***“TERM DEPOSIT MARKETING ANALYSIS USING DECISION TREE AND CLUSTERING TECHNIQUES”***

***Course: TOPICS IN DATA SCIENCE (CS-59000)***

*Under the Guidance of,*

***Gokarna Aryal, Ph.D.,   
Department of Statistics***

*Team Includes:*

***Vaishak Balachandra- 037831852***

***Tanisha Majumder – 037132310***

*Purdue University Northwest   
Department of Computer Science   
2025-26*

# *TERM DEPOSIT MARKETING ANALYSIS USING DECISION TREE AND CLUSTERING TECHNIQUES*

## 1. Introduction

The banking industry relies heavily on effective marketing strategies to promote their products and services. This report analyzes a dataset from a Portuguese banking institution’s direct marketing campaigns conducted between May 2008 and November 2010. The campaigns primarily used phone calls to determine whether clients would subscribe to a term deposit.

The objectives of this project are:

1. To identify key factors that influence a customer’s decision to subscribe to term deposit.
2. To segment customers based on their demographic and behavioral patterns.
3. To optimize marketing strategies for different customer segments.
4. To determine the most effective contact method for each segment.

Decision trees and clustering techniques were chosen for this analysis due to their complementary nature. Decision trees provide explicit classification rules, while clustering helps identify natural groupings within the customer base.

## 2. Data Description

The dataset comes from the UCI Machine Learning Repository, created by Paulo Cortez and Sérgio Moro in 2012. It contains 45,211 instances representing direct marketing campaigns of a Portuguese banking institution.

**Link:** *https://archive.ics.uci.edu/dataset/222/bank+marketing*

The dataset includes 16 input features and 1 target variable:

**Categorical features:** - job, marital, education, default: Has credit in default? (yes, no), housing: Has housing loan? (yes, no), loan: Has personal loan? (yes, no), contact, month: Last contact month of year (jan-dec), poutcome (success, failure, other, unknown)

**Numerical features:** - age, balance, day: Last contact day of the month, duration: Last contact duration in seconds, campaign: Number of contacts during this campaign, pdays: Days since last contact in previous campaign (-1 means never contacted), previous: Number of contacts before this campaign

**Target variable:** - y: Has the client subscribed to a term deposit? (yes, no)

The target variable shows significant class imbalance, with only 11.6% of customers subscribing to a term deposit.

***R CODE***:

> ## Read the data

> ## Assuming dataset file is available in working directory

> data <- read.csv("bank-full.csv", sep=";")

> head(data, 5)

age job marital education default balance housing loan contact day month duration campaign

1 58 management married tertiary no 2143 yes no unknown 5 may 261 1

2 44 technician single secondary no 29 yes no unknown 5 may 151 1

3 33 entrepreneur married secondary no 2 yes yes unknown 5 may 76 1

4 47 blue-collar married unknown no 1506 yes no unknown 5 may 92 1

5 33 unknown single unknown no 1 no no unknown 5 may 198 1

pdays previous poutcome y

1 -1 0 unknown no

2 -1 0 unknown no

3 -1 0 unknown no

4 -1 0 unknown no

5 -1 0 unknown no

## 3. Exploratory Data Analysis (EDA)

### Data Cleaning and Preprocessing

Initial inspection of the data revealed no missing values, which simplifies our preprocessing steps. Categorical variables were converted to factors for proper analysis.

***R CODE***:

> ##### Exploratory Data Analysis (EDA)

>

> ## Dataset dimensions

> print("Dataset dimensions:")

[1] "Dataset dimensions:"

> dim(data)

[1] 45211 17

> ## Column names and structure

> print("Column names:")

[1] "Column names:"

> names(data)

[1] "age" "job" "marital" "education" "default" "balance" "housing" "loan" "contact"

[10] "day" "month" "duration" "campaign" "pdays" "previous" "poutcome" "y"

> print("Structure of the data:")

[1] "Structure of the data:"

> str(data)

'data.frame': 45211 obs. of 17 variables:

$ age : int 58 44 33 47 33 35 28 42 58 43 ...

$ job : chr "management" "technician" "entrepreneur" "blue-collar" ...

$ marital : chr "married" "single" "married" "married" ...

$ education: chr "tertiary" "secondary" "secondary" "unknown" ...

$ default : chr "no" "no" "no" "no" ...

$ balance : int 2143 29 2 1506 1 231 447 2 121 593 ...

$ housing : chr "yes" "yes" "yes" "yes" ...

$ loan : chr "no" "no" "yes" "no" ...

$ contact : chr "unknown" "unknown" "unknown" "unknown" ...

$ day : int 5 5 5 5 5 5 5 5 5 5 ...

$ month : chr "may" "may" "may" "may" ...

$ duration : int 261 151 76 92 198 139 217 380 50 55 ...

$ campaign : int 1 1 1 1 1 1 1 1 1 1 ...

$ pdays : int -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...

$ previous : int 0 0 0 0 0 0 0 0 0 0 ...

$ poutcome : chr "unknown" "unknown" "unknown" "unknown" ...

$ y : chr "no" "no" "no" "no" ...

> ## Check for missing values

> missing\_values <- colSums(is.na(data))

> print("Missing values in each column:")

[1] "Missing values in each column:"

> print(missing\_values)

age job marital education default balance housing loan contact day month

0 0 0 0 0 0 0 0 0 0 0

duration campaign pdays previous poutcome y

0 0 0 0 0 0

> print("So, no columns with NA values!!")

[1] "So, no columns with NA values!!"

### Distribution Analysis

The dataset exhibits significant class imbalance with 88.4% “no” responses and 11.6% “yes” responses to term deposit subscription, representing a ratio of approximately 7.6:1.

**Key demographic distributions:** - Age: Majority of customers are between 25-40 years old - Job: Blue-collar workers (22.5%), management (22.5%), and technicians (18.3%) form the largest groups - Education: 51.3% have secondary education, followed by tertiary (28.5%) and primary (13.9%) - Contact timing: May shows the highest volume of last contacts

***R CODE***:

> ## Distribution of the target variable

> print("Distribution of term deposit subscriptions:")

[1] "Distribution of term deposit subscriptions:"

> table(data$y)

no yes

39922 5289

> subscription\_percent <- prop.table(table(data$y)) \* 100

> print(paste("Percentage of 'yes':", round(subscription\_percent["yes"], 2), "%"))

[1] "Percentage of 'yes': 11.7 %"

> print(paste("Percentage of 'no':", round(subscription\_percent["no"], 2), "%"))

[1] "Percentage of 'no': 88.3 %"

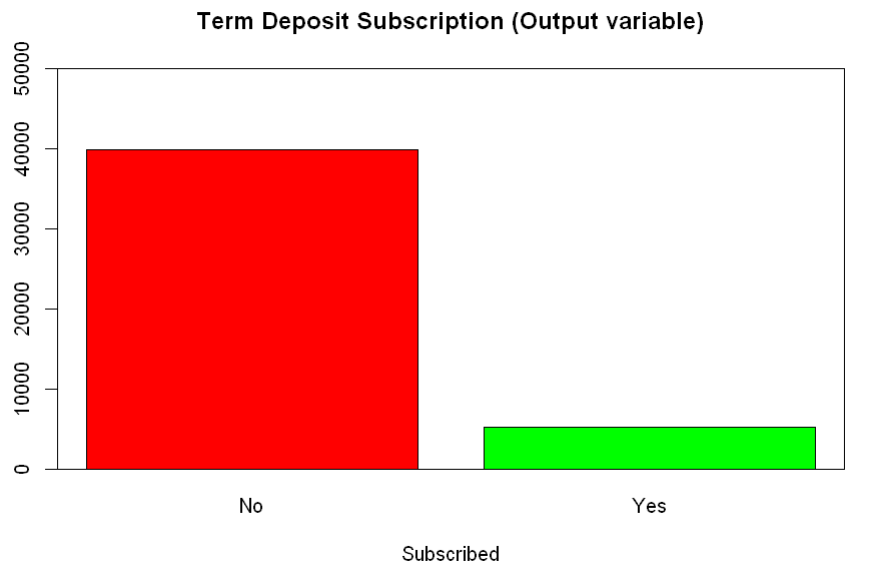
> # Print class imbalance ratio

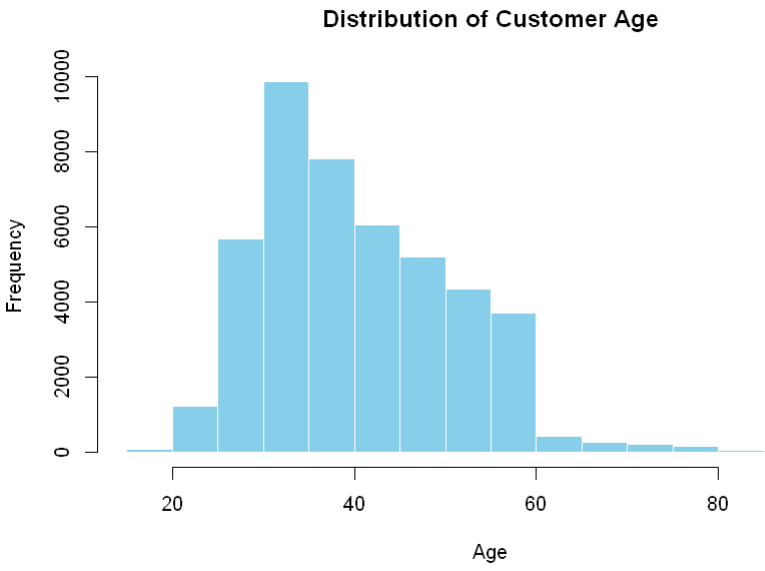
> imbalance\_ratio <- subscription\_percent["no"] / subscription\_percent["yes"]

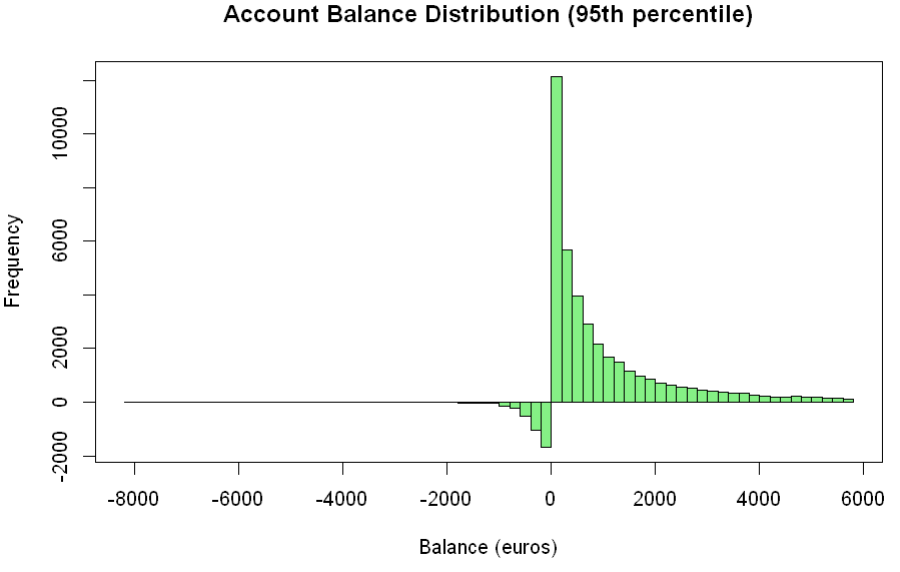
> print(paste("Class imbalance ratio (no:yes):", round(imbalance\_ratio, 2), ":1"))

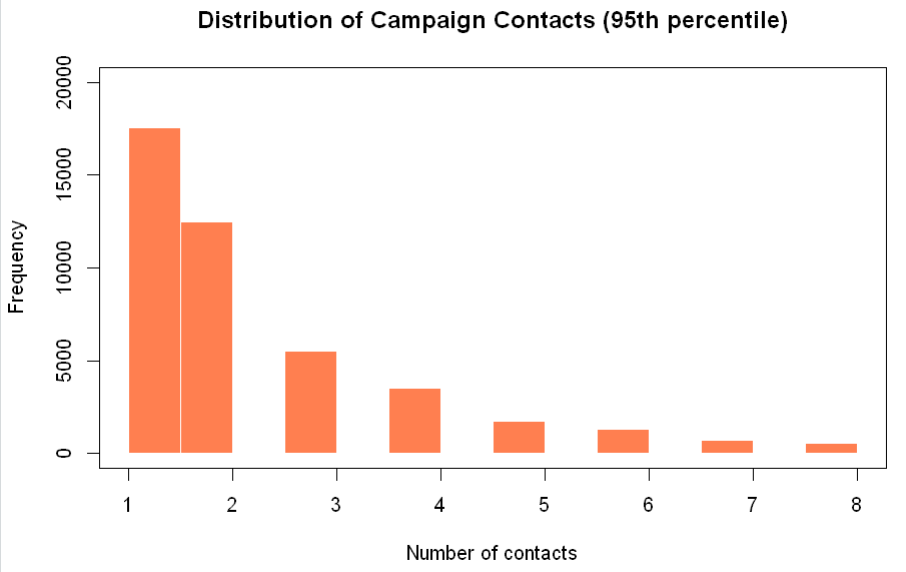
[1] "Class imbalance ratio (no:yes): 7.55 :1"

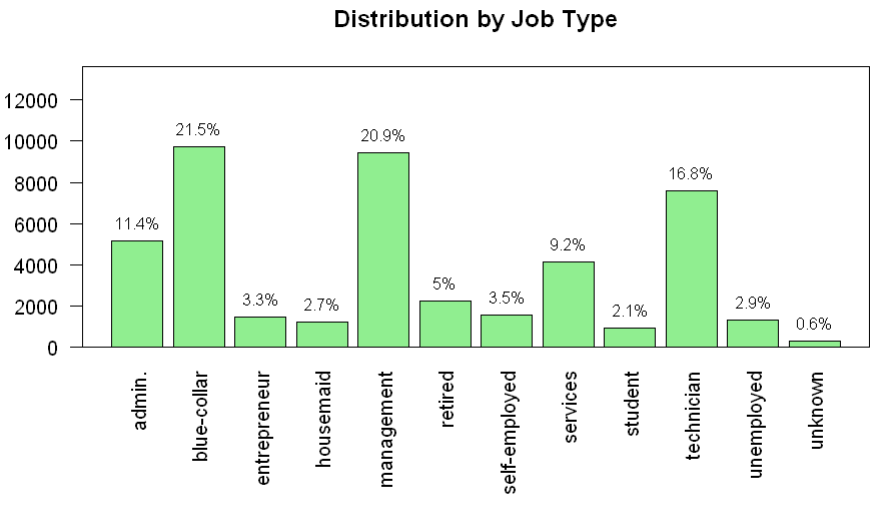
**EXPLORATORY DATA ANALYSIS: PLOTS AND OBSERVATIONS:**

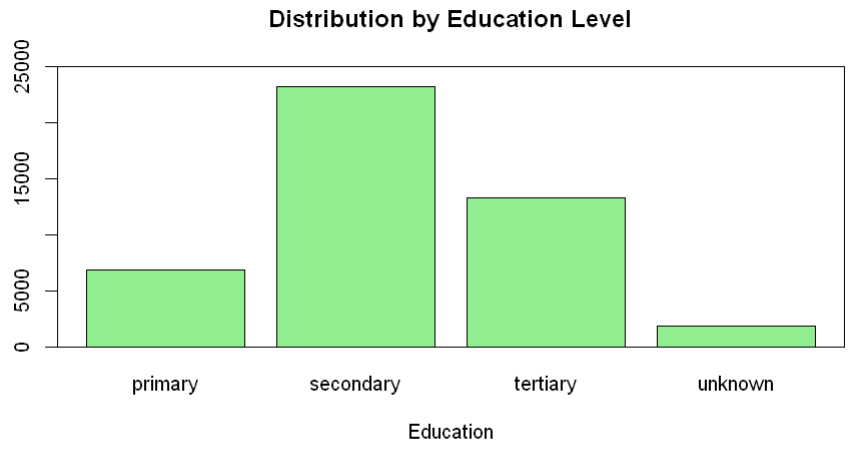
  
*Fig 1: Distribution of term deposit subscription  
(Insight: Only 11.6% of the customers agreed to opt term deposit subscription)*

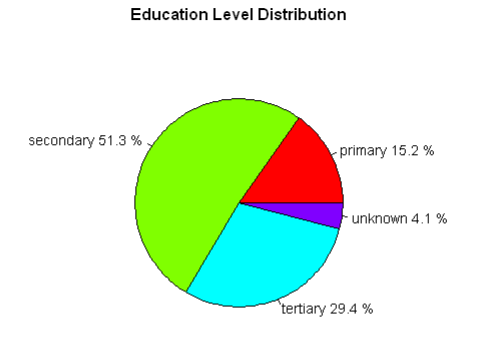
****  
*Fig 2: Distribution of Age of the customers  
(Insight: Majority of the customers were aged between 25-40 years)*

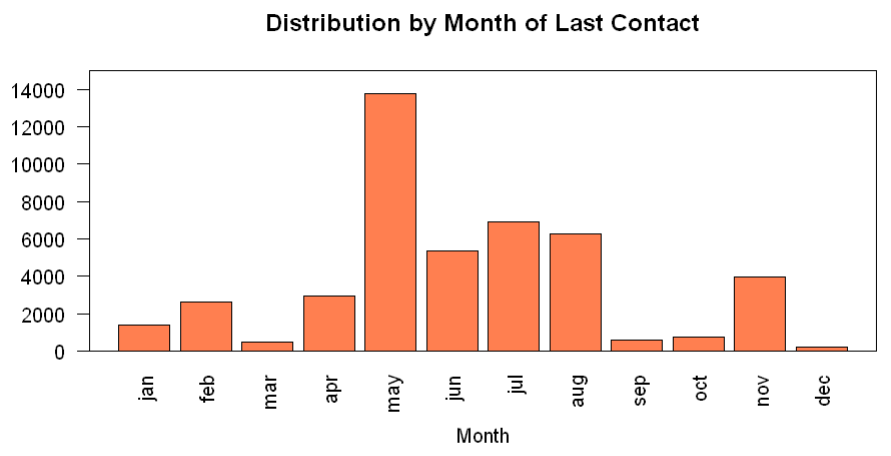
 *Fig 3: Account Balance Distribution  
(Insight: Most of the customers who reached out have a positive account balance)*

 *Fig 4: Distribution of Campaign Contacts  
(Insight: Around 66% of the customers have been contacted <= twice during the campaign)*

**** *Fig 5: Distribution of Customer’s Job Type  
(Insight: Most common job category includes Blue-collar, Management and Technician)*

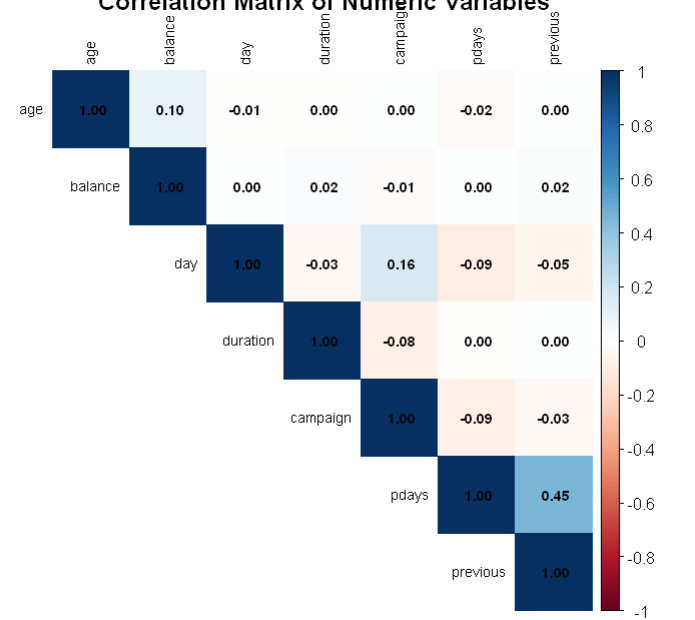
**** *Fig 6: Distribution of Education Levels  
(Insight: 51.3% of the customers have secondary education)*

  
*Fig 7: Pie Distribution of Education Levels  
(Insight: 51.3% of the customers have secondary education)*

 *Fig 8: Distribution of Age of the customers  
(Insight:* *More customers were contacted in May than in any other month. So, summertime is the time where most of the last contacts were made!!)*

### Correlation Analysis

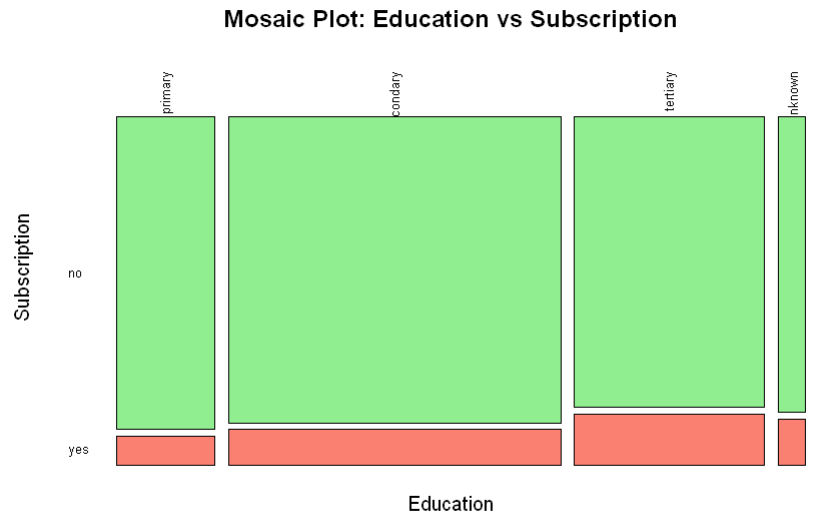
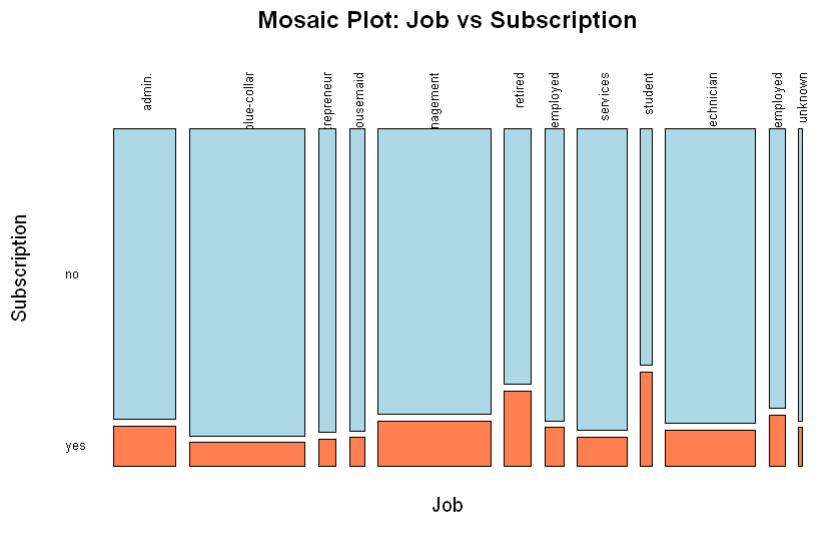
Correlation analysis revealed considerable independence among numerical variables, suggesting minimal multicollinearity issues. The strongest correlation was observed between ‘previous’ and ‘poutcome’ variables.

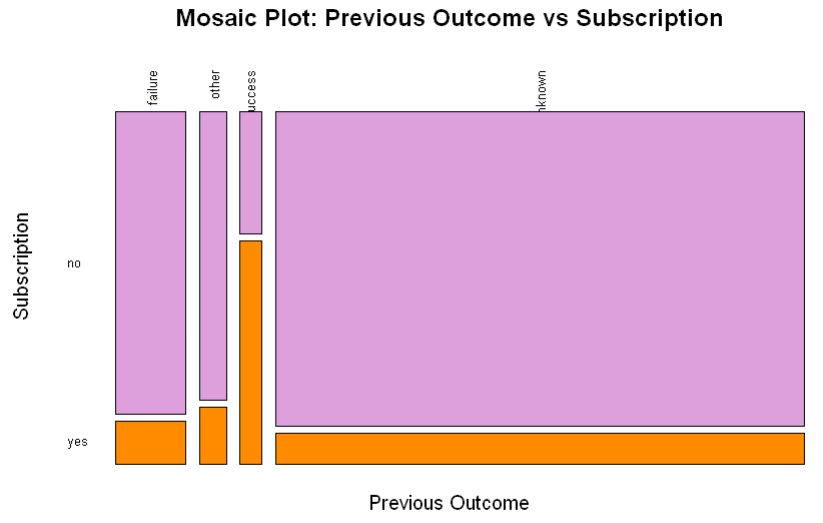
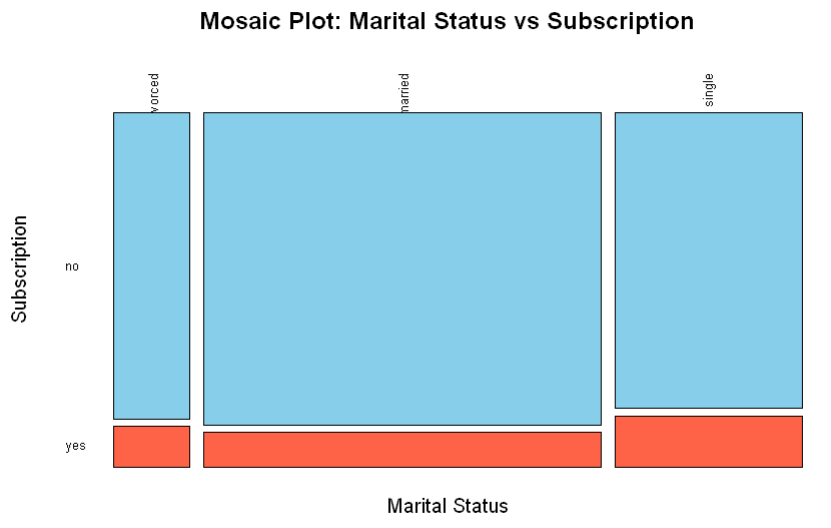
 *Fig 9: Correlation Matrix of Numerical Variables  
(Insight:* *Large Independency is observed in the dataset!!)*

### Cross-tabulation Analysis (via Mosaic Plots)

Few Cross-tabulation with the target variable revealed several interesting patterns:

* **Job vs. Subscription:** - Students (28.5%) and retired individuals (23.3%) have the highest subscription rates - Blue-collar workers (7.6%) have among the lowest subscription rates.
* **Education vs. Subscription:** - Tertiary education (14.3%) shows the highest subscription rate - Unknown education has the lowest at 8.5%.
* **Previous outcome vs. Subscription:** - Previous success (64.8%) is strongly associated with new subscriptions - Previous failure (10.5%) has a lower subscription rate than even those with no previous contact.





*Fig 10: Cross Tabulation analysis of few variables (Job, Education, Marital Status, Previous Outcome) against output variable (Subscription) via Mosaic Plots*

## Supervised Learning: Decision Tree Classification

***R CODE***: *(70% training, rest testing)*

> # Function to build and evaluate decision tree

> evaluate\_dt <- function(train\_set, test\_set, dataset\_name) {

+ print(paste("Building decision tree using", dataset\_name))

+

+ # Build the decision tree model

+ dt\_model <- rpart(y ~ age + job + marital + education + default + balance +

+ housing + loan + contact + month + duration + campaign +

+ pdays + previous + poutcome,

+ data = train\_set,

+ method = "class",

+ control = rpart.control(cp = 0.001, minbucket = 20))

+

+ # Prune the tree to prevent overfitting

+ cp\_optimal <- dt\_model$cptable[which.min(dt\_model$cptable[,"xerror"]),"CP"]

+ dt\_model\_pruned <- prune(dt\_model, cp = cp\_optimal)

+

+ # Plot the decision tree

+ rpart.plot(dt\_model\_pruned,

+ extra = 104, # Show fitted values and percentages

+ box.palette = "RdBu",

+ shadow.col = "gray",

+ nn = TRUE, # Show node numbers

+ main = paste("Decision Tree -", dataset\_name))

+

+ # Evaluate the decision tree on test data

+ dt\_predictions <- predict(dt\_model\_pruned, test\_set, type = "class")

+ dt\_confusion\_matrix <- table(Predicted = dt\_predictions, Actual = test\_set$y)

+ print(paste("Decision Tree Confusion Matrix for", dataset\_name, ":"))

+ print(dt\_confusion\_matrix)

+

+ # Calculate performance metrics

+ dt\_accuracy <- sum(diag(dt\_confusion\_matrix)) / sum(dt\_confusion\_matrix)

+

+ # Handle the case where some classes might be missing in the predictions

+ if (dim(dt\_confusion\_matrix)[1] < 2 || dim(dt\_confusion\_matrix)[2] < 2) {

+ dt\_precision <- dt\_recall <- dt\_f1 <- NA

+ } else {

+ dt\_precision <- dt\_confusion\_matrix[2,2] / sum(dt\_confusion\_matrix[2,])

+ dt\_recall <- dt\_confusion\_matrix[2,2] / sum(dt\_confusion\_matrix[,2])

+ dt\_f1 <- 2 \* dt\_precision \* dt\_recall / (dt\_precision + dt\_recall)

+ }

+

+ print(paste("Decision Tree Accuracy:", round(dt\_accuracy, 4)))

+ print(paste("Decision Tree Precision:", round(dt\_precision, 4)))

+ print(paste("Decision Tree Recall:", round(dt\_recall, 4)))

+ print(paste("Decision Tree F1 Score:", round(dt\_f1, 4)))

+

+ # Variable importance

+ dt\_var\_importance <- dt\_model\_pruned$variable.importance

+ if (length(dt\_var\_importance) > 0) {

+ dt\_var\_importance\_df <- data.frame(

+ Variable = names(dt\_var\_importance),

+ Importance = dt\_var\_importance

+ )

+ dt\_var\_importance\_df <- dt\_var\_importance\_df[order(-dt\_var\_importance\_df$Importance),]

+

+ print(paste("Decision Tree Variable Importance for", dataset\_name, ":"))

+ print(dt\_var\_importance\_df)

+

+ # Return results

+ return(list(

+ model = dt\_model\_pruned,

+ accuracy = dt\_accuracy,

+ precision = dt\_precision,

+ recall = dt\_recall,

+ f1 = dt\_f1,

+ var\_importance = dt\_var\_importance\_df,

+ predictions = dt\_predictions

+ ))

+ } else {

+ print("No variable importance available - model might be too simple")

+ return(list(

+ model = dt\_model\_pruned,

+ accuracy = dt\_accuracy,

+ precision = dt\_precision,

+ recall = dt\_recall,

+ f1 = dt\_f1,

+ var\_importance = NULL,

+ predictions = dt\_predictions

+ ))

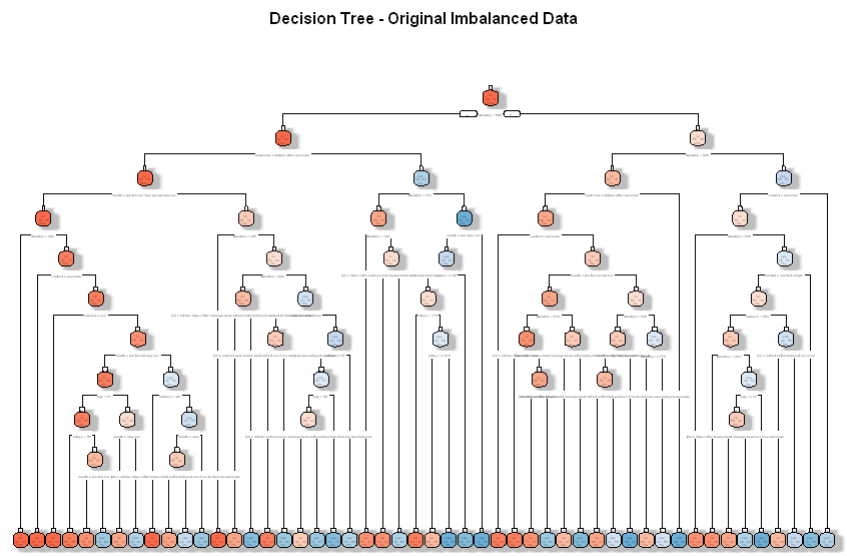
+ }

+ }

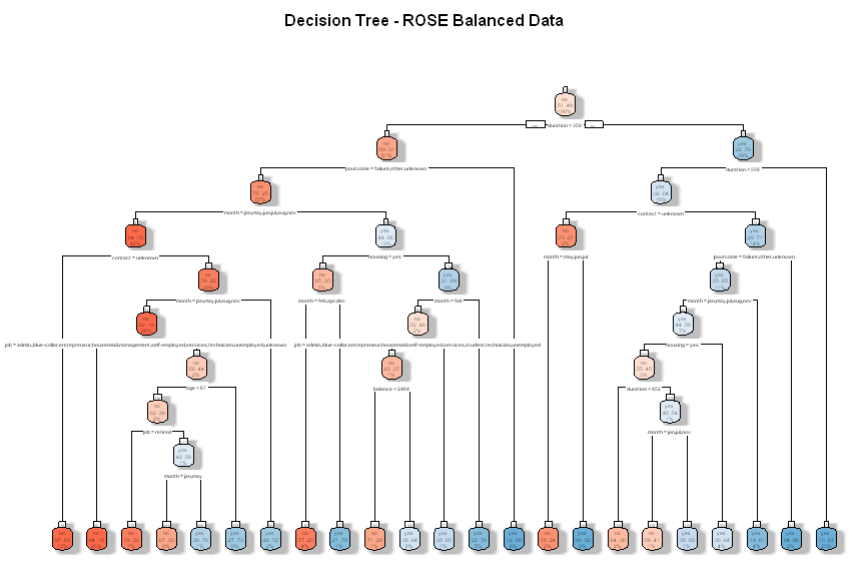
### Model Setup and Evaluation: [DUE TO SPACE CONSTRAINT, IMAGE SIZE HAS BEEN REDUCED]

To address the class imbalance issue, we tested three approaches:

1. **Original imbalanced data**



1. **ROSE balanced data**



1. **Under-sampled data (reducing majority class)**

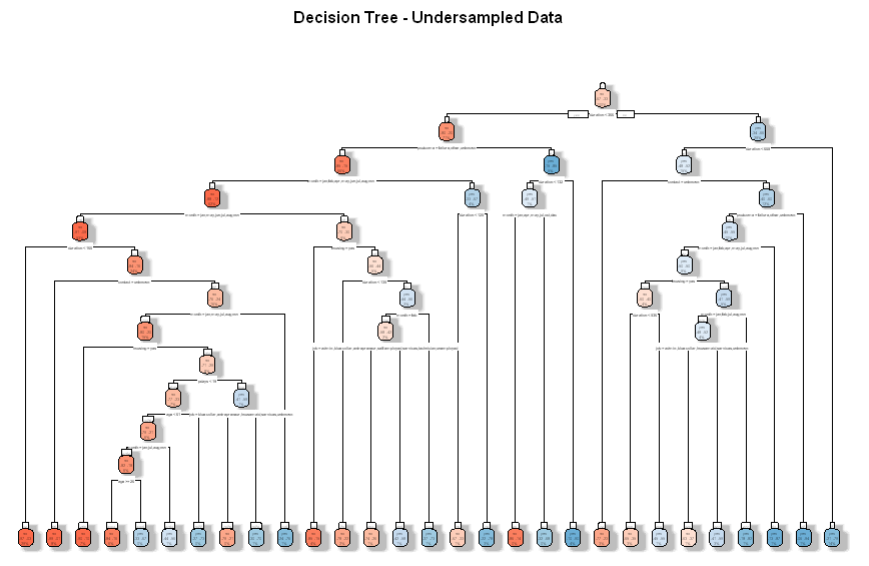


Fig 10: Decision Tree Plots of 3 different cases

***R CODE***:

> # ------------------------

> # Method 1: ROSE (Random Over-Sampling Examples)

> # ------------------------

> library(ROSE)

> set.seed(123)

> rose\_train <- ROSE(y ~ ., data = train\_data, seed = 123)$data

> cat("ROSE class distribution:\n")

ROSE class distribution:

> print(table(rose\_train$y))

no yes

16022 15625

> # ------------------------

> # Method 2: Random Undersampling

> # ------------------------

> set.seed(123)

> yes\_idx <- which(train\_data$y == "yes")

> no\_idx <- which(train\_data$y == "no")

> no\_keep <- min(length(no\_idx), length(yes\_idx) \* 2)

> undersampled\_train <- train\_data[c(yes\_idx, sample(no\_idx, no\_keep)), ]

> cat("Undersampling class distribution:\n")

Undersampling class distribution:

> print(table(undersampled\_train$y))

no yes

7446 3723

### Variable Importance Analysis

The decision tree model identified the following variables as most important for predicting term deposit subscription:

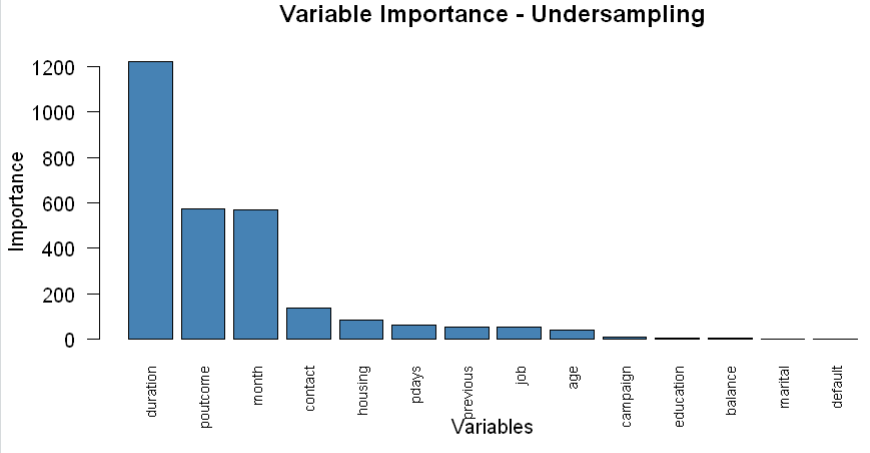
1. Duration (of last contact)

2. Previous outcome

3. Month

4. Age

5. Balance

 *Fig 11: Variable importance in the decision tree model*

The decision tree structure reveals that clients with longer contact durations (>441 seconds) and previous successful outcomes are highly likely to subscribe. Meanwhile, shorter contacts (<155 seconds) typically result in non-subscription.

***R CODE***:

> print(model\_comparison)

Model Accuracy Precision Recall F1\_Score

1 Original 0.9057800 0.6128527 0.4993614 0.5503167

2 ROSE 0.8278531 0.3862089 0.8333333 0.5278059

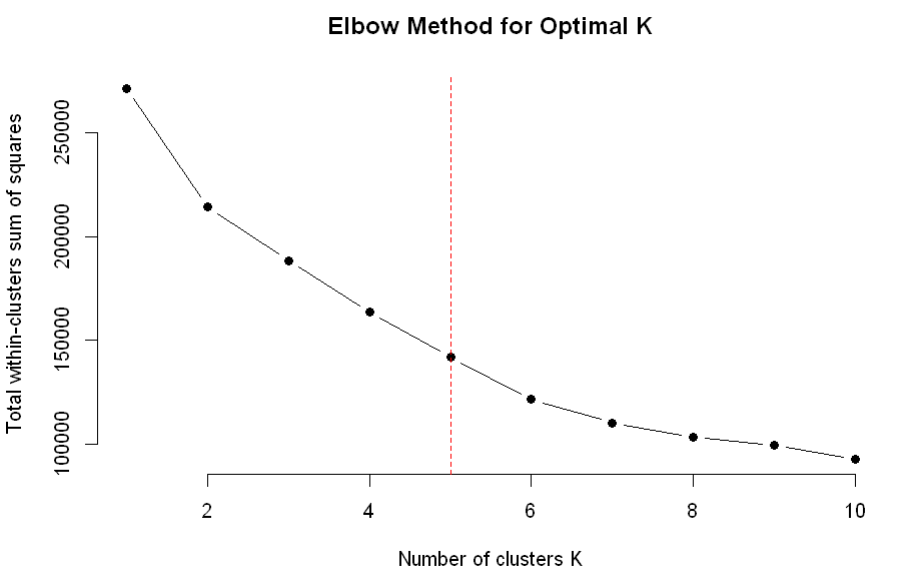
3 Undersampling 0.8650840 0.4519651 0.7931034 0.5757997

**Inference**: Best model based on F1 score: ***Under-sampling****,* with a compromise in accuracy.

## 5. Unsupervised Learning: Clustering Analysis

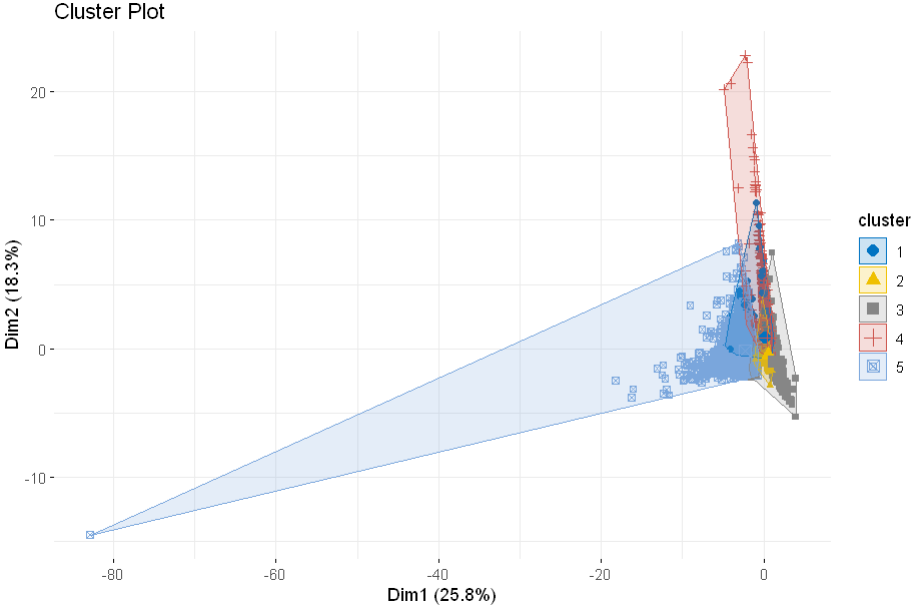
### Cluster Determination

Clustering analysis was performed using K-means on the numerical features: age, balance, duration, campaign, pdays, and previous. The optimal number of clusters was determined to be 5 using the elbow method.

  
*Fig 12: Elbow method for determining optimal number of clusters*

### Cluster Characteristics

1. **Cluster 1** has the highest avg duration and the highest subscription rate (y\_numeric = 0.469).
2. **Cluster 2** shows high campaign exposure but very low subscription rate (y\_numeric = 0.025).
3. **Cluster 3** contains older individuals with high balances but a low subscription rate.
4. **Cluster 4** had prior contacts (high previous and pdays) and a moderate subscription rate.
5. **Cluster 5** consists of younger individuals with low balances and a low subscription rate.

 *Figure 13: Optimal k=5 clustering*

***R CODE***: *(Subscription rate by cluster)*

> # Cluster subscription rates

> cluster\_subscription <- aggregate(y\_numeric ~ cluster, data = data, FUN = mean)

> print("Subscription Rate by Cluster:")

[1] "Subscription Rate by Cluster:"

> print(cluster\_subscription)

cluster y\_numeric

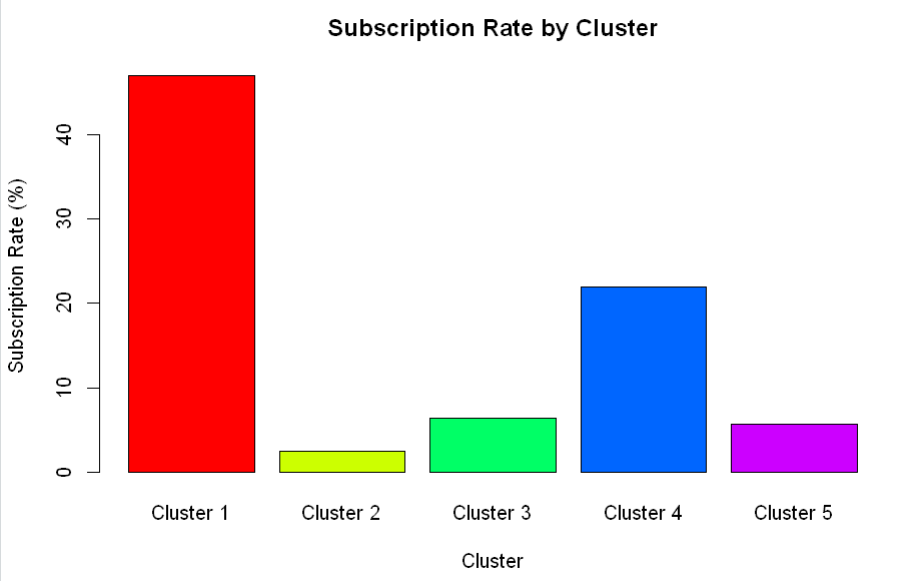
1 1 0.46946108

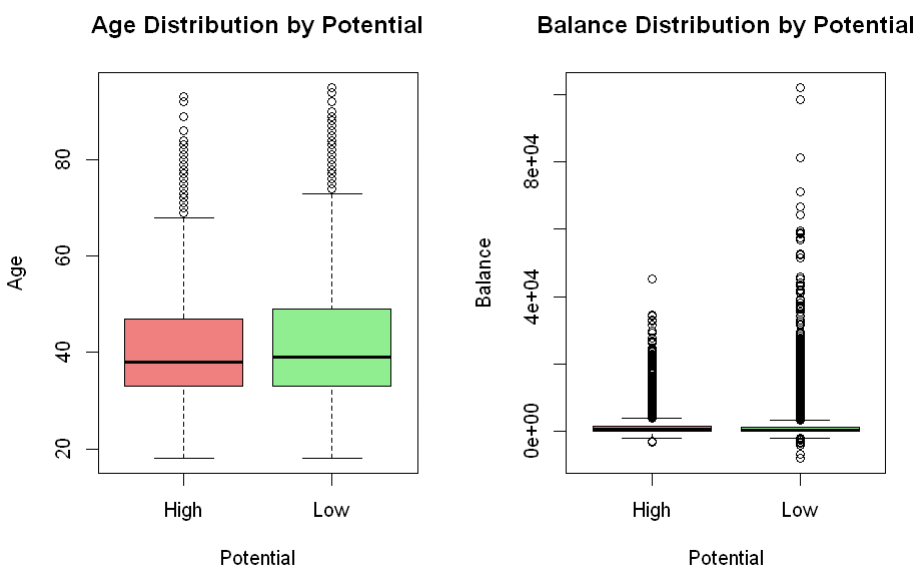
2 2 0.02490660

3 3 0.06439959

4 4 0.21876974

5 5 0.05755799

 *Figure 14: Subscription rates by cluster*

*Figure 15: Age/Balance Distribution by Potential*

## 6. Conclusion and Key Findings

This analysis has yielded several actionable insights for optimizing the bank’s term deposit marketing campaigns:

1. **Key Predictors of Subscription:**
   * Duration of contact is the strongest predictor, with longer conversations (>7 minutes) strongly associated with subscription
   * Previous campaign success is highly influential, with 64.8% of previously successful clients subscribing again
   * Contact timing matters, with May, July, and August showing higher subscription rates
   * Older clients and those with higher balances are more receptive
2. **Effective Segmentation Strategy:**
   * Focus marketing resources on high-potential segments (Clusters 2 and 3)
   * For low-potential segments, use less intensive contact strategies or different products
3. **Optimized Contact Approach:**
   * Use cellular contact for high-potential clients
   * Match contact timing with client availability patterns
   * Allocate more time for conversations with high-potential clients (aim for >7 minutes)
   * Consider previous interaction outcomes when planning new contacts
4. **Demographic Targeting:**
   * Focus on retired individuals, students, and management positions
   * Clients with tertiary education show higher subscription rates
   * Clients with successful previous interactions should be prioritized

These findings can help the bank significantly improve its marketing efficiency by focusing resources on segments with higher subscription probability and using contact methods tailored to each segment’s characteristics.

## 8. Appendix

### A. Dataset Background and Source

The dataset used in this project is the ‘Bank Marketing’ dataset, made publicly available through the UCI Machine Learning Repository. It was created by Paulo Cortez and Sérgio Moro in 2012. The dataset is derived from direct marketing campaigns of a Portuguese banking institution, conducted between May 2008 and November 2010. The campaigns primarily used phone calls to determine if clients would subscribe to a term deposit.

**Citation:** Moro et al., 2011. S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM’2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

**Link:** *https://archive.ics.uci.edu/dataset/222/bank+marketing*