Nesly Cesar

# Final Report:

# Predictive Modeling for Bank Marketing Campaign Success

## Problem Statement

In the increasingly competitive financial services sector, effective marketing strategies are essential for maintaining and growing a customer base. One of the main challenges banks face is determining which clients are most likely to subscribe to financial products, such as term deposits, when approached through marketing campaigns. This project aims to use predictive modeling techniques to improve the success of bank marketing campaigns by accurately predicting term deposit subscriptions.

The key research questions guiding this project are:

1. How can predictive modeling techniques improve the success of bank marketing campaigns in predicting term deposit subscriptions?
2. What factors most significantly influence a client's likelihood to subscribe to a term deposit, and how can these insights be used to optimize marketing strategies?
3. Which machine learning algorithm provides the most accurate predictions for client subscription to term deposits in a bank marketing campaign?
4. How can direct marketing strategies be enhanced using data mining to increase the rate of term deposit subscriptions?

## Methodology

To answer the research questions, the project follows these steps:

### Data Wrangling and Overview

The dataset used in this study comes from direct marketing campaigns conducted by a Portuguese banking institution. The data contains a variety of client-related attributes (e.g., age, job, balance), campaign-related attributes (e.g., number of contacts, last contact duration), and

macroeconomic features. The target variable is binary, indicating whether a client subscribed to a term deposit (‘yes’ or ‘no’).

- **Number of Instances**: 45,211
- **Number of Attributes**: 16 features + 1 target variable
- **Main Features**:
    - **Client Data**: age, job, marital status, education, balance
    - **Campaign Data**: number of contacts, last contact duration
    - **Macroeconomic Data**: month, previous campaign outcome

Also, the dataset came almost cleaned and no cleaning or tidiness issues had to be corrected.

## Exploratory Data Analysis (EDA)

Aside the basic statistical summary of the dataset explored, two other kinds of EDA were performed on the dataset which are Univariate and Bivariate explorations. The univariate exploration involves the distribution plots of individual variables. The target class showed a 88.3-11.7 distribution of each class (Figure 1) which showed that the dataset is imbalanced and should be considered for the model development.

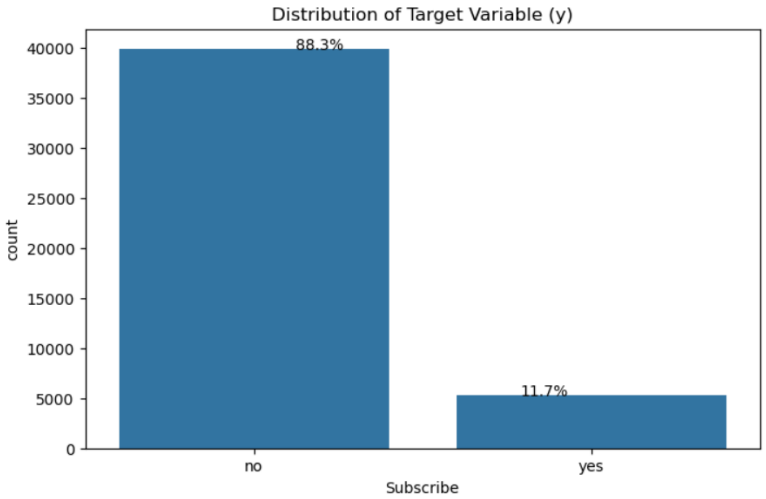


Figure 1: Target class distribution

The univariate exploration also showed all the continuous variable to contain outliers (Figure 2).

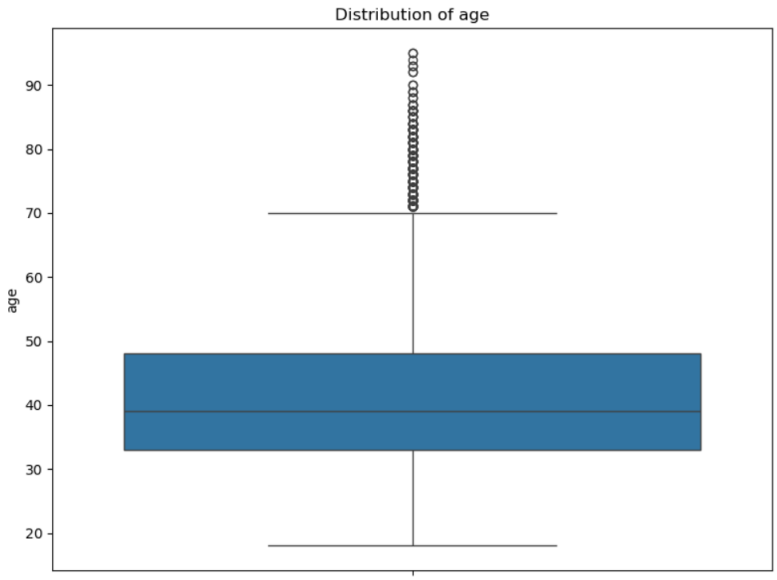


Figure 2: Example of Continuous Variables indicating the presence of outliers (Age Distribution)

To find correlation among attributes attributes, bivariate exploration was performed which involved proportionate distribution plots of independent variables with respect to the target variable. Also, Pearson correlation was used to quantify the correlations. Figure 3 shows no multicolinearity among the continuous variables.

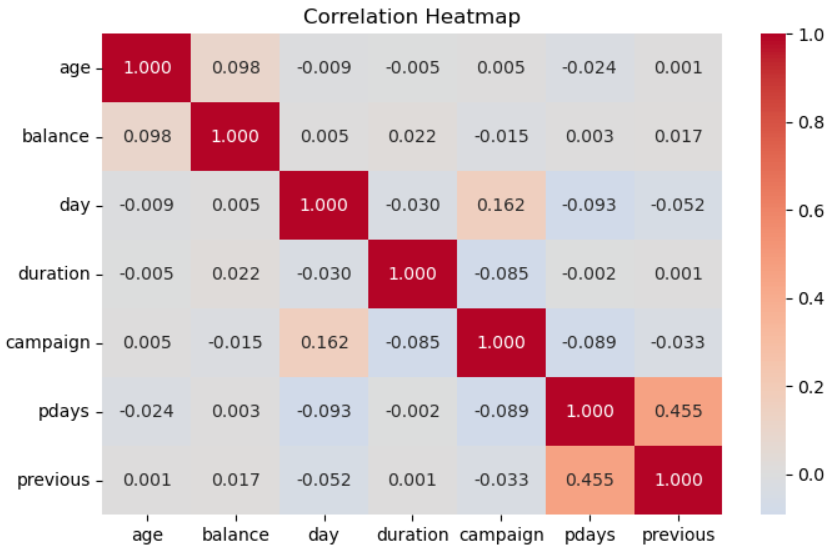


Figure 3: Correlation heatmap of the continuous variables

To summarize the EDA, the following are key insights:

- Call duration by subscription outcome (Figure 4) shows that the longer the call, the higher the likelihood of subscription.

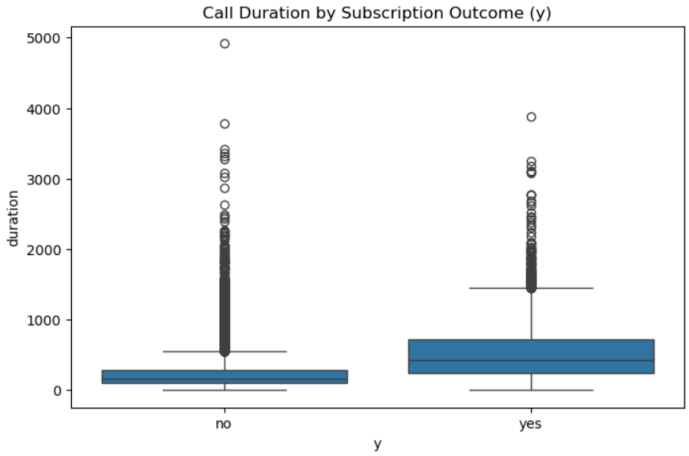


Figure 4: Call Duration by Subscription Outcome

- The proportion plot of the housing loan status against the target variable (Figure 5) reveals that clients without housing loans are more likely to subscribe to a term deposit than those with housing loans. This suggests that clients without the financial burden of a housing loan may have more disposable income available for investments, making them a more favorable target for term deposit campaigns. On the other hand, clients with housing loans tend to subscribe at lower rates, likely due to existing financial commitments.

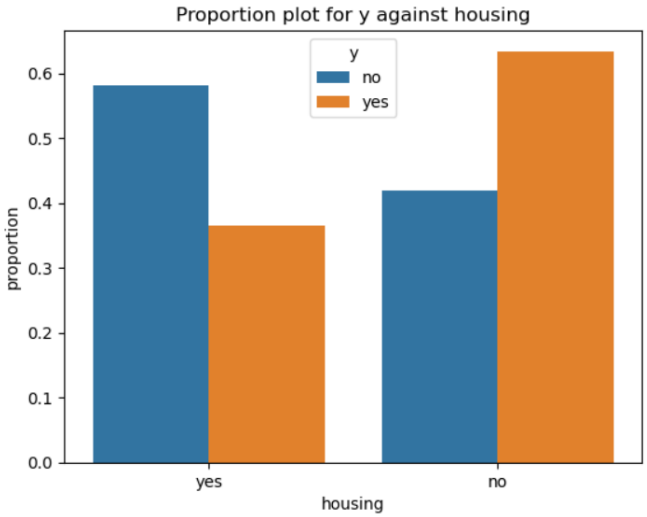


Figure 5: Proportion plot of the housing loan status against the target variable

## Data Preprocessing and Feature Engineering

The data preprocessing steps carried out include:

- **Handling Outliers**: As said earlier, the continuous variables contained outliers, which were handled by using the Interquartile Range (IQR) method. Values falling below the lower bound (Q1 - 1.5IQR) or above the upper bound (Q3 + 1.5IQR) were capped at these thresholds. This ensured that extreme values, which could distort the model's performance, were mitigated without removing important data points.
- **Categorical Variables Encoding**: Categorical variables were transformed using one-hot encoding and label encoding (for ordinal variables like education).
- **Dropping Redundant Columns**: Columns like 'previous' and 'pdays', which had many zero values, were dropped.
- **Standardization**: All continuous numeric variables were standardized to ensure uniformity.
- **Data Splitting**: The dataset was split into an 80% training set and 20% test set for model evaluation.

## Modeling Techniques

- Two machine learning algorithms were tested, including Logistic Regression (LR) and Gradient Boosting Classifier (GBC).
- Hyperparameter tuning was performed using "GridSearchCV" for both Logistic Regression and Gradient Boosting, optimizing for "F1 score" to balance precision and recall.
- Model performance was evaluated on the test set using key metrics: accuracy, precision, recall, and F1 score

### Feature Importance

The Gradient Boosting Classifier provided feature importance, helping to identify which factors most significantly influence a client's likelihood to subscribe to a term deposit.

# Results

## Model Performance Comparison

After testing multiple models, the Tuned Gradient Boosting Classifier (GBC) emerged as the best-performing model.

Table 1: Analysis of various models

| S/N | Model Name | Accuracy Score | F1 Score | Precision Score | Recall Score |
|---|---|---|---|---|---|
| 1 | LR Model | 0.901 | 0.459 | 0.633 | 0.360 |
| 2 | GBC Model | 0.904 | 0.487 | 0.645 | 0.390 |
| 3 | Tuned LR Model | 0.901 | 0.461 | 0.635 | 0.362 |
| 4 | Tuned GBC Model | 0.901 | 0.476 | 0.623 | 0.386 |

Also, as evident in Figure 6, the Gradient Boosting Classifier achieved the best balance between precision and recall, providing the highest F1 score. This model was therefore selected as the best predictive model for determining term deposit subscriptions.

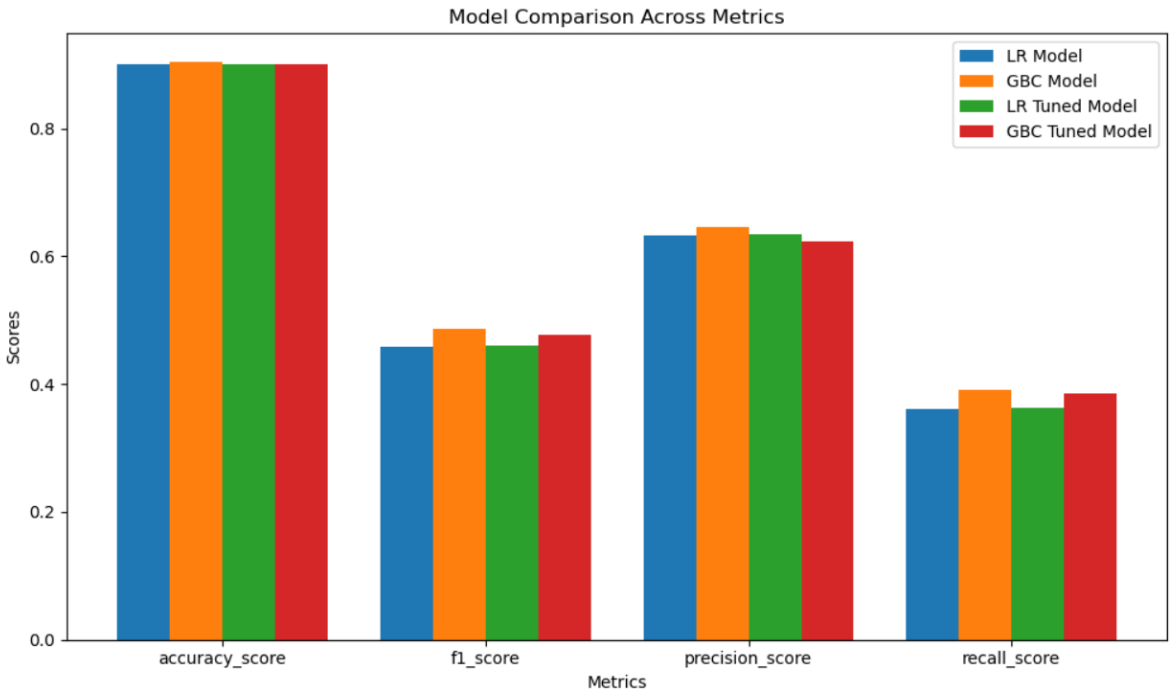


Figure 6: Model Comparison across Metrics

## Key Factors Influencing Client Subscriptions

As seen in Figure 7 below, the feature importance analysis from the Gradient Boosting Classifier revealed the most significant factors influencing a client's likelihood to subscribe to a term deposit:

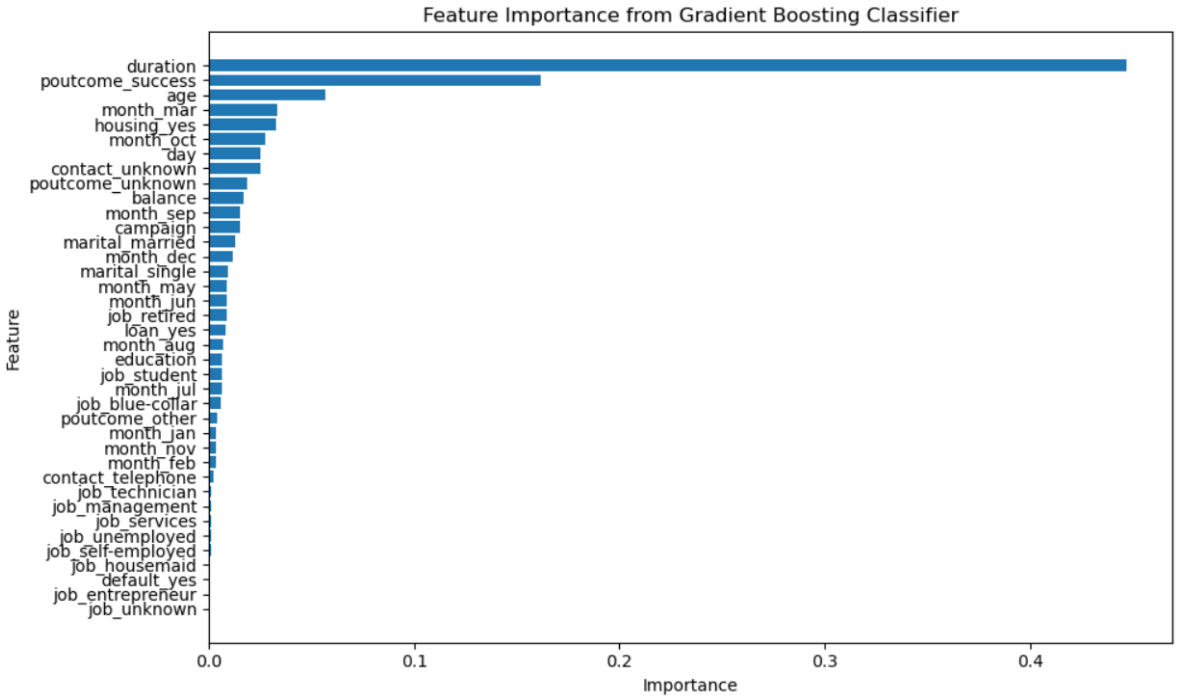


Figure 7: Feature Importance from Gradient Boosting Classifier

- **Call Duration**: The length of the last phone call had the greatest influence. Clients who engage longer in a conversation are more likely to subscribe.
- **Previous Campaign Success**: Clients who had successfully subscribed in previous campaigns were more likely to subscribe again.
- **Age**: Older clients were more likely to subscribe than younger clients.
- **Month of Contact**: Certain months (e.g., March, October) saw higher subscription rates, likely due to seasonal trends.
- **Housing Loan**: Whether a client had a housing loan also impacted their likelihood to subscribe, with certain financial profiles being more inclined to invest in term deposits.
- **Balance**: Clients with a higher bank balance were more likely to subscribe, suggesting that disposable income plays a role in term deposit investments.

## Discussion

The results from the Gradient Boosting model offer valuable insights for optimizing the bank's marketing strategy:

1. **Target High-Engagement Clients**: Clients with longer call durations show a higher likelihood of subscribing, so efforts should be focused on clients who engage during calls. Additionally, re-targeting clients who previously subscribed can yield high returns.
2. **Segmented Marketing by Age and Financial Status**: Age and balance are strong predictors of subscription likelihood. Marketing strategies should be personalized to cater to older clients and those with higher balances, potentially offering tailored deposit plans.
3. **Timing of Campaigns**: Certain months are more successful for marketing campaigns. Identifying and focusing efforts on high-yield months (e.g., March, October) can significantly boost the campaign's success rate.
4. **Cost-Effective Contact Strategy**: By knowing which clients are more likely to subscribe (based on past behavior and financial status), the bank can reduce unnecessary contact with unlikely subscribers, optimizing marketing resources.

## Conclusion

This project demonstrates how predictive modeling can significantly improve the success of bank marketing campaigns by accurately predicting client subscription behavior. The Gradient Boosting Classifier was found to be the best model for this task, and the analysis revealed key factors, such as call duration, previous campaign success, and age, that influence a client's decision to subscribe to a term deposit.

These insights can be directly applied to optimize marketing strategies, improve resource allocation, and boost subscription rates by focusing on high-likelihood clients, targeting appropriate times, and tailoring messages based on client profiles. Additionally, using predictive models like Gradient Boosting allows banks to make data-driven decisions that can enhance the overall effectiveness of their direct marketing efforts.

## Future Work

Future research could explore the following areas:

- **Additional Features**: Incorporating more macroeconomic indicators (e.g., inflation rate, interest rates) could further refine the model's predictions.
- **Time-Series Analysis**: Analyzing the temporal aspect of client behavior over multiple campaigns could provide deeper insights into long-term trends.
- **Improved Imbalance Handling**: Techniques like oversampling (SMOTE) or undersampling could be explored to address the class imbalance more effectively.

By continuously refining these predictive models, the bank can enhance its ability to convert more potential clients into term deposit subscribers while making marketing campaigns more cost-effective.