Term Deposit Subscription Prediction Report

Author: Kezia Agyemang-Saahene

Introduction

This project aims to develop a machine learning model that predicts whether a bank client will subscribe to a term deposit. The dataset, bank-additional-full.csv, was provided by a Portuguese banking institution and includes both client information and historical marketing outcomes.

The final objective is to assist the bank in identifying potential clients likely to subscribe, helping improve campaign targeting and efficiency.

Tools and Libraries Used

- Pandas and NumPy for data handling and manipulation
- Matplotlib and Seaborn for visualizations
- Scikit-learn for machine learning modeling and evaluation
- Logistic Regression and Random Forest for classification

Data Overview

- The dataset contains 41,188 rows and 21 columns including client attributes (e.g., age, job, education), contact information, previous campaign details, and the target variable “y” indicating subscription status.
- A copy of the dataset was created to preserve the original data during cleaning and transformation.

Data Cleaning and Preparation

a. Duplicates and Missing Values

- Duplicates found: 12 rows were removed from the dataset.
- Missing values: None were detected in the dataset.

b. Outlier Detection

Using the IQR method, outliers were identified in the following numeric columns:

- age, duration, campaign, pdays, previous

  Outliers in the campaign column were removed to reduce skewness and improve model learning.

c. Feature Engineering

An age_group column was created by binning age into categories:
Young, Young Adult, Adult, Middle-aged, Senior, Retired

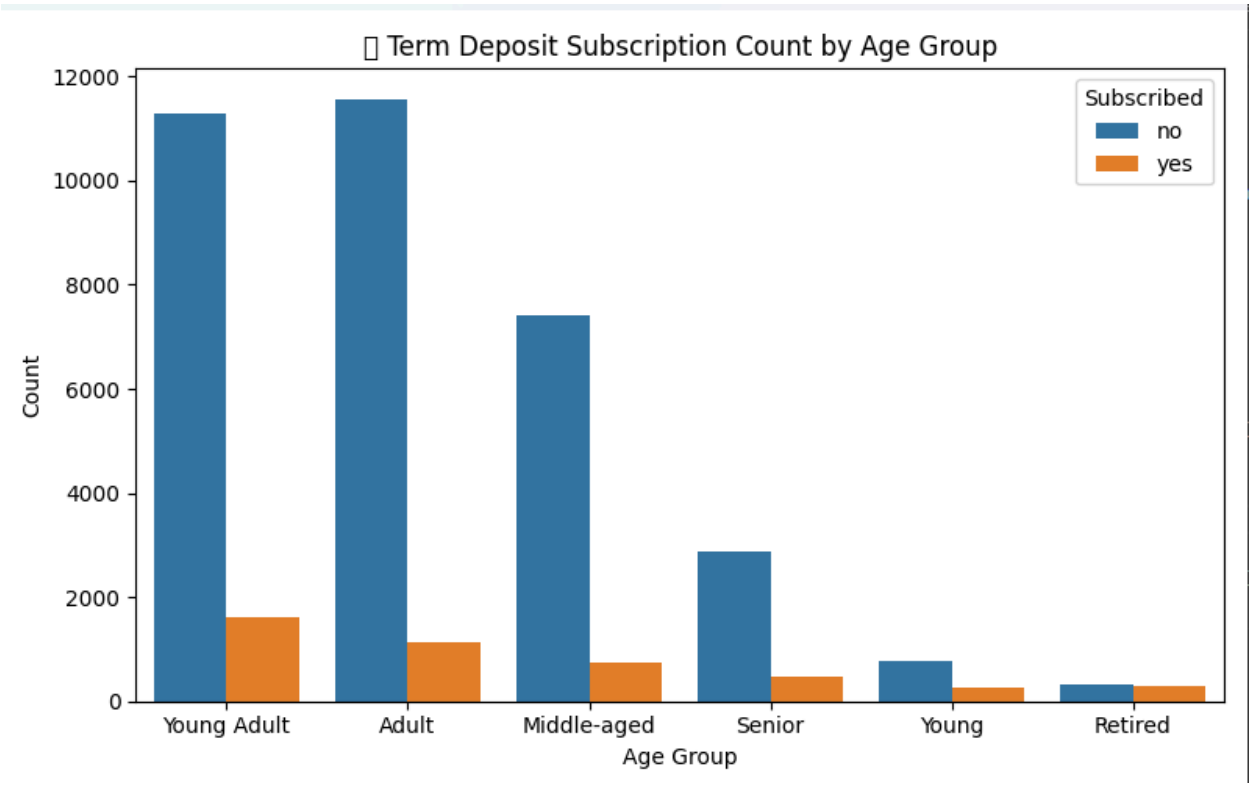
Exploratory Data Analysis (EDA)

Visualizations were generated to understand client behavior and how it relates to subscription outcomes.

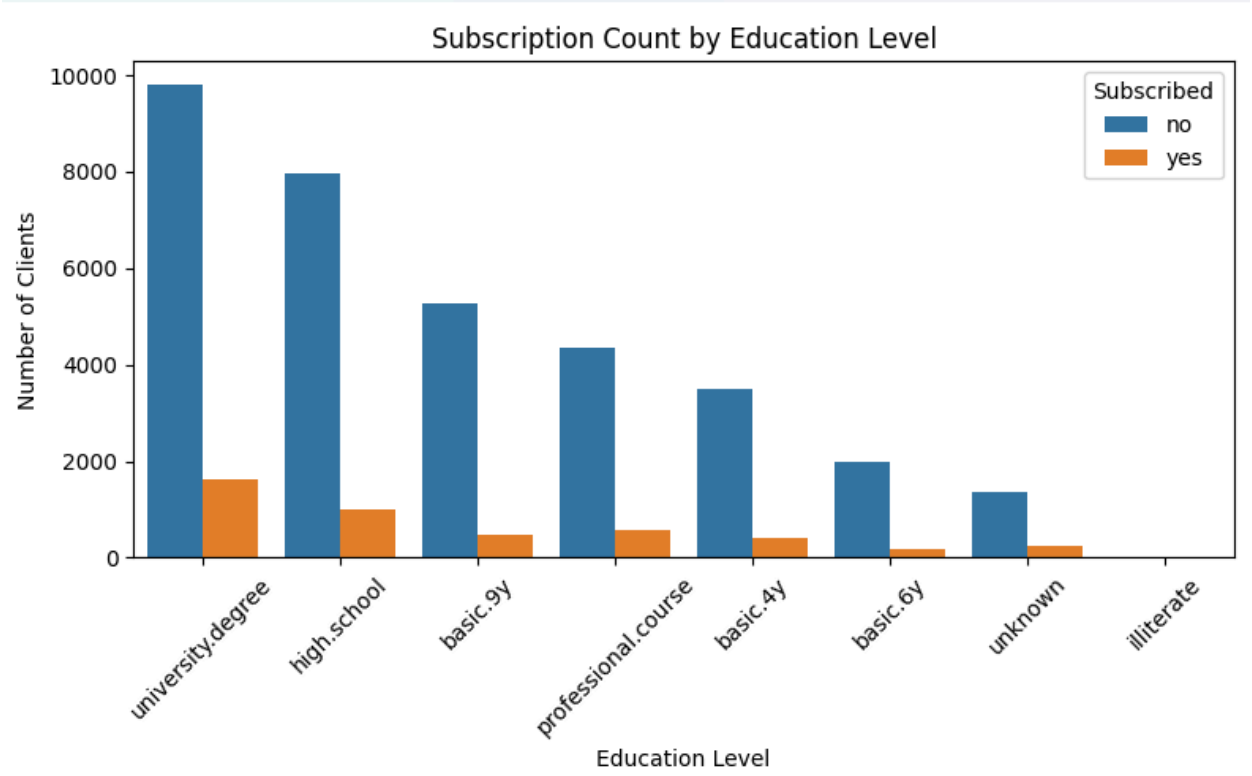
a. Subscription Distribution

- Only ~11.5% of clients subscribed to a term deposit.

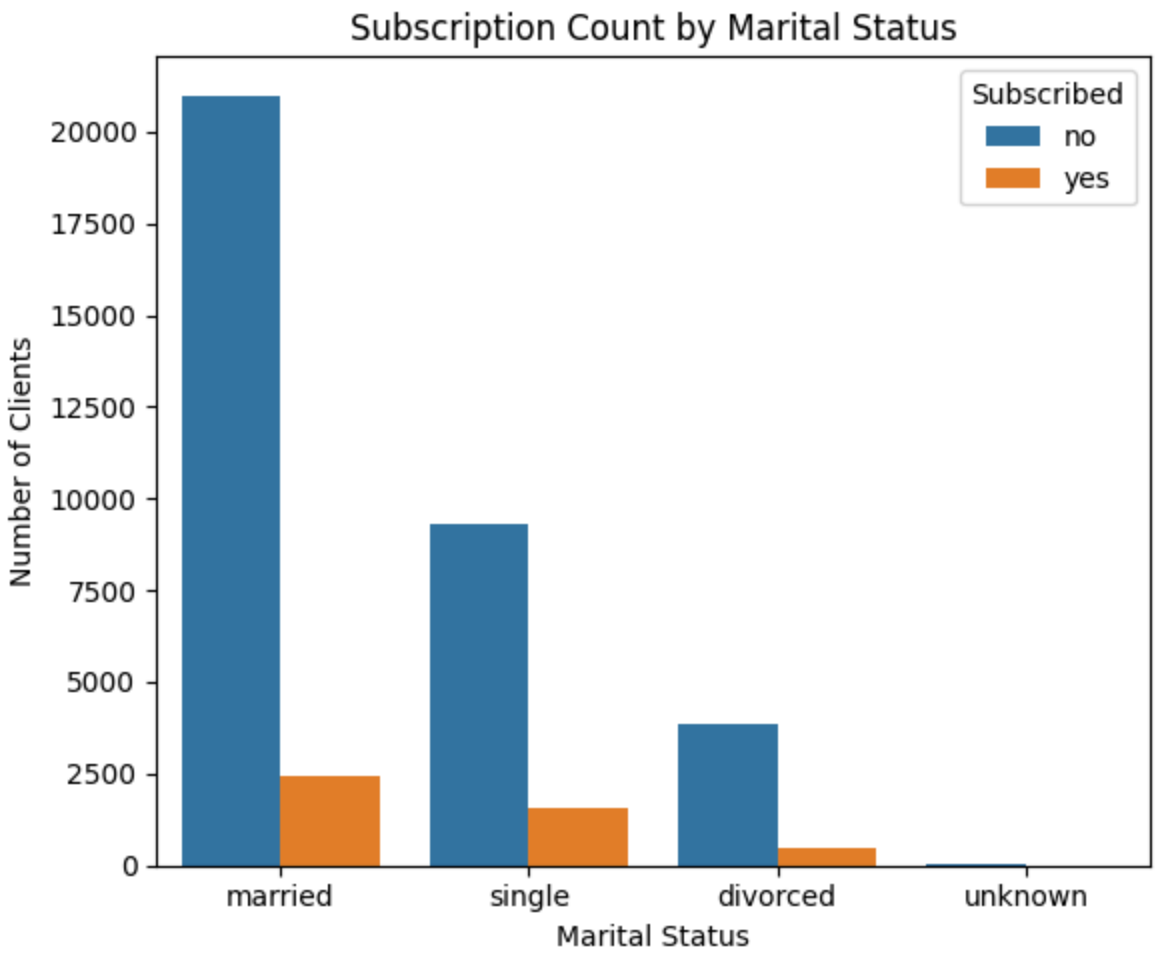
b. Insights by Features:



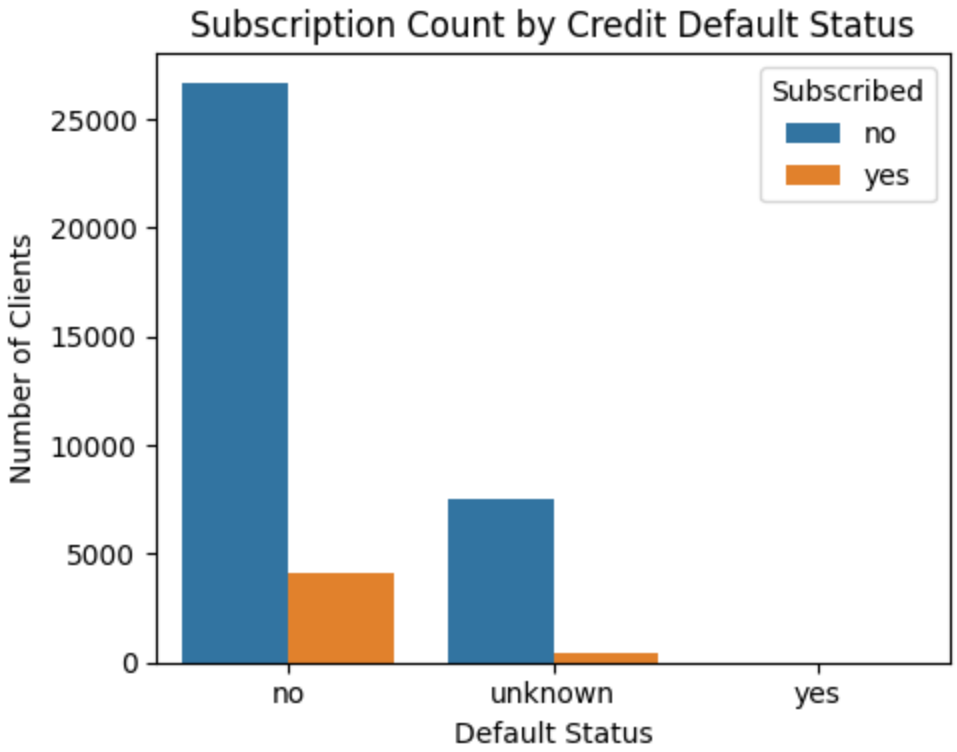
- Age Group: Young Adult clients had the highest subscriber counts.



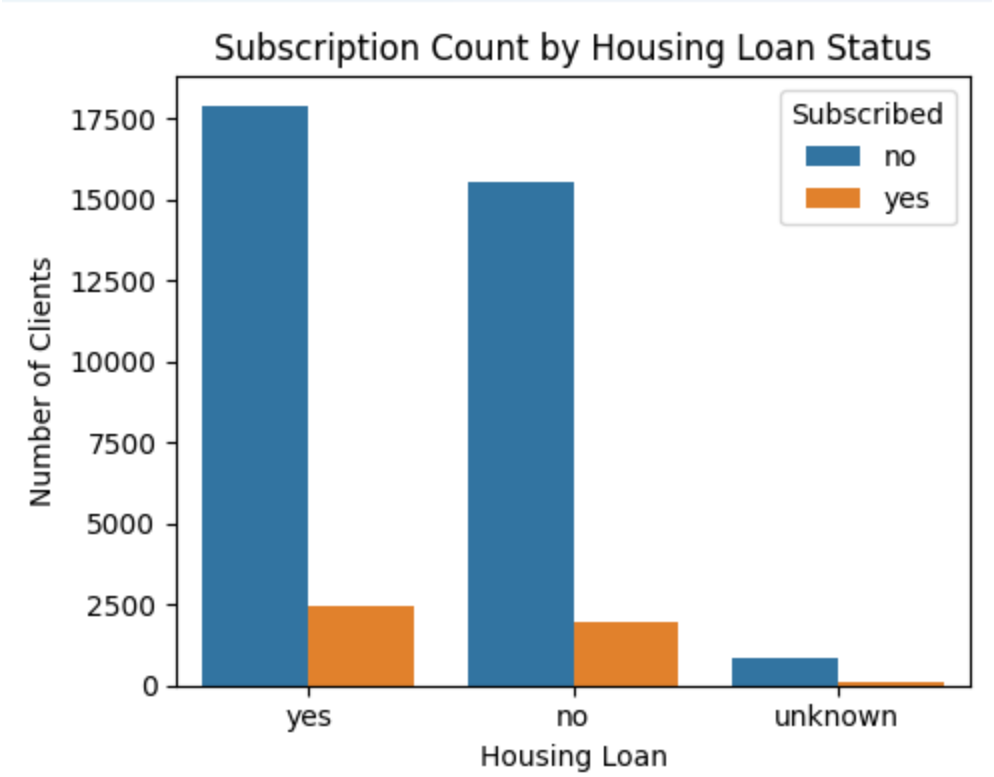
- Education: Clients with university degrees subscribed most often.



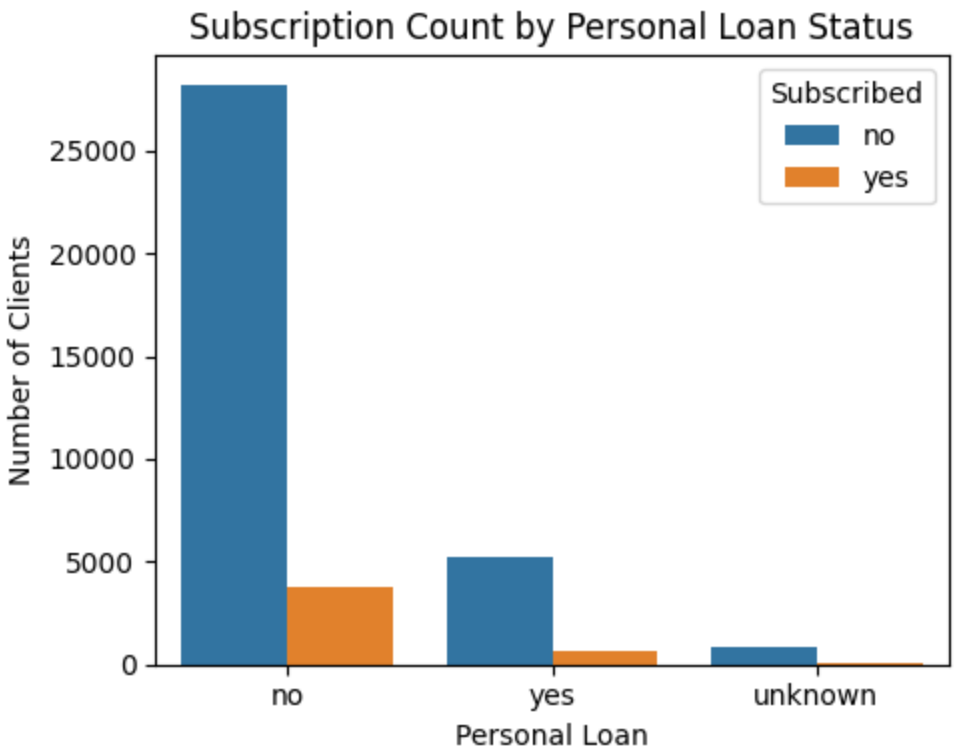
- Marital Status: Married clients formed the majority of subscribers.



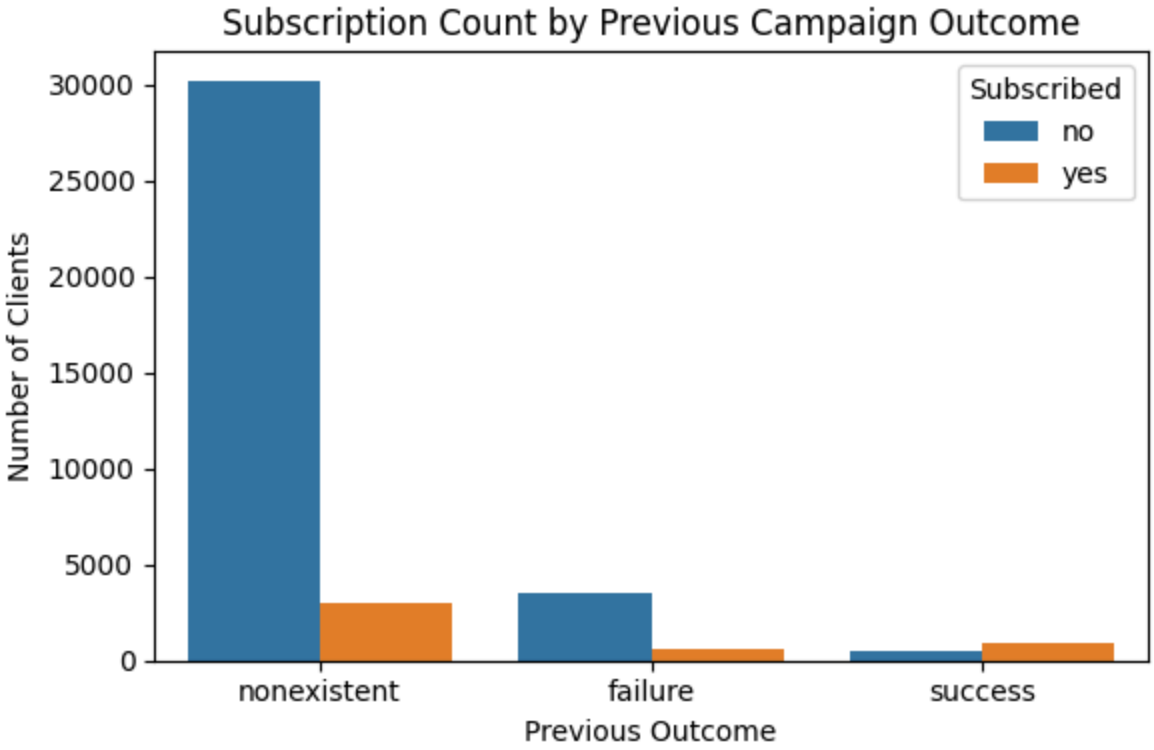
- Credit default Status: Clients that do not have defaulted credit formed the majority of subscribers



- Housing Loan Status: Clients that have housing loans formed the majority of subscribers.



- Personal Loan Status: Clients that do not have personal loans formed the majority of subscribers.



- Previous Campaign Outcome: Positive outcomes from previous campaigns did not lead to higher subscriptions.

Data Encoding and Correlation Analysis

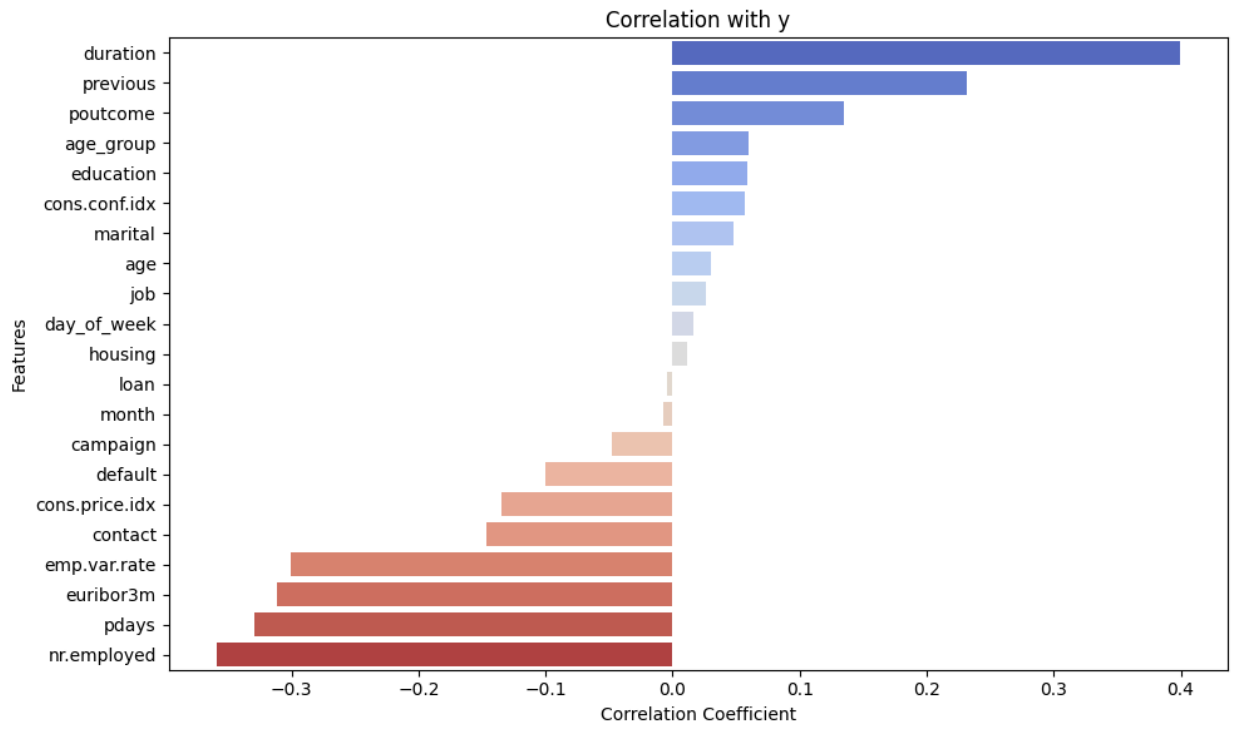
a. Label Encoding

Categorical variables were encoded using LabelEncoder.

b. Target Variable Mapping

Y was mapped from "yes"/"no" to 1 and 0.

c. Feature Correlation



Top features that positively correlated with subscription:

Duration, previous and poutcome.

Feature Selection and Modeling

- marital was removed due to potential ethical concerns and bias in decision-making.
- the loan column was dropped as it showed low correlation with the target.
- Target variable: y (1 if subscribed, 0 if not)

a. Train/Test Split

The data was split into 80% training data and 20% test data.

Model Training and Evaluation

Model 1: Logistic Regression

- Trained using class_weight='balanced' to handle class imbalance.
- Achieved an **accuracy of 85%** and a **ROC AUC score of 0.93.**
- **Recall** for subscribed clients (**88%)** was very high, indicating strong ability to identify actual subscribers.
- However, **precision** was low (**43%**), meaning many clients were wrongly predicted as subscribers.
- **F1-score (Subscribed):** 58%

Logistic regression performed reasonably well in identifying most subscribers but at the cost of some false positives.

Model 2: Random Forest Classifier (Default Threshold - 0.5)

- Delivered the highest **accuracy** of **91%** and a **ROC AUC** of **0.94**.
- **Precision improved to 65%**, reducing false positives.
- **Recall dropped to 50%**, meaning many actual subscribers were missed.
- **F1-score:** 56%

Random Forest (Threshold - 0.3**)**

- Maintained high accuracy (**90%**) and ROC AUC (**0.94**).
- **Recall improved significantly to 79%**, capturing more actual subscribers.
- **Precision stood at 56%**, offering a good balance.
- This model offers the **best** balance between precision and recall and is well-suited for this model as it captures as many potential subscribers as possible.
- Threshold Adjustment (0.3):
  Precision: 0.56
  Recall: 0.79
  F1-Score: 0.35

Lowering the threshold increased recall significantly, making this configuration ideal as the goal is to capture as many potential subscribers as possible.

Key Findings from EDA

- Only **11.5%** of clients subscribed to a term deposit, indicating **class imbalance**.
- **Young Adult** clients subscribed more than other age groups.
- Clients with a **university degree** had the highest subscription rate.
- **Married** clients formed a large portion of subscribers.
- Clients who **did not default on credit** and **did not have personal loans** were more likely to subscribe.

- Surprisingly, a **positive outcome in a previous campaign** didn't strongly predict future subscription.

Actionable Recommendations for Marketing Team

1. **Prioritize Clients Based on Call Duration**
   Focus on clients who remain engaged during calls – longer duration indicates interest.

2. **Target by Education and Age**
   Invest in campaigns focused on **university-educated, young to middle-aged adults**.

3. **Avoid Over-targeting Previous Campaign Positives**
   Prior campaign success doesn't always translate; rely on fresh engagement signals.

4. **Use the Random Forest Model (Threshold 0.3)**
   This model maximizes recall (~79%), capturing more potential subscribers even if some false positives occur – ideal for marketing scenarios.

5. **Avoid Bias-Prone Features**
   Ethically exclude features like **marital status** to prevent unfair decision-making and bias.

Conclusion

Both logistic regression and random forest models were able to identify subscription patterns. However, the random forest model with threshold tuning achieved the best balance between capturing actual subscribers (recall) and maintaining overall prediction accuracy.