**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Strategic Thinking* |
| **Assessment Title:** | CA 3 Final Submission |
| **Lecturer Name:** | James Garza |
| **Student Full Name:** | Federico Ariton |
| **Student Number:** | sba22090 |
| **Assessment Due Date:** | *10/05/2024* |
| **Date of Submission:** | *10/05/2024* |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Contenido

[**Introduction** 2](#_Toc165844778)

[Objective 2](#_Toc165844779)

[**Business Description** 2](#_Toc165844780)

[**Hypothesis** 2](#_Toc165844781)

[Age 2](#_Toc165844782)

[Education Level 3](#_Toc165844783)

[**General Goal** 3](#_Toc165844784)

[**Success Criteria/Indicators** 3](#_Toc165844785)

[**Structure of the project** 4](#_Toc165844786)

[**Models and machine learning algorithms** 4](#_Toc165844787)

[**Libraries** 4](#_Toc165844788)

[**Dataset** 5](#_Toc165844789)

[**Attributes** 5](#_Toc165844790)

[**Dimensions** 5](#_Toc165844791)

[**Source** 5](#_Toc165844792)

[**Descriptive statistics** 6](#_Toc165844793)

[**Duration:** 6](#_Toc165844794)

[**Pdays:** 6](#_Toc165844795)

[**Emp.var.rate:** 6](#_Toc165844796)

[**Cons.price.idx:** 6](#_Toc165844797)

[**Cons.conf.idx:** 6](#_Toc165844798)

[**Euribor3m:** 6](#_Toc165844799)

[**Nr.employed:** 6](#_Toc165844800)

[**Y:** 6](#_Toc165844801)

[**Findings** 6](#_Toc165844802)

[**Data visualization** 7](#_Toc165844803)

[**EDA** 7](#_Toc165844804)

[**Histogram** 7](#_Toc165844805)

[**Age** 8](#_Toc165844806)

[**Duration** 8](#_Toc165844807)

[**Campaign** 8](#_Toc165844808)

[**Pdays** 8](#_Toc165844809)

[**Previous** 8](#_Toc165844810)

[**Economic Indicators (emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed)** 8](#_Toc165844811)

[**Bar chart** 8](#_Toc165844812)

[**Job** 9](#_Toc165844813)

[**Marital** 9](#_Toc165844814)

[**Education** 9](#_Toc165844815)

[**Default** 9](#_Toc165844816)

[**Housing** 9](#_Toc165844817)

[**Loan** 9](#_Toc165844818)

[**Contact** 9](#_Toc165844819)

[**Month** 9](#_Toc165844820)

[**Day of Week** 10](#_Toc165844821)

[**Poutcome** 10](#_Toc165844822)

[**Target Variable (y)** 10](#_Toc165844823)

[Insights and Observations 10](#_Toc165844824)

[**Age vs y** 11](#_Toc165844825)

[**Duration vs y** 11](#_Toc165844826)

[**Campaign vs y** 12](#_Toc165844827)

[**Pdays vs y** 12](#_Toc165844828)

[**Previous vs y** 12](#_Toc165844829)

[**Emp.var.rate vs y** 12](#_Toc165844830)

[**Cons.conf.idx vs y** 12](#_Toc165844831)

[**Euribor3m vs y** 12](#_Toc165844832)

[**Nr.employed vs. y** 12](#_Toc165844833)

[**Summary of Observations** 12](#_Toc165844834)

[**Subscription Trend** 12](#_Toc165844835)

[**Age Influence** 12](#_Toc165844836)

[**Economic Indicators** 12](#_Toc165844837)

[**Past Campaigns** 12](#_Toc165844838)

[**Data preparation** 13](#_Toc165844839)

[**Machine Learning Algorithm.** 22](#_Toc165844840)

[Logistic Regression 22](#_Toc165844841)

[KNN 23](#_Toc165844842)

[Decision tree 24](#_Toc165844843)

[Gaussian Naïve Bayes 25](#_Toc165844844)

[Model Comparison 26](#_Toc165844845)

[**Conclusion and next steps** 28](#_Toc165844846)

[**Reference** 29](#_Toc165844847)

[**Data Dictionary** 30](#_Toc165844848)

# **Introduction**

**In today's competitive banking environment, effective marketing strategies are crucial for attracting and retaining customers** **(www.driveresearch.com, n.d.). One of the prominent methods used by banks is direct marketing through phone calls. This report focuses on the analysis of data from direct marketing campaigns conducted by a Portuguese banking institution. The goal is to predict whether a customer will subscribe to a term deposit (denoted by the variable y) based on a range of demographic, social-economic, and campaign-related features.**

# Objective

Develop a predictive model capable of forecasting customer behavior concerning term deposit subscriptions. This model aims to accurately classify whether a customer will subscribe to a term deposit based on their demographic profile and interaction history with the bank

# **Business Description**

**The dataset originates from a bank's marketing department, which has been conducting direct marketing campaigns to promote term deposit products. Term deposits are a critical product for banks as they represent a stable funding source** **(Investopedia, n.d.). The marketing strategies employed typically involve direct phone calls to existing customers, offering them the opportunity to invest in term deposits at competitive interest rates.**

# **Hypothesis**

## Age

**Clients in older age groups are more likely to subscribe to a term deposit due to their propensity for long-term financial planning.**

**Null Hypothesis (H0): There is no significant difference in term deposit subscription rates between different age groups.**

**Alternative Hypothesis (H1): Older age groups have a significantly higher rate of term deposit subscription compared to younger age groups.**

## Education Level

**Clients with higher education levels are more inclined to subscribe to term deposits as they may have a better understanding of financial products.**

**Null Hypothesis (H0): There is no significant difference in term deposit subscription rates across different education levels.**

**Alternative Hypothesis (H1): Clients with higher education levels have a significantly higher rate of term deposit subscription compared to those with lower education levels.**

# **General Goal**

**The general goal of this project is to:**

**Analyze the data to identify the key factors that influence a customer’s decision to subscribe to a bank term deposit.**

**Develop a predictive model that can accurately forecast whether a customer will subscribe to a term deposit based on their profile and economic context.**

# **Success Criteria/Indicators**

**Success for this project will be measured by:**

**The ability to identify and quantify the impact of various factors on the decision to subscribe to a term deposit.**

**The accuracy and predictive performance of the model, as measured by metrics such as the Area Under the Receiver Operating Characteristic, accuracy, precision, recall, and F1-score.**

**Practical recommendations for the bank’s marketing team on how to tailor their campaigns based on the findings, potentially leading to higher conversion rates and more efficient use of marketing resources.**

# **Structure of the project**

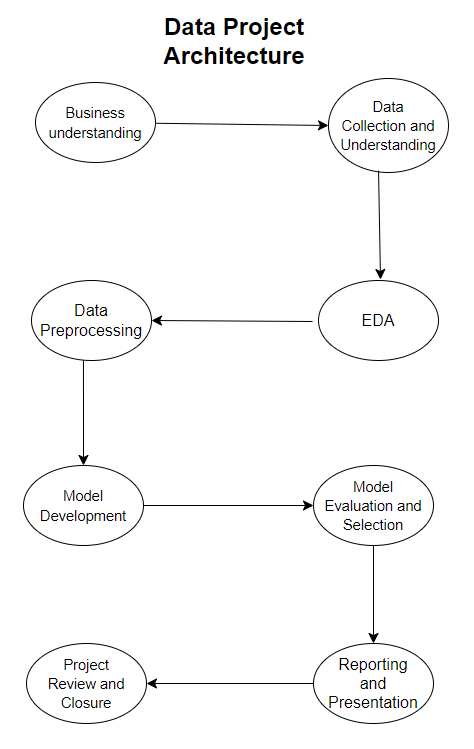


Figure 1 Data project phases

# **Models and machine learning algorithms**

**Machine learning modeling, a variety of classification algorithms were implemented from the Scikit-learn library, including RandomForestClassifier, KNeighborsClassifier, SVC (Support Vector Classifier), GaussianNB (Naive Bayes), DecisionTreeClassifier, and Logistic Regression. Additionally, cross\_val\_score helped in assessing the models' performance through cross-validation. Model evaluation was carried out using metrics like confusion\_matrix, accuracy\_score, f1\_score, and classification\_report to obtain a comprehensive understanding of model effectiveness.**

# **Libraries**

**Different Python libraries were used to facilitate data manipulation, visualization, and machine learning model development. Pandas and NumPy were employed for data handling and numerical operations, while Matplotlib and Seaborn were utilized for creating insightful visualizations. To suppress unnecessary warnings and maintain a clean workflow, we used the warnings library.**

# **Dataset**

**The data for the bank marketing campaign analysis has been used for the study, containing 21 variables (or columns) and 41,188 observations (or rows). Out of these, 7 variables are numerical, while 14 variables are categorical. The dataset can help identify the effectiveness of various marketing campaigns, based on the number of contacts and the outcome of previous campaigns. By analyzing categorical variables like job, marital status, education, and loan history, we can profile customers who are more likely to subscribe to a term deposit.**

# **Attributes**

**In machine learning, attributes (or features) are the input variables used to predict the target variable. In this dataset, the target variable is "y", which represents whether the client has subscribed to a term deposit. The other columns are the attributes that can be used as predictors**

# **Dimensions**

**The dataset contains 41,188 entries with 21 features, including customer demographics, economic indicators, and campaign details. Key features include age, job, marital status, education, housing, default, loan, contact, month,day of the week, duration, campaign, Pdays, previous, Poutcome, Emp.var.rate, cons.price.idx, cons.confi.idx, Euribor3m, Nr.emplyed, Y.**

# **Source**

**The dataset was obtained from the UC Irvine Machine Learning Repository, a online resource that hosts a variety of datasets commonly used for academic and research purposes in machine learning.**

**https://archive.ics.uci.edu/dataset/222/bank+marketing**

# **Descriptive statistics**

**In our dataset, while we are familiar with the rows and columns, it's crucial to also understand the significance of each column. Some columns have obvious meanings, but others can be more complex and require detailed analysis. Let's analysis to the complexities columns for have a better understand of our data:**

**Duration: The duration of the last contact, in seconds**

**Pdays: The number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted).**

**Emp.var.rate: Employment variation rate ,quarterly indicator this rate indicate the influence customers financial stability.**

**Cons.price.idx: Consumer price index(CPI) ,monthly indicator that measures the average change over time in the prices paid by consumers for a market basket of consumer goods and services. (www.cso.ie, n.d.)**

**Cons.conf.idx: Consumer confidence index(CCI) ,monthly indicator that reflects the degree of confidence individual households have in the performance of the economy. (Investopedia, n.d.)**

**Euribor3m: Euribor 3 month rate ,daily indicator is the interest rate that refers to various financial products, including mortgages and savings accounts.**

**“Euribor is the acronym for the Euro Interbank Offered Rate. This is the interest rate at which credit institutions lend money to each other, which is often referred to as “the price of money” (Bankinter, n.d.).**

**Nr.employed: Number of employees.**

**Y: The target variable indicating whether the client has subscribed to a term deposit (binary: 'yes','no').**

## **Findings**

**Longer call durations and recent previous contact are associated with higher subscription rates. Positive employment variation rates, higher consumer price and confidence indices, and lower Euribor 3-month rates also correlate with increased likelihood of subscription. Additionally, higher employment levels are linked to more subscriptions. However, the significant imbalance in the target variable necessitates data balancing techniques like SMOTE for accurate predictive modeling.**

# **Data visualization**

## **EDA**

**Proceeding to create the next steps for we can understand the data:**

**- Histograms for numerical variables to understand their distribution.**

**- Bar charts for categorical variables to see the distribution of different categories.**

**- A correlation matrix to observe any potential correlations between numerical variables.**

## **Histogram**

**The histograms provide (matplotlib.org, n.d.) insights into the distribution of the numerical variables:**

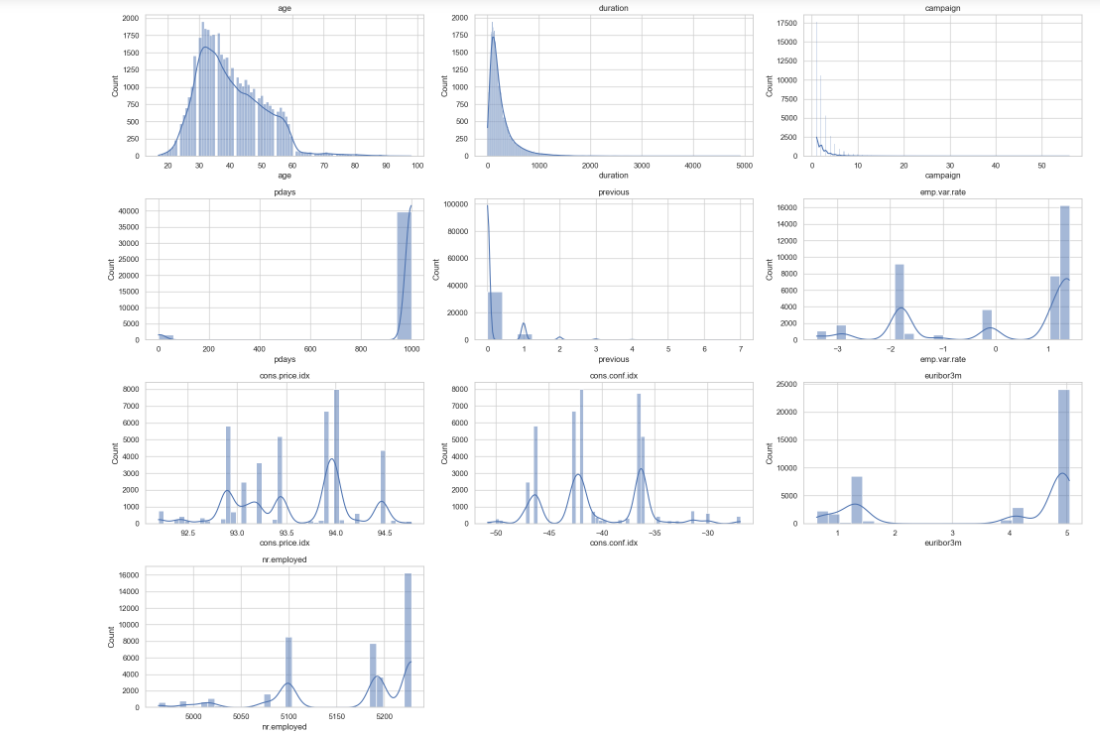
****

Figure 2 Histogram of the distribution of the numeric values

**Age: Most customers are in the 30-40 age range, with a right-skewed distribution.**

**Duration: This shows a right-skewed distribution, with most calls being relatively short.**

**Campaign: Most customers were contacted a few times, with the distribution heavily skewed to the right.**

**Pdays: There's a spike at 999, which likely represents a 'not contacted' category. This needs to be considered in any analysis.**

**Previous: Similar to pdays, most customers were not previously contacted, as indicated by the peak at 0.**

**Economic Indicators (emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed): These show various distributions, some are multimodal, reflecting different economic conditions during the data collection period.**

## **Bar chart**

**Next, let's plot bar charts for the key categorical variables to understand their distribution. We will focus on 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day\_of\_week', 'poutcome', and the target variable 'y'.**

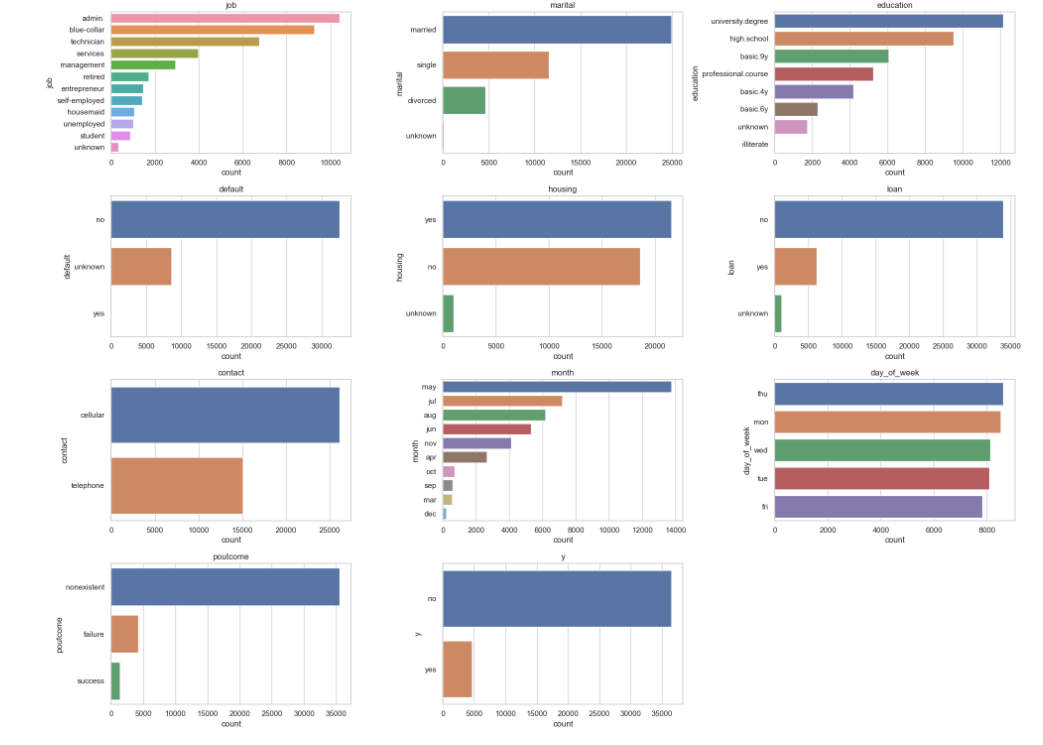
****

Figure 3 bar chart of the distribution of the categorical variables

**The bar charts provide a visual representation of the distribution of categorical variables:**

**Job: The dataset includes a variety of job categories, with 'admin.', 'blue-collar', and 'technician' being the most common.**

**Marital: Most customers are married, followed by single and divorced.**

**Education: The highest frequency is for university degree, followed by high school education.**

**Default: Majority of customers are recorded as 'no' for default, with a significant portion marked as 'unknown'.**

**Housing: A fairly even distribution among 'yes', 'no', and 'unknown'.**

**Loan: Most customers do not have a personal loan.**

**Contact: 'Cellular' is the most common method of contact.**

**Month: There's a clear seasonality in contact, with May being the most common month.**

**Day of Week: Distribution is relatively even across different days of the week.**

**Poutcome: Majority of outcomes from the previous marketing campaign are 'nonexistent', indicating no prior contact.**

**Target Variable (y): There are significantly more 'no' responses than 'yes', indicating a lower subscription rate.**

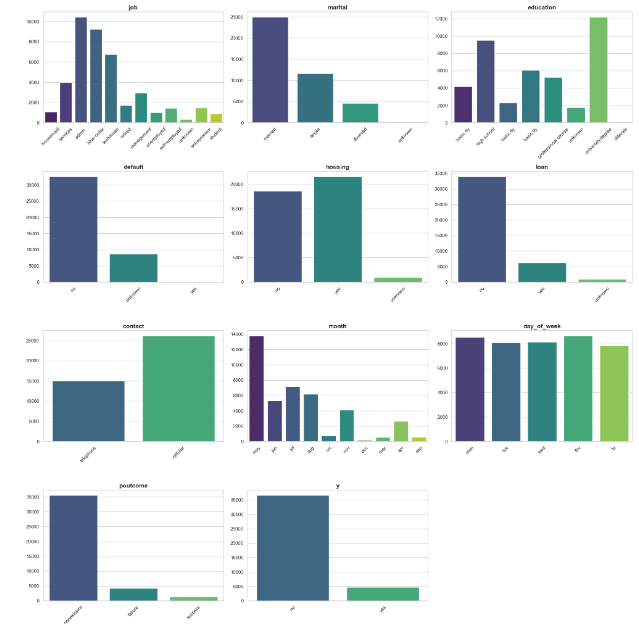
****

Figure 4 New bar chart of the distribution of the categorical variables

Insights and Observations**:**

**The dataset shows significant class imbalance in the target variable, which will need to be addressed in modeling.**

**There are clear differences in customer characteristics across job, education, and marital status, highlighting the need for tailored marketing strategies.**

**Seasonal effects and campaign timing appear to play a crucial role, especially given the high contact rates in May and June.**

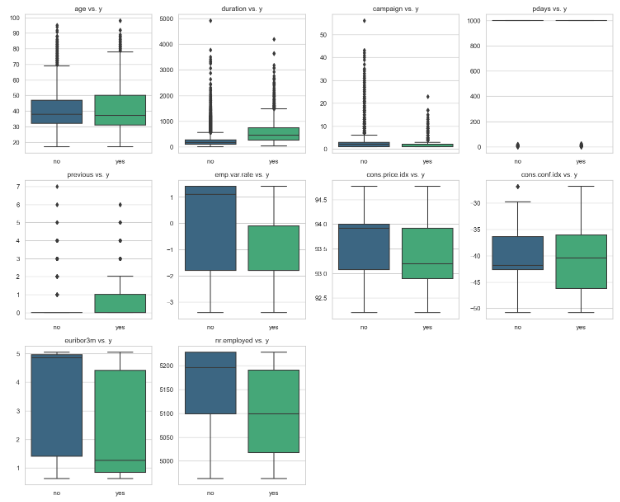
****

Figure 5 Box plot relationship between the numerical and the target variable y

**This** **set of plots illustrates the relationship between the numerical features in the bank marketing campaign dataset and the target variable "y" (whether the client subscribed to a term deposit). Here's an analysis of each plot:**

**Age vs y: clients who subscribed to a term deposit (y = 'yes') appear to be older on average compared to those who did not subscribe (y = 'no').**

**There is a broad age range in both groups, with significant overlap.**

**Duration vs y: clients with longer call durations are more likely to subscribe to a term deposit.**

**Calls that resulted in a subscription ('yes') tend to have much longer durations than those that did not ('no').**

**Campaign vs y: the number of contacts in the current campaign shows that clients who subscribed generally had fewer contacts.**

**The majority of clients who did not subscribe were contacted more than once.**

**Pdays vs y: most clients who subscribed have pdays values of -1, indicating that they were not previously contacted in earlier campaigns.**

**Higher pdays values are rare, suggesting that past campaigns did not heavily target the clients.**

**Previous vs y :clients who subscribed have slightly higher median values for the number of previous contacts.**

**This suggests that clients who were contacted previously are more likely to subscribe.**

**Emp.var.rate vs y: clients who subscribed tend to have been contacted during periods of negative employment variation rates.**

**The median value is higher (more positive) for clients who did not subscribe.**

**Cons.conf.idx vs y: The consumer confidence index does not show a significant difference between clients who subscribed and those who did not.**

**However, those who did subscribe generally have higher median confidence index values.**

**Euribor3m vs y: Clients who subscribed generally have lower Euribor 3-month rates compared to those who did not.**

**There is a significant difference in the median Euribor values between the two groups.**

**Nr.employed vs. y : clients who subscribed tend to have been contacted when the number of employees was relatively lower.**

**The median value for the 'yes' group is significantly lower than the 'no' group.**

**Summary of Observations:**

**Subscription Trend: Clients with longer call durations and fewer campaign contacts are more likely to subscribe to a term deposit.**

**Age Influence: Older clients tend to subscribe more frequently than younger clients.**

**Economic Indicators: Economic factors such as employment variation rate, consumer price index, and Euribor rate significantly influence the likelihood of subscription.**

**Past Campaigns: Clients contacted in past campaigns are more likely to subscribe, indicating that familiarity and past positive experiences impact subscription rates.**

# **Data preparation**

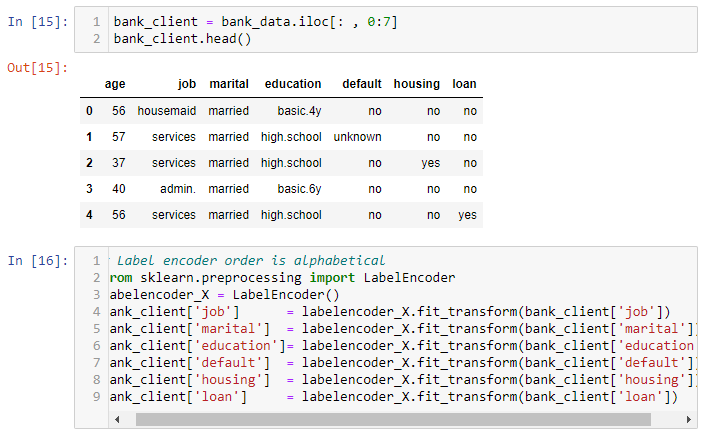
****

Figure 6 Label encoder for the first 7 attributes

**The initial data preparation involved subsetting the dataset to include only the first seven client-related columns: age, job, marital, education, default, housing, and loan. Each categorical feature (job, marital, education, default, housing, and loan) was then encoded into numerical format using Scikit-learn's LabelEncoder. For instance, the job categories like 'admin.' and 'technician' were converted to numerical labels such as 0 and 8, respectively. This encoding ensured that all features were in numerical form and suitable for machine learning model development.**

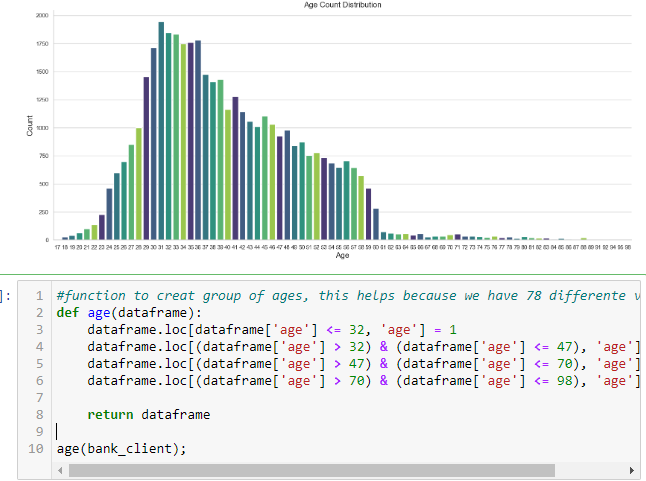
****

Figure 7 bar plot of age and encoding

**A bar plot was generated to visualize the distribution of customer ages using the Seaborn library. The plot, created with a specified color palette ("viridis") and custom figure size (20x8 inches), illustrates the age count distribution among the bank clients. The x-axis represents customer ages, while the y-axis shows the count of clients in each age group.**

**After examining the plot, the age distribution was categorized into four distinct groups using a custom age() function:**

**Group 1: Clients aged 32 and below**

**Group 2: Clients aged between 33 and 47**

**Group 3: Clients aged between 48 and 70**

**Group 4: Clients aged between 71 and 98**

**The function replaces the actual age values with their corresponding group numbers, simplifying the analysis by reducing the number of unique age values.**

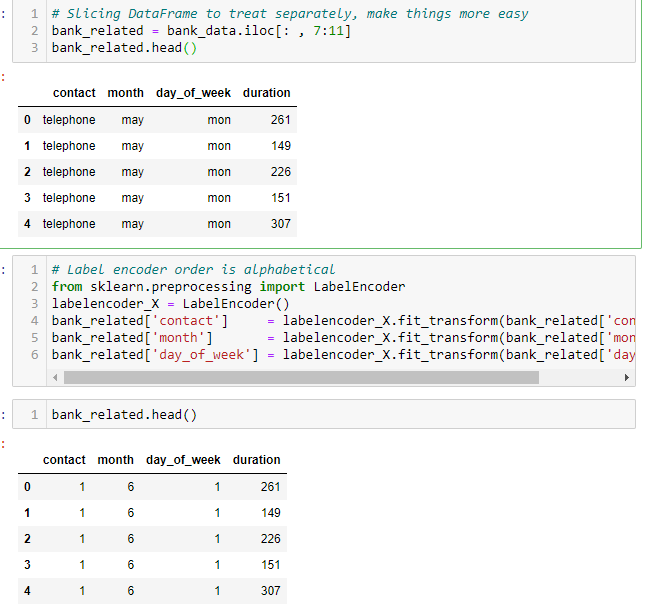
****

Figure 8 Encoding the last contact

**A subset of features from the bank marketing dataset related to communication methods and campaign timing was extracted into a separate DataFrame named bank\_related. This subset includes the features contact, month, day\_of\_week, and duration Contact: Converts communication types such as 'cellular' and 'telephone' into numerical labels.**

**Month: Converts the last contact month (e.g., 'jan', 'may') into corresponding numerical labels.**

**Day\_of\_week: Converts the last contact day of the week (e.g., 'mon', 'fri') into numerical labels.**

****

Figure 9 Visualization of duration and encoding

**To analyze call durations in the bank marketing dataset, a box plot and histogram were created, showing that the distribution is right-skewed with many outliers. To simplify analysis, the call durations were grouped into five categories**

**Group 1: Calls with a duration of 102 seconds or less.**

**Group 2: Calls with a duration between 102 and 180 seconds.**

**Group 3: Calls with a duration between 180 and 319 seconds.**

**Group 4: Calls with a duration between 319 and 644.5 seconds.**

**Group 5: Calls longer than 644.5 seconds.**

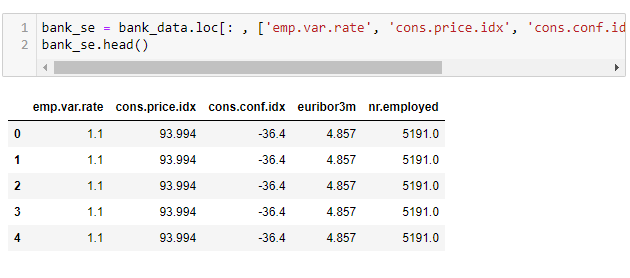
****

Figure 10 Storing the columns in a variable

**We stored a subset of economic features in the variable bank\_se, which includes emp.var.rate (employment variation rate), cons.price.idx (consumer price index), cons.conf.idx (consumer confidence index), euribor3m (Euribor 3-month rate), and nr.employed (number of employees). No label encoding was required as these features are already in numerical format.**

****

Figure 11 Encoding the columns related to marketing campaigns

**We stored features related to previous marketing campaigns in the variable bank\_o, including campaign, pdays, previous, and poutcome. The poutcome feature was encoded to numerical values: 'nonexistent' (1), 'failure' (2), and 'success' (3).**

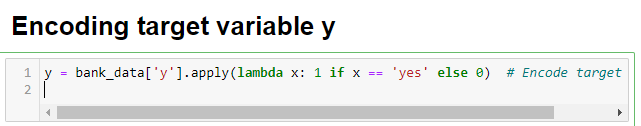
****

Figure 12 Encoding the target variable y

**The target variable y was encoded into numerical format, where 'yes' was mapped to 1 and 'no' to 0. This binary encoding allows for effective classification modeling.**

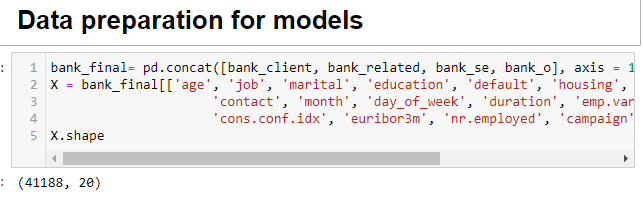
****

Figure 13 Combining all the variable in a feature

**We combined all relevant DataFrames (bank\_client, bank\_related, bank\_se, and bank\_o) into a single DataFrame, bank\_final, to ensure that all necessary predictor variables were included in the final feature set X. This comprehensive feature set provides a unified basis for building predictive models effectively.**

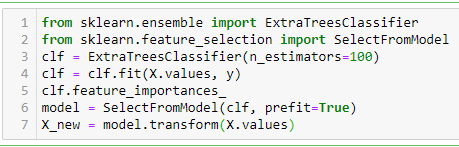
****

Figure 14 Feature Engineering

**We used an ExtraTreesClassifier** (scikit-learn.org, n.d.) **to identify and select the most important features for predicting term deposit subscriptions. This process is crucial because it reduces the feature set to only the most relevant predictors, improving model performance and interpretability while reducing overfitting risk. The selected features were stored in a new feature set, X\_new.**

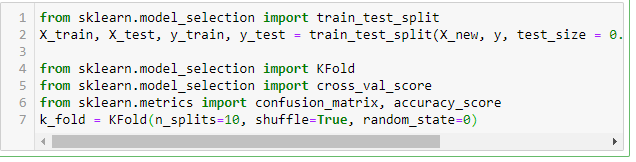
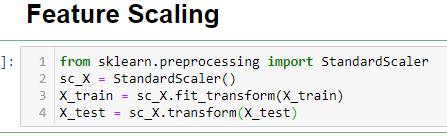
****

Figure 15 Splitting the data set

**We split the dataset into training and test sets using train\_test\_split, with 80% of the data used for training and 20% for testing. The stratify parameter (scikit-learn, 2018) ensures that the target variable, y, is proportionally represented in both sets. Additionally, a 10-fold cross-validation scheme was set up with KFold to evaluate model performance more robustly.**

****

**To ensure all features are on a comparable scale, we standardized the training and test sets using StandardScaler. This transformation centers the data around zero with unit variance, which can improve the performance of many machine learning algorithms.**

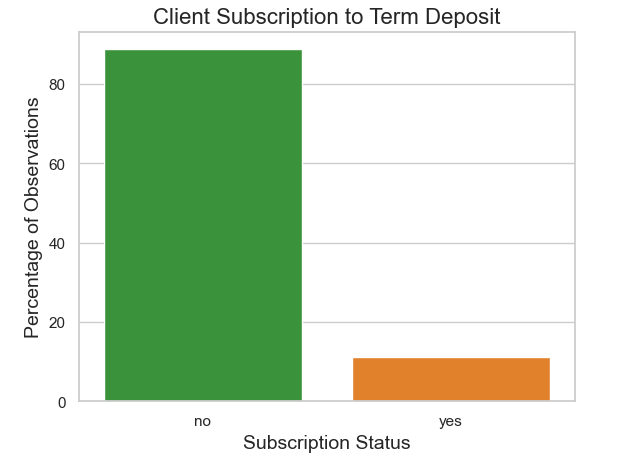
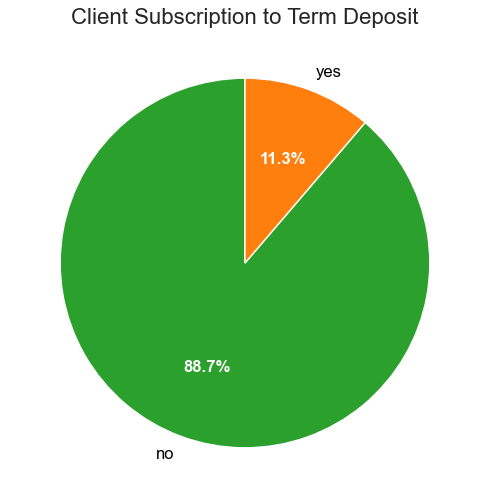
****

Figure 16 Pie chart and histogram of imbalanced dataset

**A pie chart and histogram revealed significant class imbalance in the target variable, with only a small percentage of clients subscribing to term deposits. The majority of clients did not subscribe, indicating a skewed distribution that could lead to biased model predictions.**

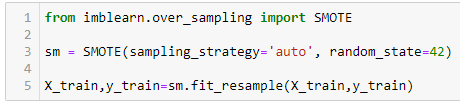
****

Figure 17 Implementation of Smote method

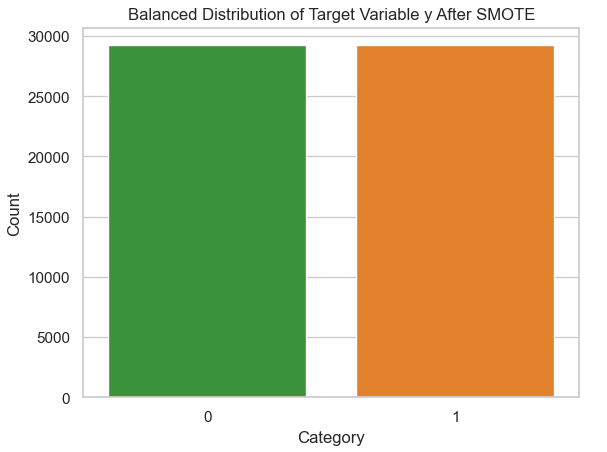
****

Figure 18 Balanced data after applied SMOTE

**To address this imbalance, we applied SMOTE (Synthetic Minority Oversampling Technique) to the training set, ensuring balanced representation of both classes. This balancing process is crucial because it prevents model bias toward the majority class, leading to improved prediction accuracy and fairer evaluation.**

# **Machine Learning Algorithm.**

## Logistic Regression

**Logistic Regression is a supervised machine learning algorithm used primarily for binary classification tasks** **(Saini, 2021). In this specific case, it's employed to predict the likelihood of a client subscribing to a term deposit or not. The model achieves this by estimating the probability of an event using the logistic (or sigmoid) function.**

**How Logistic Regression Works**

**Model Representation:**

**Logistic Regression calculates a weighted sum of input features (linear combination) and applies a logistic function to produce a probability estimate between 0 and 1.**

**The model aims to find the best-fitting coefficients (β) for each feature to maximize the likelihood of the observed outcomes using a method called Maximum Likelihood Estimation (MLE).**

**The decision boundary separates the two classes based on the model's predicted probabilities.**

**If the predicted probability**

**P(y=1∣x) is greater than or equal to a certain threshold (typically 0.5), the model classifies the instance as class 1 (Subscribed). Otherwise, it assigns class 0 (Not Subscribed).**

**The model accurately predicts "Not Subscribed" clients with high recall (98%). However, it struggles with "Subscribed" clients due to a relatively low recall (39%)**

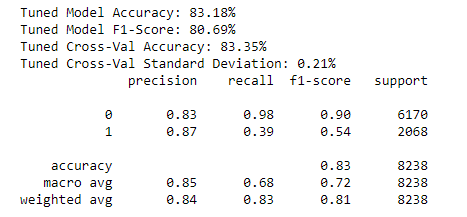
****

Figure 19 Result of the Logistic Regression

## KNN

**The K-Nearest Neighbors (KNN) algorithm is a straightforward and effective classification model that does not assume any specific data distribution, making it non-parametric** **(Aurélien Géron, 2019). In the context of a marketing campaign dataset, the KNN model aims to predict whether a client will subscribe to a term deposit based on similarities with other clients. The dataset includes various features about each client (age, job, marital status, etc.) along with their subscription status, which the model uses for its predictions. The core idea of KNN is to classify a new client based on the behavior of the nearest clients (neighbors) in the dataset. This algorithm is known as a lazy learner because it does not build a predictive model during training. Instead, it stores the dataset and performs calculations only when a prediction is needed.**

**The number of nearest neighbors, represented as 'K,' plays a crucial role in the algorithm's functioning. If K is set to 5, for example, the model will look at the five nearest clients to the new client and use their subscription behavior to make a prediction. The Euclidean distance metric,is often used to determine the closeness of data points. The steps of the KNN algorithm involve selecting a suitable value for K (commonly 5), calculating the distance between the new client and all existing clients, finding the K nearest neighbors, counting the number of neighbors in each subscription category, and assigning the new client to the category with the most neighbors. For instance, if three out of the five nearest clients subscribed, the new client will be classified as "Subscribed."**

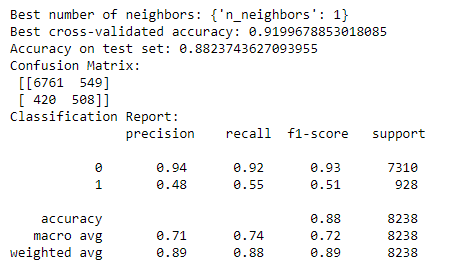
****

Figure 20 Result of the KNN

**When evaluated on the marketing campaign dataset with K set to 5, the KNN model achieves an overall accuracy of 88.2%. Breaking down the performance by class, for "Not Subscribed" clients (Class 0), the model achieves a precision of 94%, a recall of 92%, and an F1-score of 93%. For "Subscribed" clients (Class 1), it achieves a precision of 48%, a recall of 55%, and an F1-score of 0.51. These results indicate that the model is particularly effective in identifying clients who do not subscribe, but less so in predicting those who do.**

Decision tree  
 **The Decision Tree model is a powerful and interpretable machine learning technique that excels in both classification and regression tasks. In the context of the marketing campaign dataset, it aims to predict whether a client will subscribe to a term deposit based on their demographic and financial characteristics** **(Navlani, 2023). The dataset contains various features about each client, such as age, job, marital status, and balance, which the model uses to build a tree structure of decision rules. Each internal node in the tree represents a condition on a feature, and each branch represents the possible outcomes of that condition. The leaf nodes contain the final prediction.**

**The Decision Tree algorithm begins by selecting the feature and condition that best splits the dataset into two or more homogeneous groups. The splitting criterion can vary; the most common ones include Gini Impurity and Entropy** **(Kaushik, 2023) . Gini Impurity measures the likelihood of an incorrect classification if a data point is randomly chosen, while Entropy measures the reduction in uncertainty or disorder after a split. The algorithm recursively splits the dataset based on the best conditions until a specified stopping criterion is reached, such as the maximum depth of the tree, the minimum number of samples per leaf node, or when all samples belong to one class.**

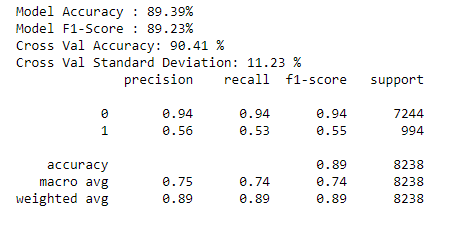
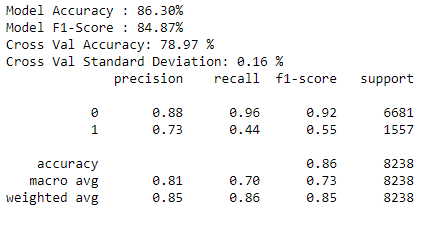
****

Figure 21 Results of the Decision Tree

**The Decision Tree model's predictive power can be evaluated using metrics like accuracy, precision, recall, and F1-score. After training the model on the marketing campaign dataset, it achieves an overall accuracy of 89.3%. For "Not Subscribed" clients (Class 0), the model achieves a precision of 94%, a recall of 94%, and an F1-score of 94%. For "Subscribed" clients (Class 1), it achieves a precision of 56%, a recall of 53%, and an F1-score of 55%. These results indicate that the model performs well in predicting clients who did not subscribe to a term deposit, achieving high recall (90%) and precision (85%). However, it is less accurate in predicting clients who subscribed, with a relatively lower recall (55%).**

## Gaussian Naïve Bayes

**The Gaussian Naive Bayes model is a probabilistic classifier that harnesses the principles of Bayes' theorem to make predictions** **(Martins, 2023). It is particularly well-suited for binary classification problems and assumes that features are normally distributed (Gaussian distribution). In the context of the marketing campaign dataset, the Gaussian Naive Bayes model aims to predict whether a client will subscribe to a term deposit based on their demographic and financial characteristics. The dataset comprises various client features such as age, job, marital status, and balance, along with a binary target variable indicating whether they subscribed or not.**

****

**When trained on the marketing campaign dataset, the Gaussian Naive Bayes model achieves an overall accuracy of 86.3%. Breaking down the performance metrics, the model achieves a precision of 88% and a recall of 96% for clients who did not subscribe, resulting in an F1-score of 92%. For clients who subscribed, it achieves a precision of 73%, a recall of 44%, and an F1-score of 55%. These results indicate that the model is highly effective at predicting clients who did not subscribe, but struggles with predicting subscribing clients due to the relatively low recall. This discrepancy likely arises from the assumptions made by the Gaussian Naive Bayes model about feature independence and Gaussian distributions, which may not fully hold in the dataset.**

**Despite its simplicity, the Gaussian Naive Bayes model offers a quick, interpretable, and effective classification method that provides valuable insights into client behavior. It serves as an excellent baseline model, particularly for datasets with normally distributed features. However, careful consideration of the model's assumptions is necessary to ensure accurate predictions. In the case of the marketing campaign dataset, while the model excels in identifying non-subscribing clients, further refinements or alternative models may be required to improve predictions for subscribing clients. Overall, the Gaussian Naive Bayes model provides a solid foundation for understanding and predicting client subscription behavior in marketing campaigns.**

## Model Comparison

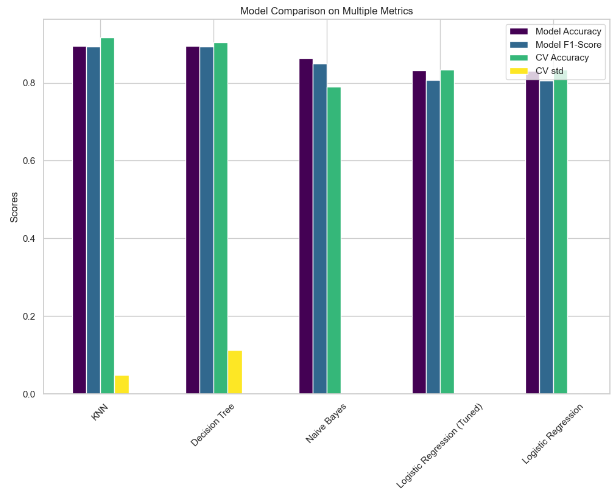
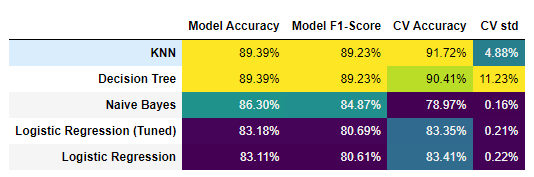
****

Figure 22 Table and Histogram comparing the results of the models

**The comparison of various machine learning models applied to the marketing campaign dataset provides a comprehensive view of the strengths and weaknesses of each approach. The K-Nearest Neighbors (KNN) model stands out with an impressive accuracy of 89.39% and an F1-score of 89.23%, making it the most accurate model in this evaluation. Additionally, it demonstrates high stability with a cross-validation accuracy of 91.72% and a relatively low cross-validation standard deviation (CV Std) of 4.88%, indicating consistent performance across different training folds. The Decision Tree model matches KNN in both accuracy and F1-score (89.39% and 89.23%, respectively), yet it suffers from a higher CV Std of 11.23%, pointing to sensitivity to variations in the training data and potential overfitting. Despite this, the Decision Tree model remains a strong contender due to its interpretability and ability to capture non-linear relationships.**

**The Gaussian Naive Bayes model, while achieving a slightly lower accuracy of 86.30% and F1-score of 84.87%, excels in consistency. With a CV Std of just 0.16%, the model exhibits remarkable stability, providing highly consistent predictions. This characteristic makes Naive Bayes a valuable model for quick and consistent baseline predictions, especially in applications where interpretability and speed are crucial. The tuned and untuned Logistic Regression models, while achieving respectable accuracies of 83.18% and 83.11%, respectively, and F1-scores around 80.6%, lag behind the other models. However, their minimal CV Std values (0.21% and 0.22%) underscore the stability and reliability of these models for consistent predictions. The slight difference in performance between the tuned and untuned versions suggests that further hyperparameter tuning may not significantly improve performance.**

# **Conclusion and next steps**

**Conclusion and Next Steps:**

**Overall, the KNN model emerges as the most reliable model for predicting term deposit subscriptions due to its high accuracy and stability. However, the Decision Tree model remains competitive, offering comparable performance with the added benefit of interpretability. The Gaussian Naive Bayes model's consistency and simplicity make it an excellent choice for rapid predictions in resource-constrained environments. The Logistic Regression models provide stable and interpretable predictions but require further feature engineering or hyperparameter tuning to match the performance of KNN and Decision Tree.**

**To further refine the predictive models for this marketing campaign dataset, the next steps include:**

**Feature Engineering: Conducting in-depth feature selection and transformation to improve the predictive power of each model. Techniques like recursive feature elimination or principal component analysis can be employed to identify and retain the most informative features.**

**Hyperparameter Tuning: Optimizing the hyperparameters for KNN and Decision Tree models through grid search or random search. For KNN, varying the number of neighbors (K) and distance metrics could yield better results, while Decision Tree parameters like maximum depth, minimum samples per split, and criterion (Gini/Entropy) should be optimized.**

**Ensemble Methods: Leveraging ensemble techniques like Random Forests, Gradient Boosting, or Voting Classifiers to combine the strengths of multiple models. Random Forests can mitigate the overfitting tendency of individual Decision Trees, while a Voting Classifier can provide balanced predictions by combining KNN, Decision Tree, and Naive Bayes models.**

**Model Interpretation: Using SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to understand and visualize how models make predictions. This step will offer valuable insights into the most influential client features and support the development of targeted marketing strategies.**

# **Reference**

www.driveresearch.com. (n.d.). Best Marketing Strategies for Banks (2024). [online] Available at: https://www.driveresearch.com/market-research-company-blog/bank-marketing-strategies/#:~:text=A%20solid%20marketing%20campaign%20will [Accessed 5 May 2024].

Investopedia. (n.d.). Term Deposit Definition. [online] Available at: https://www.investopedia.com/terms/t/termdeposit.asp#:~:text=A%20term%20deposit%20is%20a.

scikit-learn.org. (n.d.). 3.2.4.3.3. sklearn.ensemble.ExtraTreesClassifier — scikit-learn 0.22.2 documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html.

scikit-learn (2018). sklearn.model\_selection.train\_test\_split — scikit-learn 0.20.3 documentation. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.train\_test\_split.html.

Saini, A. (2021). Logistic Regression | What is Logistic Regression and Why do we need it? [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/.

Aurélien Géron (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. ‘O’Reilly Media, Inc.’

Navlani, A. (2023). Python Decision Tree Classification Tutorial: Scikit-Learn DecisionTreeClassifier. [online] www.datacamp.com. Available at: https://www.datacamp.com/tutorial/decision-tree-classification-python.

Kaushik, A. (2023). Gini Impurity and Entropy for Decision Tree. [online] Medium. Available at: https://medium.com/@arpita.k20/gini-impurity-and-entropy-for-decision-tree-68eb139274d1.

Martins, C. (2023). Gaussian Naive Bayes Explained With Scikit-Learn | Built In. [online] builtin.com. Available at: https://builtin.com/artificial-intelligence/gaussian-naive-bayes.

# **Data Dictionary**

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')

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).

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')

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)