# Portugese Bank Marking

## 1. Introduction

Banks rely heavily on marketing campaigns to promote financial products such as term deposits. These campaigns require significant investment in terms of time, manpower, and cost. Therefore, it is essential for banks to identify potential customers who are most likely to subscribe to the product.

This project focuses on analyzing the **Portuguese Bank Marketing dataset**, which contains information from direct marketing campaigns conducted by a Portuguese banking institution. The goal is to understand customer behavior and build a predictive system that helps the bank improve marketing efficiency and decision-making.

---

## 2. Problem Statement

The main objective of this project is to predict whether a customer will subscribe to a **term deposit** based on demographic, financial, and campaign-related attributes.

### Business Problems Addressed:

- Identifying customers who are likely to respond positively to marketing campaigns
- Reducing unnecessary marketing calls and operational costs
- Improving campaign success rate
- Enhancing customer targeting strategies

This is a **binary classification problem**, where the target variable indicates whether the client subscribed to a term deposit (`yes` or `no`).

---

## 3. Dataset Description

### Dataset Size

**41188 rows × 21 columns**

### Feature Overview

The dataset contains customer demographic information, financial details, campaign-related variables, and macroeconomic indicators.

#### Customer Demographics

- `age`: Age of the client
- `job`: Type of job
- `marital`: Marital status
- `education`: Education level

**Financial Information**

- `default`: Has credit in default
- `housing`: Has housing loan
- `loan`: Has personal loan

**Campaign Details**

- `contact`: Type of contact communication
- `month`: Last contact month
- `day_of_week`: Day of last contact
- `duration`: Duration of last contact (in seconds)
- `campaign`: Number of contacts during the current campaign
- `pdays`: Days passed since last contact in a previous campaign
- `previous`: Number of contacts before this campaign
- `poutcome`: Outcome of the previous campaign

**Macroeconomic Indicators**

- `emp.var.rate`: Employment variation rate
- `cons.price.idx`: Consumer price index
- `cons.conf.idx`: Consumer confidence index
- `euribor3m`: Euribor 3-month rate
- `nr.employed`: Number of employees

**Target Variable**

- `y`: Client subscribed to a term deposit (yes/no)

---

# 4. Data Preprocessing

## 4.1 Data Cleaning

- Checked for missing and unknown values in categorical features
- Handled `unknown` values by either retaining them as a separate category or replacing them where appropriate

## 4.2 Encoding Categorical Variables

- Converted categorical variables into numerical format using **Label Encoding** and **One-Hot Encoding**

## 4.3 Feature Scaling

- Applied **StandardScaler** to normalize numerical features such as age, duration, and economic indicators

## 4.4 Handling Class Imbalance

- The dataset shows imbalance between subscribers and non-subscribers

- Applied techniques such as:
- Class weight balancing
- Oversampling methods like SMOTE

---

## 5. Exploratory Data Analysis (EDA)

### 5.1 Univariate Analysis

- Most customers belong to the age group of 30–50 years
- Majority of customers work in administrative, technician, and blue-collar jobs
- A smaller proportion of customers subscribed to the term deposit

### 5.2 Bivariate Analysis

- Longer call durations are strongly associated with successful subscriptions
- Customers without housing or personal loans show higher subscription rates
- Previous successful campaign outcomes significantly increase the likelihood of subscription

### 5.3 Key Insights from EDA

- Call duration is one of the most influential features
- Repeated calls in the same campaign reduce customer interest
- Past customer interaction history is a strong predictor

---

## 6. Model Building

### 6.1 Train-Test Split

- Training data: 80%
- Testing data: 20%

### 6.2 Machine Learning Algorithms Used

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier

### 6.3 Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC score

---

# 7. Model Performance and Results

Among all the models tested, **Random Forest** and **Gradient Boosting** performed the best, providing high accuracy and balanced precision-recall scores.

### Best Performing Model:

- **Gradient Boosting Classifier**
- High predictive performance on unseen data
- Robust handling of nonlinear relationships

---

# 8. Feature Importance Analysis

The most important features influencing customer subscription are:

1. `duration`
2. `poutcome`
3. `campaign`
4. `age`
5. `housing`
6. `euribor3m`

These features play a critical role in predicting customer behavior.

---

# 9. Business Insights and Recommendations

### Business Insights

- Customers who had successful past interactions are more likely to subscribe
- Longer and meaningful conversations increase conversion rates
- Excessive calling negatively impacts customer response

### Recommendations

- Focus marketing efforts on high-probability customers identified by the model
- Reduce unnecessary repeated calls
- Improve agent training to optimize call duration
- Use predictive analytics before launching campaigns

---

# 10. Conclusion

This project demonstrates how data analysis and machine learning can significantly improve banking marketing strategies. By predicting customer subscription behavior, banks can reduce costs, increase efficiency, and enhance customer experience.

The developed model can be integrated into CRM systems to assist marketing teams in real-time decision-making.

## 11. Future Scope

- Deploy the model using Flask or FastAPI
- Build a real-time dashboard using Power BI or Streamlit
- Perform cost-benefit analysis of campaigns
- Experiment with deep learning techniques

## 12. Tools and Technologies Used

- Python (Pandas, NumPy, Scikit-learn)
- Matplotlib, Seaborn
- Jupyter Notebook
- Machine Learning Algorithms