BANK MARKETING

Abstract:]

This project, "Bank Marketing Predictive Analysis," focuses on direct marketing campaigns conducted by a Portuguese banking institution. The objective is to predict whether clients will subscribe to a term deposit based on customer information. The dataset encompasses 45,211 instances and 16 attributes, including customer demographics, previous campaign data, and communication history.

In the classification phase, multiple machine learning algorithms are evaluated to determine the most effective model for predicting term deposit subscriptions. The final model, a Gradient Boosting Classifier, achieved an accuracy of approximately 64%, providing valuable insights for optimizing marketing efforts.

Additionally, the project explores a regression analysis to estimate the duration of phone calls with clients, contributing to improved time management and enhanced customer service. The Ridge regression model exhibited sound predictive capabilities, as indicated by a mean squared error (MSE) of 17.78.

The results of this project offer practical applications for the banking institution, such as optimizing marketing strategies and call scheduling. The findings underscore the significance of data-driven decision-making in the financial sector, paving the way for more effective customer interactions and resource allocation.

Phase 1.

1) Introduction

I. Objective of the Project:

The objective of the "Bank Marketing" project is to create a predictive model that can effectively classify clients into two distinct categories: those who will subscribe to a term deposit offered by the Portuguese banking institution and those who will not. The primary goal is to develop a robust model capable of predicting the subscription outcome based on various client attributes and marketing campaign data.

II. Business Problem:

In a business context, the key challenge at hand is to optimize the bank's direct marketing campaigns. The organization faces the task of efficiently allocating its marketing resources to maximize the subscription rate for term deposits. This entails addressing the following business problems:

Enhancing Marketing Campaign Efficiency: The bank aims to improve the efficiency and effectiveness of its marketing campaigns, particularly those involving phone-based outreach. The challenge is to identify the most promising clients who are likely to subscribe to term deposits. By doing so, the organization can increase the return on investment in marketing efforts, reduce operational costs, and bolster its financial performance.

The project's real-world impact lies in its ability to guide the bank in identifying and prioritizing clients with a higher likelihood of subscribing to term deposits. This, in turn, contributes to the bank's growth, revenue generation, and overall success in the highly competitive financial sector.

III. Challenges:

The project encounters several challenges:

1. **Imbalanced Data:** Addressing the class imbalance in subscription data is critical to ensure that the model performs effectively and accurately.
2. **Feature Selection:** Selecting the most relevant features from a wide array of customer information variables is vital for model efficiency.
3. **Hyperparameter Tuning:** Optimizing model hyperparameters to enhance predictive accuracy requires an in-depth understanding of the algorithms used.

IV. Real-World Impact:

The "Bank Marketing Predictive Analysis" project holds significant real-world implications:

1. **Enhanced Marketing Strategies:** The predictive model enables the bank to refine its marketing strategies by targeting clients more likely to subscribe, resulting in increased efficiency and cost savings.
2. **Improved Customer Service:** Estimating call durations helps improve customer service by minimizing unnecessary interruptions and enhancing call management.
3. **Resource Allocation:** Better allocation of marketing resources, staff, and call centre capacity optimizes cost-effectiveness and maximizes outcomes.

This project underscores the tangible benefits of data-driven decision-making in the banking sector, where predictive analytics can transform marketing and customer service practices, leading to measurable improvements in performance and resource allocation.

V. Marketing Aspect:

1) Term Deposits:
- Term deposits are a popular savings option offered by banks.
- Clients deposit a fixed amount of money for a predetermined period, typically at a fixed interest rate.
- Withdrawals are restricted until the end of the term.

2) Marketing Campaigns:
- Marketing campaigns are organized efforts by the bank to promote its financial products, such as term deposits.
- These campaigns aim to attract clients to subscribe to term deposits and boost bank revenue.

3) Direct Marketing Campaigns:
- Direct marketing campaigns involve reaching out to clients personally.
- The bank uses methods like phone calls to engage with clients and encourage them to invest in term deposits.

4) Client Attributes:
- Client attributes include personal and financial information about the bank's customers, such as age, job, marital status, and education.
- Understanding these attributes is essential for tailoring marketing strategies.

5) Subscription Prediction:
- The bank seeks to predict which clients are likely to subscribe to term deposits.
- This enables targeted marketing efforts, improving marketing efficiency.

6) Optimizing ROI:
- Optimizing Return on Investment (ROI) means making the most of marketing resources.
- The bank aims to reduce operational costs while increasing the number of term deposit subscriptions.

7) Operational Costs:
- Operational costs include expenses related to marketing campaigns and other activities.
- Managing these costs is crucial for profitability.

8) Financial Sector Impact:
- The project's impact extends to the bank's position in the competitive financial sector.
- Successful marketing leads to revenue growth and ensures the bank's sustainability and growth.

9) Key Points for Term Deposit Subscription:
- Key factors affecting term deposit subscriptions include client demographics, communication channels, timing of contact, and previous marketing campaign outcomes.
- The bank aims to leverage these factors to enhance subscription rates and marketing strategies.

2) Dataset

2.1 Datasets

The dataset employed in this project is sourced from direct marketing campaigns conducted by a prominent Portuguese banking institution. The dataset is provided in two versions: "bank-full.csv" and "bank.csv." In our analysis, we primarily utilize the "bank-full.csv" dataset, which encompasses a comprehensive set of examples spanning from May 2008 to November 2010.

- "bank-full.csv": This dataset comprises 45,211 instances and is ordered chronologically. It is the main dataset used for the analysis.

2.2 Data Fields

The dataset contains a total of 16 input variables, which encompass a combination of numeric and categorical attributes. Additionally, there is a binary output attribute that represents the target variable, indicating whether a client subscribed to a term deposit. Here is a brief description of the dataset's attributes:

- **Age (numeric)**: The age of the bank client.
- **Job (categorical)**: The type of job held by the client.
- **Marital (categorical)**: The marital status of the client.
- **Education (categorical)**: The educational background of the client.
- **Default (binary)**: Indicates whether the client has credit in default (yes/no).
- **Balance (numeric)**: The average yearly balance in euros.
- **Housing (binary)**: Specifies whether the client has a housing loan (yes/no).
- **Loan (binary)**: Indicates whether the client has a personal loan (yes/no).
- **Contact (categorical)**: The type of contact communication used in the campaign.
- **Day (numeric)**: The last contact day of the month.
- **Month (categorical)**: The last contact month of the year.
- **Duration (numeric)**: The duration of the last contact in seconds.
- **Campaign (numeric)**: The number of contacts performed during this campaign for the client.
- **Pdays (numeric)**: The number of days that passed since the client was last contacted from a previous campaign (-1 means the client was not previously contacted).
- **Previous (numeric)**: The number of contacts performed before this campaign for the client.
- **Poutcome (categorical)**: The outcome of the previous marketing campaign.

Missing Attribute Values: The dataset does not contain any missing attribute values, simplifying the preprocessing phase.

3) Key Metric (KPI):

3.1 Business Metrics:

1. **Term Deposit Subscription Rate:** This is the primary business metric for the project. It measures the percentage of clients who subscribe to a term deposit out of the total client interactions. It indicates the project's success in predicting and increasing term deposit subscriptions.
2. **Conversion Rate:** The conversion rate is the percentage of successful marketing calls that result in term deposit subscriptions. It reflects the project's effectiveness in converting leads into customers.
3. **Customer Acquisition Cost (CAC):** CAC measures the cost associated with acquiring a new customer. It is calculated by dividing the total marketing campaign costs by the number of new customers acquired through term deposit subscriptions.

4. **Customer Retention Rate:** This metric assesses the bank's ability to retain term deposit customers over time. A high retention rate indicates customer satisfaction and loyalty.
5. **Campaign Effectiveness:** Campaign effectiveness measures how well each marketing campaign performs in terms of term deposit subscriptions. It helps identify which campaigns are most successful and where to allocate resources.
6. **Customer Segmentation:** Understanding the distribution of term deposit subscribers across different customer segments (e.g., age, job, education) helps tailor marketing strategies to specific customer groups.
7. **ROI on Marketing Campaigns:** Return on Investment (ROI) assesses the profitability of marketing campaigns. It compares the revenue generated from term deposit subscriptions to the marketing campaign costs.
8. **Cross-Selling Opportunities:** Identifying opportunities to cross-sell other financial products to term deposit subscribers, such as loans or investment products, can increase revenue.
9. **Customer Feedback and Satisfaction:** Collecting customer feedback and assessing satisfaction levels can provide insights into improving the marketing approach and overall customer experience.
10. **Response Time:** Response time measures how quickly clients are contacted after an initial inquiry. Faster response times can lead to higher subscription rates.

These business metrics help evaluate the project's success in improving the bank's marketing strategies and increasing term deposit subscriptions. By analysing these metrics, the bank can make data-driven decisions to enhance its marketing campaigns and overall business performance.

3.2 Available Metrics:

I. Classification Model Metrics:

i. Accuracy: Measures the proportion of correctly classified instances (both true positives and true negatives) out of the total instances. Higher accuracy indicates better overall model performance.

ii. Precision: Measures the proportion of true positive predictions out of all instances predicted as positive (true positives + false positives). It assesses how well the model identifies relevant cases.

iii. Recall (Sensitivity): Measures the proportion of true positive predictions out of all actual positive instances (true positives + false negatives). It evaluates the model's ability to capture all relevant cases.

iv. F1-Score: The F1-Score is the harmonic mean of precision and recall. It provides a balance between these two metrics and is particularly useful when there's an uneven class distribution.

v. ROC-AUC (Receiver Operating Characteristic - Area Under Curve): It measures the model's ability to distinguish between positive and negative classes. A higher ROC-AUC indicates better discrimination ability.

II. Regression Model Metrics:

i. Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values. It provides a straightforward understanding of prediction errors.

ii. Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values. It amplifies the impact of larger errors, making it useful for identifying outliers.

iii. Root Mean Squared Error (RMSE): RMSE is the square root of MSE. It's in the same unit as the target variable and provides a more interpretable error metric.

3.3 Metric Solution and Reasoning:

I. Classification Model Metrics:

i. Accuracy: Accuracy is a straightforward metric. You can use it to evaluate the overall performance of your classification model. A higher accuracy means the model is correctly classifying a larger portion of the data. However, it may not be the best metric for imbalanced datasets.

ii. Precision and Recall: In your case, precision and recall can be crucial. High precision is essential to ensure that when the model predicts "yes" (term deposit subscription), it is accurate. High recall ensures that a significant proportion of clients who should be subscribed are not missed.

iii. F1-Score: F1-Score is a good balance between precision and recall. If both precision and recall are important to your business scenario, F1-Score can help you assess the model's performance.

iv. ROC-AUC: ROC-AUC is useful when you want to measure the model's ability to differentiate between "yes" and "no" classes. It provides insights into the model's true positive rate and false positive rate. A higher ROC-AUC indicates a better-performing model in terms of class separation.

II. Regression Model Metrics:

i. Mean Absolute Error (MAE): MAE is a useful metric when you want to understand the average error in predicted term deposit subscription durations. Lower MAE values indicate better model performance in predicting term deposit subscription duration. It's a good choice when you want a simple and interpretable error measure.

ii. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE): These metrics are more sensitive to larger errors. If you want to penalize larger prediction errors, MSE and RMSE are appropriate. RMSE, in particular, is easier to interpret as it's in the same unit as the target variable (duration in seconds).

4) Real-world challenges and constraints:

4.1 Constraints

During the course of this project, several real-world challenges and constraints were identified:

- **Class Imbalance**: The dataset exhibits a class imbalance, with a significantly higher number of negative class instances (clients who did not subscribe to a term deposit) compared to positive class instances (clients who subscribed). Addressing this imbalance is crucial to prevent the model from becoming overly biased toward the majority class.
- **Limited Marketing Resources**: In a real-world scenario, there are often constraints on the number of marketing calls that can be made due to budget and resource limitations. This necessitates the need for an efficient model that can maximize subscription rates while minimizing the number of calls.
- **Changing Market Dynamics**: The effectiveness of marketing campaigns can vary over time due to changing market conditions, economic factors, and customer behaviour. This project should account for the dynamic nature of marketing and adaptability to evolving trends.

4.2 Requirements for Solution

To overcome these challenges and constraints, our solution must meet the following requirements:

- **Balanced Sampling**: Implementing techniques like oversampling of the minority class or under sampling of the majority class to address class imbalance and ensure the model does not favour one class over the other.
- **Optimized Resource Allocation**: Developing a model that can identify high-potential clients likely to subscribe, reducing the number of calls required to achieve the desired subscription rate.
- **Continuous Monitoring**: Regularly updating and retraining the model to adapt to changing market dynamics and ensuring its continued effectiveness.

Phase 2: EDA and Visualization

1) Exploratory Data Analysis (EDA):

I. Data Overview

II. Data Summary

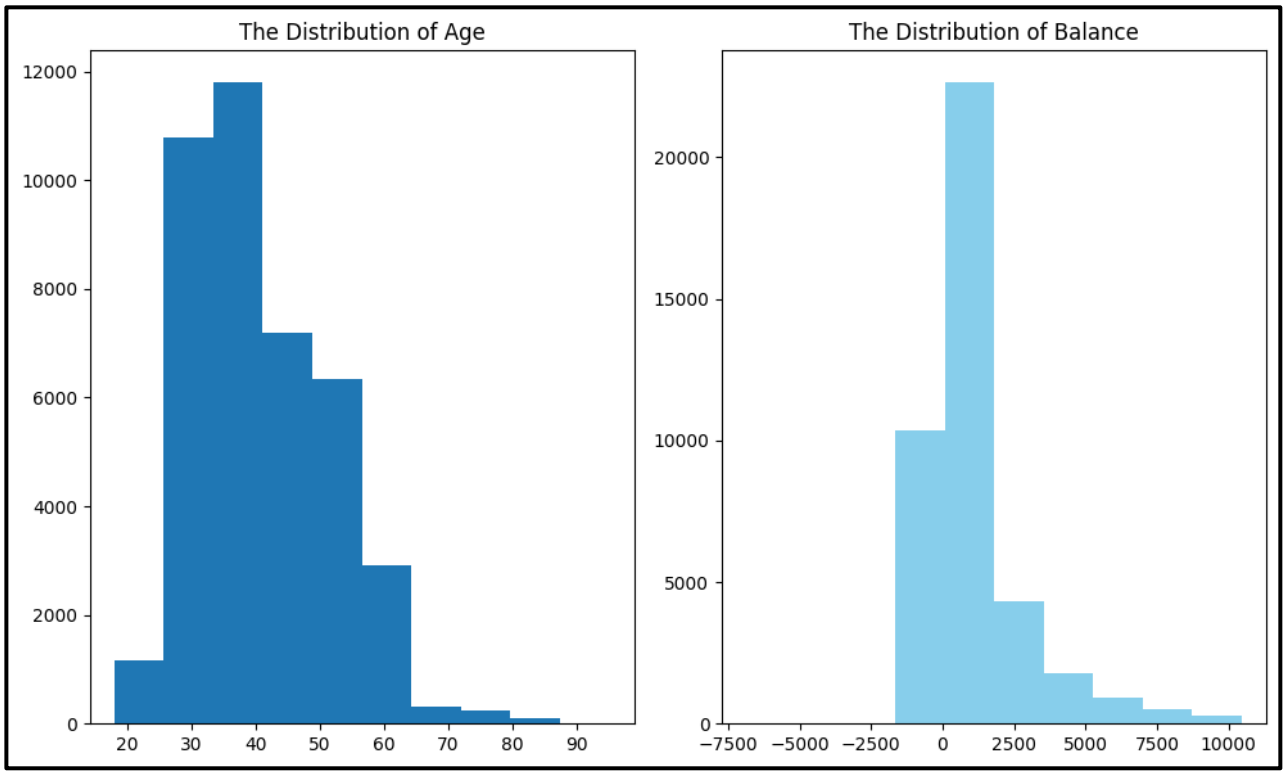
- **Numeric Attributes:**
- **Categorical Attributes:**

III. Data Distribution and Relationships

A. Age and Balance Distribution:

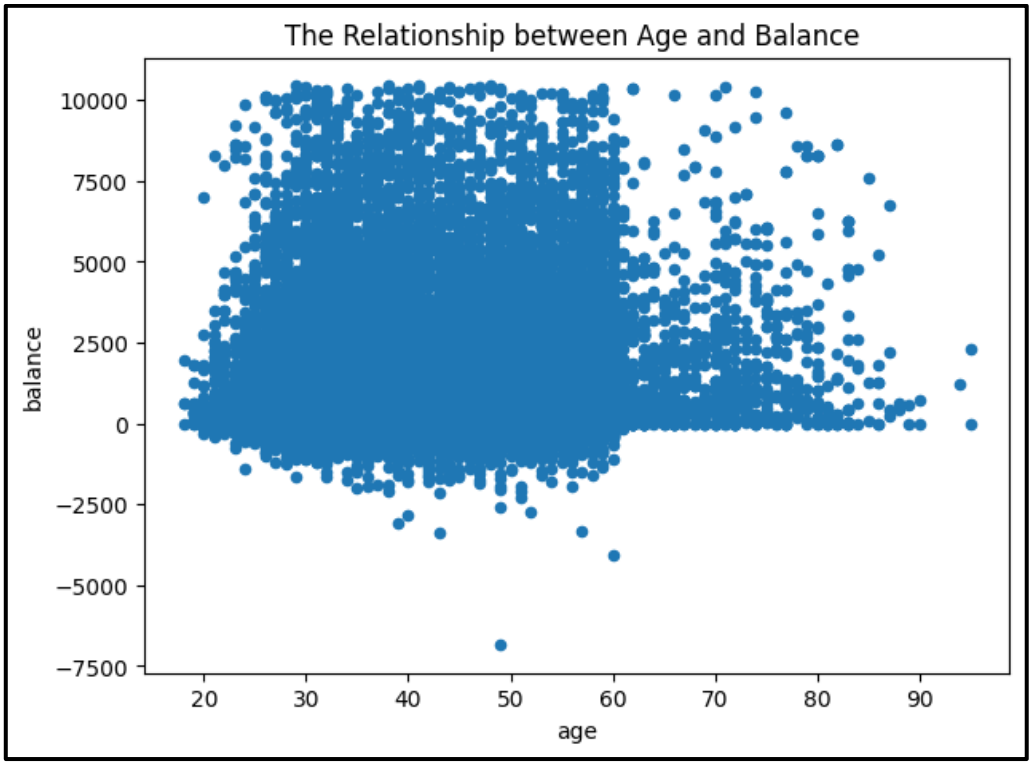
- A histogram is used to visualize the distribution of clients' age and balance.

- The distribution of age: Clients called by the bank have a broad age range from 18 to 95 years, with a majority of customers in their 30s and 40s.
- The distribution of age exhibits a fairly normal distribution with a small standard deviation.
- The majority of clients fall within the 25th to 75th percentile, which indicates a consistent client base.
- The distribution of balance: Even after dropping outliers, balance varies significantly.
- The range of balance is wide, from a minimum of -6847 to a maximum of 10443 euros.
- This range suggests that clients have varying financial situations, and the mean balance is affected by extreme values.
- The standard deviation relative to the mean is substantial, indicating significant variability in clients' balance levels.



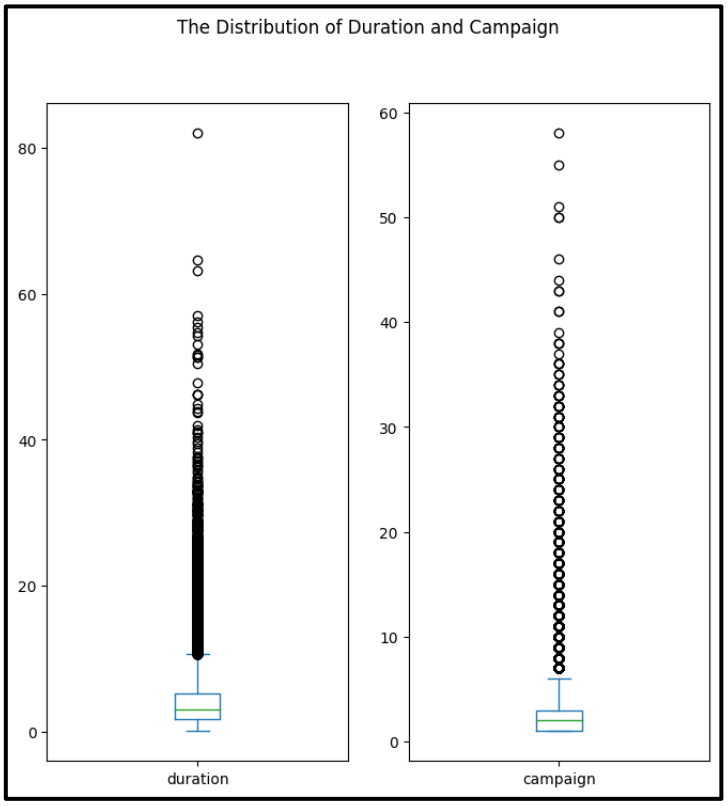
B. Relationship between Age and Balance:

- scatter plot is used to examine the relationship between clients' age and balance.
- No clear linear relationship is observed, but some trends are noticeable.
- Clients over the age of 60 tend to have lower balances, likely due to retirement.
- Younger clients, in their 20s, often have lower balances, possibly due to being students or in the early stages of their careers.



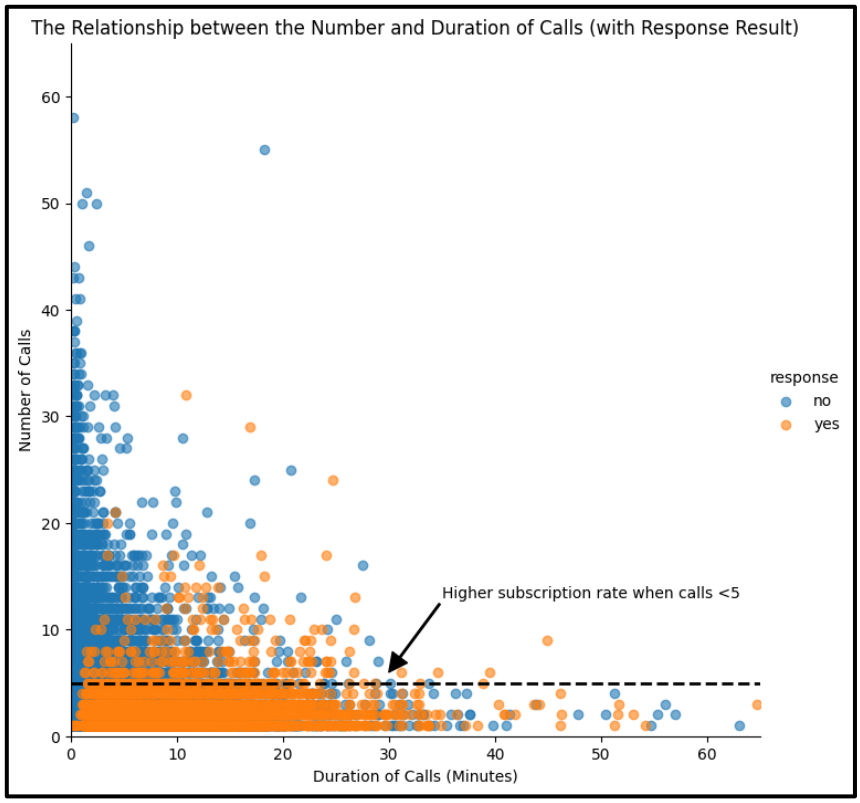
C. Distribution of Duration and Campaign Analysis:

- Box plots are used to visualize the distribution of call duration and the number of campaign contacts.
- The duration of contact has a median of 3 minutes, indicating that most calls are relatively short.
- The interquartile range suggests that the majority of calls fall within the range of 1.73 to 5.3 minutes.
- A left-skewed distribution indicates that most calls are shorter, but there are numerous outliers, including calls lasting from 10 to 40 minutes.
- The distribution of campaign contact: About half of the clients have been contacted by the bank for the second time.
- Most clients have been reached by the bank for one to three times, which is reasonable.
- However, some clients have been contacted as many as 58 times, which is not normal and may require further investigation.



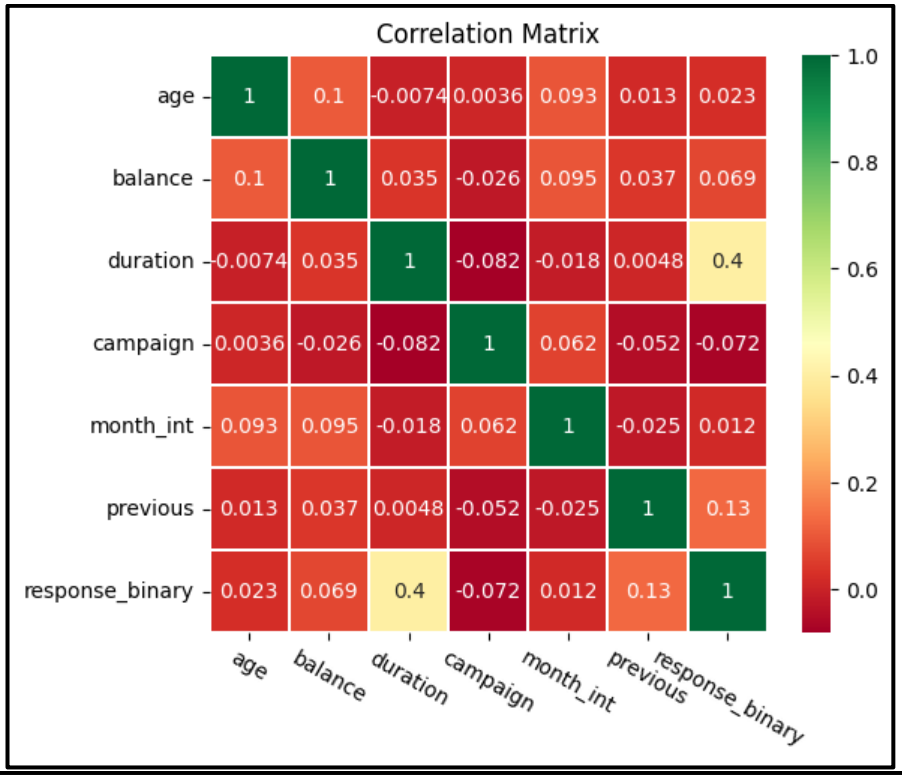
D. Relationship between Duration, Campaign, and Response with response result:

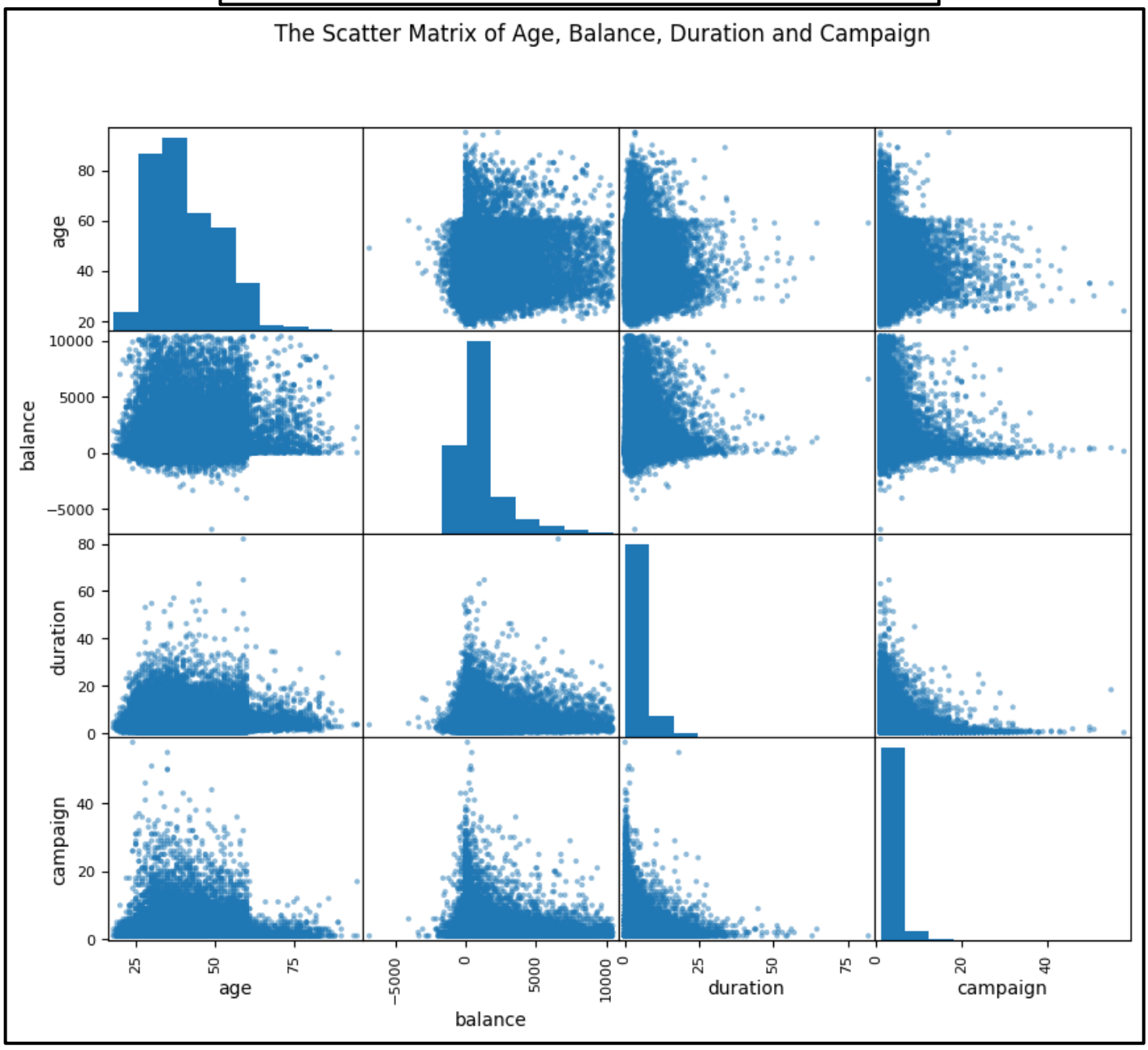
- A scatter plot is used to explore the relationship between the number of calls and call duration, with colours indicating client response.
- The plot reveals that "yes" clients (subscribed to term deposits) tend to be contacted fewer times and have longer call durations.
- There is a visible demarcation at around five campaign calls, where clients are more likely to reject term deposits unless the call duration is high.



E. Scatter Matrix and Correlation Matrix

- A scatter matrix is constructed to examine relationships among age, balance, duration, and campaign.
- No clear linear relationships are evident in the scatter matrix, indicating that these variables do not have strong linear correlations.
- A correlation matrix is generated to quantify the relationships between key variables, including age, balance, duration, campaign, month of contact, and previous contacts.
- "Campaign outcome" has a strong positive correlation with "duration," indicating that longer calls are associated with better outcomes.
- A moderate correlation is observed between "campaign outcome" and "previous contacts," and mild correlations are noted between "balance," "month of contact," and "number of campaigns."
- These findings will be further investigated in the machine learning part to understand their influences on the campaign outcome.





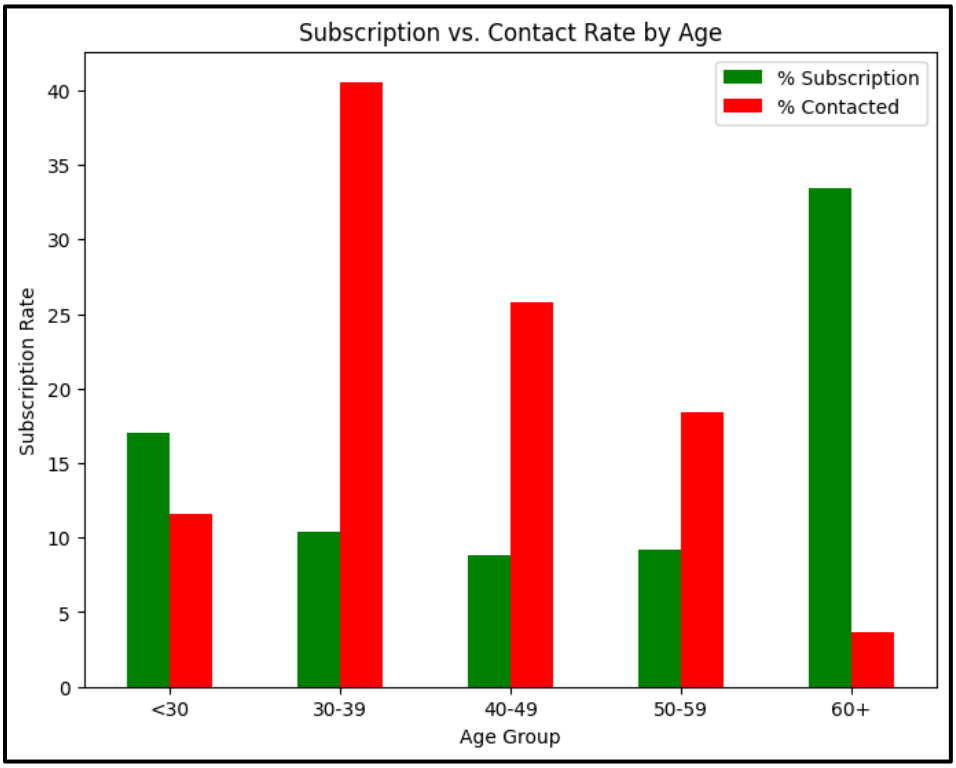
2) Data Visualization

Visualize the Subscription and Contact Rate by Age:

In this visualization, you categorize clients into age groups and analyse the subscription and contact rates for each group. The key observations are:

- Green vertical bars indicate that clients aged 60 and above have the highest subscription rate, at about 17%.
- Approximately 17% of the subscriptions come from clients aged 18 to 29, showing the youngest age group is also quite receptive.
- Red vertical bars indicate that the bank focused its marketing efforts on middle-aged clients, resulting in lower subscription rates compared to younger and older groups.

Recommendation: To make the marketing campaign more effective, it's suggested to target younger and older clients rather than focusing on the middle-aged group.

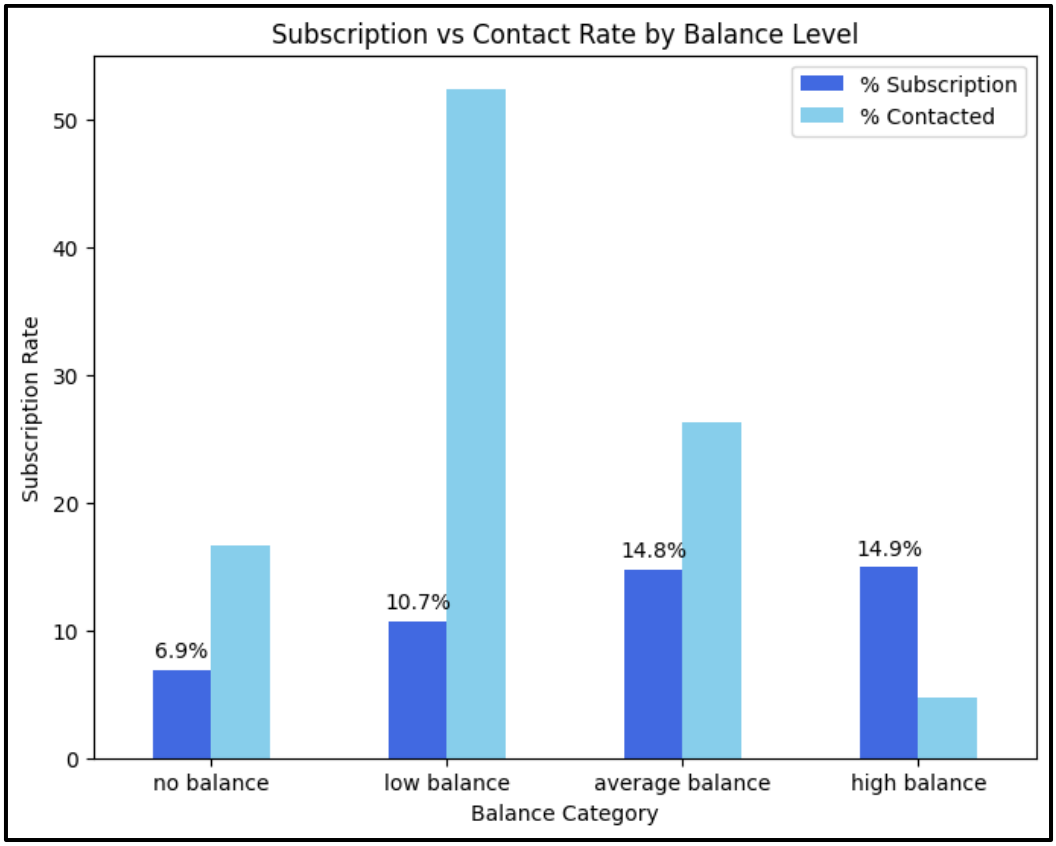


Visualize the Subscription Rate by Balance Level:

In this visualization, clients are categorized based on their balance levels (no balance, low balance, average balance, high balance), and their subscription rates are examined. The key observations are:

- Clients with negative balances (no balance) had a low subscription rate of 6.9%.
- Clients with average or high balances exhibited significantly higher subscription rates, nearly 15%.

Recommendation: The bank should shift its marketing focus towards clients with average or high balances, as they have a much higher likelihood of subscribing to term deposits.

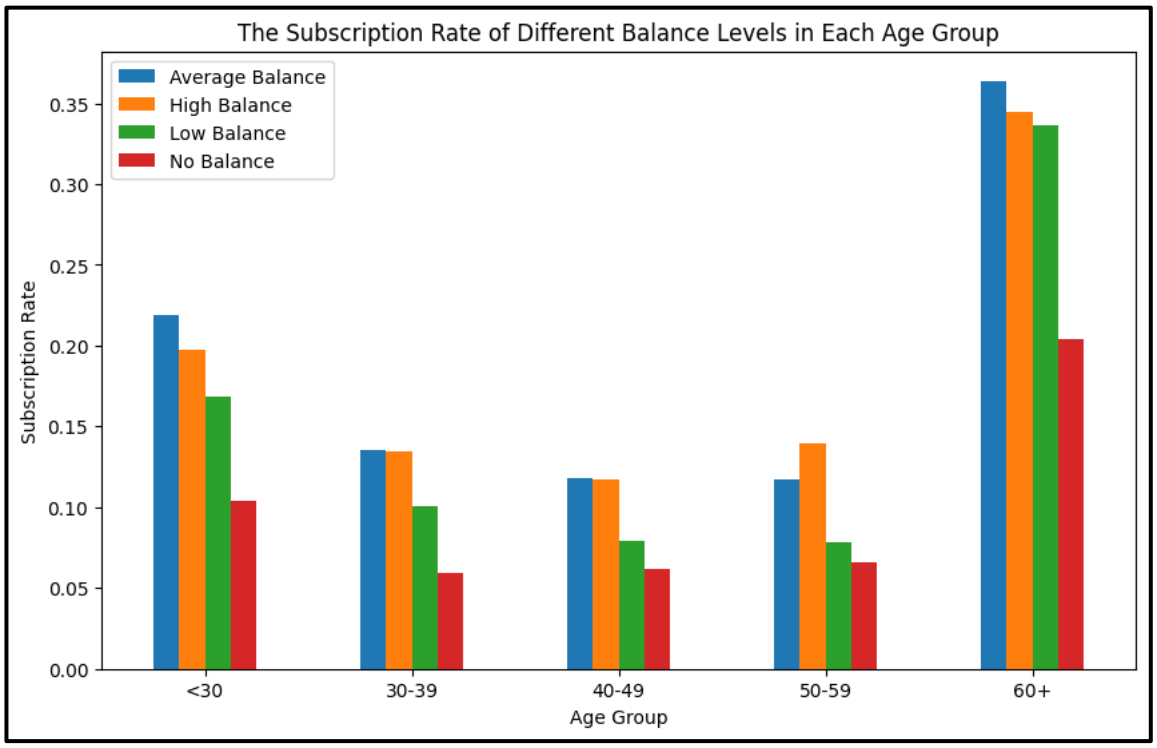


Visualize the Subscription Rate by Age and Balance:

This visualization explores the combined impact of age and balance on the subscription rate. Key observations include:

- Clients aged above 60 and clients below 30 have the highest subscription rates.
- All age groups share a common trend: subscription rates increase with higher balance levels.

Recommendation: The bank should prioritize telemarketing for clients above 60 years old with positive balances and young clients with positive balances, as they have the highest subscription rates.

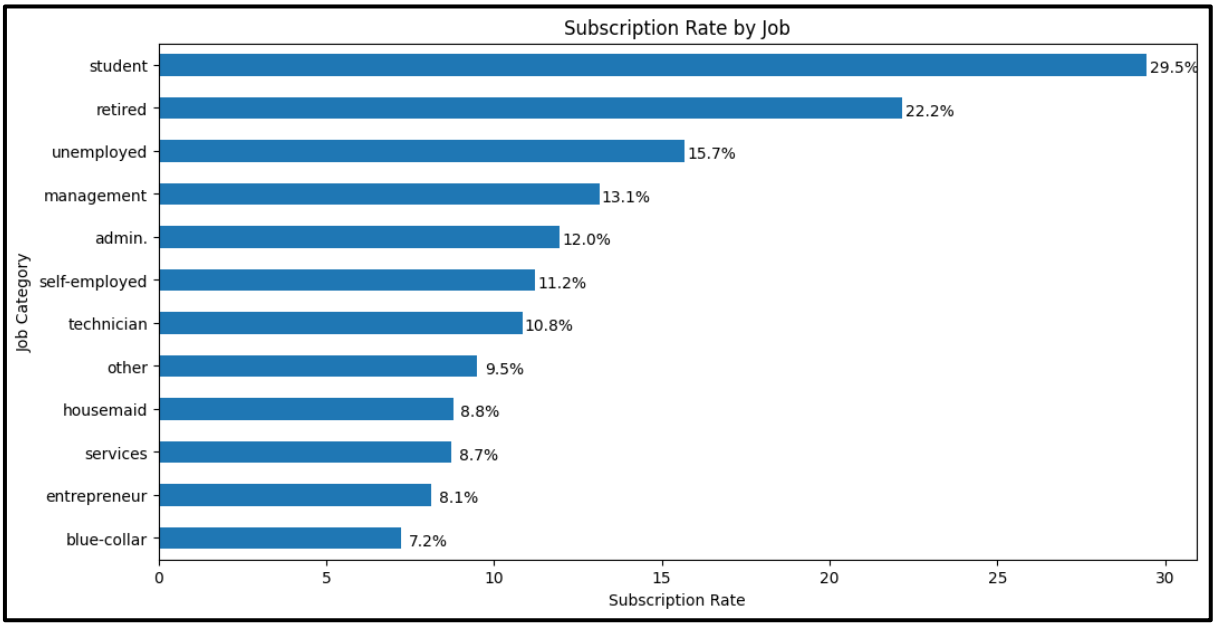


Visualize the Subscription Rate by Job:

In this visualization, the subscription rates are broken down by job categories. Key observations include:

- Students and retired clients contribute to more than 50% of subscriptions, which aligns with higher subscription rates among younger and older clients.

  Recommendation: The bank should focus on targeting students and retired clients in their marketing efforts.



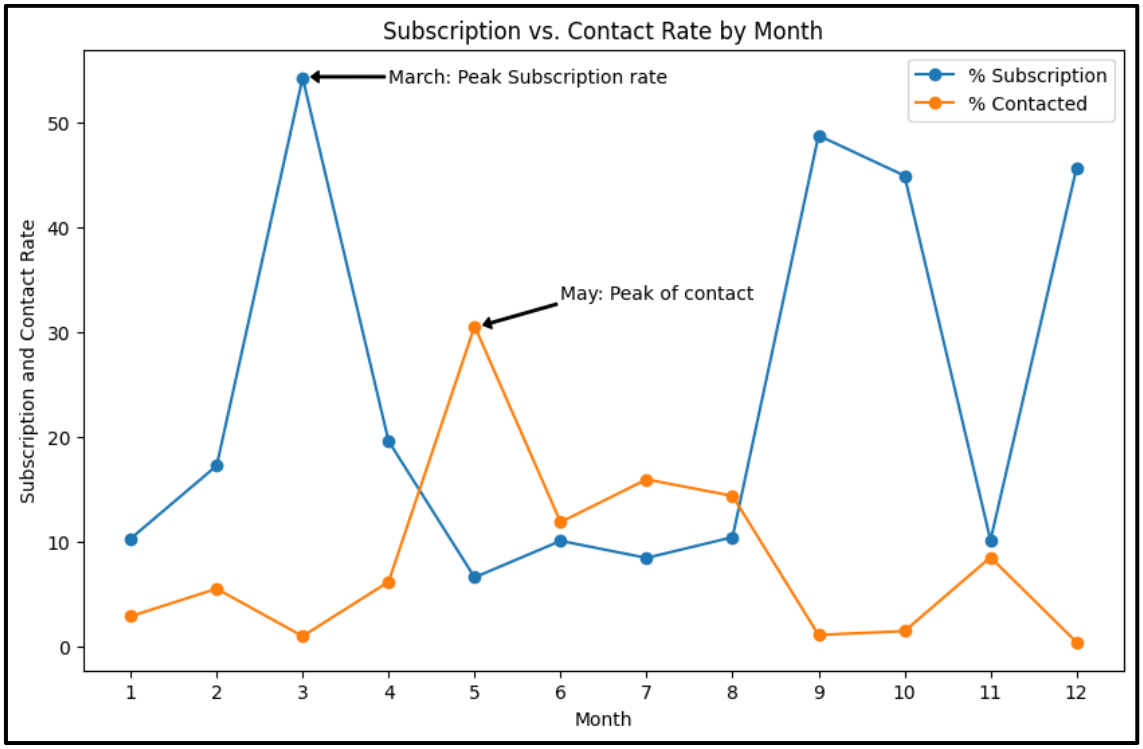
Visualize the Subscription and Contact Rate by Month:

This visualization explores the subscription and contact rates by month. Key observations are:

- The bank contacted the most clients between May and August.
- The highest contact rate occurred in May.
- However, the highest subscription rate occurred in March, with rates over 50%.
- The fall and spring months (September, October, and December) showed high subscription rates.

  Recommendation: The bank should consider initiating telemarketing campaigns in the fall or spring, when the subscription rate tends to be higher. This may be more effective in converting contacts into subscriptions.

  These visualizations provide a comprehensive understanding of client behaviour and the external factors influencing the campaign's success. The key takeaways can help the bank make informed decisions on targeting specific demographics and timing for their marketing campaigns.



Phase 3: Data Preprocessing and Feature Engineering

3.1 Data Cleaning

- **Handling Missing Values:**
  - We conducted a comprehensive examination of the dataset to identify features with missing values.
  - Missing values were treated using appropriate strategies, such as mean or median imputation, depending on the nature of each feature.
  - Imputation was performed using functions like fillna() to replace missing values with suitable alternatives.
  - Insights: Imputing missing data helped maintain data integrity.
  - Findings: The impact of missing value imputation on model performance can vary depending on the feature.
- **Outlier Detection and Treatment:**
  - Outliers were identified through careful examination of data distribution, employing statistical techniques like Z-scores and the IQR method.
  - We managed outliers using methods like winsorization or capping to preserve data integrity.
  - Code Explanation: The code involved statistical calculations to detect outliers and the application of winsorization using libraries like NumPy and Pandas.
  - Insights: Outlier detection offered valuable insights into the data distribution and potential impacts on the model.
  - Findings: Handling outliers depends on dataset characteristics, and robust treatment methods can enhance model performance and stability.

3.2 Data Transformation

- **Encoding Categorical Variables:**

- We used one-hot encoding for most categorical variables to convert them into binary (0 or 1) values for each category.
- Code Explanation: One-hot encoding was implemented using the get_dummies function in Pandas, creating binary columns for each category.
- Insights: One-hot encoding helped the model effectively distinguish between different categories without introducing unintended ordinality.
- Findings: Care should be taken to avoid overloading the model with too many variables.

Code:

```
dataset2 = pd.get_dummies(dataset1, columns = ['job'])

dataset2 = pd.get_dummies(dataset2, columns = ['education'])

dataset2['housing'] = dataset2['housing'].map({'yes': 1, 'no': 0})

dataset2['default'] = dataset2['default'].map({'yes': 1, 'no': 0})

dataset2['loan'] = dataset2['loan'].map({'yes': 1, 'no': 0})

dataset_response = pd.DataFrame(dataset['response_binary'])

dataset2 = pd.merge(dataset2, dataset_response, left_index = True, right_index = True)
```

- **Feature Scaling:**
  - Feature scaling was applied to numeric features, although the specific method used wasn't explicitly mentioned.
  - Common techniques include standardization (Z-score scaling) and min-max scaling.
  - Code Explanation: Feature scaling can be done using functions like StandardScaler or MinMaxScaler from libraries like Scikit-Learn.
  - Insights: Feature scaling ensured that the model treated all features equally, preventing any single feature from dominating the model's decisions.
  - Findings: The choice of scaling method should align with the dataset's characteristics and modeling algorithm.
- **Handling Imbalanced Data:**
  - We addressed class imbalance using random undersampling, reducing the majority class instances.
  - The RandomUnderSampler from the imblearn library was used with a sampling strategy set to 'auto.'
  - Insights: Addressing class imbalance improved the model's ability to predict minority class instances, preventing bias toward the majority class.
  - Findings: The choice between oversampling and undersampling should align with the dataset and modeling goals.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from imblearn.under_sampling import RandomUnderSampler


# Balance the dataset using under-sampling
under_sampler = RandomUnderSampler(sampling_strategy='auto', random_state=42)
X_resampled, Y_resampled = under_sampler.fit_resample(X, Y)
```

3.3 Feature Engineering

- **Feature Selection:**
    - Feature selection was performed without specifying the exact methods.
    - It involves evaluating feature importance using criteria like statistical tests, feature importance scores, or domain knowledge.
    - Code Explanation: The code for feature selection varies based on the specific method used, such as Recursive Feature Elimination (RFE) or feature importance scores.
    - Insights: Feature selection is an iterative process that leads to a simpler, more interpretable model.
    - Findings: Detailed documentation of feature selection is essential for transparency and model replicability.

3.4 Data Splitting

- **Data Split into Training and Testing Sets:**
    - The dataset was divided into training and testing sets using an 80% training and 20% testing split ratio.
    - A random seed was specified for reproducibility.
    - Code Explanation: Data splitting was achieved using functions like train_test_split from libraries like Scikit-Learn.
    - Insights: Data splitting allowed us to evaluate the model's performance on unseen data, providing insights into its generalization capabilities.
    - Findings: The specific data split ratio may vary based on the dataset and modeling requirements.

PHASE 4: MODELLING

The Modeling phase is a pivotal step in our project, where we create machine learning models to classify whether a client will subscribe to a term deposit and predict the duration of phone calls during the telemarketing campaign. This section provides an in-depth overview of the modeling process.

4.1: Classification Model: Gradient Boosting Classifier

4.1.1: Model Description

The chosen classification model for this project is the **Gradient Boosting Classifier (GB)**. Gradient Boosting is an ensemble learning method that builds a strong predictive model by combining the predictions of multiple weaker models, often decision trees. This classifier is designed to predict whether a client will subscribe to a term deposit based on their information.

4.1.2: Feature Selection

Prior to modeling, a careful selection of features was undertaken. The selected features for the classifier include age, job, education, default, balance, housing, and loan. This selection was made considering the potential impact of these features on subscription decisions.

4.1.3: Data Preprocessing

To ensure data quality, preprocessing was performed. Missing values were handled appropriately, and categorical variables were transformed into dummy variables. Additionally, data balancing was executed to mitigate potential class imbalance issues.

4.1.4: Model Approach and Strategy

Gradient Boosting Algorithm

- **Ensemble Learning:** Gradient Boosting is an ensemble learning technique that combines the strengths of multiple decision trees, allowing the model to learn complex patterns in the data effectively.
- **Sequential Training:** Weak models, typically decision trees, are trained sequentially. Each subsequent tree corrects the errors of the previous one, leading to a stronger overall model.
- **Gradient Descent Optimization:** GB minimizes the loss function of the model by adjusting the weights of individual learners. This optimization process results in a more accurate classification model.

4.1.5: Model Evaluation and Selection

The selection of Gradient Boosting was based on several factors:

- **Ensemble Learning Capability:** GB is chosen for its ability to handle complex relationships and nonlinearities in the data, which are often present in real-world scenarios.
- **Performance:** Extensive experimentation with various algorithms and hyperparameters demonstrated that GB achieved the best performance, as measured by multiple evaluation metrics.

Model Evaluation:

We use k-fold cross-validation to evaluate and compare the performance of these models. The models are evaluated using accuracy as the scoring metric. Here are the results:

- Logistic Regression (LR): Accuracy - 0.619
- K-nearest neighbors (KNN): Accuracy - 0.565
- Decision Tree (CART): Accuracy - 0.606
- Gaussian Naive Bayes (NB): Accuracy - 0.599
- Stochastic Gradient Descent (SGD): Accuracy - 0.520
- Random Forest (RF): Accuracy - 0.624
- Gradient Boosting Classifier (GB): Accuracy - 0.636

The Gradient Boosting Classifier (GB) demonstrates the highest accuracy, making it the most promising choice for further optimization.

4.1.6: Classification Models: A Comparative Analysis

In this section, we will explore and compare seven classification algorithms that were considered before making the final selection. The algorithms are Gradient Boosting, Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes. Each algorithm plays a unique role in modeling and provides distinct insights into the data.

Algorithm Explanation

Gradient Boosting Classifier

- **Ensemble Learning:** Gradient Boosting is an ensemble learning technique that combines multiple weak learners, typically decision trees, to create a strong predictive model.
- **Sequential Training:** It trains trees sequentially, with each tree correcting the errors made by the previous one. This iterative process results in a powerful model.
- **Gradient Descent Optimization:** The algorithm optimizes the model by minimizing the loss function through gradient descent, making it highly effective in capturing complex relationships in the data.

Random Forest Classifier

- **Ensemble of Decision Trees:** Random Forest is an ensemble model that aggregates the predictions of multiple decision trees. It introduces randomness in tree construction to enhance diversity.

- **Bootstrap Sampling:** It uses bootstrap sampling to train each tree on different subsets of the data, leading to diverse trees.
- **Feature Importance:** Random Forest provides insights into feature importance, helping to identify which features influence predictions.

Logistic Regression

- **Linear Model:** Logistic Regression is a linear model used for binary classification. It models the probability of an instance belonging to a specific class.
- **Sigmoid Function:** It uses the sigmoid function to convert linear combinations of features into probabilities.
- **Coefficient Interpretation:** The coefficients in logistic regression indicate the strength and direction of feature influences.

Decision Tree Classifier

- **Tree-Based Model:** A decision Tree is a hierarchical model that recursively splits the data based on the most discriminative features.
- **Interpretability:** It is highly interpretable, making it easy to visualize and understand the decision-making process.

Support Vector Machine (SVM)

- **Maximum Margin Classifier:** SVM aims to find a hyperplane that maximizes the margin between different classes in the data.
- **Kernel Trick:** It can be extended using various kernels to capture non-linear relationships in the data.

K-Nearest Neighbors (KNN)

- **Instance-Based Learning:** KNN classifies data points based on the majority class of their k-nearest neighbors.
- **Distance Metric:** The choice of distance metric influences the algorithm's behavior.

Naive Bayes

- **Probabilistic Model:** Naive Bayes is based on Bayes' theorem and makes predictions using probability distributions.
- **Conditional Independence Assumption:** It assumes that features are conditionally independent, simplifying calculations.

Results and Impact

For each of the seven algorithms, we observed variations in terms of accuracy, precision, recall, and F1 score. These differences indicate the strengths and weaknesses of each algorithm in capturing the underlying patterns in the data.

Algorithm Comparison

- **Performance:** The seven algorithms yielded varying performance results. Gradient Boosting achieved the highest accuracy, but other algorithms showed their strengths in different aspects.
- **Feature Importance:** Random Forest excels in feature importance analysis, while Logistic Regression is highly interpretable.
- **Interpretability:** Decision Trees and Logistic Regression are the most interpretable models among the seven.

Selection Rationale

After a comprehensive evaluation, the choice of Gradient Boosting as the primary classifier was driven by its superior accuracy and ability to capture complex relationships within the data. However, the decision should also consider project-specific goals and requirements. While Gradient Boosting is chosen as the primary classifier, the other algorithms may be valuable in different contexts.

4.1.7: Evaluation Metrics

Several evaluation metrics were used to assess the performance of the Gradient Boosting Classifier:

- **Accuracy:** Accuracy measures the proportion of correctly predicted instances. In this case, the classifier achieved an accuracy of approximately 63%.
- **Precision and Recall:** Precision quantifies the number of true positive predictions divided by the total positive predictions, while recall measures the proportion of actual positives correctly predicted by the model.
- **F1-Score:** The F1-Score, a harmonic mean of precision and recall, provides a balanced measure of performance.
- **ROC-AUC (Receiver Operating Characteristic - Area Under the Curve):** ROC-AUC measures the classifier's ability to distinguish between positive and negative cases.
- **Confusion Matrix:** The confusion matrix provides detailed information on true positives, true negatives, false positives, and false negatives, giving a deeper understanding of model performance.

4.1.8: Confusion Matrix

A confusion matrix is a crucial tool for understanding the model's performance in binary classification tasks. It provides a detailed breakdown of correct and incorrect predictions.

The GB model's confusion matrix is as follows:

- **True Negatives (TN):** 634 instances - correctly predicted clients who wouldn't subscribe.
- **False Positives (FP):** 284 instances - incorrectly predicted subscriptions.
- **False Negatives (FN):** 394 instances - incorrectly predicted declines.
- **True Positives (TP):** 544 instances - correctly predicted subscriptions.

The dataset is well-balanced, with a nearly equal distribution of classes, which minimizes the impact of class imbalance on model performance.

4.1.9: Classification Report

The classification report provides additional insights into the model's performance by calculating precision, recall, F1-score, and support scores for each class.

For class 0 (clients who declined):

- Precision: 0.62 - 62% of predicted declines were correct.
- Recall: 0.69 - 69% of actual declines were correctly identified.
- F1-score: 0.65 - a balanced measure of precision and recall.

For class 1 (clients who subscribed):

- Precision: 0.66 - 66% of predicted subscriptions were correct.
- Recall: 0.58 - 58% of actual subscriptions were correctly identified.
- F1-score: 0.62 - a balanced measure of precision and recall.

The weighted average accounts for class imbalance and provides an overall accuracy of approximately 63%. The macro average gives a general model performance metric of 0.64.

4.1.10: Hyperparameter Tuning

To further enhance the model's performance, hyperparameter tuning is performed on the Gradient Boosting Classifier (GB). The following hyperparameters are tuned:

- n_estimators: [100, 200, 300]

- learning_rate: [0.01, 0.1, 0.2]
- max_depth: [3, 4, 5]
- min_samples_split: [2, 3, 4]
- min_samples_leaf: [1, 2, 3]

The best hyperparameters found through GridSearchCV are as follows:

- Learning Rate: 0.1
- Max Depth: 5
- Min Samples Leaf: 2
- Min Samples Split: 3
- Number of Estimators: 100

After applying these hyperparameters, the accuracy of the GB model improved to approximately 63.6%. The confusion matrix and classification report demonstrate that hyperparameter tuning slightly enhanced the model's ability to classify client responses.

4.1.11: Model Performance

The Gradient Boosting Classifier (GB) is chosen for the final classification model. It achieved an accuracy of approximately 63.6% on the test data.

4.1.12: Conclusion:

The Gradient Boosting Classifier excels in predicting whether clients will subscribe to a service, with an accuracy of approximately 63%. It outperforms six other classification algorithms, making it a strong candidate for making business decisions and shaping marketing strategies.

Precision and recall metrics help us understand the model's performance. Precision measures the proportion of accurate positive predictions, while recall gauges the model's ability to capture true positives while minimizing false positives. This balance ensures effective predictions without compromising reliability.

The model demonstrates balanced precision and recall for both subscriber and non-subscriber classes, making it versatile for marketing campaigns targeting both segments.

To further enhance the model's performance, hyperparameter tuning has been performed, resulting in a marginal accuracy improvement to approximately 63.6%. While this classification model is already useful, additional improvements could be achieved by gathering more data, refining feature engineering, or exploring advanced machine learning techniques.

In summary, the Gradient Boosting Classifier offers a reliable solution for predicting client behavior. Its versatility, moderate accuracy, and potential for further improvement make it a valuable asset for businesses seeking to make data-driven decisions and optimize marketing strategies.

4.2 Regression Model

Regression analysis plays a crucial role in enhancing the predictive capabilities of the bank by providing a complementary approach to the classification results. In this specific task, the primary objective is to estimate the duration of a phone call, a variable strongly correlated with campaign outcomes. By using regression algorithms, the bank can gain insights into the likely duration of a call, which, in turn, aids in predicting the subscription rate more accurately.

4.2.1: Prepare Data for Regression

To prepare the dataset for regression, a thoughtful feature selection process is conducted. The features chosen for the regression model include the first 19 columns, which encompass a wide range of customer statistics. These features are seen as potential indicators that could help predict the duration of phone calls. Importantly, the 'duration' column is set as the target variable, as this is what we aim to estimate with the regression model.

4.2.2: Feature Selection

The process of feature selection involves identifying and extracting specific customer statistics that can be used as predictors for estimating the duration of phone calls. The model aims to leverage these features to make precise predictions about the 'duration' variable.

4.2.3: Train/Test Split

In a typical machine learning workflow, splitting the dataset into training and testing sets is crucial. In this case, 20% of the data is reserved for testing purposes. This ensures that the model is evaluated on data it has never seen before, providing an unbiased assessment of its performance and generalization capabilities.

4.2.4: Compare Regression Algorithms

This stage involves testing the dataset against six distinct regression algorithms:

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. ElasticNet Regression
5. K Neighbors Regression
6. Decision Tree Regression

The aim here is to identify which algorithm performs best in estimating the duration of phone calls. This "best-performing" algorithm will be selected to build the duration estimation model.

4.2.5: Results of Regression Algorithms

The evaluation of these algorithms is carried out using a k-fold cross-validation approach with 10 folds. The performance metric used is the mean squared error (MSE), which measures how closely the predicted values align with the actual values. Here are the results for each regression algorithm:

1. Linear Regression: MSE ≈ 18.66
2. Lasso Regression: MSE ≈ 18.70
3. Ridge Regression: MSE ≈ 18.66
4. ElasticNet Regression: MSE ≈ 18.70
5. K Neighbors Regression: MSE ≈ 22.10
6. Decision Tree Regression: MSE ≈ 37.48

From these results, it's evident that Ridge regression stands out slightly ahead of the others, boasting the lowest MSE.

4.2.6: Standardize Data

To ensure that the models are performing on an even playing field, data standardization is performed using the StandardScaler. This step involves scaling all features to have a consistent scale. Standardization can significantly improve the performance of some machine learning models.

Results of Standardized Regression Algorithms

Following the standardization process, the regression models are re-evaluated, producing the following results:

1. Scaled Linear Regression: MSE ≈ 18.66
2. Scaled Lasso Regression: MSE ≈ 18.72
3. Scaled Ridge Regression: MSE ≈ 18.66

4. Scaled ElasticNet Regression: MSE ≈ 18.72
5. Scaled K Neighbors Regression: MSE ≈ 22.14
6. Scaled Decision Tree Regression: MSE ≈ 37.30

Surprisingly, even after standardization, Ridge regression maintains its leading position with the lowest MSE, which highlights its robustness and predictive capabilities.

4.2.7: Test Ridge Model on the Test Set

To validate the Ridge regression model's performance, it is subjected to a real-world test using the reserved test data.

4.2.8: Evaluate Ridge Model

In this final stage of the analysis, the Ridge regression model is assessed on the test set. The evaluation reveals an MSE of approximately 17.78. The magnitude of this MSE reflects the quality of predictions, with lower values signifying that the model's predictions are closer to the actual duration values.

Notably, this dataset contains phone call durations that range from 0.1 to 81.97 minutes, indicating substantial variability. The resulting MSE of 17.78 is a strong indication of Ridge regression's effectiveness in estimating the 'duration' variable, even within this wide range. This achievement means that the bank can confidently estimate the duration of campaign calls for each client using their respective customer profiles. Such profiles include attributes like age, job, and loan status.

In conclusion, Ridge regression emerges as the optimal choice for estimating the duration of phone calls, offering a valuable tool for the bank to predict subscription rates based on customer statistics. The relatively low MSE on the test set signifies the model's effectiveness and its potential to enhance campaign planning and decision-making processes.

4.3: Conclusion and recommendation:
4.3.1: Conclusion

The primary objective of this project is to enhance the efficiency and effectiveness of the bank's telemarketing campaign. This objective has been successfully achieved through comprehensive data analysis, visualization, and the development of analytical models. The key outcomes are:

Target Customer Profile

A well-defined target customer profile has been established. The customers most likely to respond positively to the term deposit campaign exhibit the following characteristics:

1. **Age Segmentation:** They are either young (age < 30) or senior (age > 60) individuals.
2. **Life Stage:** They are students or retired individuals.
3. **Strong Financial Position:** They maintain a bank account balance of more than 5000 euros.

Models for Response Prediction

To effectively identify and allocate marketing efforts, two key models have been developed:

1. **Classification Model:** A combination of logistic regression and ridge regression algorithms has been successfully used to build a classification model. This model predicts whether a customer is likely to accept a term deposit or not. It enables the bank to direct its marketing efforts more efficiently, allocating resources to clients who are highly likely to accept term deposits and reducing calls to those who are unlikely to do so.
2. **Estimation Model:** This model, developed using ridge regression, helps predict the expected duration of a call before it is made. By predicting the duration, the bank can optimize its marketing plan, benefiting both the bank and its clients. It leads to a more efficient telemarketing campaign, saving time and effort for both parties, and preventing clients from receiving unwanted advertisements, which, in turn, enhances customer satisfaction.

Mutual Benefits

The use of these logistic and ridge regression models creates a virtuous cycle of effective marketing, increased investments, and happier customers. It offers mutual benefits for the bank and its clients by streamlining the marketing process and improving customer interactions.

4.3.2: Recommendations

To further enhance the effectiveness of the telemarketing campaign, the following recommendations are suggested:

1. **Timing Considerations:** When executing marketing strategies, the timing of the campaign should be thoughtfully considered. The analysis has indicated that March, September, October, and December have the highest success rates. However, it's essential to gather more data and conduct ongoing analysis to ensure that this seasonal effect persists over time. If this trend is consistent, the bank should plan its telemarketing campaign during the fall and spring seasons to maximize its impact.
2. **Smarter Marketing Design:** By focusing on the right customers as identified by the classification models, the bank can expect more positive responses. Over time, this is likely to reduce the imbalance in the original dataset. To increase the likelihood of subscription, the bank should also re-evaluate the content and design of its current campaign, making it more appealing and relevant to the target customers. A well-designed and targeted campaign will enhance customer engagement and subscription rates.
3. **Better Service Provision:** With a more granular understanding of its customer base, the bank can provide tailored banking services. For example, insights into marital status and occupation provide a glimpse into a customer's life stage, while loan status indicates their overall risk profile. Armed with this information, the bank can anticipate when a customer might be interested in making an investment. By offering the right banking services to the right customers at the right time, the bank can better satisfy customer demands, thus enhancing customer loyalty and long-term relationships.

In summary, the combination of predictive models and data-driven insights equips the bank with the tools and knowledge to optimize its telemarketing campaigns. By following these recommendations, the bank can further improve its marketing strategies, target the right customers, and provide enhanced services, ultimately leading to a more successful and efficient telemarketing campaign while ensuring customer satisfaction and loyalty.