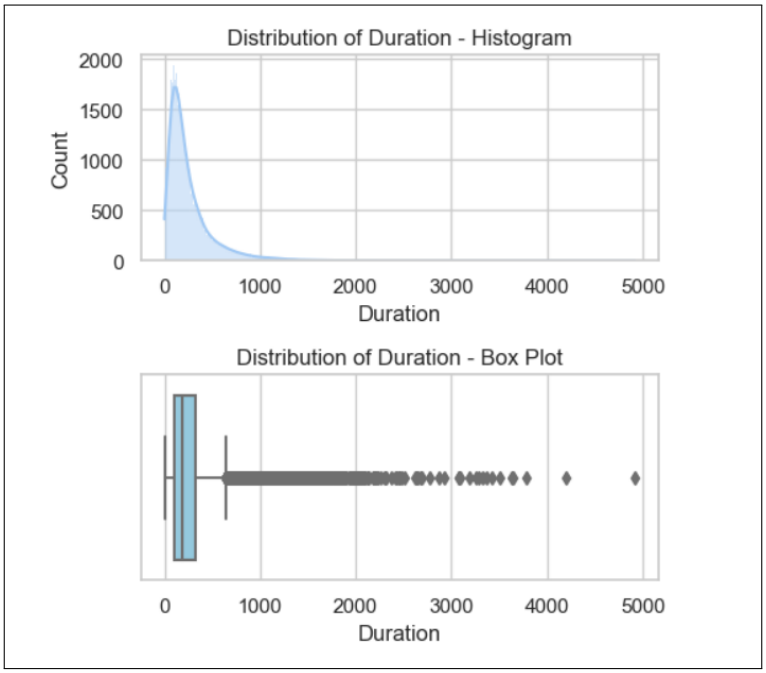


FIGURE 5 – Distribution of Duration

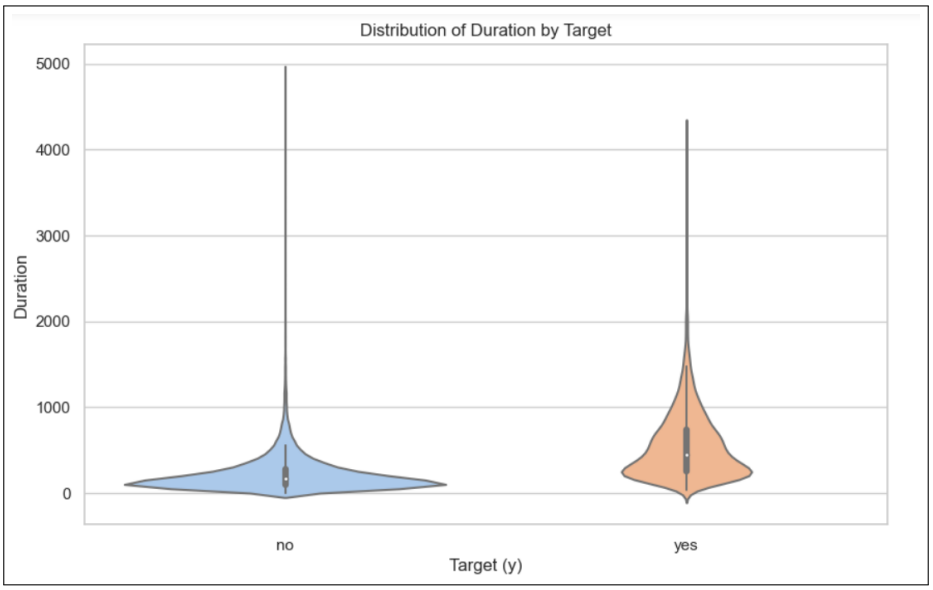


FIGURE 6 – Distribution of Duration by Target

Number of instances where the call duration is equal to 0 : 4. The values of 'previous' for these instances are all 0, indicating that these clients were not contacted previously. Given their minimal presence in the dataset, comprising only 4 rows, they are unlikely to impact the analysis results. If our goal is to evaluate the marketing campaign's effectiveness, we should prioritize a larger subset of instances where both the call duration and 'previous' are 0. Therefore, we we proceeded by dropping these rows.
For the 'contact' feature we have the distribution in figure 7. We conducted a chi-square test to assess its association with the target variable 'y'. The chi-square statistic yielded a value of 862.18, indicating a substantial difference between the observed and expected frequencies in the contingency table. Furthermore, the calculated p-value, approximately 1.64e-189, is significantly lower than the typical significance threshold of 0.05. This exceptionally low p-value suggests that the observed results are highly improbable under the assumption of independence between the 'contact' and 'y' variables. Consequently, we reject the null hypothesis of independence, concluding that there exists a statistically significant association between these two variables.

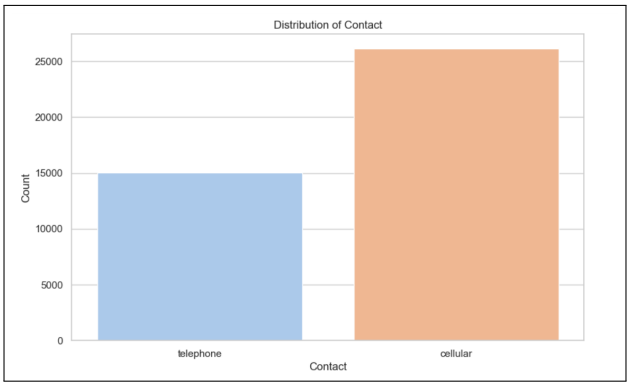


FIGURE 7 – Distribution of Contact

Considering the 'campaign' feature, the distribution varies based on whether 'y' is equal to 'no' or 'yes', indicating that the number of contacts during the campaign affects the outcome variable 'y'. Moreover, there is a notable presence of outliers when 'y' is equal to 'no'.

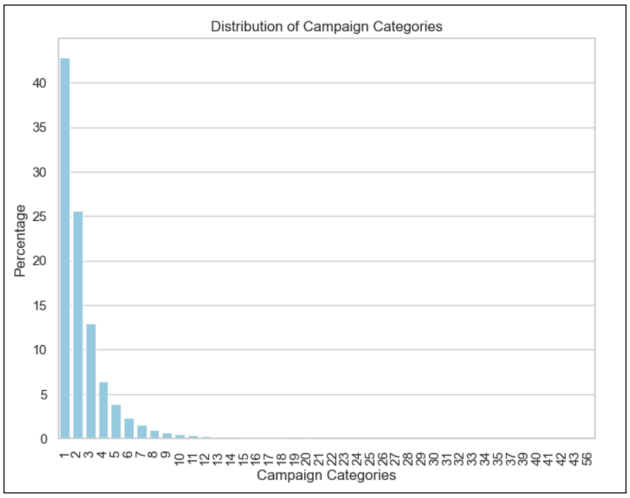


FIGURE 8 – Distribution of Campaign Categories

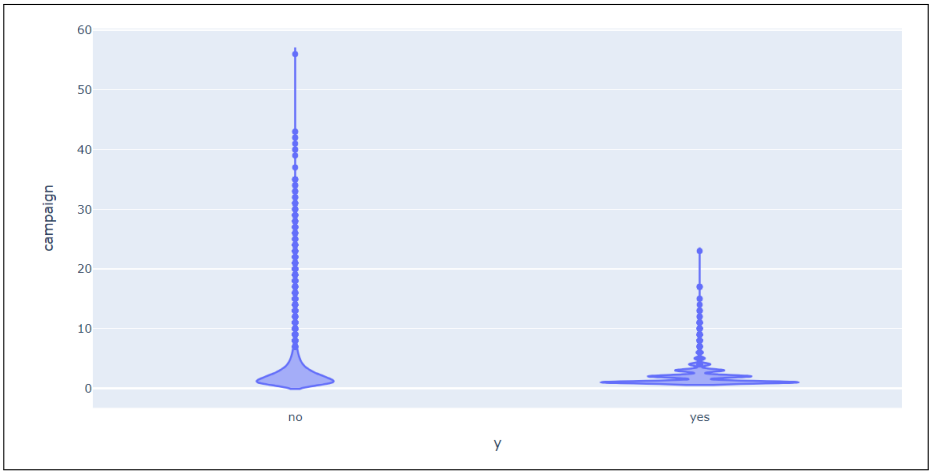


FIGURE 9 – Violon Plot for Campaign Feature

For 'poutcome', as indicated in the figure below, when the outcome is 'nonexistent,' the proportion of 'no' increases significantly, jumping from 73% to 91%. Therefore, we cannot delete this column, as the missing values have a noticeable effect on the results.

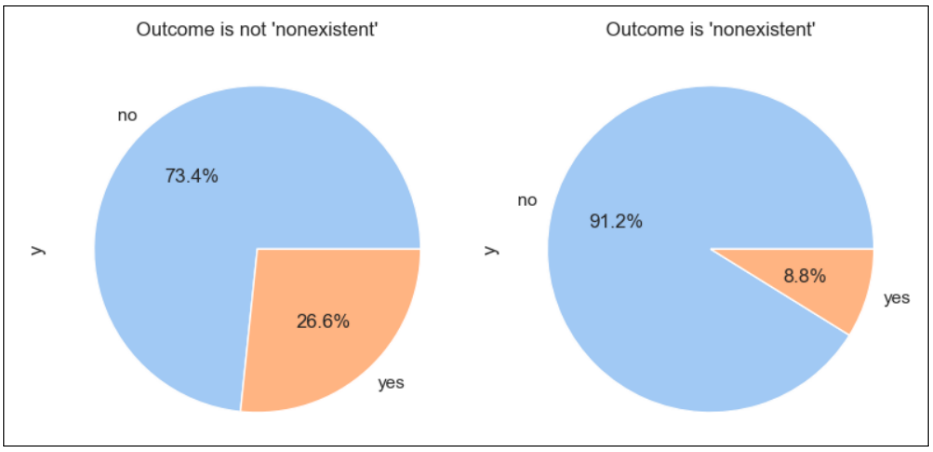


FIGURE 10 – Pie Chart for poutcome Feature

Let's now see the feature 'previous'. This is its distribution :

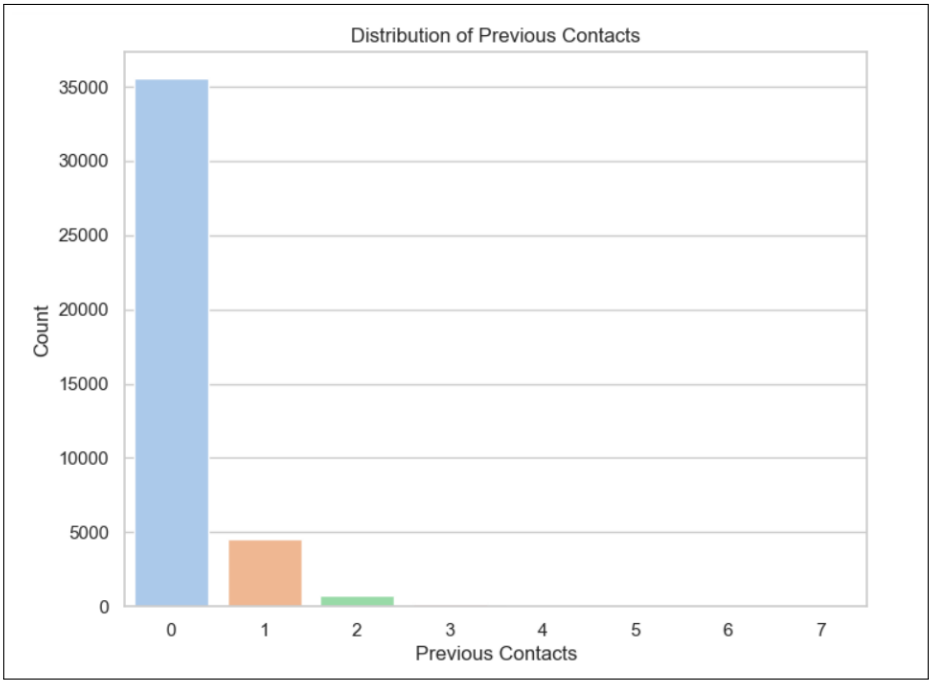


FIGURE 11 – Distribution of previous

A significant portion of customers did not have previous calls. Therefore, it is crucial to ascertain whether the number of previous calls affected the results or not to determine whether we should retain or remove the column.

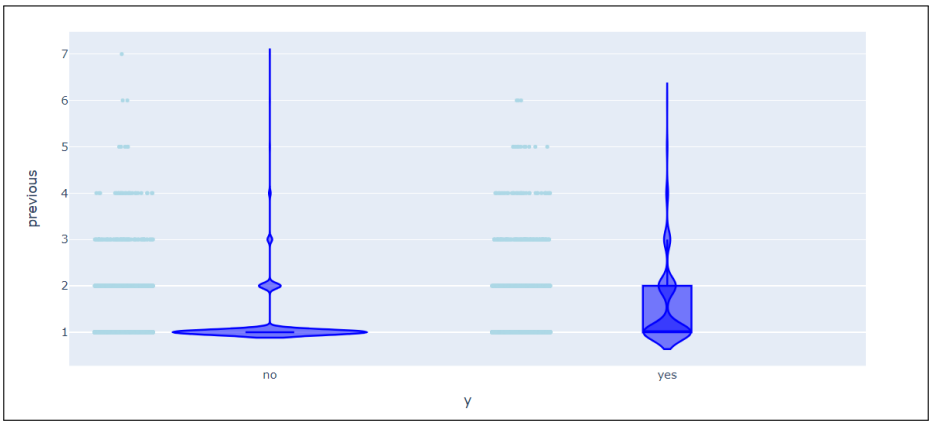


FIGURE 12 – Violin Plot of Target Variable vs. customers who had previous calls

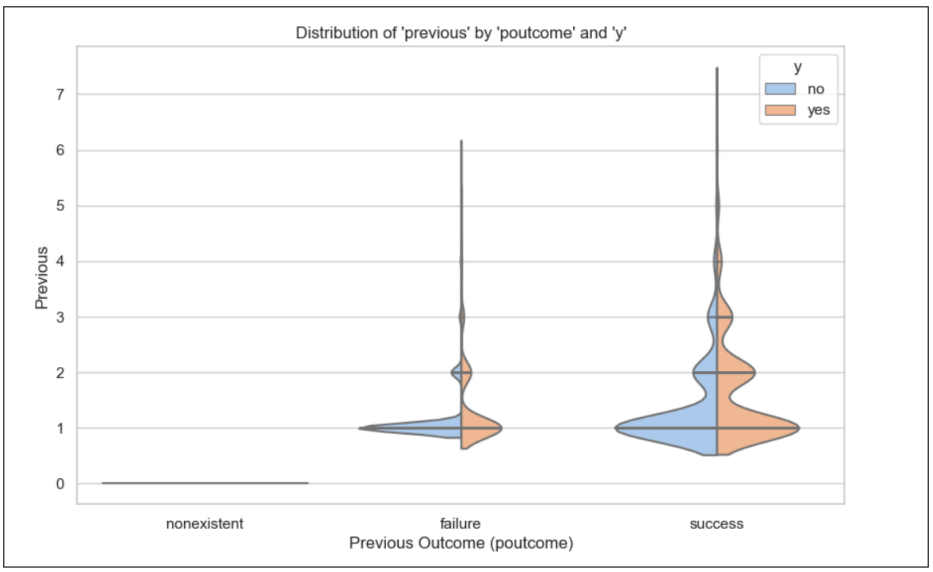


FIGURE 13 – Distribution of 'previous' by 'poutcome' and 'y'

The ratio of 'yes' to 'no' outcomes for clients with no previous calls is approximately 0.36, while for clients with previous calls, this ratio drops to around 0.097. This disparity suggests that the variable 'previous' has an impact on the outcome 'y', indicating that individuals who have had more contacts before the campaign are more likely to subscribe to the service. Retaining the column 'previous', which represents the number of contacts performed before the current campaign, could be valuable in predicting subscription outcomes or understanding customer behavior and preferences.This can be understood thanks to the figure 13.

III Explatory Data Analysis

The distribution of our target variable is depicted in Figure 14. We observe that our data is imbalanced, with approximately 88.73% of the observations having 'y' equal to 'no'. This indicates the need to employ methods to balance our data before training our models.

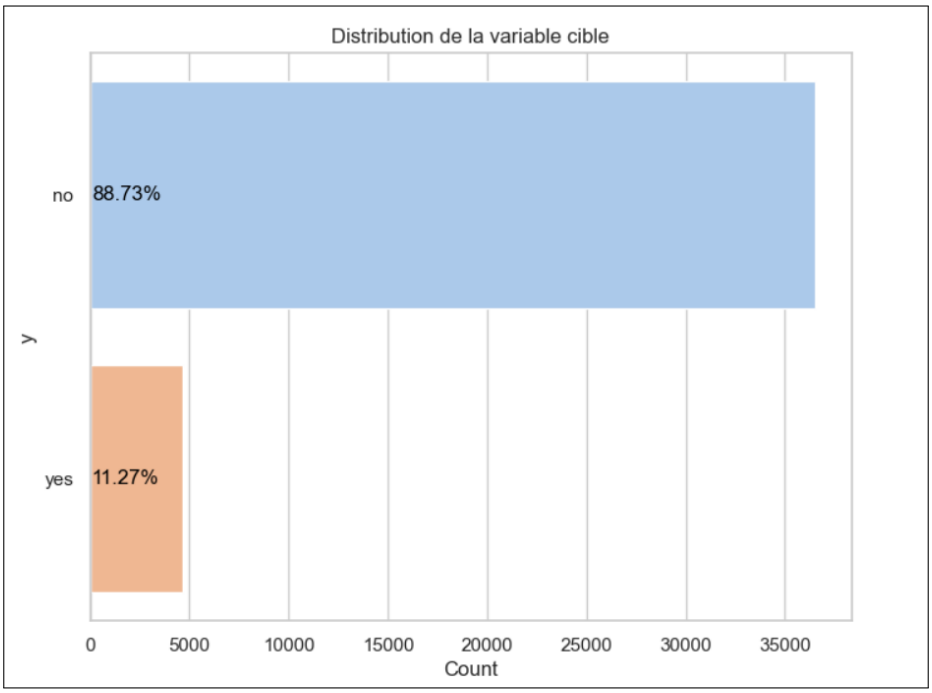


FIGURE 14 – Distribution of Target Variable

III.1 Exploring Correlated Numerical Features

To observe the correlation between numerical features and the target variable 'y', it's necessary to convert 'y' into a numerical variable. In binary classification tasks, 'y' typically represents two classes, labeled as '0' and '1'. The Pearson correlation measures the strength of the linear relationship between two variables, ranging from -1 to 1. A value of -1 indicates a total negative linear correlation, 0 signifies no correlation, and +1 denotes a total positive correlation.

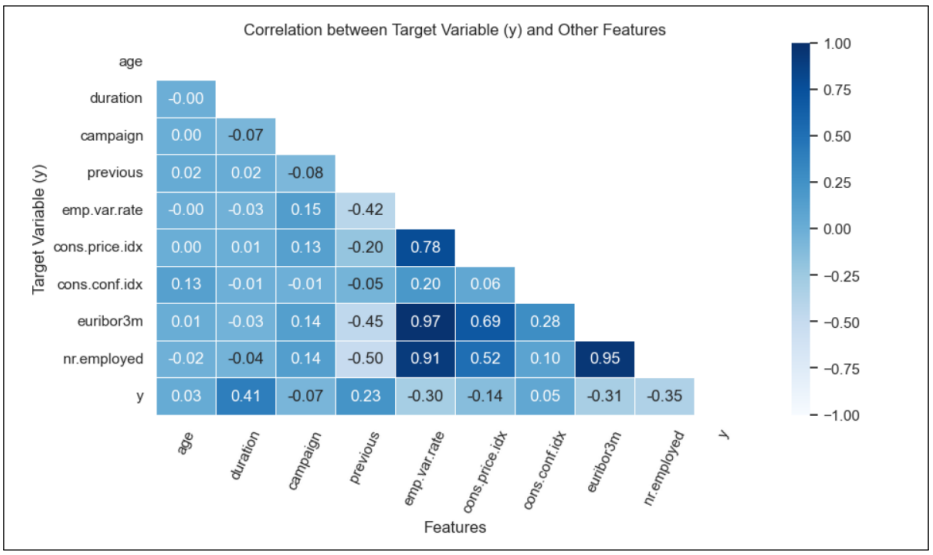


FIGURE 15 – Correlation between Target Variable (y) and Other Features

Numerical Feature	Correlation with 'y'
duration	0.405285
nr.employed	0.354688
euribor3m	0.307768
emp.var.rate	0.298313
previous	0.230192

The analysis reveals that the feature with the highest correlation with 'y' is duration. This underscores the significance of this feature and prompts us to explore its relationship with other features. Additionally, it suggests the potential importance of time-related features in our analysis. Social and economic context features also hold significance, which is logical given that the data was collected during a period of financial crisis.

Feature 1	Feature 2	Correlation
duration	y	0.405285
emp.var.rate	cons.price.idx	0.775290
emp.var.rate	euribor3m	0.972241
emp.var.rate	nr.employed	0.906945
cons.price.idx	emp.var.rate	0.775290
cons.price.idx	euribor3m	0.688167
cons.price.idx	nr.employed	0.521931
euribor3m	emp.var.rate	0.972241
euribor3m	cons.price.idx	0.688167
euribor3m	nr.employed	0.945146
nr.employed	emp.var.rate	0.906945
nr.employed	cons.price.idx	0.521931
nr.employed	euribor3m	0.945146
y	duration	0.405285

Considering the period of financial crisis during which the data was collected, it's logical that social and economic context features also hold significance.
The global financial crisis that began in 2008 can be summarized into three main phases :

- **2008 : Banking System Seizure** - The crisis escalated on September 15, 2008, when Lehman Brothers, a major investment bank, declared bankruptcy, challenging the notion that all banks were "too big to fail." Governments responded by injecting substantial capital into banks to avert collapse, but the global economy plunged into a period of freefall.
- **2009 : Coordinated Global Response** - In early 2009, amid the global financial crisis, G20 nations collaborate to forestall recession from escalating into a depression. A pivotal moment occurs on April 2, 2009, during the London G20 summit, where world leaders pledge to implement a $5 trillion fiscal expansion, enhance resources for international institutions, and undertake bank reforms. However, this period also marks the onset of diminishing international cooperation as countries increasingly prioritize individual agendas over collective action.
- **2010 : Sovereign Debt Crisis** - On May 9, 2010, attention transitions from the private sector to the public sector as governments grapple with escalating budget deficits. Greece's receipt of financial aid from the IMF and EU underscores apprehensions regarding government solvency. Consequently, austerity measures gain prominence, influencing policy decisions across multiple nations.

These phases illustrate the progression of the crisis from its origins in the banking system to broader economic impacts and concerns about sovereign debt. Since our data is ordered by date (from May 2008 to November 2010), we can create a new column named "year" because the year may influence in this case.

III.2 Analyzing Bank Client Data

Considering the various plots we analyzed in the notebook, several observations emerge. Retirees and students exhibit the highest proportions of deposit subscriptions, despite constituting a smaller client base. Conversely, blue-collar workers, characterized by manual labor or skilled trades, show the lowest subscription rates. Administrators stand out as the most contacted group during the campaign. While married individuals engage the most with the campaign, singles demonstrate a higher propensity to subscribe to deposits. Surprisingly, illiterate individuals exhibit the highest deposit subscription rate, albeit being a minority. Moreover, there's a clear trend indicating that higher education levels correlate with increased subscription rates. Interestingly, the presence of a housing loan doesn't notably affect subscription rates. Regarding age distribution, it skews towards the right, with the majority falling within the 30 to 50 age bracket, along with noticeable outliers.
Let's employ Cramer's Correlation. Cramer's V correlation measures the association between two attributes, with its value ranging from 0 (indicating no relationship between the attributes) to 1 (signifying complete association between variables). A value of 1 is achieved only when one attribute entirely determines the other attribute.

$$V = \sqrt{\frac{\varphi^2}{\min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

Cramer's V Formula

where attributes are as follows,

- φ is the phi coefficient.
- χ^2 is derived from Pearson's chi-squared test
- n is the grand total of observations and
- k being the number of columns.
- r being the number of rows.

FIGURE 16 – Enter Caption

Using this correlation method, we obtained the following results :

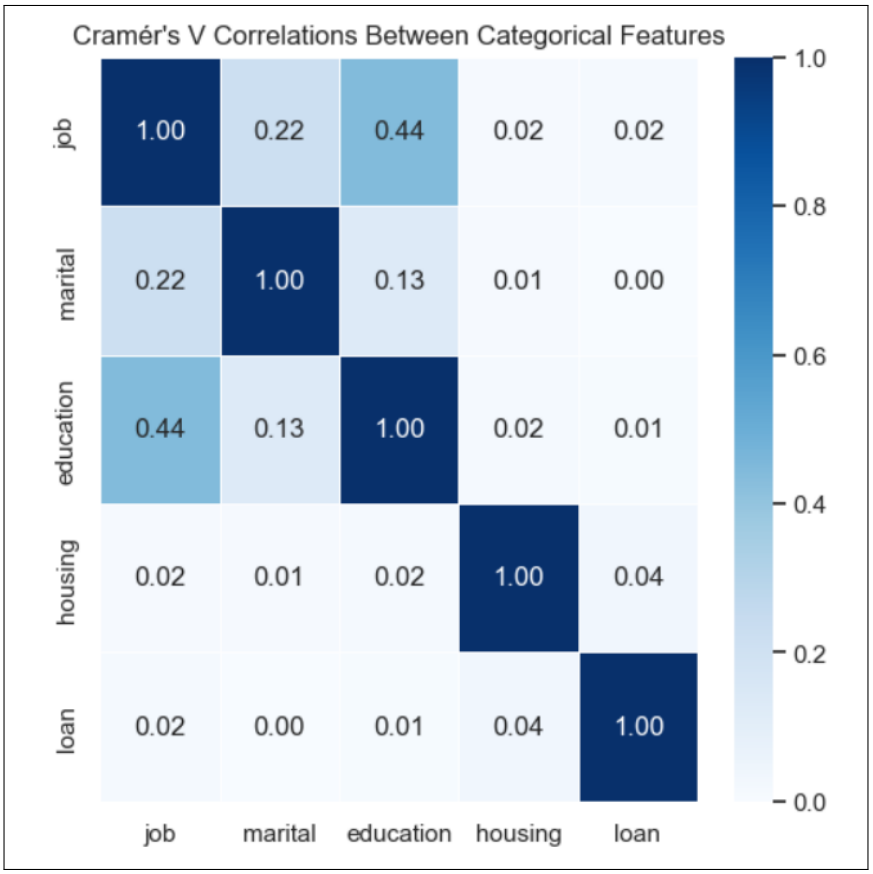


FIGURE 17 – Cramér's V Correlations Between Categorical Features

As anticipated, we found a considerable correlation between job and education, consistent with our observations. However, we didn't observe any significant correlations between the other features.

III.3 Analyzing Related with the last contact of the current campaign Data

Considering the years, we observe nearly four times as many individuals making a deposit in 2009 compared to 2008. In terms of months, May stands out with the highest number of contacts, although it yields poor results for deposits. Conversely, December, October, and September exhibit significantly lower contacts, with almost 50% making a deposit, which is suspicious. Similarly, when considering the days of the week, there is no significant difference observed. However, it's noteworthy that the majority of contacts were made using cellular communication.

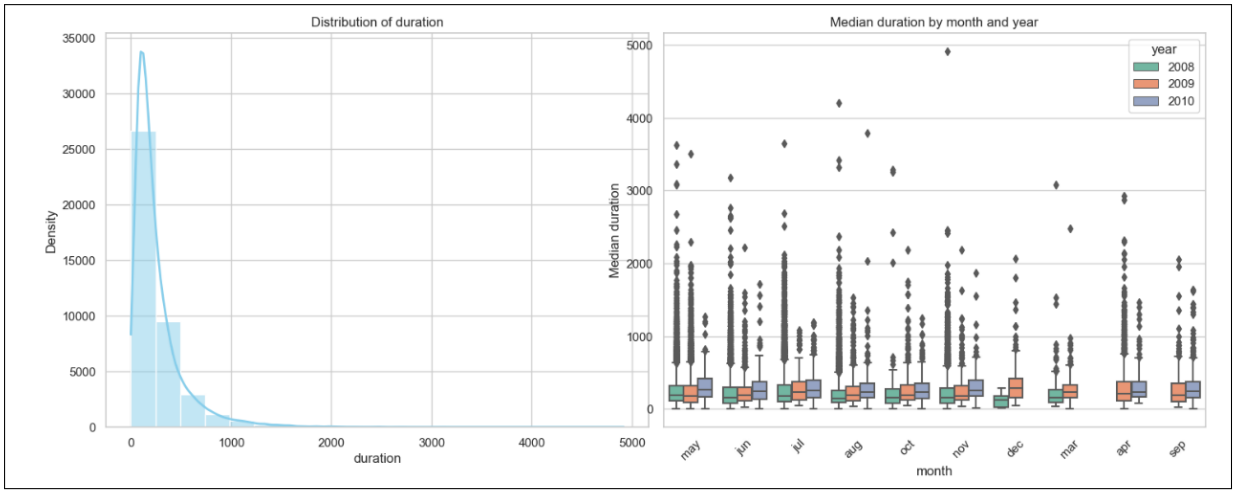


FIGURE 18 – Distribution of Time Related Features

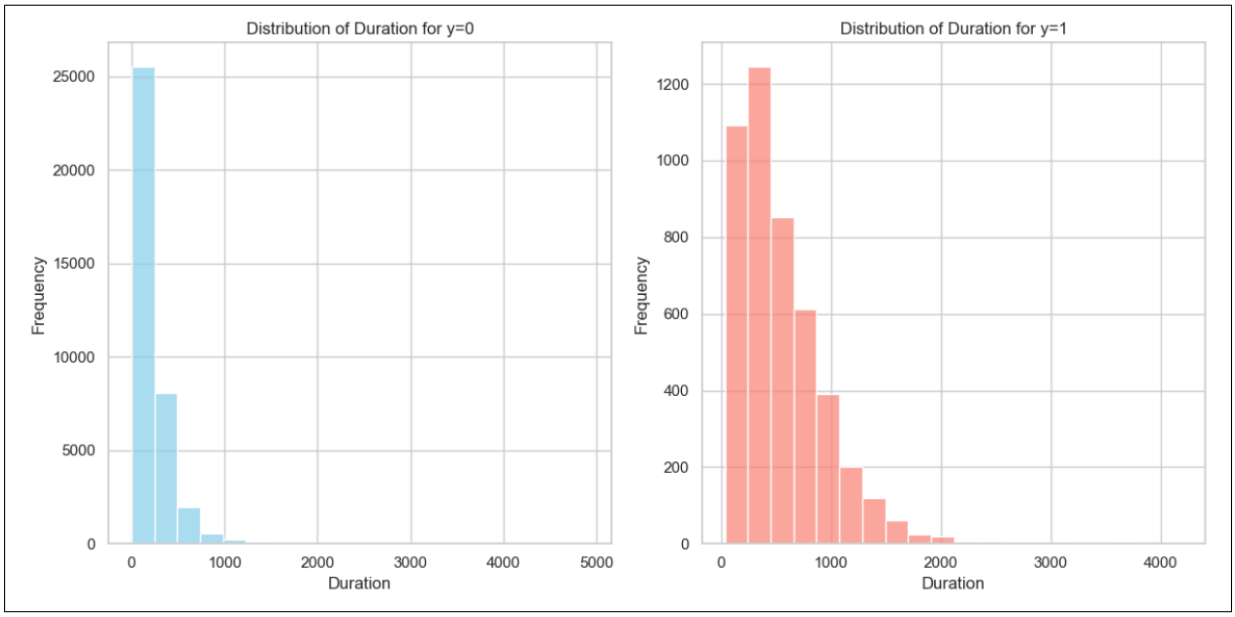


FIGURE 19 – Distribution of Duration

We see clearly tahnks to figure 18 and 19 that the longer the call duration, the higher the likelihood of securing a deposit. Extended conversations may signify increased client engagement or interest, often resulting in successful outcomes such as a deposit.

III.4 Analyzing Other Attributes

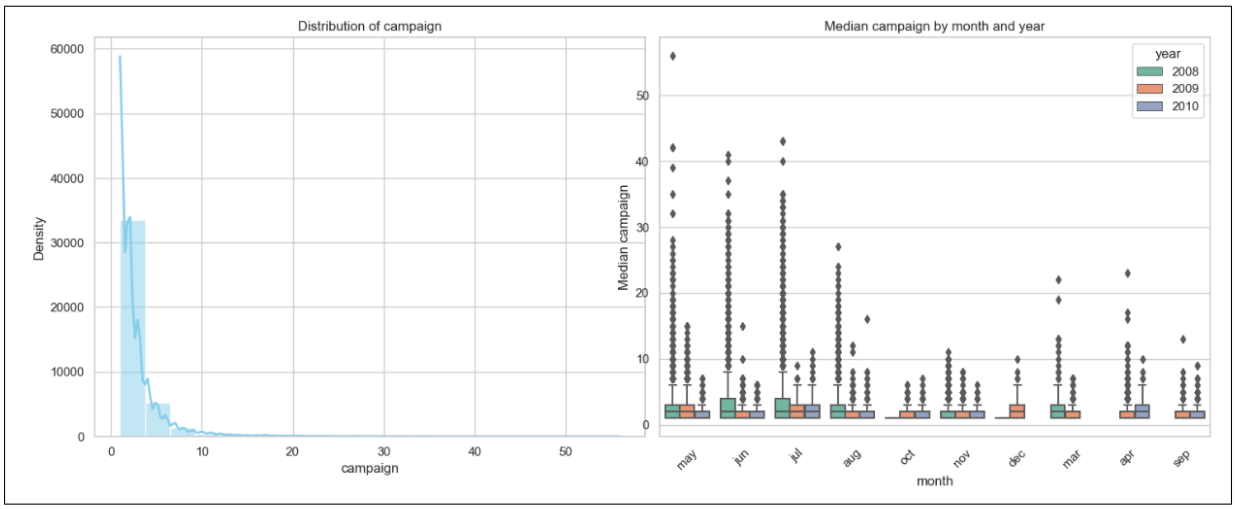


FIGURE 20 – Analyzing the Feature Campaign

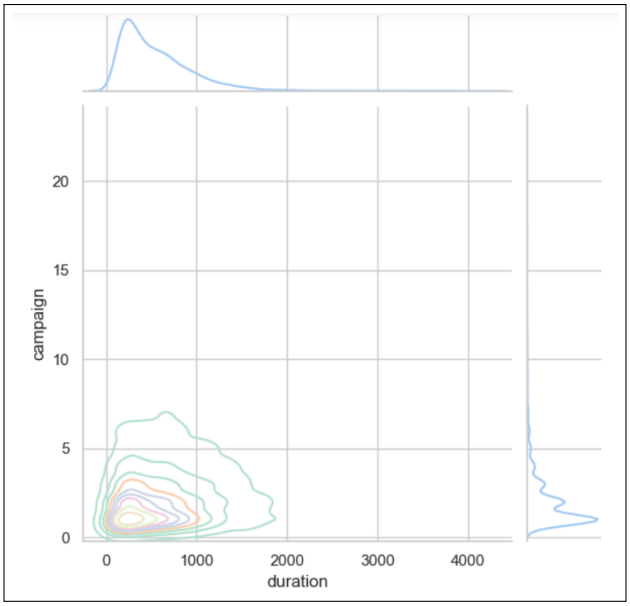


FIGURE 21 – Campaign vs Duration Distribution

Figures 20 and 21 display skewed right distributions, indicating a positive skewness in the data, with few instances of very high values compared to the majority of lower values. The jointplot illustrates that both call duration and the number of campaign contacts peak at the same point, suggesting a positive correlation between the two variables. Clients contacted more frequently during the campaign tend to have longer call durations. The skewed right distribution for call duration implies that while most calls have shorter durations, some outliers exhibit significantly longer durations. If we filter the data where

y=0, we observe similar results, albeit with lower values for both features. From Figure 22, we note the

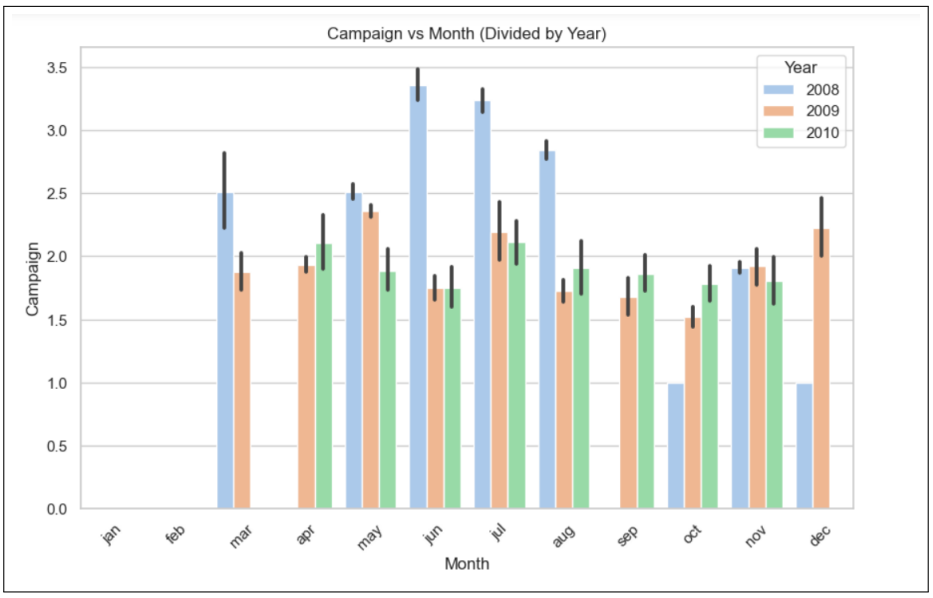


FIGURE 22 – Campaign vs month

highest number of contacts occurring in June, July, and August, especially in 2008. However, during this period in 2008, only 5% of contacts resulted in a positive outcome ($y = 1$). Given the presence of outliers, it's essential to investigate whether these outliers correspond to individuals who ultimately made a deposit. For features 'poucome' and 'previous' :

Percentage Matrix	**poutcome=failure**	**poutcome=success**
$y = 0$	0.857714	0.348871
$y = 1$	0.142286	0.651129

Overall, clients who had a successful outcome in the previous campaign are more likely to succeed in the current campaign, while those who had a failure outcome in the previous campaign are less likely to succeed. However, for success, the symmetrical curves suggest that the number of contacts was not the influencing factor for the deposit variable or the outcome. On the other hand, for failure, it seems that if you had a previous outcome of failure, you are also more likely to experience another failure. Perhaps, this indicates a slight improvement, suggesting that some clients who previously failed may have responded positively in the subsequent campaign.

III.5 Analyzing Social and Economic Context Attributes

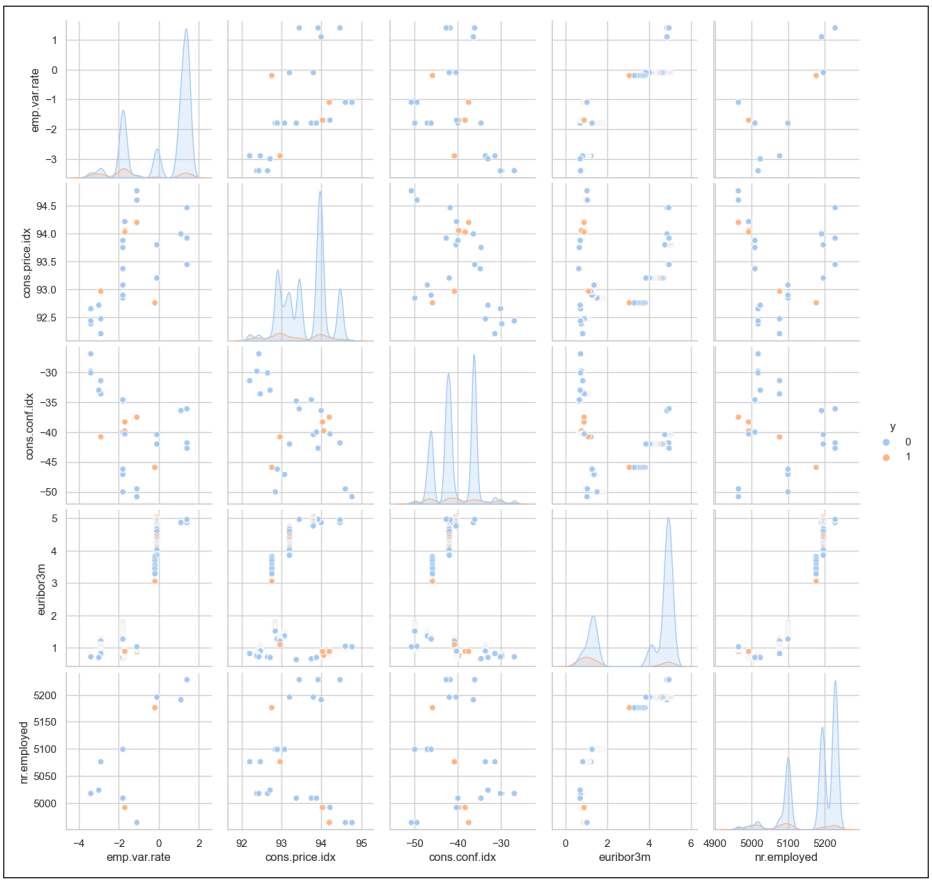


FIGURE 23 – Pairplot for Social and Economic Context Attributes

From Figure 23, we observe similar trends in the plots of euribor vs. (cons.price and cons.conf). As euribor increases, the employment rate also tends to increase. Additionally, higher index values are associated with a higher likelihood of making a deposit. When the 'emp.var.rate' is negative, customers are more inclined to respond affirmatively. Although this variable is beyond the bank's control, it may signify something significant as it demonstrates a discernible trend. Hence, it warrants inclusion in the model.

The social and economic factors exhibit significant differences concerning whether clients respond positively or negatively. This implies their potential importance for predicting responses. They display diverse patterns and do not distribute evenly. Furthermore, with the exception of consumer confidence, these factors are closely correlated. This suggests that we may only need one of them in our model, as they convey similar information.

IV Feature Engineering

IV.1 Handling Outliers

Statistical outlier detection involves using statistical methods to identify exceptional data points within a dataset. By converting these outliers into **z-scores**, you can determine how many standard deviations they deviate from the mean. When a data point's z-score is significantly high or low, it is typically flagged as an outlier. A commonly used threshold is a z-score exceeding 3 or falling below -3.
To apply statistical outlier detection using z-scores to your data, you can follow these steps :

1. **Calculate Z-Scores :** Compute the z-scores for each data point in your dataset. The z-score of a data point measures how many standard deviations it is away from the mean.

$$Z = \frac{(X - \mu)}{\sigma}$$

   Where :

   X is the value of the data point.

   μ is the mean of the data.

   σ is the standard deviation of the data.

2. **Identify Outliers :** Any data point with a z-score greater than 3 or less than -3 is typically considered an outlier.

In our analysis, we discovered that the DataFrame "outliers" is empty, indicating that no outliers were detected based on the specified threshold (z-score > 3). This implies that the data points in the DataFrame fall within a reasonable range and do not display extreme deviations from the mean. Hence, there are no observations in the dataset that meet the criteria for being considered outliers.
However, it's worth noting that outliers were previously identified in variables such as age, duration, and campaign. To delve deeper into this issue, we will employ the Interquartile Range (IQR) method for further investigation.

IQR Method : The interquartile range (IQR) tells the range of the middle half of your dataset. It can be used to create "fences" around data and then define outliers as any values that fall outside those fences. To detect outliers using the Interquartile Range (IQR) method, follow these steps :

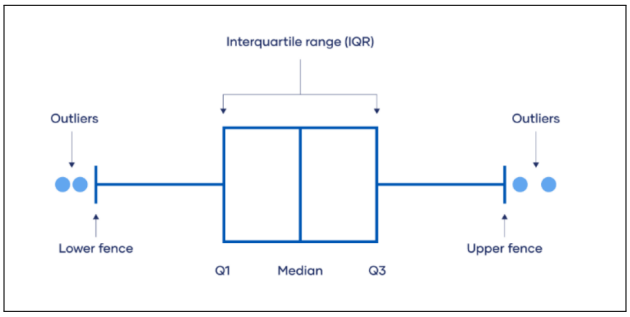


FIGURE 24 – IQR Method

1. Calculate the first quartile (Q1) and third quartile (Q3) of the data.
2. Compute the interquartile range (IQR) as the difference between Q3 and Q1 : $IQR = Q3 - Q1$.
3. Define a threshold to identify outliers. Typically, values that fall below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ are considered outliers.
4. Identify any data points that fall outside of the defined threshold as outliers.

In our case we used this method for the features 'age', 'duration' and 'campaign'.

IV.2 Encoding

Converting Categorical Variables into Numerical Format

Converting categorical variables into a numerical format is an essential preprocessing step in preparing data for machine learning tasks. Categorical data, consisting of non-numeric values like text or categories, must be encoded into numerical representations to ensure compatibility with machine learning algorithms.

Mapping / Ordinal Encoding

This encoding technique converts categorical variables with a natural ordering or hierarchy into a numerical format. For example, months and days of the week have an inherent order, and assigning numerical values to them can capture this order, enabling machine learning algorithms to interpret it.

In this case, we used ordinal encoding because :

- Months and days of the week have a clear order, which can be captured by assigning numerical values.
- It simplifies the feature space by converting categorical variables into numerical ones, making it easier for machine learning algorithms to process.
- It preserves the ordinal relationship between categories, which can be important for certain algorithms to learn patterns effectively.

Mapping for Housing and Loan :

- **Conversion to Binary Representation :** The original categorical variables ('housing' and 'loan') have three categories each ('yes', 'no', and 'unknown'). By mapping 'yes' to 1 and 'no' to 0, we convert them into binary variables, simplifying their interpretation in predictive models.
- **Handling of 'Unknown' Category :** The 'unknown' category represents missing or unknown values. By mapping it to -1, we create a distinct numerical value to represent this category, retaining information about the presence of missing data while converting the variable into numerical format.
- **Simplicity and Clarity :** Mapping provides a simple and clear way to convert categorical variables into numerical ones, making it a practical choice for this encoding task.

Dummy Variables

This encoding technique, also known as one-hot encoding or dummy variable encoding, involves creating binary dummy variables for each category of a categorical variable, where each variable represents the presence (1) or absence (0) of a category.

We chose one-hot encoding for the 'contact' and 'poutcome' variables for these main reasons :

- One-hot encoding eliminates the need for arbitrary numerical assignments to categories, avoiding the imposition of any artificial ordinality on the data.

— It preserves all the information present in the original categorical variables, ensuring that no information is lost during the encoding process.

Frequency Encoding

In frequency encoding, categorical variables are encoded based on the frequency of each category in the dataset.

Here's how it works :

— First, the frequency of each category in the original categorical variables ('job' and 'education') is calculated using the `value_counts()` function.

— The frequencies are then converted into dictionaries (`bank_job` and `bank_ed`) for each variable, where the category names are the keys and their respective frequencies are the values.

— Each category in the original categorical variables is replaced with its corresponding frequency using the `map()` function, transforming the categorical variables ('job' and 'education') into numerical representations based on the frequency of each category.

Frequency encoding is chosen for these reasons :

— Frequency encoding is particularly useful for categorical variables with high cardinality, reducing dimensionality and potentially avoiding overfitting.

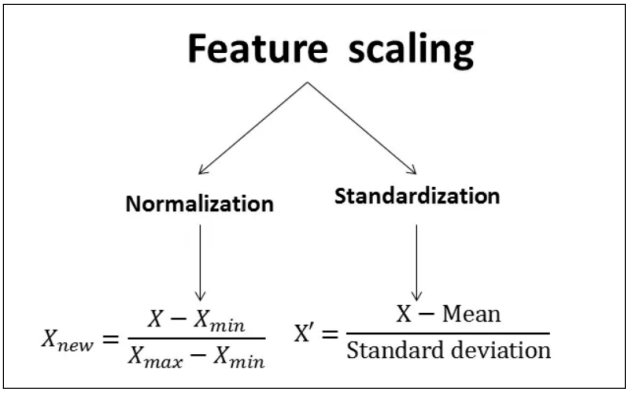
— It helps to retain the information present in the original categorical variables while converting them into numerical format.

Target Guided Ordinal Encoding

In target-guided ordinal encoding, ordinal labels are assigned based on the relationship between the categorical variable and the target variable.

This technique is useful because it can potentially improve the predictive performance of the model. The resulting ordinal labels provide insight into the relationship between marital status and the target variable.

IV.3 Scaling



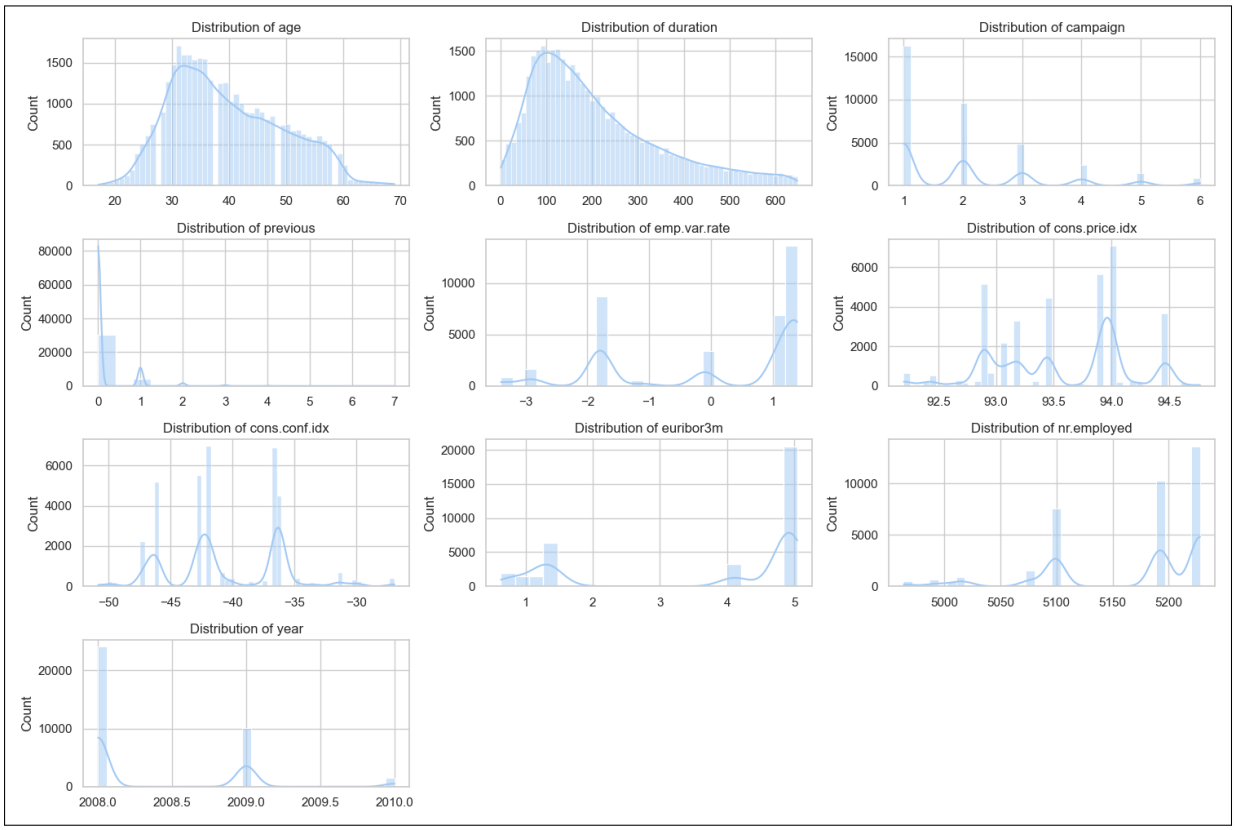


FIGURE 25 – Distributions for Numerical Features

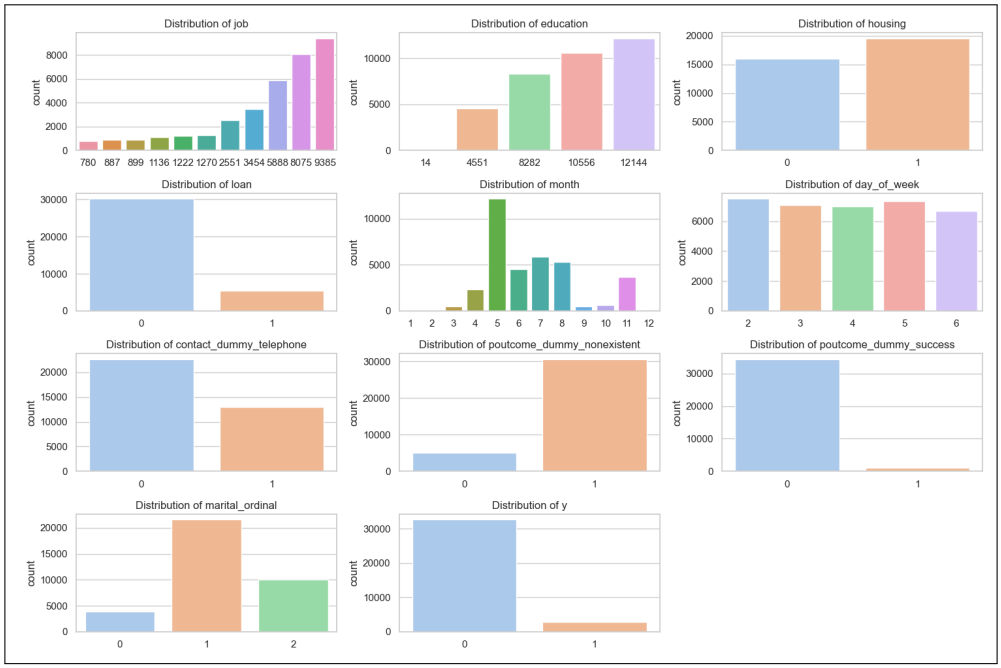


FIGURE 26 – Distributions for Categorical Features

Standardizing Features

Standardizing features is a preprocessing step in machine learning. It ensures scale consistency by transforming features to have a mean of 0 and a standard deviation of 1, facilitating the comparison of their relative importance within the model. For distance-based algorithms like K-Nearest Neighbors and Support Vector Machines, standardizing features is essential to prevent those with larger scales from unduly influencing distance calculations. Lastly, standardization preserves the interpretability of the model by keeping the relationship between features and the target variable unchanged while placing them on a consistent scale for easier coefficient interpretation.

Normalization

Normalization is a method of scaling a variable to a fixed range, typically between 0 and 1. It is used when the scale of a variable is not known or when the variable has a non-uniform distribution. This method helps to bring all the variables on the same scale, making it easier for the model to converge.

Determining if Data is Standardized

To determine if our data is standardized, we need to check the following criteria :

- The mean of the variable should be zero, and the standard deviation should be one.
- The distribution of the standardized data should follow a normal distribution.
- Variables should have a minimum value of -3 and a maximum of 3 if the data is standardized using the z-score method.

There are built-in functions in some libraries like scikit-learn, which can be used to standardize the data. Functions like `StandardScaler()` or `MinMaxScaler()` can be used for this purpose.

In our case, we are utilizing the `StandardScaler()` function to standardize features within our dataset. This process involves removing the mean and scaling the features to have unit variance.

The standard score, or z-score, of a sample (x) is calculated using the formula : $z = \frac{x-\mu}{\sigma}$, where μ represents the mean of the training samples (or zero if `with_mean=False`), and σ represents the standard deviation of the training samples (or one if `with_std=False`).

Standardization involves centering and scaling each feature independently by computing the relevant statistics on the samples in the training set. The mean and standard deviation are then stored to be used for later data transformations using the `transform()` function. This ensures consistency in the scaling process across different datasets or during model deployment. Standardization (or z-score normalization) is often preferred when the features have varying scales and the data does not follow a strict uniform distribution, which is our case. Additionally, certain machine learning algorithms, such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) with radial basis function (RBF) kernels, are sensitive to feature scale. Standardizing the features can help these algorithms perform better by making the feature space more isotropic.

IV.4 Feature Selection

The Extra Trees Classifier is an ensemble learning method that utilizes multiple decision trees to generate predictions. By amalgamating the outcomes of numerous weak learners, the Extra Trees Classifier can offer more resilient and consistent feature importance estimations. This helps mitigate the possibility of selecting features that are only deemed important by happenstance. The Extra Trees Classifier is a po-

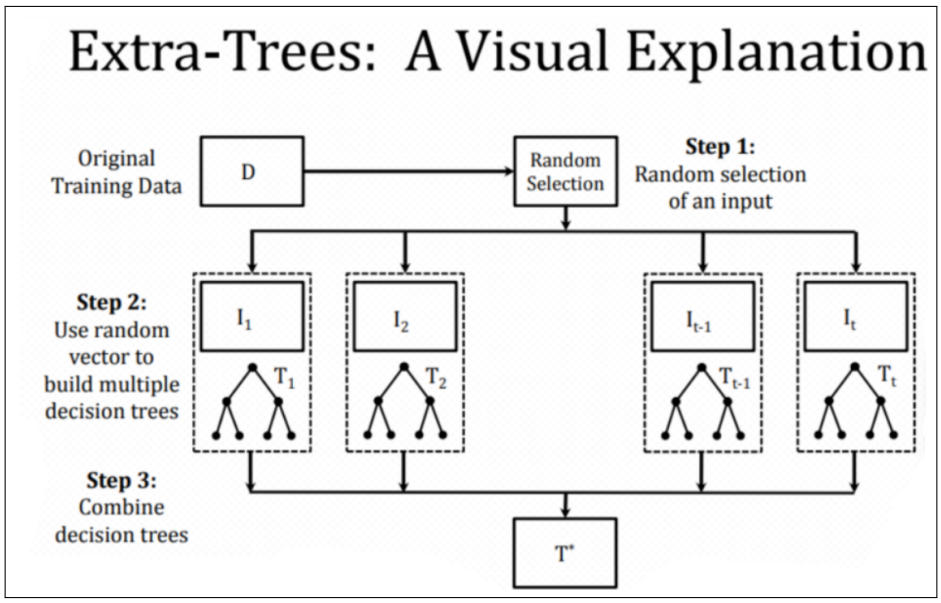


FIGURE 27 – Extra Trees Classifier

werful tool for feature selection, offering several advantages. It excels in robustness, as it is less sensitive to noise and irrelevant features thanks to its use of multiple decision trees and feature importance scores. Additionally, it boasts computational efficiency, making it suitable for large datasets. The classifier also helps reduce bias by employing random selection of subsets and splitting points. It provides feature ranking based on importance scores, allowing for informed decision-making in feature selection. Moreover, it handles multicollinearity effectively by utilizing random selection and splits. Its flexibility enables adaptation of feature inclusion thresholds as per specific requirements. Finally, the Extra Trees Classifier enhances model generalization and provides interpretable insights into influential features, thereby aiding in better understanding and interpretation of the predictive model.

Applying this to our data we obtain : The bar plot showcases the significance of each feature in influen-

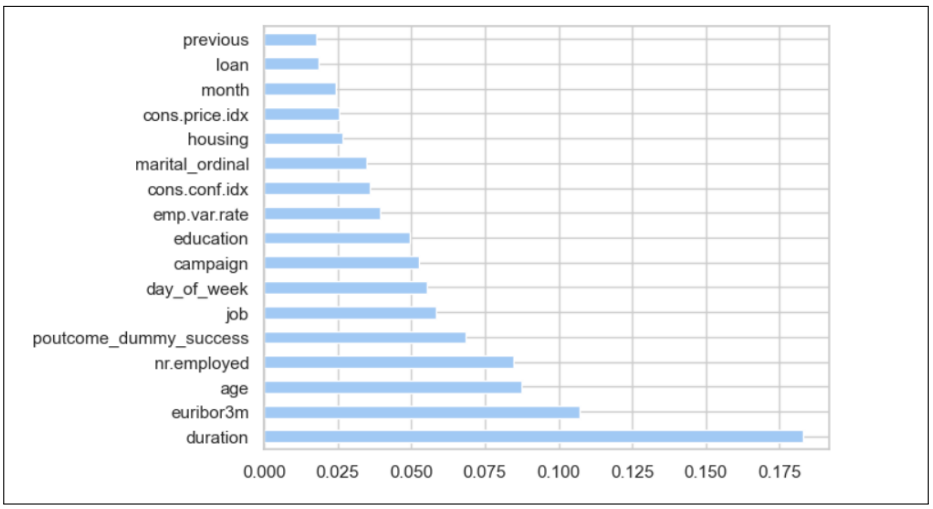


FIGURE 28 – Feature Selection

cing the output, with emphasis on the top 15 features.

V Modeling Data

After partitioning the dataset into training and testing sets, we observed that the training set consists of 28,437 samples with 15 input features, while the test set comprises 7,110 samples. The target variable, representing whether a deposit is made, is binary, with values of either 0 or 1. This binary nature indicates the suitability of classification models for our predictive task, as they are designed to handle such categorical outcomes effectively. Our goal is to utilize these models to forecast deposit outcomes based on the provided feature values.

V.1 Model Selection

To identify the optimal classification models, we utilize cross-validation instead of assessing multiple models individually. Cross-validation partitions the dataset systematically, training the model on one subset and evaluating its performance on another. Repeating this process with different partitions yields multiple accuracy estimates, enabling comprehensive model performance comparisons. By selecting the model with the highest accuracy across these iterations, we ensure robustness and reliability in our model selection process. This approach optimizes decision-making and resource allocation, enhancing the effectiveness of our predictive modeling efforts.

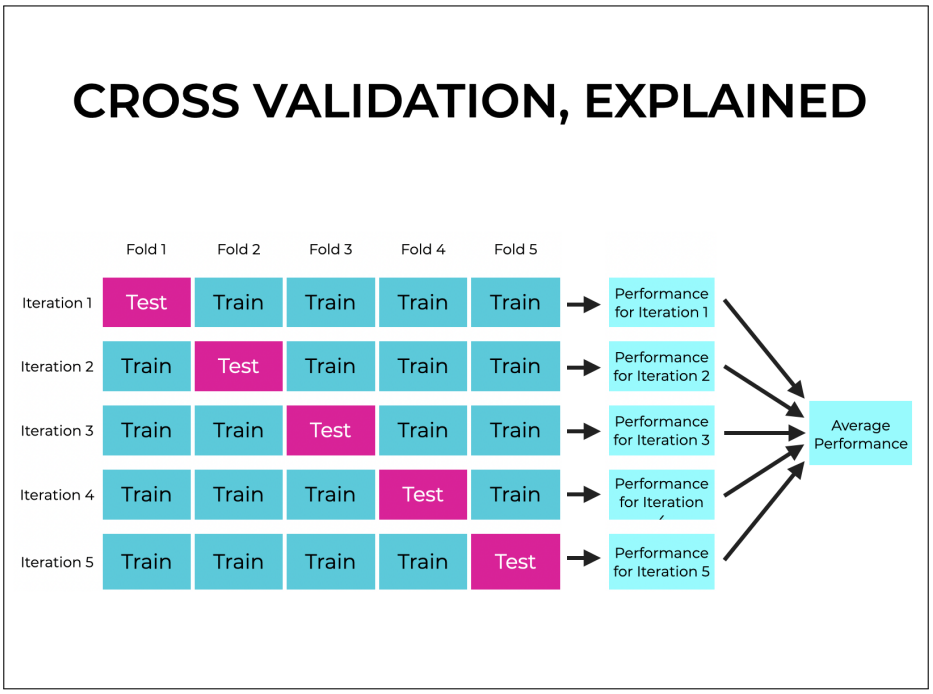


FIGURE 29 – Cross Validation

Our data is imbalanced. This can be managed through resampling techniques or by utilizing models specifically designed to handle such scenarios. Here are some machine learning models adept at handling imbalanced data in classification tasks :

- **Random Forest :** An ensemble learning method that combines multiple decision trees, Random Forest excels at capturing intricate relationships within imbalanced datasets.
- **Gradient Boosting Machines (GBM) :** GBM sequentially builds trees, focusing on correcting previous mistakes, making it well-suited for improving performance on both classes in imbalanced data.
- **Support Vector Machines (SVM) :** SVM, a robust classification algorithm, can be tailored to handle imbalanced data by adjusting class weights or employing cost-sensitive learning methods.
- **AdaBoost :** This ensemble learning method assigns higher weights to misclassified minority class instances, effectively addressing imbalanced data challenges.
- **XGBoost :** Renowned for its speed and performance, XGBoost adapts to class imbalance by adjusting learning objectives or incorporating class weights.
- **Logistic Regression with Class Weights :** Logistic Regression, a straightforward yet powerful algorithm, effectively manages imbalanced data by adjusting class weights or employing penalized techniques.
- **Nearest Neighbors with Weighted Voting :** k-Nearest Neighbors (kNN) can be adapted for imbalanced data by utilizing weighted voting, where closer neighbors contribute more significantly to classification decisions.

Based on the test results in the notebook, we observe that the Support Vector Classifier (SVC) demonstrates the highest accuracy, followed by Logistic Regression. Let's explore them deeper.

V.1.1 Logistic Regression

Hyperparameter tuning

We'll train the logistic regression model with optimized parameters and evaluate its accuracy to gauge its performance. We perform hyperparameter tuning for a logistic regression model using grid search with cross-validation.

Hyperparameters :

Hyperparameters are external settings chosen by the user before training the model. They influence the model's behavior and performance but are not learned from the data. Techniques like grid search can help optimize hyperparameters by evaluating different combinations to enhance model performance.

Grid search :

To find the best hyperparameters for optimal model performance, methods like Grid Search offer systematic exploration of different combinations. Unlike manual search, Grid Search evaluates various hyperparameter values exhaustively. By performing cross-validation, it assesses model performance across different parameter settings. However, this exhaustive approach can be time-consuming and resource-intensive as the number of hyperparameters increases.

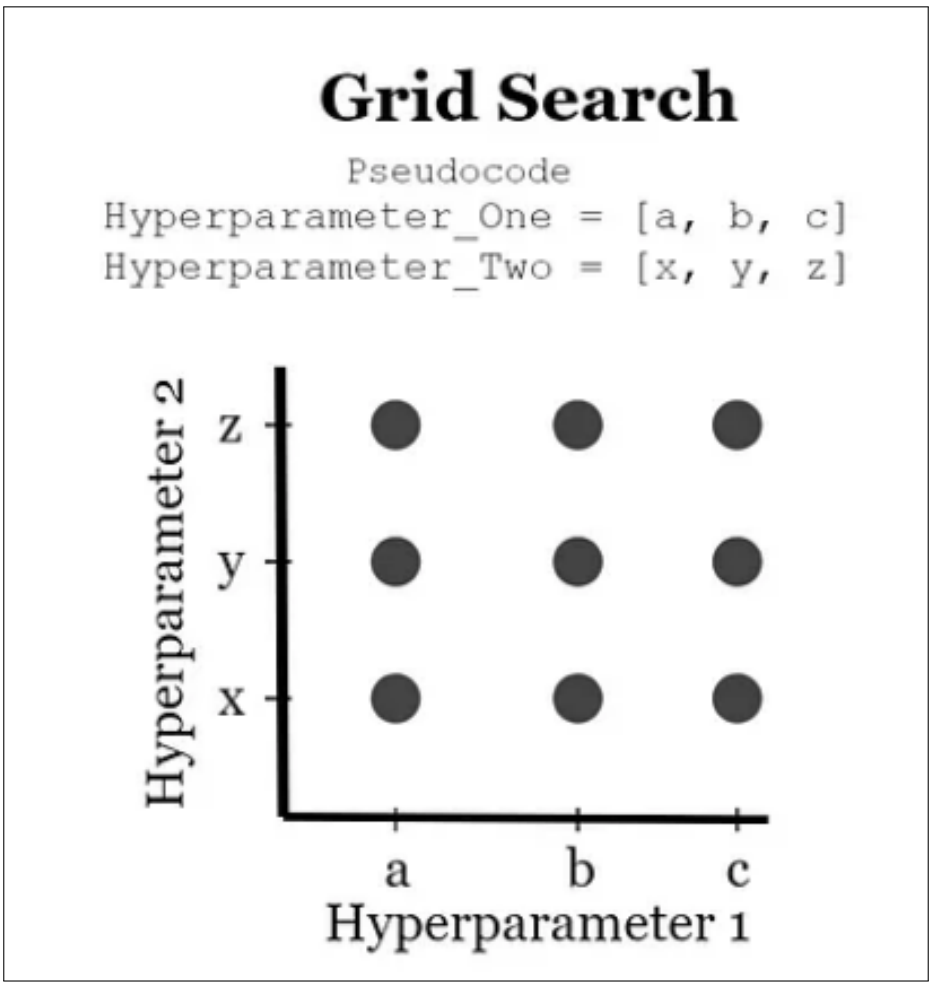


FIGURE 30 – Grid Search

We've obtained the optimal parameters for the model, achieving a mean accuracy of 93%. We obtain the following confusion matrix : The classification report provides a comprehensive assessment of the

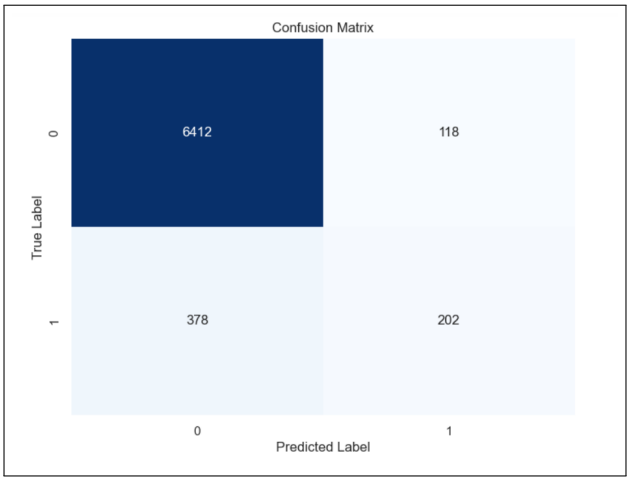


FIGURE 31 – Confusion Matrix

model's performance across various metrics. The precision, representing the accuracy of positive and negative predictions, indicates a high precision of 94% for the negative class, ensuring that instances labeled as negative are accurate. However, the precision drops to 63% for the positive class, suggesting some misclassification of positive instances. The recall metric highlights the model's ability to correctly identify actual positive and negative instances, with a high recall of 98% for the negative class but a lower recall of 35% for the positive class. The F1-score, balancing precision and recall, demonstrates a favorable balance for the negative class (96%) but a lower score of 45% for the positive class, indicating a trade-off between precision and recall. The overall accuracy of the model is 93%, with macro and weighted averages reflecting performance across both classes. Despite strong performance in predicting the negative class, the model struggles with the positive class, particularly in terms of recall and F1-score.

ROC Curve :

A Receiver Operating Characteristic (ROC) curve visually represents a binary classifier's diagnostic ability. It plots the true positive rate (TPR) against the false positive rate (FPR). TPR indicates the proportion of correct positive predictions, while FPR represents the proportion of incorrect positive predictions. In medical testing, TPR reflects the correct identification of individuals with a disease. For probabilistic classifiers, the ROC curve is generated by adjusting the score threshold, allowing us to assess the model's performance across various thresholds.

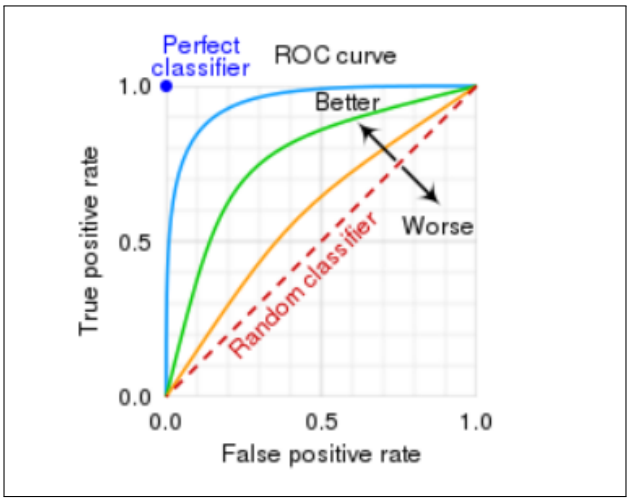


FIGURE 32 – ROC curve

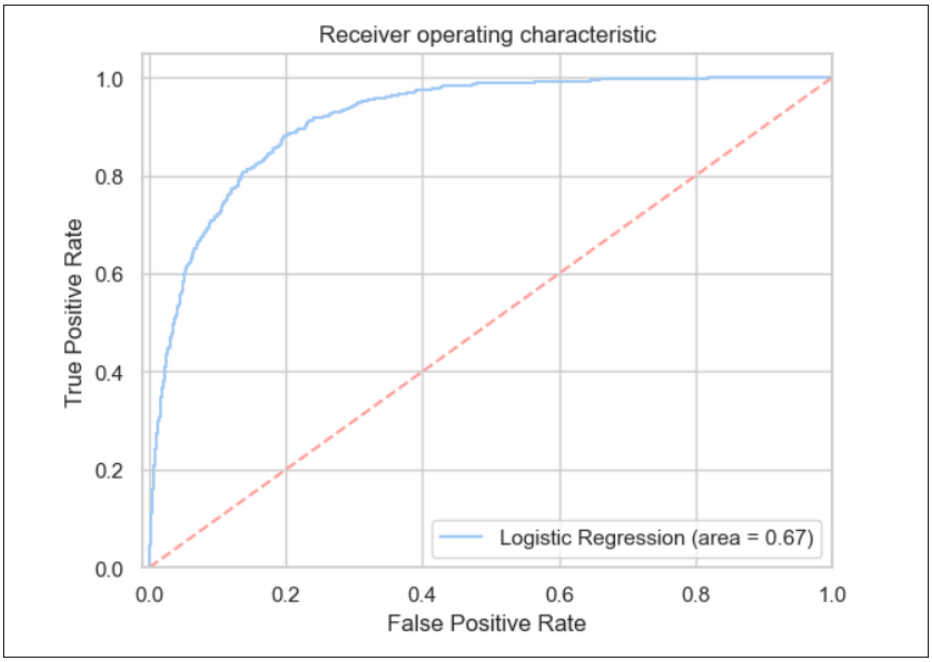


FIGURE 33 – ROC curve for Logistic Regression

The ROC curve depicted in Figure 33 provides insight into the logistic model's ability to discriminate between true positives and false positives. A curve positioned closer to the top-left corner signifies superior model performance. In our analysis, selecting a threshold value between 0.8 and 0.9 can optimize true positive outcomes, balancing sensitivity and specificity effectively.

V.1.2 Support Vector Classifier

Accuracy : 0.9184247538677919

Class	Precision	Recall	F1-Score	Support
0	0.92	1.00	0.96	6530
1	0.00	0.00	0.00	580

The accuracy of the Support Vector Classifier (SVC) model is approximately 91.84%. However, the model performs poorly in predicting the minority class (class 1), as indicated by the low precision, recall, and F1-score for class 1. This suggests that the model struggles to correctly identify instances of the minority class, resulting in a low F1-score overall.

Conclusion

After an exploration of the dataset, we identified key features that significantly influence the likelihood of making a deposit. Our analysis highlighted the importance of factors such as duration of calls, employment variation rate, and the number of contacts during the campaign in predicting subscription outcomes. Specifically, we observed that longer call durations and lower employment variation rates were associated with higher probabilities of making a deposit. Furthermore, our examination of categorical variables revealed intriguing insights. We found that certain demographic factors, such as marital status and education level, played a crucial role in subscription outcomes. For instance, married individuals were more engaged with the campaign, while higher education levels were positively correlated with subscription rates. In terms of modeling, we experimented with various classification algorithms, including Support Vector Classifier (SVC), Logistic Regression, and Random Forest. While SVC demonstrated the highest overall accuracy, it struggled with predicting the minority class accurately. Logistic Regression, on the other hand, provided a more balanced performance across both classes, albeit with slightly lower accuracy. Overall, our analysis underscores the importance of considering the interplay between different features and their impact on subscription outcomes. Going forward, utilizing these insights to fine-tune feature engineering techniques and select the most suitable models will be imperative for enhancing predictive accuracy and achieving favorable campaign results.