Project Proposal

Predicting Term Deposit Subscription in Portuguese Banking Campaigns

---

Team Members

1. Tushar Ahuja
2. Srinivas Allanki
3. Mihir

Research Questions

1. What client demographic and financial attributes are most predictive of a term deposit subscription?

Approach: Use logistic regression and decision tree classifiers to assess the impact of variables like age, job, marital status, and balance on subscription likelihood.

Goal: Identify client characteristics that strongly influence subscription rates, enabling targeted outreach based on client profiles.

2. Which campaign-related factors (e.g., contact type, duration, timing) are associated with higher subscription rates?

Approach: Conduct exploratory data analysis (EDA) and feature importance analysis using random forests to evaluate the effectiveness of campaign attributes (contact type, last contact duration, month of last contact).

Goal: Discover optimal campaign strategies and timing to increase subscription rates, refining marketing tactics based on campaign characteristics.

3. How do previous campaign interactions (e.g., past contacts, outcomes of previous campaigns) affect the likelihood of a successful subscription?

Approach: Analyze past campaign metrics (e.g., previous contacts, days since last contact, previous campaign outcomes) and apply predictive modeling (logistic regression and support vector machines).

Goal: Assess the impact of prior client interactions on the current campaign's success, informing strategies for follow-up or re-engagement.

Project Details

- Cases: Individual clients contacted during the bank's marketing campaigns.
- Sample Size: 45,211 records in the bank.csv data set (with 17 input features).

Data Collection Method

- Description: Data is derived from phone-based marketing campaigns conducted by a Portuguese bank between 2008 and 2010, targeting clients for term deposit subscriptions.
- Purpose: Gather information on client and campaign attributes to study factors influencing term deposit subscriptions.

Type of Study

- Observational Study: This is an observational data set, as it records outcomes without any experimental manipulation or random assignment.

Data Source

- Source: Obtained from the UCI Machine Learning Repository.
- Citation: Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.
- UCI Repository link: **https://archive.ics.uci.edu/ml/datasets/bank+marketing**

Response Variable

- Response Variable: y (whether a client subscribed to a term deposit).
- Type: Categorical (Binary: "yes" or "no").

Explanatory Variables

- Client Attributes: Age (Numerical), Job (Categorical), Marital Status (Categorical), Education (Categorical), Balance (Numerical), Default (Binary), Housing Loan (Binary), Personal Loan (Binary).
- Campaign Details: Contact Type (Categorical), Last Contact Day (Numerical), Last Contact Month (Categorical), Last Contact Duration (Numerical).
- Additional Attributes: Number of Contacts (campaign, Numerical), Days Since Last Contact (pdays, Numerical), Contacts in Previous Campaign (previous, Numerical), Outcome of Last Campaign (poutcome, Categorical).

Summary Statistics

```
library(tidyverse)
library(readr)
library(dplyr)
```

Acquiring Data set

```
data = read_delim("https://raw.githubusercontent.com/SrinivasA06/Python/refs/heads/main/bank.csv"
                  , delim = ";")
nrow(data)
```

```
## [1] 4521
```

```
glimpse(data)
```

```
## Rows: 4,521
## Columns: 17
## $ age       <dbl> 30, 33, 35, 30, 59, 35, 36, 39, 41, 43, 39, 43, 36, 20, 31, ~
## $ job       <chr> "unemployed", "services", "management", "management", "blue-~
## $ marital   <chr> "married", "married", "single", "married", "married", "singl~
## $ education <chr> "primary", "secondary", "tertiary", "tertiary", "secondary",~
## $ default   <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no", ~
## $ balance   <dbl> 1787, 4789, 1350, 1476, 0, 747, 307, 147, 221, -88, 9374, 26~
## $ housing   <chr> "no", "yes", "yes", "yes", "yes", "no", "yes", "yes", "yes",~
## $ loan      <chr> "no", "yes", "no", "yes", "no", "no", "no", "no", "no", "yes~
## $ contact   <chr> "cellular", "cellular", "cellular", "unknown", "unknown", "c~
## $ day       <dbl> 19, 11, 16, 3, 5, 23, 14, 6, 14, 17, 20, 17, 13, 30, 29, 29,~
## $ month     <chr> "oct", "may", "apr", "jun", "may", "feb", "may", "may", "may~
## $ duration  <dbl> 79, 220, 185, 199, 226, 141, 341, 151, 57, 313, 273, 113, 32~
## $ campaign  <dbl> 1, 1, 1, 4, 1, 2, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 5, 1, 1, 1, ~
## $ pdays     <dbl> -1, 339, 330, -1, -1, 176, 330, -1, -1, 147, -1, -1, -1, -1,~
## $ previous  <dbl> 0, 4, 1, 0, 0, 3, 2, 0, 0, 2, 0, 0, 0, 0, 1, 0, 0, 2, 0, 1, ~
## $ poutcome  <chr> "unknown", "failure", "failure", "unknown", "unknown", "fail~
## $ y         <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no", ~
```

Numerical Variables

```
# List of numerical columns
numerical_columns <- c("age", "balance", "day", "duration", "campaign", "pdays", "previous")
summary(data[numerical_columns])
```

```
##       age            balance            day            duration
##  Min.   :19.00   Min.   :-3313   Min.   : 1.00   Min.   :   4
##  1st Qu.:33.00   1st Qu.:   69   1st Qu.: 9.00   1st Qu.: 104
##  Median :39.00   Median :  444   Median :16.00   Median : 185
##  Mean   :41.17   Mean   : 1423   Mean   :15.92   Mean   : 264
##  3rd Qu.:49.00   3rd Qu.: 1480   3rd Qu.:21.00   3rd Qu.: 329
##  Max.   :87.00   Max.   :71188   Max.   :31.00   Max.   :3025
##     campaign          pdays            previous
##  Min.   : 1.000   Min.   : -1.00   Min.   : 0.0000
##  1st Qu.: 1.000   1st Qu.: -1.00   1st Qu.: 0.0000
##  Median : 2.000   Median : -1.00   Median : 0.0000
##  Mean   : 2.794   Mean   : 39.77   Mean   : 0.5426
##  3rd Qu.: 3.000   3rd Qu.: -1.00   3rd Qu.: 0.0000
##  Max.   :50.000   Max.   :871.00   Max.   :25.0000
```

Key Takeaways

- Age: The dataset covers a broad age range, with a slight skew towards middle-aged individuals.
- Balance: The presence of negative balances and a few very high balances suggests varied financial situations among the respondents.
- Campaign Engagement: Most individuals were contacted once or twice during the campaign, with only a few receiving extensive engagement.
- Pdays and Previous Campaign: A large proportion of individuals have -1 (no prior contact) values in both the pdays and previous columns, suggesting limited prior engagement.
- Duration: The call duration varies significantly, with most calls being relatively short, but a few extending much longer.

Count of Missing Values

```
# Count missing values per column
missing_counts <- sapply(data, function(x) sum(is.na(x)))
missing_counts
```

```
##       age       job   marital education   default   balance   housing      loan
##         0         0         0         0         0         0         0         0
##   contact       day     month  duration  campaign     pdays  previous  poutcome
##         0         0         0         0         0         0         0         0
##         y
##         0
```

Frequency Table

```
# List of categorical columns
categorical_columns <- setdiff(names(data), numerical_columns)
lapply(data[categorical_columns], table)
```

```
## $job
##
##        admin.   blue-collar  entrepreneur     housemaid    management
##           478           946           168           112           969
##       retired self-employed      services       student    technician
##           230           183           417            84           768
##    unemployed       unknown
##           128            38
##
## $marital
##
## divorced  married   single
##      528     2797     1196
##
## $education
##
##   primary secondary  tertiary   unknown
##       678      2306      1350       187
##
## $default
##
```

```
##   no  yes
## 4445   76
##
## $housing
##
##   no  yes
## 1962 2559
##
## $loan
##
##   no  yes
## 3830  691
##
## $contact
##
##  cellular telephone   unknown
##      2896       301      1324
##
## $month
##
##  apr  aug  dec  feb  jan  jul  jun  mar  may  nov  oct  sep
##  293  633   20  222  148  706  531   49 1398  389   80   52
##
## $poutcome
##
## failure   other success unknown
##     490     197     129    3705
##
## $y
##
##   no  yes
## 4000  521
```

Key Takeaways

- Most respondents are **married** with **secondary education** and are not in loan situations.
- The **cellular contact** method is the most common, and **May** appears to be a peak contact month.