
| | |
|---|---|
| **Student Name:** | **Nathania Mbeshi** |
| **Student ID:** | **301477619** |
| **Course Code:** | **BA723** |
| **Course Name:** | **Business Analytics Capstone** |

## Contents

## Content Tables



## Table of Figures

# Executive Summary

## 0.1. Executive Introduction

This project analyzes the telemarketing campaign data to identify patterns and factors influencing customer subscriptions to term deposits. Using historical customer demographics, behavioral data, and previous campaign interactions, we applied advanced machine learning models, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting. The aim was to uncover actionable insights to improve campaign effectiveness, enhance targeting accuracy, and increase conversion rates.

The analysis revealed key predictors of subscription, such as customer age, job type, marital status, contact communication type, and duration of the last call. Models were evaluated using precision, recall, F1-score, and accuracy, with the Gradient Boosting Classifier achieving the best overall performance. These findings provide a data-driven foundation for optimizing future marketing strategies and improving return on investment.

## 0.2. Executive Objective

This analysis was conducted to address the following key objectives:

1. **Identify Key Drivers** – Determine the most influential features affecting a customer's decision to subscribe to a term deposit.
2. **Predictive Modeling** – Assess the feasibility of accurately predicting subscription outcomes using historical client data combined with macroeconomic indicators.
3. **Customer Profiling** – Develop a clear profile of clients with a high likelihood of subscribing.
4. **Conversion Optimization** – Increase term deposit conversion rates by gaining insights into both client attributes and campaign characteristics.
5. **Targeting Enhancement** – Identify predictive variables to refine future targeting strategies and improve marketing effectiveness.

## 0.3. Executive Model Description

Three predictive modeling approaches were evaluated to identify the most effective solution:

- **Logistic Regression** – Selected for its interpretability and ability to provide clear insights into the relationship between independent variables and the target outcome.
- **Decision Tree** – Applied for its capability to segment the dataset into distinct, interpretable decision rules, enabling straightforward pattern recognition.
- **Random Forest** – Chosen for its superior predictive performance and feature selection through ensemble learning. The model was further optimized by running

iterations with varying numbers of features: 15, 20, and 10. This would facilitate a comparative feature selection analysis.

## 0.4. Executive Recommendations

- **Update and Expand Data Sources:** Acquire more recent datasets to ensure the model remains relevant and reflective of current market conditions. Broaden the modeling scope to incorporate advanced machine learning algorithms such as Gradient Boosting and XGBoost.
- **Enhance Customer Profiling:** Integrate additional customer attributes including behavioral, financial, geographic, and psychographic variables to improve the model's predictive accuracy and segmentation capabilities.
- **Targeted Campaign Strategies:** Prioritize outreach to clients exhibiting longer call durations and recent contact history, as these segments demonstrate higher likelihoods of subscription.
- **Demographic-Based Targeting:** Focus marketing efforts on specific age groups and occupational categories that have historically shown elevated subscription rates.
- **Contact Frequency Management:** Avoid repeated outreach during periods of economic downturn to mitigate customer fatigue and maintain brand goodwill.
- **Model Deployment:** Implement the finalized model into operational systems to support real-time decision-making and targeted campaign execution.

# Introduction

## 1.0. Background

Telemarketing has long been an important channel for banks to promote financial products, particularly term deposits. A strong deposit base not only provides stable funding but also reassures customers about the institution's financial health, reducing concerns about liquidity and solvency (Liao, Chen, & Hsieh, 2011). In response to stricter capital requirements, banks increasingly rely on direct marketing strategies—including mail, digital, and telemarketing campaigns—to attract deposits efficiently and cost-effectively (Ivashina & Scharfstein, 2010).

Among these strategies, telemarketing has been widely used for its ability to deliver personalized engagement, real-time communication, and immediate feedback (Yan et al., 2020). Unlike digital campaigns, telemarketing allows banks to explain complex product features directly, address customer questions on the spot, and build trust through human interaction. It also complements multi-channel marketing strategies by providing flexible, adaptive engagement.

However, recent industry analyses highlight the increasing cost of telemarketing relative to its returns. According to The Telemarketing Company (TTMC, 2021), while telemarketing remains effective for lead generation and nurturing high-quality prospects, companies often face high operational and human resource costs, especially when campaigns are managed in-house or outsourced without proper optimization. Rising expenses, coupled with inconsistent conversion rates, make it critical for banks to identify and target customers more efficiently.

In this context, machine learning (ML) techniques offer a solution by predicting which clients are most likely to subscribe to term deposits, thereby optimizing resource allocation and improving campaign effectiveness (Moro et al., 2014; Feng et al., 2022; Ghatasheh et al., 2020; Yan et al., 2020). This study examines telemarketing campaign data from a Portuguese bank collected between 2008 and 2010, including client demographics, financial status, and previous campaign interactions. Using ML models such as Logistic Regression, Decision Trees, and Random Forest, we aim to forecast subscription likelihood and identify the most influential factors in campaign success.

By combining data-driven targeting with the personal engagement of telemarketing, banks can enhance conversion rates, reduce costs, and ensure that telemarketing continues to be a viable and strategic channel for customer acquisition.

## 2.0. Problem Statement

The Portuguese bank faces challenges in optimizing its telemarketing campaigns for term deposits due to the high costs and low efficiency of cold-calling approaches. Many calls fail to convert potential clients, resulting in wasted resources and suboptimal campaign performance.

To address this, the bank aims to leverage data-driven techniques to better target prospective customers. Specifically, the objective is to develop a binary classification model that predicts whether a client is likely to subscribe to a term deposit (y = yes/no), using information from client profiles, previous contact history, and relevant historical subscriptions.

By accurately identifying clients with a high probability of subscription, the bank can improve the effectiveness of its marketing campaigns, reduce operational costs associated with unproductive calls, and increase overall conversion rates. This approach will enable the bank to allocate resources more strategically and maximize the return on investment from telemarketing efforts.

## 3.0. Objectives & Measurement

The objectives of this project were:

- To apply and demonstrate the knowledge and skills acquired in predictive modeling and machine learning.
- To address the bank's business problem by developing a reliable model that predicts the likelihood of a client subscribing to a term deposit.

The effectiveness of the predictive models will be evaluated using the following metrics:

- **Accuracy:** The overall proportion of correct predictions.
- **Recall (Sensitivity):** The ability of the model to correctly identify clients who will subscribe.
- **F1 Score:** The harmonic mean of precision and recall, providing a balanced measure for imbalanced classes.
- **ROC-AUC**: A measure of the model's ability to distinguish between subscribing and non-subscribing clients.

## 4.0. Assumptions and Limitations

The following were the assumptions made for the sake of the project analysis

1. The dataset under analysis includes only clients who were contacted during the telemarketing campaigns and does not represent the entire customer base.
2. Economic conditions and market factors have evolved since 2010, which may influence customer behavior and the generalizability of the findings.
3. The report is intended for shareholders and focuses on evaluating the performance and strategic relevance of the bank's telemarketing segment, including considerations for potentially discontinuing this channel.

## Data Sources

The analysis in this project is based on the bank-full.csv dataset obtained from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/222/bank+marketing). This dataset pertains to direct marketing campaigns conducted via phone calls by a Portuguese banking institution, with the classification objective of predicting whether a client subscribes to a term deposit.

## 5.0. Data Set Introduction

Dataset Overview:

- Number of Instances: 45,211
- Number of Features: 17 input variables
- Data Type: Multivariate, including categorical and numerical attributes
- Missing Values: None

The bank-full.csv dataset is an older version of the dataset with fewer input variables compared to bank-additional-full.csv. The dataset is ordered chronologically and captures campaigns from May 2008 to November 2010, similar to the data analyzed in Moro et al. (2014). It provides comprehensive information about client demographics, financial status, and historical contact details relevant to telemarketing campaigns.

Input Features:

- Client Data: age, job, marital, education, default, balance, housing, loan
- Contact Details: contact, day, month, duration
- Campaign History: campaign, pdays, previous, poutcome

Target Variable:

- y – indicates whether the client subscribed to a term deposit (binary: "yes" or "no").

This dataset allows the application of predictive modeling techniques to identify potential subscribers and improve the efficiency of telemarketing campaigns. Its combination of demographic, financial, and campaign-related variables provide a rich basis for analyzing client behavior and developing machine learning models to forecast subscription likelihood.

| | age | job | marital | education | default | balance | housing | loan | contact | day | month | duration | campaign | pdays | previous | poutcome | y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 58 | management | married | tertiary | no | 2143 | yes | no | unknown | 5 | may | 261 | 1 | -1 | 0 | unknown | no |
| 1 | 44 | technician | single | secondary | no | 29 | yes | no | unknown | 5 | may | 151 | 1 | -1 | 0 | unknown | no |
| 2 | 33 | entrepreneur | married | secondary | no | 2 | yes | yes | unknown | 5 | may | 76 | 1 | -1 | 0 | unknown | no |
| 3 | 47 | blue-collar | married | unknown | no | 1506 | yes | no | unknown | 5 | may | 92 | 1 | -1 | 0 | unknown | no |
| 4 | 33 | unknown | single | unknown | no | 1 | no | no | unknown | 5 | may | 198 | 1 | -1 | 0 | unknown | no |

**Figure 1: First few columns of the dataset**

## 6.0. Exclusions

During preprocessing, certain modifications were made to optimize model performance and interpretability:

- The duration variable was excluded from modeling due to its high correlation with the target variable (y), which could otherwise lead to data leakage and artificially inflated performance.
- The contact variable was also removed, as contemporary telemarketing campaigns predominantly utilize cellular communication, rendering the original contact type less informative for predictive modeling.

These exclusions ensured that the dataset was relevant, and appropriate for building robust machine learning models to predict term deposit subscription likelihood.

## 6.1. Initial Data Cleansing or Preparation

The dataset utilized in this study was complete, containing no explicit missing values and no duplicate records. Initial preprocessing steps were undertaken to ensure data quality, feature relevance, and suitability for modeling.

**Duplicate Check and Removal:**

The dataset was examined for duplicate records. No duplicate rows were identified, confirming the dataset's uniqueness.

**Handling 'Unknown' Values:**

- All object-type columns were reviewed to identify unique values, revealing the presence of 'unknown' entries in several features.
- For job, marital, and education, 'unknown' values were replaced with the mode of each respective column to maintain data quality.
- The poutcome variable retained its 'unknown' category, as it provides meaningful information on previous campaign outcomes.

The figures below demonstrate the change after the imputation:

```
Columns with 'unknown' values and their counts:
job            288
education     1857
contact      13020
poutcome     36959
dtype: int64
```

Figure 2: Columns containing unknown values and their respective counts

```
Columns with 'unknown' values and their counts:
poutcome    36959
dtype: int64
```

Figure 3: Dataset after unknown values were imputed and categorical labels standardized

**Feature Exclusions:**

- The duration variable was excluded due to its strong correlation with the target variable (y), which could introduce data leakage.
- The contact variable was removed, as in modern telemarketing all contacts are assumed to occur via cellular communication, rendering the distinction obsolete.

**Missing Value Assessment**:

A completeness check confirmed the absence of explicit null values.

**Variable Classification:**

To support appropriate preprocessing, features were grouped as follows:

- Numerical Variables: age, balance, day, campaign, pdays, previous
- Categorical Variables: job, marital, education, default, housing, loan, month, poutcome

**Categorical Variables:**

- Binary categorical variables (default, housing, loan) were converted into numerical binary values (0 and 1).
- The target variable (y) was also mapped to binary values (yes = 1, no = 0).

These preprocessing steps ensured that the dataset was ready for further analysis and machine learning model training.

## 7.0. Data Dictionary

| Variable | Encoding / Transformation | Notes |
|---|---|---|
| job | One-hot encoding | Each job type becomes a binary column (e.g., job_admin., job_blue-collar) to handle categorical data for machine learning algorithms. |
| marital | One-hot encoding | Divorced' includes both divorced and widowed individuals. Missing or unknown values are imputed using the mode. |
| education | One-hot encoding | Grouping is done: basic (basic.4y, basic.6y, basic.9y), high. School, tertiary (professional. Course, university. Degree), illiterate and unknown values are imputed using the mode. |
| default, housing, loan | Binary encoding | 'yes' = 1, 'no' = 0. |
| contact | One-hot encoding | Includes 'cellular', 'telephone', 'unknown'. |
| month | One-hot encoding | Each month converted into a separate binary variable. |
| poutcome | One-hot encoding | Categories: 'unknown', 'other', 'failure', 'success'. |

| | | |
|---|---|---|
| duration | Numeric | Dropped; note that it is highly correlated with subscription likelihood and should not be used as a predictive feature if simulating real-world unseen data (post-contact). |
| pdays | Numeric / Transformation | -1 indicates client was not previously contacted; |
| previous | Numeric | Used as-is. |
| campaign | Numeric | Dropped. Number of contacts in the current campaign; can be treated as continuous. |
| balance | Numeric / Transformation | Euros. |
| age | Numeric / Transformation | Used as-is or binned into categories. |
| y | Binary encoding | Target variable: 'yes' = 1, 'no' = 0. |

**Table 1: Data dictionary**

## Data Exploration

## 8.0. Data Exploration Techniques

A comprehensive exploratory data analysis (EDA) was conducted to understand the characteristics, distributions, and relationships within both numerical and categorical variables.

**Numerical Variables**

The numerical features were examined using multiple techniques to capture central tendency, dispersion, and potential relationships with the target variable. The following steps were undertaken:

- Descriptive Statistics: The describe function was applied to generate count, mean, standard deviation, minimum, maximum, and quartile values for all numerical columns.
- Distribution Analysis: Histograms were plotted for each numerical variable to visualize their frequency distributions.
- Skewness: The skewness of each numerical feature was calculated to assess the degree of distribution asymmetry.

- Target Relationship Analysis: Boxplots were generated to compare the distribution of numerical features across the target variable (y).
- Correlation Analysis: A heatmap was created to visualize linear relationships between numerical variables, including binary columns converted to float for consistency.

**Categorical Variables**

The categorical features were analyzed to understand category frequencies and their potential influence on the target variable:

- Descriptive Summary: The describe(include='object') method was used to obtain counts, unique category values, the most frequent category, and its frequency for each categorical variable.
- Category Distribution: Bar charts were generated to display the count of each category within each categorical feature.
- Target Relationship Analysis: plot visualizations were used to examine the relationship between categorical features and the target variable (y).

**Target Variable Distribution**

- The percentage distribution of both classes in the target variable (yes and no) was calculated and displayed.

## 8.1. Data Visualization

**A. Numerical Variables**

The following section presents the visualizations and corresponding explanations of EDA process for the numerical variables:

- **Descriptive Statistics:**
  - **Age:** Mean ~41 years (range 18–95), with the majority between 33 and 48.
  - **Balance:** Mean €1,362, median €448; highly right-skewed (range -€8,019 to €102,127) with notable outliers**.**
  - **Campaign:** Mean 2.76 contacts (range 1–63); most clients contacted ≤3 times.
  - **Pdays:** Mean 40 days since last contact; -1 indicates no prior contact; 75th percentile -1, max 871**.**
  - **Previous:** Mean 0.58 prior contacts; 75th percentile 0; max 275, indicating few prior interactions for most clients.

- **Default, Housing, Loan: Binary variables;** ~1.8% defaulted, 55.6% have housing loans, 16% have personal loans.

| | age | balance | campaign | pdays | previous | default | housing | loan |
|---|---|---|---|---|---|---|---|---|
| count | 45211.000000 | 45211.000000 | 45211.000000 | 45211.000000 | 45211.000000 | 45211.000000 | 45211.000000 | 45211.000000 |
| mean | 40.936210 | 1362.272058 | 2.763841 | 40.197828 | 0.580323 | 0.018027 | 0.555838 | 0.160226 |
| std | 10.618762 | 3044.765829 | 3.098021 | 100.128746 | 2.303441 | 0.133049 | 0.496878 | 0.366820 |
| min | 18.000000 | -8019.000000 | 1.000000 | -1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 33.000000 | 72.000000 | 1.000000 | -1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 50% | 39.000000 | 448.000000 | 2.000000 | -1.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 |
| 75% | 48.000000 | 1428.000000 | 3.000000 | -1.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 |
| max | 95.000000 | 102127.000000 | 63.000000 | 871.000000 | 275.000000 | 1.000000 | 1.000000 | 1.000000 |

Figure 4: Descriptive statistics of the numerical variables

- **Distribution Analysis:**
  - **Age** – The age distribution exhibits a peak in the late 30s to early 40s, with fewer clients in the younger and older age ranges. The distribution is moderately right-skewed.

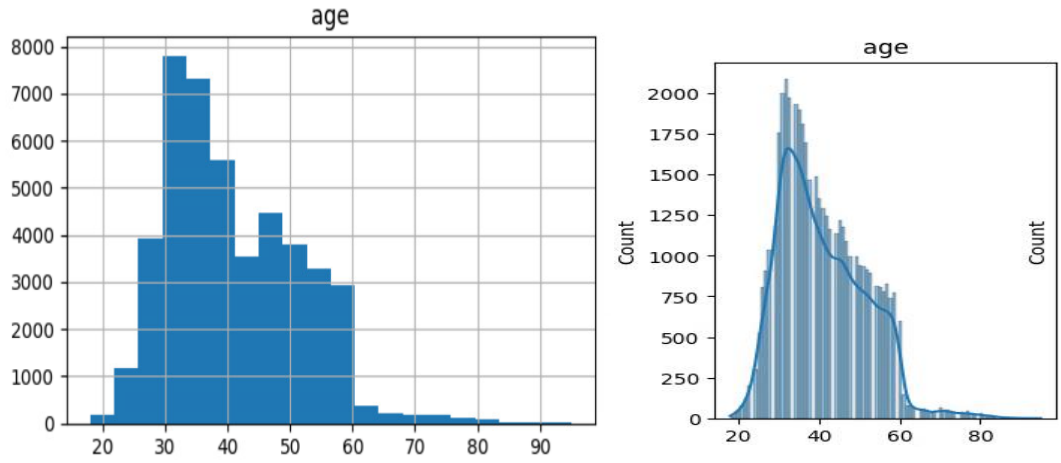


Figure 5: Age distribution

- **Balance** – The balance distribution is strongly right-skewed. Most clients maintain a balance near zero or a modest positive amount, while a small number of clients have very high balances, forming a long right tail and indicating potential outliers.

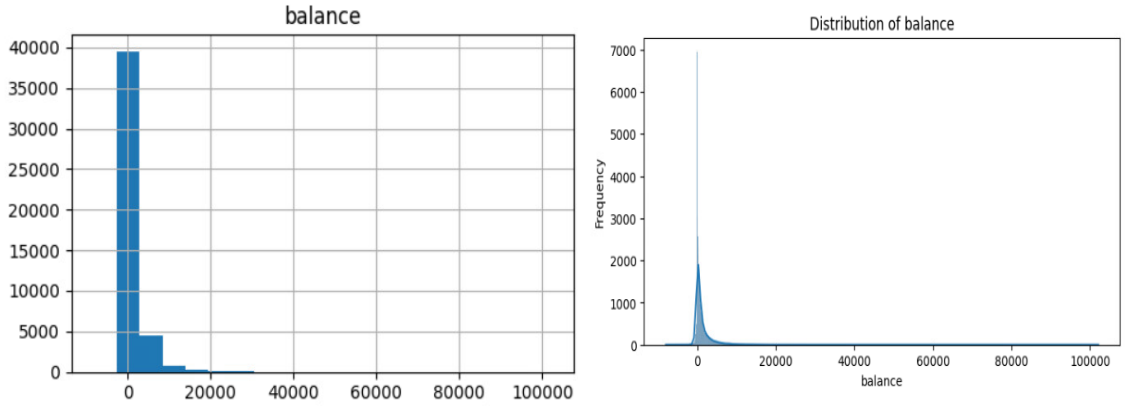


Figure 6: Balance distribution

- **Campaign** – Similar to balance, the number of contacts per campaign is right-skewed. The majority of clients were contacted only a few times, while a small subset experienced a much higher number of contacts, represented by the long tail

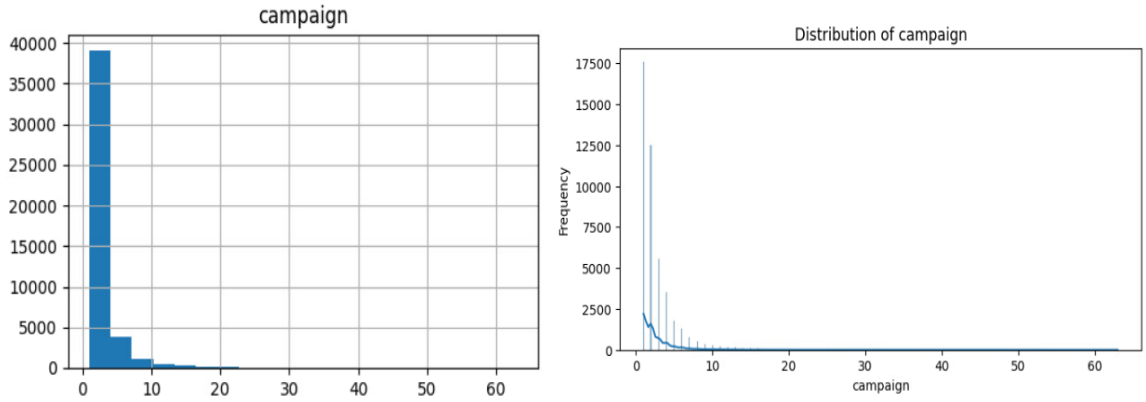


Figure 7: Campaign distribution

- **Pdays** – The distribution is dominated by a large proportion of clients with a value of -1, indicating no prior contact. For clients previously contacted, the number of days since the last contact shows a right-skewed distribution.

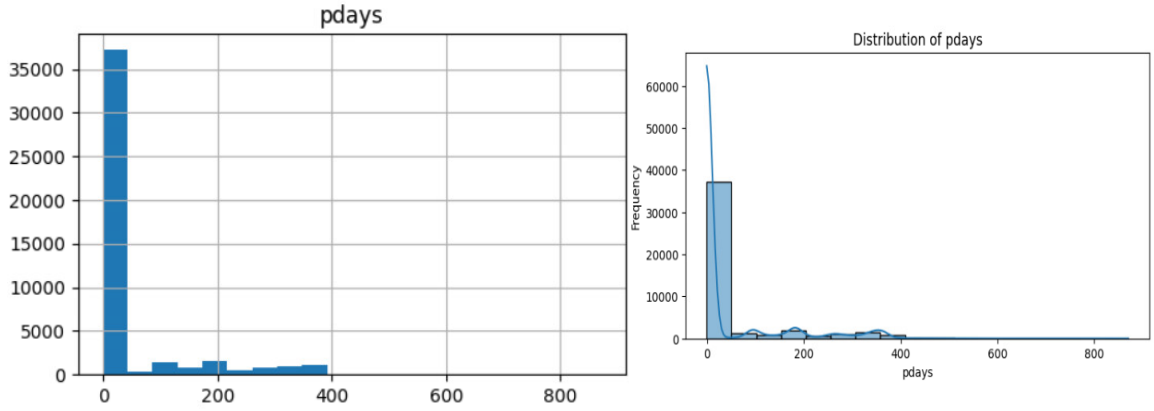


Figure 8: Number of days that passed by after the client was last contacted

- **Skewness:**
  The skewness values indicate that most numerical variables are right-skewed, with previous (41.85), balance (8.36), and campaign (4.90) showing particularly strong positive skewness. Only housing exhibits slight left skew (-0.22), suggesting its distribution is nearly symmetric.

| Variable | Skewness |
|---|---|
| age | 0.684795 |
| balance | 8.360031 |
| campaign | 4.898488 |
| pdays | 2.615629 |
| previous | 41.845066 |
| default | 7.245135 |
| housing | -0.224759 |
| loan | 1.852556 |

Table 2: Numerical variables skewness figures

- **Target Relationship Analysis**
  - **Subscription by age:** Median ages are similar, though slightly higher for subscribers, indicating older clients may be marginally more likely to subscribe.

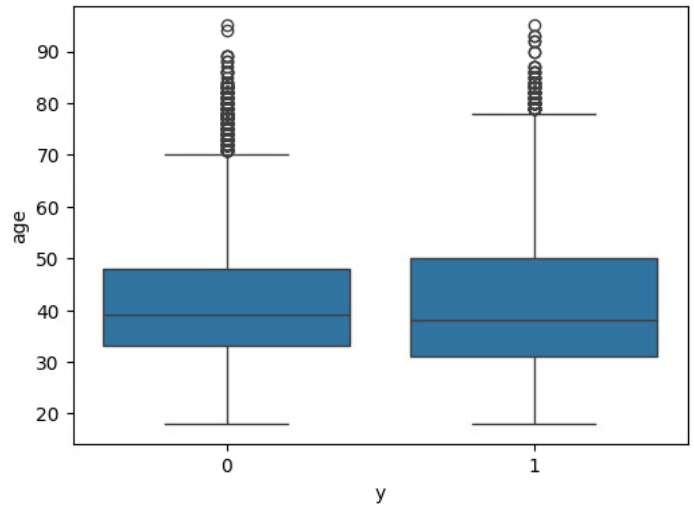


Figure 9: Subscription by age

- **Subscription by account balance:** Higher balances are associated with subscription, as the `Yes' group shows more extreme positive outliers.

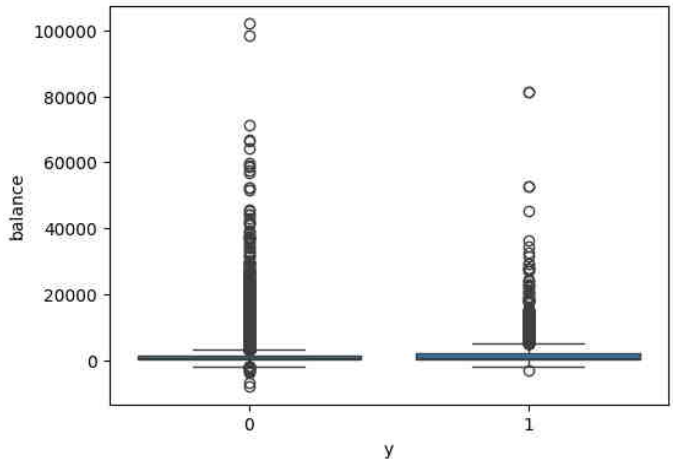


Figure 10: Subscription by account balance

- **Subscription by number of campaigns:** Most clients cluster at low campaign counts, suggesting the number of current contacts is not a strong predictor.

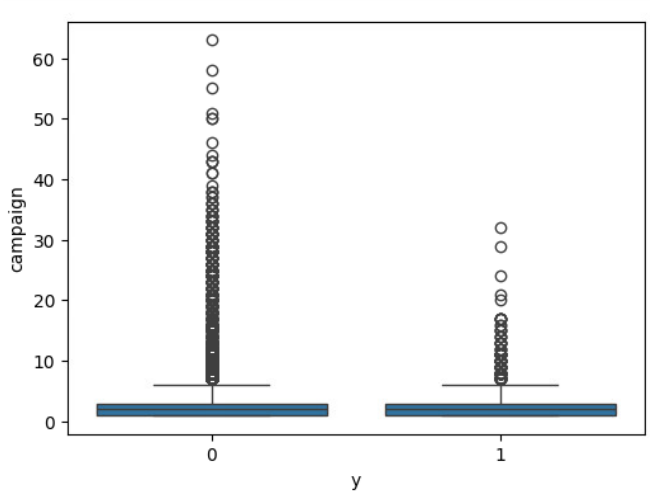


Figure 11: Subscription by number of campaigns

- **Subscription by previous & pdays:** Subscribers tend to have been contacted before, with higher pdays and previous values, implying prior engagement increases subscription likelihood**.**

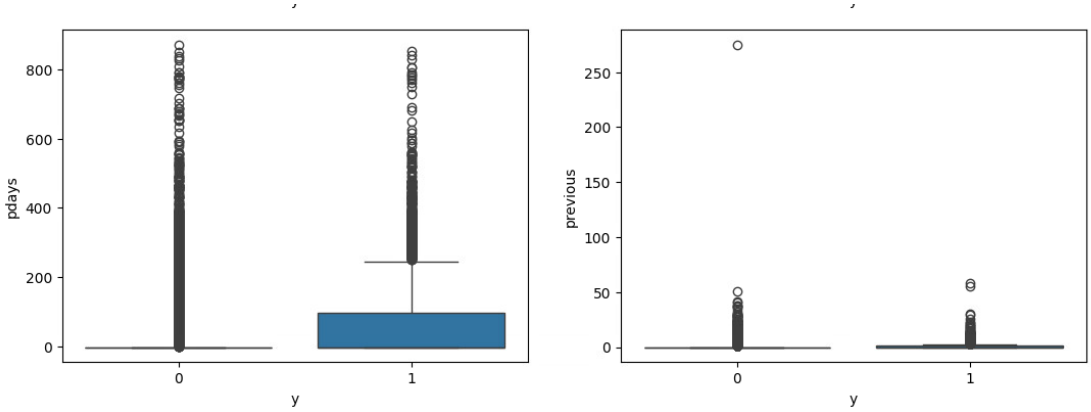


Figure 12: Subscription by previous & pdays

- **Default:** Very few clients who defaulted subscribe; most are non-defaulters**.**

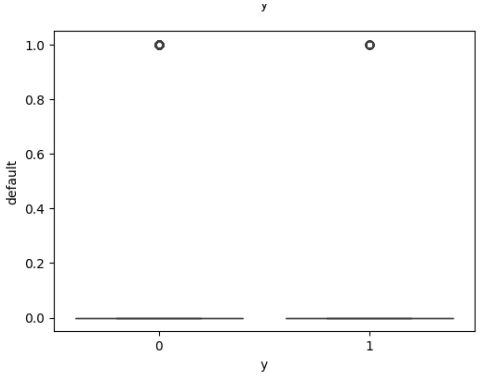


Figure 13: Subscription by rate of default

- **Housing & Loan:** Clients without housing or personal loans are more likely to subscribe, while those with loans are concentrated in the `No' group

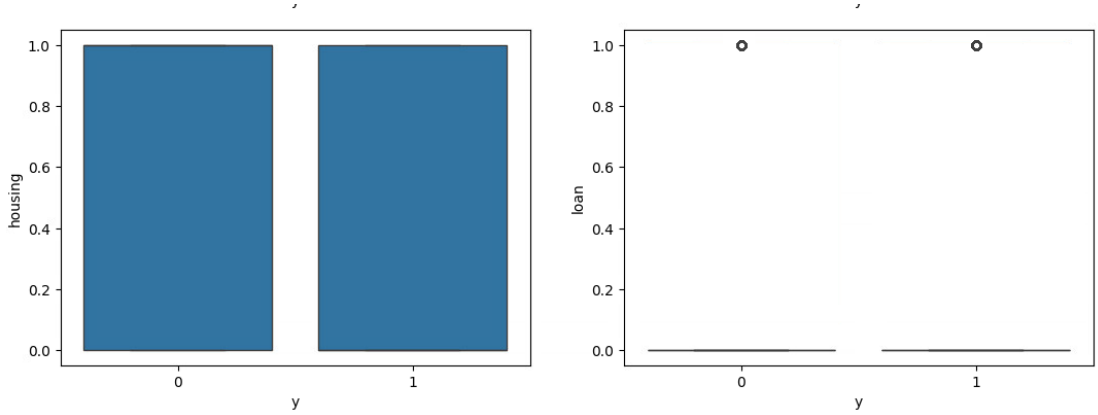


Figure 14: Subscription if client has housing and/or personal loans

- **Correlation Analysis:**
  - **Pdays & Previous (0.45):** Shows the strongest positive correlation, indicating that clients with more previous contacts tend to have a higher number of days since the last contact.
  - **Housing & Age (-0.19):** Weak negative correlation, suggesting younger clients are slightly more likely to have housing loans.
  - **Balance & Age (0.10):** Very weak positive correlation, indicating older clients may have marginally higher balances.
  - **Campaign & Previous (0.04):** Very weak positive correlation between current campaign contacts and prior contacts.
  - **Other Variables:** Most other numerical pairs exhibit negligible correlations, implying minimal linear dependence.

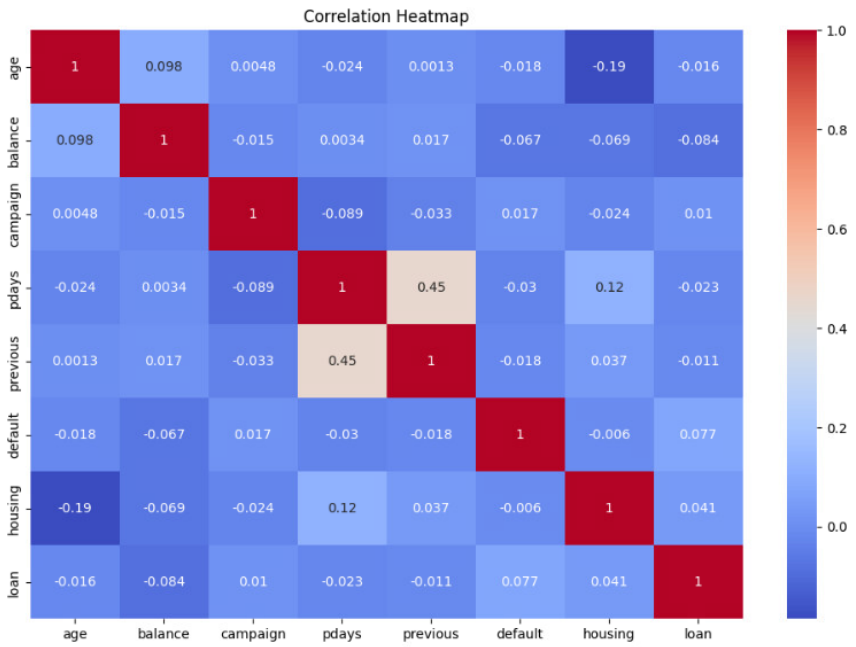


**Figure 15: Correlation matrix of the numerical variables**

## B. Categorical Variables

The following section presents the visualizations and corresponding explanations of EDA process for the categorical variables.

- **Descriptive Summary:**
  - **Job:** 'Blue-collar' is the most common job type. Management, technician, and admin roles have high counts, while students and retirees show a relatively higher subscription rate.
  - **Marital:** 'Married' is most frequent. Single and divorced clients exhibit a slightly higher proportion of subscriptions compared to married clients.
  - **Education:** 'Secondary' education is most common. Clients with tertiary education show a marginally higher subscription rate.
  - **Day:** Contact distribution varies across days, with minor variations in subscription rates; no strong daily pattern is evident.
  - **Month:** 'May' has the highest contact volume, but months like March, April, September, October, and December show a higher proportion of subscriptions, highlighting the impact of timing.
  - **Poutcome**: Dominated by 'unknown'. Previous campaign 'success' strongly predicts subscription, whereas 'failure' and 'other' are associated with non-subscription, emphasizing the importance of prior campaign outcomes.

| | job | marital | education | day | month | poutcome |
|---|---|---|---|---|---|---|
| **count** | 45211 | 45211 | 45211 | 45211 | 45211 | 45211 |
| **unique** | 11 | 3 | 3 | 31 | 12 | 4 |
| **top** | blue-collar | married | secondary | 20 | may | unknown |
| **freq** | 10020 | 27214 | 25059 | 2752 | 13766 | 36959 |

Figure 16: Descriptive summary of the categorical variables

- **Category Distribution and Target Relationship Analysis:**
  - **Job:** 'Blue-collar' is the most common job type. Management, technician, and admin roles have high counts, while students and retirees show a relatively higher subscription rate**.**

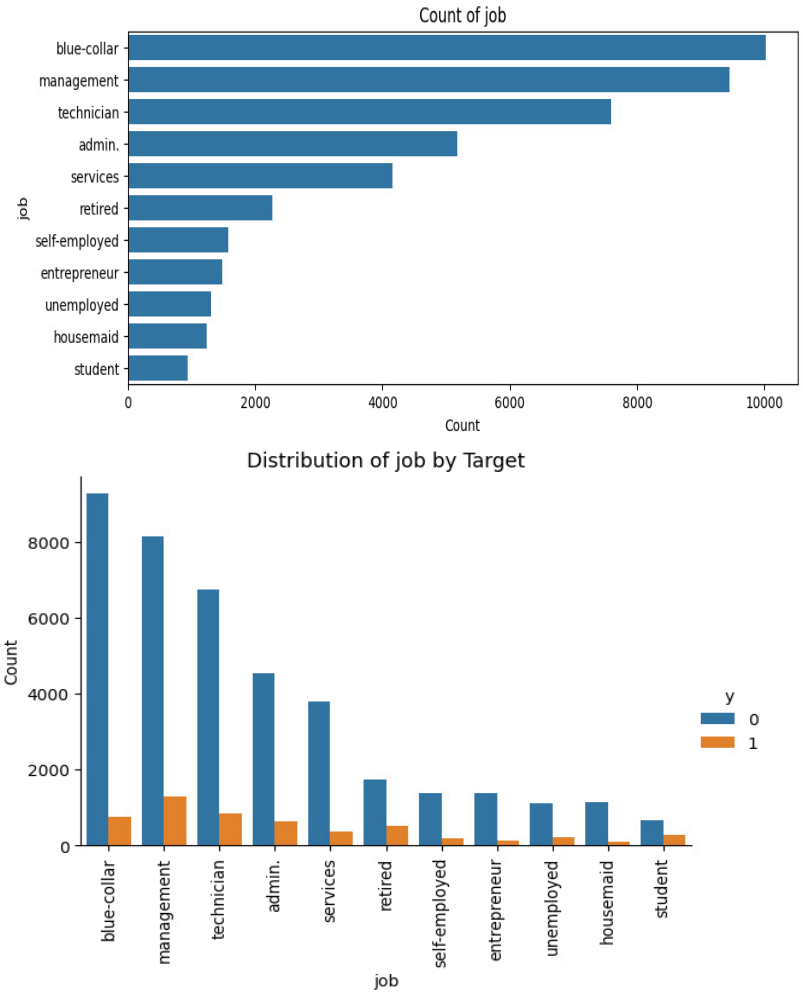


Figure 17: Job distribution and relationship with target

- **Marital:** 'Married' is most frequent, but single and divorced clients have a slightly higher likelihood of subscribing**.**

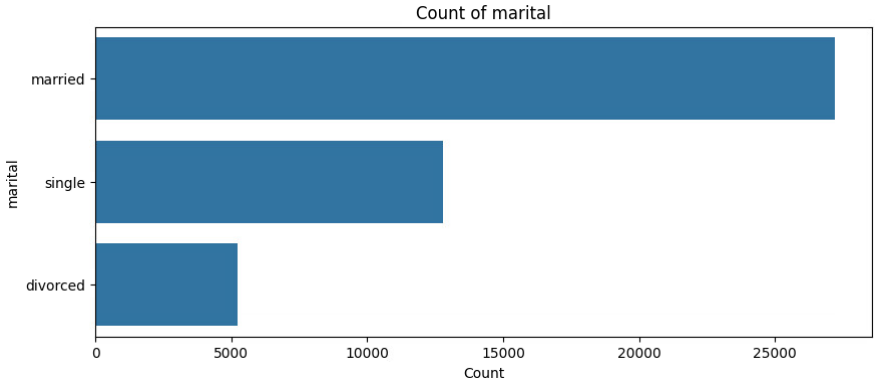


**Figure 18: Marital status count**

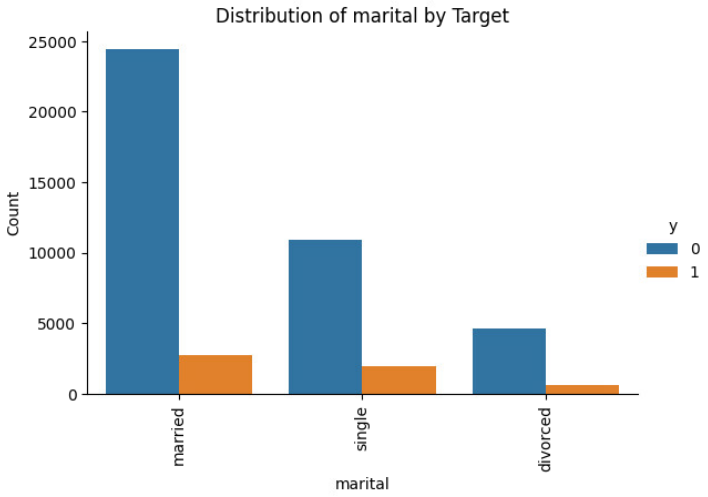


**Figure 19: Relationship between marital status and target**

- **Education:** 'Secondary' education is most common; clients with tertiary education show a marginally higher subscription rate**.**

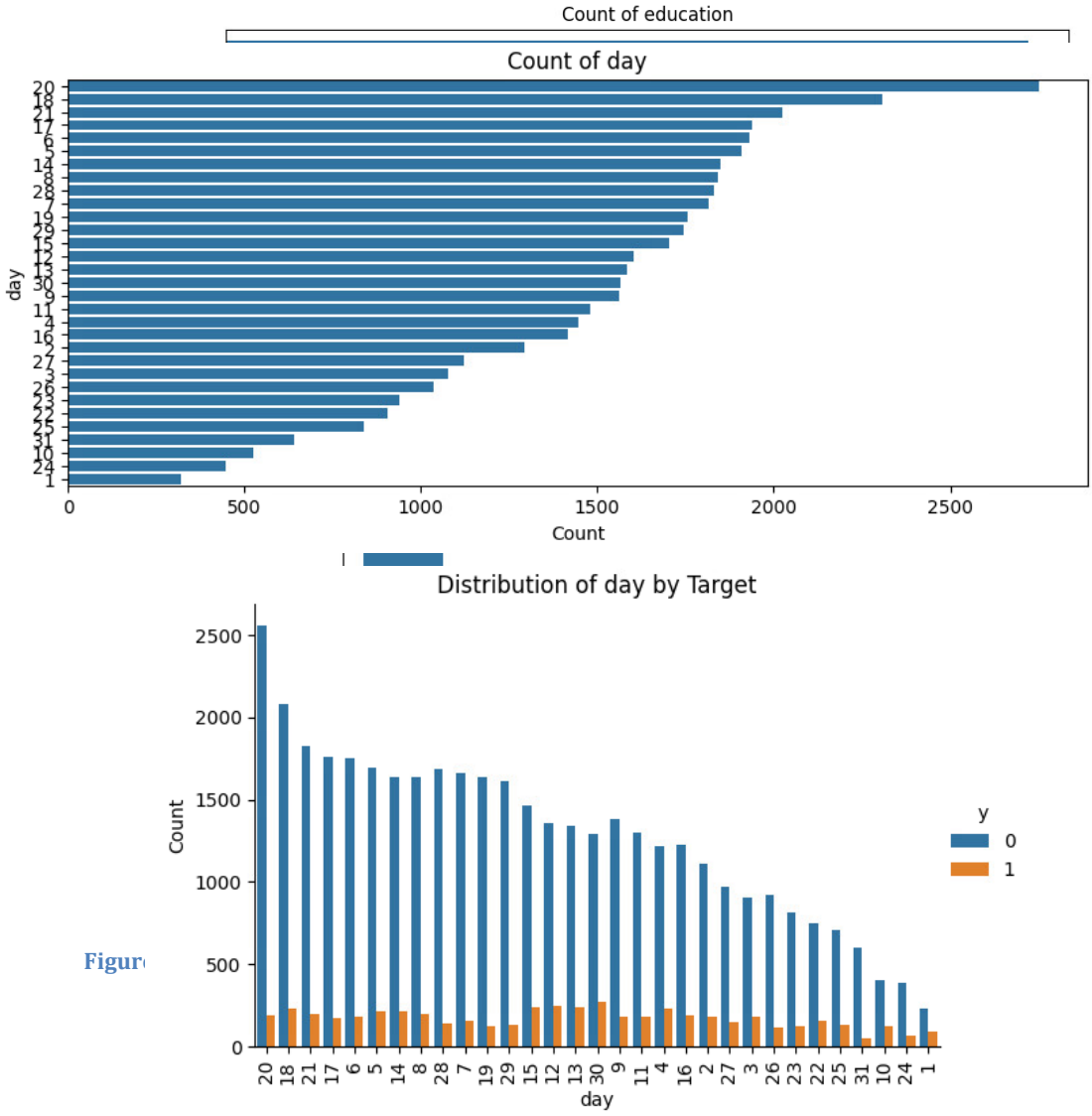


**Figure 21: Distribution of day of week contacted and the relationship with target variable**

- **Day:** Contact distribution is higher closer to the end of the month and more specifically on the 20th day**.**

- **Month:** 'May' has the highest contact volume, and months like March, April, September, October, and December exhibit lower subscription proportions, indicating timing matters**.**

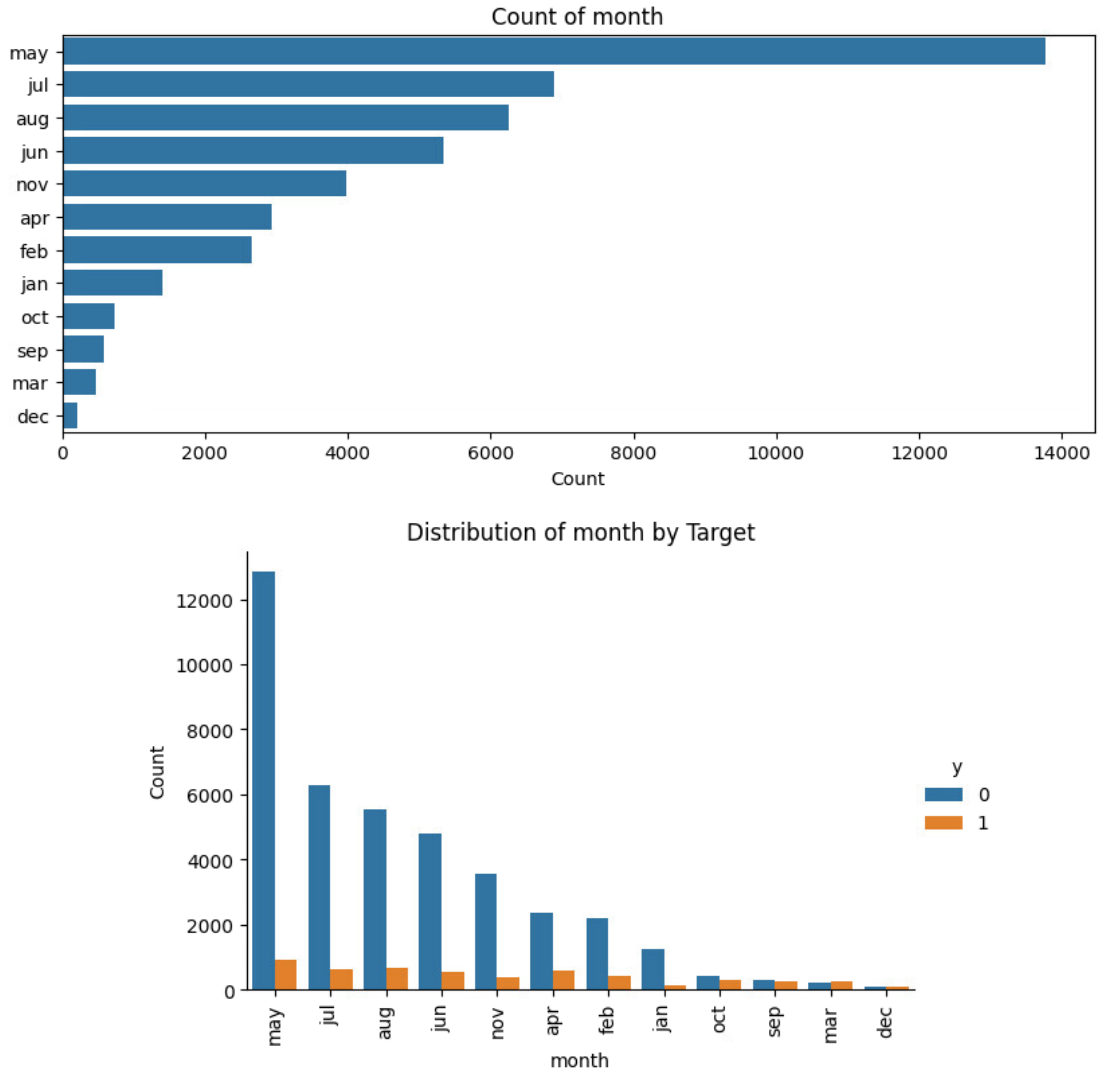


Figure 22: Distribution of month by target and relationship with target variable

- **Poutcome:** Dominated by 'unknown'. Previous campaign success strongly predicts subscription, whereas failure or other outcomes are linked to non-subscription.

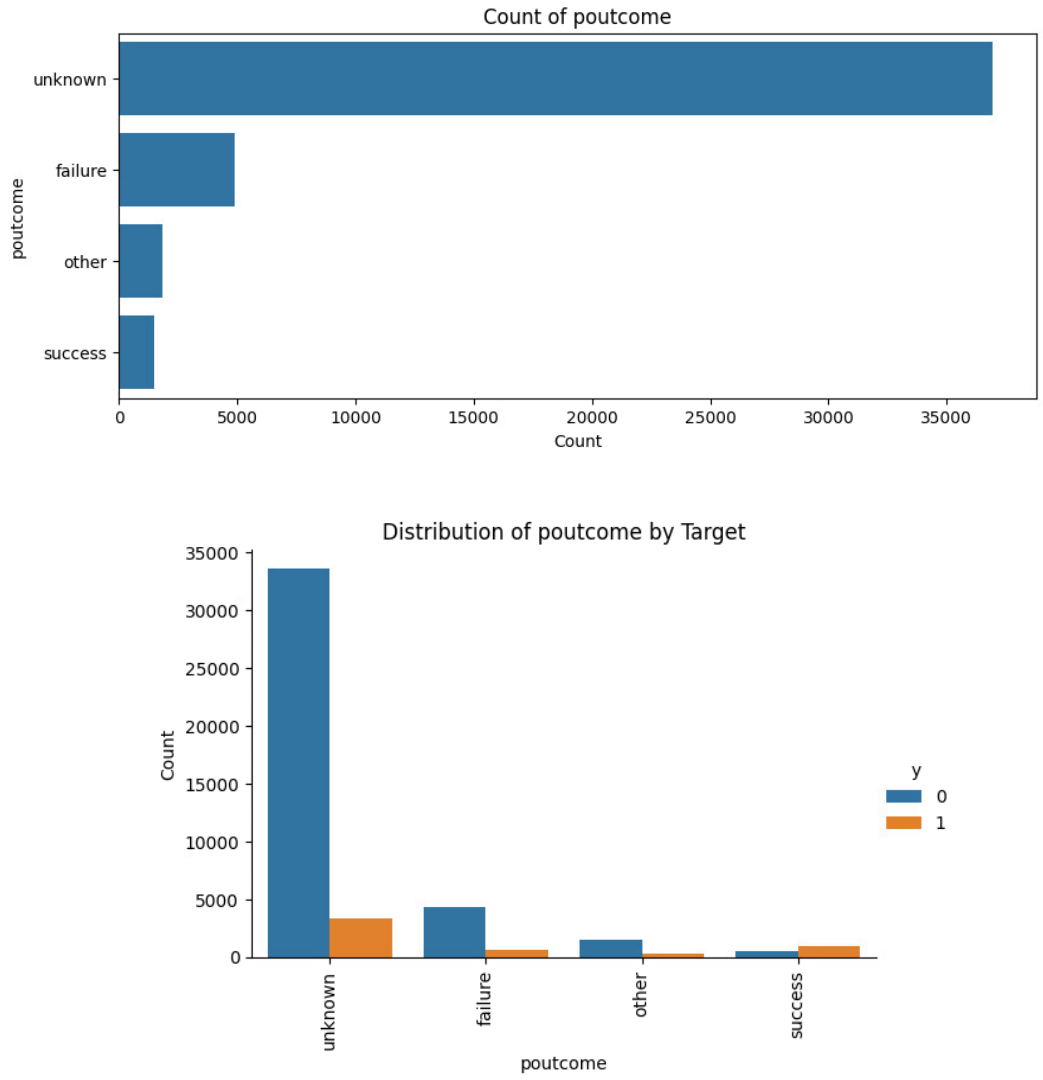


Figure 23: poutcome distribution and relationship with target variable

## Target Variable Distribution

- The target variable is highly imbalanced, with 'No' responses comprising approximately 88% of the dataset. This imbalance indicates that resampling or balancing techniques should be applied prior to model development to ensure reliable and unbiased predictions.

```
percentage of NO and YES
 y
0    88.30152
1    11.69848
Name: count, dtype: float64
```

Figure 24: Percentage distribution of target variable

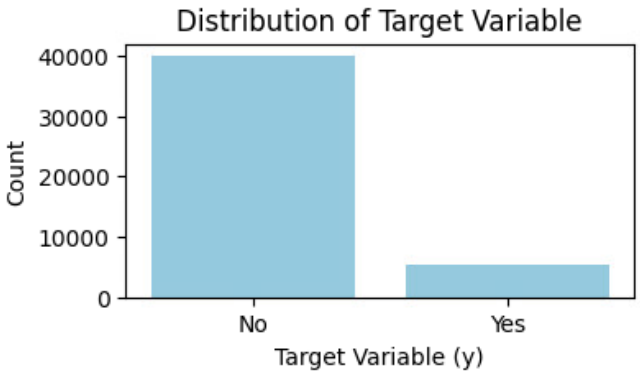


Figure 25: Distribution of target variable

## 9.0. Summary

The dataset has been thoroughly explored and cleaned, and it is now ready for transformation and feature engineering.

- **Numerical Variables:**
  - **Age:** Most clients are between 33 and 48 years, with a mean of 41. The distribution is moderately right-skewed, with a few older clients representing outliers.
  - **Balance**: Highly right-skewed, with the majority of clients having small positive balances. Some extreme outliers exist, reflecting very high or negative balances.
  - **Campaign**: The number of contacts per campaign is generally low (1–3), though some clients were contacted many times. This distribution is right-skewed.
  - **Pdays & Previous**: Most clients were not contacted in previous campaigns (pdays = -1, previous = 0). Those who were previously contacted show higher pdays and previous values, suggesting prior engagement increases subscription likelihood.
  - **Binary Variables (Default, Housing, Loan):** Most clients have not defaulted; slightly more than half have housing loans, and a smaller proportion have personal loans.
- **Categorical Variables:**
  - **Job, Marital, Education:** Blue-collar, married, and secondary education are the most frequent categories. Students and retired individuals, single and divorced clients, and those with tertiary education have slightly higher subscription rates relative to their group totals.
  - **Day & Month:** Client contacts are predominantly concentrated toward the end of the month (17th, 18th, 20th, and 21st). While May records the highest contact volume, months such as March, April, September, October, and December exhibit lower subscription rates, indicating that the timing of contacts may influence client responses.

- **Poutcome:** Heavily dominated by unknowns. Among known outcomes, prior campaign success is strongly associated with subscription, highlighting the importance of previous engagement.

- **Target Variable**
  - The target variable is highly imbalanced, with 'No' responses comprising approximately 88% of the dataset.
  - This imbalance indicates that resampling or balancing techniques should be applied prior to model development to ensure reliable and unbiased predictions.

The dataset exhibits clear patterns in both numerical and categorical features relevant for predicting subscription behavior. The target variable y is highly imbalanced, with 88% of clients classified as 'No' and 12% as 'Yes', highlighting the need for resampling techniques such as SMOTE or under sampling before model development. These insights provide a foundation for feature engineering, encoding, and model training to improve predictive performance.

## Data Preparation and Feature Engineering

### 10.0. Data Preparation

Exploration of the target variable *y* revealed a significant imbalance, with approximately 88% of instances labeled 'No' and only 12% labeled 'Yes'. Such imbalance can cause predictive models to be biased toward the majority class, often resulting in poor performance for the minority class. To address this, resampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) were applied. SMOTE generates synthetic examples of the minority class by interpolating between existing minority instances, effectively increasing their representation in the dataset.

This allows the model to better learn patterns associated with the 'Yes' class, improving overall predictive accuracy, recall, and F1-score, and reducing the risk of the model ignoring minority cases. By balancing the classes, we provide the model with a fairer, more informative dataset, enabling it to make more reliable predictions for both classes.

```
Value counts before SMOTETomek:
y
0    31937
1     4231
Name: count, dtype: int64

Value counts after SMOTETomek:
y
0    31139
1    23154
Name: count, dtype: int64
```

**Figure 26: Balancing dataset with SMOTETomek**

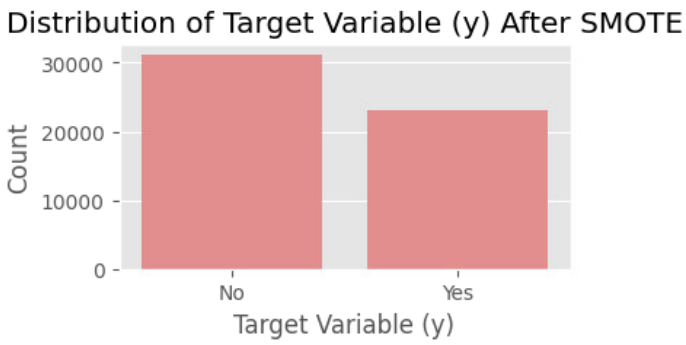


**Figure 27: Target variable distribution after balancing dataset**

# Model Exploration

## 11.0. Modeling Approach/Introduction

The dataset was divided into training and testing sets. Allocating 20% of the data to the testing set and 80% to the training set. Then initializing the random number generator so that the split is consistent across runs.

**Purpose of Splitting:**

Dividing the dataset allows the model to learn patterns from the training data and then be evaluated on unseen data. This approach provides an unbiased estimate of model performance and helps prevent overfitting, ensuring that the model generalizes well to new, real-world data.

**Result of Splitting**

Following the train–test split, the dataset dimensions were as follows:

- x_train: (36,168, 14) — 80% of the original dataset (45,211 rows), containing 14 feature columns after removing duration and contact.
- x_test: (9,043, 14) — 20% of the dataset, maintaining the same feature set as x_train.
- y_train: (36,168,) — Target variable for the training set, aligned with x_train rows.
- y_test: (9,043,) — Target variable for the test set, aligned with x_test rows.

After preprocessing and applying SMOTETomek resampling to address class imbalance:

- x_sm: (54,293, 42) — The row count increased due to SMOTE oversampling the minority class, while the column count increased from 14 to 42 because one-hot encoding expanded categorical variables into multiple binary features.

- y_sm: (54,293,) — Target variable for the resampled training set, matching the row count of x_sm.

This confirms that:

- The train–test split preserved class proportions via stratification.
- One-hot encoding substantially increased feature dimensionality.
- SMOTETomek effectively balanced the target variable by increasing minority class samples, ensuring that the model receives a more representative distribution during training.

## 12.0. Model Technique #1 – Decision Tree

A Decision Tree is a supervised learning method that predicts outcomes by splitting data into branches based on decision rules derived from input features. Each split represents a question about the data (e.g., *Is age > 40?*), and the process continues until a final decision is reached at a leaf node.

For the initial model, a Decision Tree Classifier was trained on the resampled training data to address class imbalance, enabling the model to learn patterns from both majority and minority classes. This approach achieved an overall accuracy of 83%, but the ability to identify the minority class ("Yes") remained limited, with an F1-score of only 0.32, indicating a bias toward predicting the majority class.

To improve performance, hyperparameter tuning with cross-validation explored different split criteria, tree depths, and minimum sample requirements for splits and leaves. The best configuration used the Gini criterion, a maximum depth of 30, and default minimum sample thresholds.

The full decision tree generated under these parameters was too large to display, so a simplified version was produced. For this version, the maximum depth was reduced to 3, retaining the most important splits while improving interpretability and preserving a balance between accuracy and clarity.

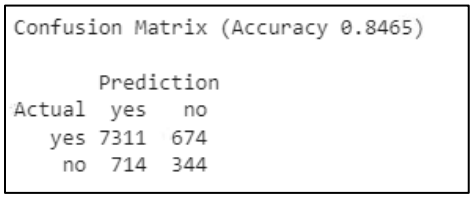


**Figure 28: Decision tree confusion matrix**

## 12.1. Results and Interpretation

The optimized Decision Tree achieved an overall accuracy of 85%, with a marginal improvement in the minority class F1-score from 0.32 to 0.33. While this gain was modest, it indicated a slight enhancement in the model's ability to correctly identify 'Yes' responses.

The confusion matrix revealed that predictions for the 'No' class were largely accurate, whereas a considerable proportion of 'Yes' cases were still misclassified as 'No'. This reinforces a known limitation of Decision Trees on imbalanced datasets that they tend to favor the majority class unless further strategies, such as cost-sensitive learning or advanced ensemble methods, are applied.

To improve interpretability, a simplified Decision Tree with a maximum depth of 3 was plotted. This compact structure made it easier to follow decision paths and identify key predictive features. Each node in the tree represents a decision rule based on a specific feature, showing the impurity (Gini score), the number of samples at that node, the class distribution, and the predicted class.

For example, early splits often involved features such as

- previous campaign outcome that was successful strongly indicated a 'Yes' prediction when positive, or
- previous contact history influenced outcomes based on recency and frequency of contact.
- Other splits reflected seasonal patterns such as the month of May or client financial information such as the client's account balance

By tracing different paths, it became clear how specific combinations of feature values led the model toward predicting either 'Yes' or 'No'. For instance, clients with a successful previous outcome were predicted as 'Yes' almost regardless of other factors, while clients contacted in less favorable months with no prior positive outcome were more likely to be classified as 'No'.

This simplified tree captures the most influential decision rules without the overwhelming complexity of the full model, balancing predictive performance with interpretability.

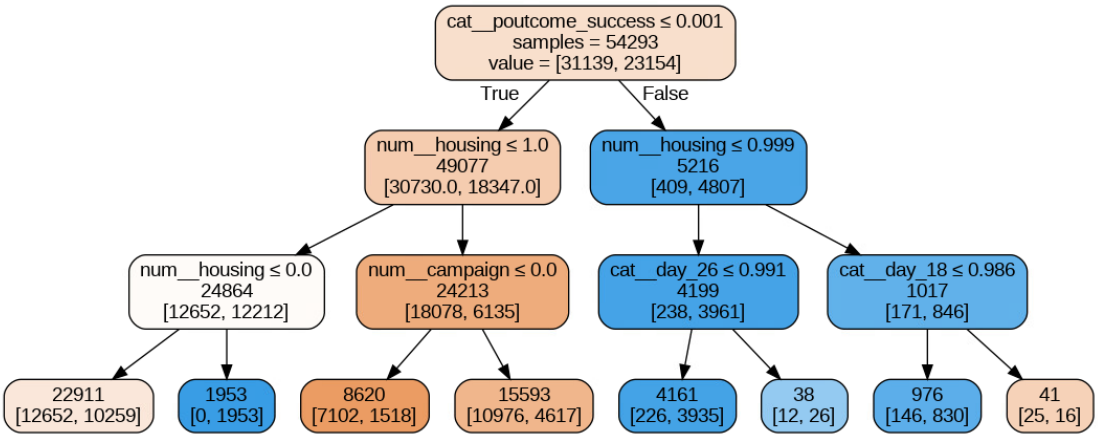


**Figure 29: Simplified decision tree**

## 13.0. Model Technique #2 – Logistic Regression

Logistic Regression is a widely used statistical method for predicting one of two possible outcomes, such as 'Yes' or 'No'. Instead of predicting the outcome directly, it estimates the probability that a given case belongs to the positive category, and then applies a threshold (commonly 0.5) to assign a class label.

In this project, Logistic Regression was trained on the resampled dataset. The model then learned patterns in the data by assigning weights to each feature, where positive weights increased the likelihood of predicting 'Yes' and negative weights reduced it.

From the fitted model, some features stood out as particularly influential. For example, having:

- A successful outcome in a previous marketing campaign strongly increased the likelihood of a current subscription, while a prior failure reduced it.
- The number of previous contacts was also a very strong predictor, with higher values linked to a much greater chance of subscription.
- Seasonal effects (such as the month of contact) and certain demographic or financial features further shaped the predictions.

By interpreting these relationships, Logistic Regression not only provided predictions but also offered insight into which factors most strongly influenced a potential subscription.

## 13.1. Results and Interpretation

On the test set, the Logistic Regression model achieved an accuracy of 72%. More importantly, it outperformed the Decision Tree in identifying the minority class ('Yes'), with a recall of 0.51 and an F1-score of 0.41, compared to the Decision Tree's recall of 0.33 and F1-

score of 0.32. This improvement means Logistic Regression was more effective at finding potential subscribers, even though it occasionally produced more false positives.

The confusion matrix for the test data revealed that the model correctly identified 535 actual subscribers, but missed 523 others, indicating that while recall improved, there is still room for better coverage of the minority class. The higher recall also shows that Logistic Regression was less biased toward predicting the majority 'No' class than the Decision Tree, a valuable property in marketing scenarios where missing potential customers is more costly than following up on a few false leads.

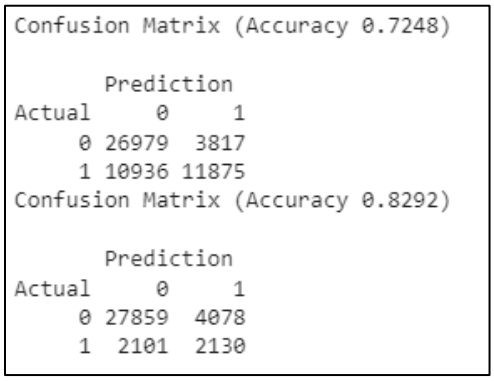


**Figure 30: Confusion matrix for the oversampled dataset vs original dataset**

### Impact of Oversampling During Training

When trained on the original imbalanced dataset, Logistic Regression achieved an accuracy of 82.92%. However, when trained on the SMOTETomek oversampled data, accuracy dropped to 72.48%. This apparent decline is expected, high accuracy on imbalanced data can be misleading because predicting the majority class for all cases yields strong accuracy but fails to identify the minority class effectively.

Oversampling had a notable impact on class detection:

- Original training data: True Positives = 2,130, False Negatives = 2,101.
- Oversampled training data: True Positives = 11,875, False Negatives = 10,936.

While the oversampled model's accuracy was lower, it demonstrated a much more balanced distribution between True Positives and False Negatives, showing that it learned to recognize the minority class ('Yes') far better. This trade-off aligns with the primary goal of handling class imbalance prioritizing recall and F1-score for the minority class over raw accuracy.

### Practical Implications

In practical marketing terms, Logistic Regression proved to be the stronger choice when the goal is to maximize the identification of potential subscribers. Its higher recall and improved minority class detection make it more suitable for campaigns were reaching as many likely

customers as possible outweighs the occasional false positive, all while maintaining competitive overall accuracy.

| | coef | odds | variable |
|---|---|---|---|
| 0 | -0.451653 | 0.636575 | num__age |
| 1 | -0.513626 | 0.598322 | num__default |
| 2 | 2.619040 | 13.722541 | num__balance |
| 3 | -0.639975 | 0.527305 | num__housing |
| 4 | -0.631780 | 0.531645 | num__loan |
| 5 | -0.185693 | 0.830529 | num__day |
| 6 | -5.029281 | 0.006544 | num__campaign |
| 7 | 0.185150 | 1.203399 | num__pdays |
| 8 | 9.024032 | 8300.177777 | num__previous |
| 9 | 0.137680 | 1.147608 | cat__job_admin. |
| 10 | 0.029130 | 1.029559 | cat__job_blue-collar |

**Figure 31: Coefficient and odds ratios of variables after logistic regression (see full table in Appendix 22.1)**

### Key Feature Insights from Logistic Regression

The model's coefficients and corresponding odds ratios highlight several important predictors of subscription likelihood:

### Strong Positive Predictors

- **Number of previous contacts** (num__previous) – The largest influence observed. Each additional contact increases the odds of subscription dramatically (odds ratio ≈ 8,300), indicating that prior engagement is a powerful driver.
- **Previous campaign success** (cat__poutcome_success) – Increases the odds by a factor of 6.84, confirming that past positive experiences strongly encourage future subscription.
- **Account balance** (num__balance) – Higher balances are linked to substantially higher subscription odds (odds ratio ≈ 13.72).
- **Certain months** – Contact in May (odds ratio ≈ 4.05), June (≈ 2.99), August (≈ 2.59), July (≈ 2.18), and retired clients (odds ratio ≈ 2.02) all increase the likelihood of a positive outcome.

### Strong Negative Predictors

**Number of contacts in current campaign** (num__campaign) – Strongly decreases odds (odds ratio ≈ 0.0065), suggesting that repeated calls in the same campaign may reduce interest.

**March (cat__month_5), October (cat__month_6), September (cat__month_7),**– All associated with lower odds, indicating fewer effective periods for outreach.

**Housing loan (num__housing) and personal loan (num__loan)** – Both slightly reduce the likelihood of subscription, possibly reflecting financial constraints.

**Previous campaign failure (cat__poutcome_failure)** – Reduces odds (odds ratio ≈ 0.58), showing that negative past experiences hinder future success.

**Moderate Effects**

- **Occupations** such as students, management, and technicians show moderate increases in subscription likelihood, while entrepreneurs and housemaids show moderate decreases.
- **Marital status** has smaller but still noticeable effects, with single clients more likely and married clients slightly less likely to subscribe.

Overall, the results suggest that timing of contact, past customer experiences, engagement history, and certain demographic and financial indicators play crucial roles in predicting subscription outcomes.

## 14.0. Model Technique #3 – Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. Each tree makes predictions, and the final output is determined by majority vote (for classification). This approach is particularly effective for handling complex relationships and interactions between features.

In this analysis, the top 20 most important features were selected based on feature importance scores. Using only these key features allows the model to focus on the variables that contribute most to predictions, improving interpretability and potentially reducing training time and overfitting.

The modeling process involved:

- **Selecting top features:** Identifying the 20 features with the highest importance scores.

- **Reducing the datasets:** Creating training and test sets containing only these selected features.

- **Training the model:** Initializing a Random Forest classifier and fitting it on the resampled training data.
- **Evaluating performance:** Making predictions on the test set and assessing accuracy, precision, recall, F1-score, and confusion matrices.]

## 14.1. Results and Interpretation

The Random Forest model trained on the top 20 features performed almost identically to the model trained on all features:

| Metric | All Features | Top Features |
|---|---|---|
| Accuracy | 0.88 | 0.88 |
| Precision (Class 1) | 0.50 | 0.51 |
| Recall (Class 1) | 0.32 | 0.32 |
| F1-score (Class 1) | 0.39 | 0.39 |

Table 3: Random Forest accuracy measurements

The confusion matrices also showed similar results, with only minor differences in true positives, false positives, and false negatives for the minority class ('Yes'). This indicates that the remaining features beyond the top 20 do not provide significant additional predictive power.

```
**Random Forest (All Features):**
Precision (Class 1): 0.50
Recall (Class 1): 0.32
F1-score (Class 1): 0.39
Accuracy: 0.88
Confusion Matrix:
[[7649  336]
 [ 723  335]]

**Random Forest (Top Features):**
Precision (Class 1): 0.51
Recall (Class 1): 0.32
F1-score (Class 1): 0.39
Accuracy: 0.88
Confusion Matrix:
[[7655  330]
 [ 721  337]]
```

Figure 32: Confusion matrix for random forest

The ROC curve and AUC analysis further supports this observation:

- Random Forest (All Features): AUC = 0.77
- Random Forest (Top Features): AUC = 0.76

Both models clearly outperform the single Decision Tree (AUC = 0.62) and slightly outperform Logistic Regression (AUC = 0.75), indicating strong overall discriminatory ability. The proximity of the ROC curves for the two Random Forest models confirms that reducing the feature set has minimal impact on performance.

**Practical Implications**

Using only the top 20 features simplifies the model without sacrificing accuracy, making it faster to train and easier to interpret.

The Random Forest model is effective at distinguishing potential subscribers ('Yes') from non-subscribers ('No'), outperforming both the Decision Tree and Logistic Regression in overall classification power.

The ensemble approach reduces the risk of overfitting, and focusing on key features can help prioritize resources for data collection and monitoring.

In marketing practice, this model can support targeting campaigns more efficiently by highlighting the most influential factors driving subscription likelihood.

## 15.0. Model Comparison

| Model | Accuracy | F1 Score (Class 1) | AUC |
|---|---|---|---|
| Random Forest (All Features) | 88% | 0.39 | 0.77 |
| Random Forest (Top Features) | 88% | 0.39 | 0.76 |
| Logistic Regression | 83% | 0.41 | 0.75 |
| Decision Tree | 85% | 0.33 | 0.62 |

**Table 4: Model Comparison**

## Model Recommendation

## 16.0 Model Selection

For this analysis, three models were selected Decision Tree, Logistic Regression and, Random Forest to provide a diverse set of approaches for addressing this binary classification problem, particularly given the imbalanced nature of the dataset. Each model was chosen for its distinct characteristics and potential contributions:

- **Decision Tree:** Decision trees are simple and highly interpretable models that allow for a clear visualization of the decision-making process based on individual features. They provide a foundational understanding of how the dataset can be segmented to predict outcomes.

- **Logistic Regression:** Logistic Regression is a linear model that estimates the probability of a binary outcome. It serves as a common baseline for classification

tasks and offers insight into the relationships between features and the target variable through interpretable coefficients.

- **Random Forest:** Random Forests are ensemble models that combine multiple decision trees to improve predictive performance. They are robust, capable of capturing complex non-linear relationships, and less prone to overfitting compared to single trees. Additionally, Random Forests provide feature importance metrics, offering valuable insight into which variables most strongly influence predictions.

By evaluating these models, we can compare their strengths and weaknesses in terms of overall performance, ability to identify the minority class, and interpretability, ultimately guiding the selection of the most suitable model for the business objective

## 17.0 Model Theory

## 17.1 Chosen Model Assumptions and Limitations

**Random Forest**

Random Forests are an ensemble learning technique that improves predictive performance by combining multiple decision trees. Unlike a single decision tree, which can be overly sensitive to the training data, Random Forests reduce variance and increase stability by training each tree on a random subset of the data (bootstrap sampling) and using a random subset of features at each split. Predictions are then aggregated across all trees using majority voting for classification tasks.

This approach allows Random Forests to capture complex relationships between features while mitigating overfitting that single trees are prone to. Although Random Forests lose the straightforward interpretability of a single tree, they provide variable importance scores to indicate which features contribute most to predictions

## 17.2 Model Assumptions and Limitations

**Assumptions:**

Random Forest does not assume feature independence; it can handle correlated features effectively (Odo, 2023).

**Limitations:**

- The model requires more computational resources compared to a single decision tree (Odo, 2023).
- Interpretability is reduced compared to a simple tree, as the ensemble produces many individual trees rather than one visualizable set of decision rules (Odo, 2023).

## 18.0 Model Sensitivity to Key Drivers

Analysis of variable importance from the Random Forest model indicates that the following features are most influential in predicting customer subscription:

- **Call duration:** longer interactions increase likelihood of conversion.
- **Contact month:** Certain months show higher conversion rates.
- **Number of previous contacts:** Past contact history significantly impacts subscription probability.

## 19.0 Additional Models to Address Business Objectives

To further improve predictive performance, advanced ensemble methods such as Gradient Boosting and XGBoost could be applied. These models often provide higher accuracy and better handling of imbalanced classes but require careful tuning to avoid overfitting.

## 20.0. Impacts on Business Problem (Scope of the recommended model)

Deploying the Random Forest model can help the business:

- Prioritize marketing calls toward clients most likely to subscribe, increasing campaign efficiency.
- Reduce wasted effort on low-probability prospects.
- Focus resources on key drivers of conversion, optimizing marketing strategy

## 21.0. Recommended Next Steps

The Random Forest model was found to be the most effective predictor among the evaluated models, demonstrating strong discriminatory ability for both the majority and minority classes. Based on these findings, the following recommendations are proposed:

- **Deployment:** Implement the Random Forest model in the production environment to support targeted marketing campaigns.
- **Retraining**: Periodically retrain the model with updated customer data to maintain accuracy and adapt to evolving market conditions.
- **Performance Monitoring**: Track key metrics such as accuracy, F1-score, and AUC to ensure ongoing effectiveness.
- **Model Enhancement**: Consider exploring boosted tree methods (Gradient Boosting, XGBoost) for potential improvements in prediction performance.

# Appendix

## 22.0 References


Breiman, L., & Cutler, A. (2001). Random forests. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Ivashina, V., & Scharfstein, D. (2010). Bank lending during the financial crisis of 2008. Journal of Financial Economics. https://www.sciencedirect.com/science/article/abs/pii/S0304405X09002396

Liao, Y., Chen, C., & Hsieh, J. (2011). Quantitative models for direct marketing: A review from systems perspective. https://www.researchgate.net/publication/221984893_Quantitative_models_for_direct_marketing_A_review_from_systems_perspective

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems.

Ghatasheh, S., et al. (2020). Modeling the telemarketing process using genetic algorithms and extreme boosting: Feature selection and cost-sensitive analytical approach. https://www.researchgate.net/publication/372178608_Modeling_the_Telemarketing_Process_using_Genetic_Algorithms_and_Extreme_Boosting_Feature_Selection_and_Cost-Sensitive_Analytical_Approach

Yan, C., Li, M., & Liu, W. (2020). Prediction of bank telephone marketing results based on improved whale algorithms optimizing S_Kohonen network. Applied Soft Computing, 92, 106259.

The Telemarketing Company (TTMC). (2021). Is telemarketing as expensive as you think? https://ttmc.co.uk/knowledge/articles/is-telemarketing-as-expensive-as-you-think

Feng, Y., et al. (2022). A dynamic ensemble selection method for bank telemarketing sales prediction. Journal of Business Research.

Odo, C. (2023). Random forest: Assumptions, advantages, disadvantages and applications. Medium. https://medium.com/@chibuike.odo.c/random-forest-assumptions-advantages-disadvantages-and-applications-2881f4ea14b6

UCI Machine Learning Repository. Bank Marketing Dataset. https://archive.ics.uci.edu/dataset/222/bank+marketing

## 22.1 Logistic Regression Results

| | coefficient | odds | variable |
|---|---|---|---|
| 0 | -0.451653 | 0.636575 | num__age |
| 1 | -0.513626 | 0.598322 | num__default |
| 2 | 2.619040 | 13.722541 | num__balance |
| 3 | -0.639975 | 0.527305 | num__housing |
| 4 | -0.631780 | 0.531645 | num__loan |
| 5 | -0.185693 | 0.830529 | num__day |
| 6 | -5.029281 | 0.006544 | num__campaign |
| 7 | 0.185150 | 1.203399 | num__pdays |
| 8 | 9.024032 | 8300.177777 | num__previous |
| 9 | 0.137680 | 1.147608 | cat__job_admin. |
| 10 | 0.029130 | 1.029559 | cat__job_blue-collar |
| 11 | -0.229928 | 0.794591 | cat__job_entrepreneur |
| 12 | -0.483389 | 0.616690 | cat__job_housemaid |
| 13 | 0.064990 | 1.067148 | cat__job_management |
| 14 | 0.701124 | 2.016018 | cat__job_retired |
| 15 | -0.259244 | 0.771635 | cat__job_self-employed |

| | | | |
|---|---|---|---|
| 16 | -0.021735 | 0.978499 | cat__job_services |
| 17 | 0.406429 | 1.501447 | cat__job_student |
| 18 | 0.063641 | 1.065709 | cat__job_technician |
| 19 | 0.093141 | 1.097616 | cat__job_unemployed |
| 20 | 0.227043 | 1.254884 | cat__marital_divorced |
| 21 | -0.025150 | 0.975164 | cat__marital_married |
| 22 | 0.299946 | 1.349786 | cat__marital_single |
| 23 | -0.036413 | 0.964242 | cat__education_0 |
| 24 | 0.161179 | 1.174896 | cat__education_1 |
| 25 | 0.377073 | 1.458010 | cat__education_2 |
| 26 | -1.008255 | 0.364855 | cat__month_1 |
| 27 | -0.204817 | 0.814796 | cat__month_2 |
| 28 | 1.398519 | 4.049200 | cat__month_3 |
| 29 | 0.281100 | 1.324586 | cat__month_4 |
| 30 | -0.731118 | 0.481371 | cat__month_5 |
| 31 | -0.455441 | 0.634168 | cat__month_6 |
| 32 | -0.363135 | 0.695492 | cat__month_7 |
| 33 | -0.577754 | 0.561157 | cat__month_8 |
| 34 | 0.780857 | 2.183343 | cat__month_9 |

| | | | |
|---|---|---|---|
| 35 | 1.094645 | 2.988122 | cat__month_10 |
| 36 | -0.666278 | 0.513617 | cat__month_11 |
| 37 | 0.953517 | 2.594819 | cat__month_12 |
| 38 | -0.521224 | 0.593793 | cat__poutcome_failure |
| 39 | -0.347268 | 0.706616 | cat__poutcome_other |
| 40 | 1.922402 | 6.837359 | cat__poutcome_success |
| 41 | -0.552070 | 0.575757 | cat__poutcome_unknown |