STAT 516 Midterm 3: Course Project

Analysis Using R

Nick Arboscello George Bujoreanu

April 17, 2025

Contents

1 Introduction

This project explores a marketing dataset from a Portuguese banking institution. The dataset was collected during a direct marketing campaign promoting term deposit subscriptions and contains client-level information. Our analysis focuses on identifying key factors that influence a customer's decision to subscribe to a term deposit, providing insights into customer behavior and marketing effectiveness. https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets?resource=download&select=test.csv

This dataset was selected due to its practical relevance to business analytics and its alignment with our field of study in Computer Information Systems. The dataset's structure, which includes both numerical and categorical variables, enables the application of a variety of statistical methods to uncover meaningful relationships.

Variables

Categorical Variables:

- job: Type of job (e.g., admin., technician, management)
- marital: Marital status (e.g., married, single, divorced)
- education: Highest education level attained (e.g., primary, secondary, tertiary)

Numerical Variables:

- age: Age of the client (in years)
- duration: Duration of last contact in seconds
- balance: Account balance (in euros)

Outcome Variable:

y: Whether the client subscribed to a term deposit (yes or no)

This subset of variables allows us to investigate how customer demographics and call characteristics relate to subscription behavior. Our analysis will address three focused research questions: (1) Is there a significant difference in account balance across different levels of education? This will be examined using a one-way ANOVA. (2) Is there an association between marital status and job type? We will explore this using a chi-square test of independence. (3) Does call duration significantly predict whether a client will subscribe to a term deposit? To answer this, we will use a logistic regression model with duration as the predictor and y (subscription) as the binary outcome.

Background

Effective marketing strategies are crucial for financial institutions aiming to convert potential clients into long-term deposit holders. Direct marketing, particularly via telephone, remains

a cost-effective yet complex channel, as it requires targeted efforts to engage the right clients at the right time. The dataset used in this project originates from a real-world marketing campaign by a Portuguese banking institution and has been the subject of academic research. Moro, Cortez, and Rita (2014) conducted a comprehensive analysis of this dataset, applying data mining techniques to develop predictive models for term deposit subscriptions. Their study, published in Decision Support Systems, demonstrated the value of using customer and call-related attributes—such as age, job type, and call duration—to forecast campaign success and improve targeting strategies. This research underscores the potential of data-driven approaches in banking and provides a foundation for our own statistical exploration using R.

Reference: Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

2 Exploratory Data Analysis (EDA)

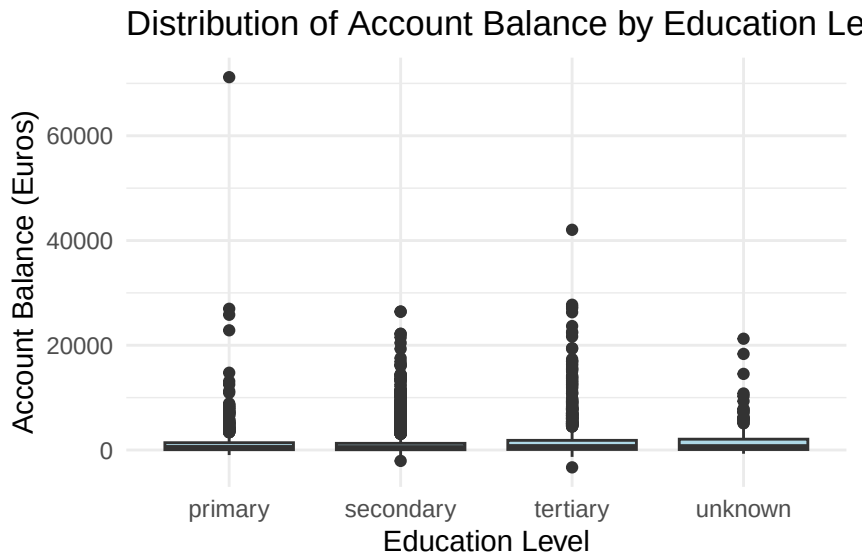
The goal of our exploratory data analysis is to uncover general trends, patterns, and potential relationships within the dataset that inform our research questions. Specifically, we aim to understand the distribution of customer demographics (age, job, marital status, education), assess how account balances and call durations vary across different groups, and identify any potential outliers or anomalies that may impact our analysis. Through summary statistics and visualizations, we hope to gain insight into how our variables may relate to account balance and marketing success.

Visualizations

```
# Boxplot: Account Balance by Education Level
library(ggplot2)
```

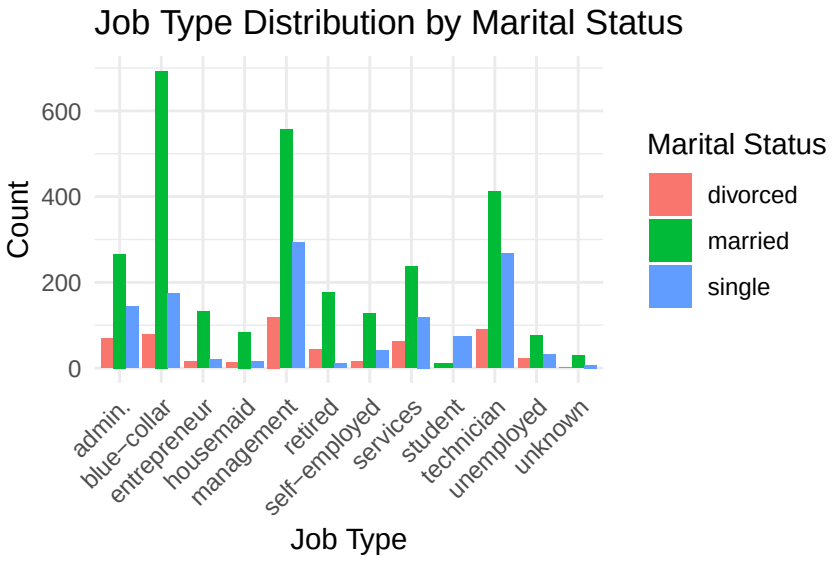
```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
ggplot(test, aes(x = education, y = balance)) +
  geom_boxplot(fill = "lightblue") +
  labs(
    title = "Distribution of Account Balance by Education Level",
    x = "Education Level",
    y = "Account Balance (Euros)"
  ) +
  theme_minimal()
```

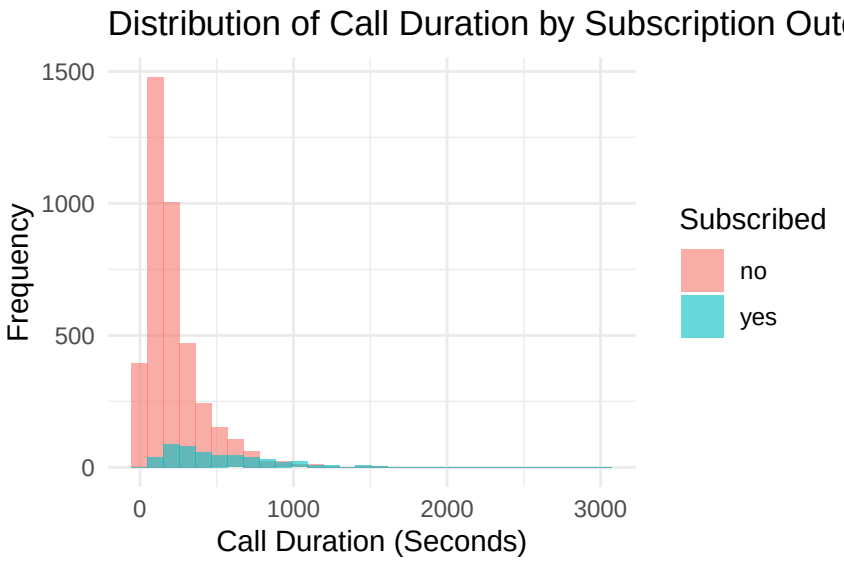


```r
# Bar Plot: Job Type by Marital Status
library(ggplot2)

ggplot(test, aes(x = job, fill = marital)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Job Type Distribution by Marital Status",
    x = "Job Type",
    y = "Count",
    fill = "Marital Status"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Histogram of Call Duration Colored by Subscription Outcome
ggplot(test, aes(x = duration, fill = y)) +
  geom_histogram(position = "identity", bins = 30, alpha = 0.6) +
  labs(
    title = "Distribution of Call Duration by Subscription Outcome",
    x = "Call Duration (Seconds)",
    y = "Frequency",
    fill = "Subscribed"
  ) +
  theme_minimal()
```




```
# Summary statistics
summary(test[, c("age", "balance", "duration")])
```

```
##       age            balance          duration
##  Min.   :19.00   Min.   :-3313   Min.   :   4
##  1st Qu.:33.00   1st Qu.:   69   1st Qu.: 104
##  Median :39.00   Median :  444   Median : 185
##  Mean   :41.17   Mean   : 1423   Mean   : 264
##  3rd Qu.:49.00   3rd Qu.: 1480   3rd Qu.: 329
##  Max.   :87.00   Max.   :71188   Max.   :3025
```

Discussion The boxplot of account balance by education level reveals high variability in balance within each education group, with tertiary education showing the widest range and the most extreme outliers. Notably, all education levels contain outliers, including some clients with negative balances, which may reflect overdrafts. This variability suggests a potential, but non-linear, relationship between education and financial behavior—supporting the motivation for our first research question involving ANOVA.

The bar plot of job type by marital status shows that certain jobs, such as "blue-collar" and "admin.," are much more common overall, especially among married individuals. Some job types (like "student" or "unemployed") appear less frequently and have skewed distributions across marital groups. These uneven frequencies may affect the chi-square test assumptions and suggest social patterns in employment and relationship status.

The histogram of call duration, separated by whether the client subscribed, shows that longer calls are generally associated with a higher chance of subscription. Most calls are short (<300 seconds), but the long tail includes calls lasting several minutes. This pattern supports our third research question and suggests that call duration may be a meaningful predictor in a logistic regression model.

Summary Statistics Table:

- Age: Mean age is ~41 years, with a wide range (19 to 95), indicating a diverse clientele.
- Balance: The mean balance is €1,423, but the high standard deviation (~€3,009) and negative minimum (-€3,313) highlight extreme variability.
- Duration: Call duration has a long right-skewed distribution, ranging from 4 to over 2,000 seconds.

3 Research Questions

In this study, we aim to investigate how client demographics and marketing-related characteristics influence financial behaviors and outcomes within the context of a direct marketing campaign conducted by a Portuguese bank. We developed three focused research questions that align with the available data and allow us to explore both behavioral and statistical relationships. These questions were chosen to guide a meaningful and data-driven analysis that connects demographic patterns to marketing effectiveness and customer decisions.

The first research question asks: Is there a significant difference in account balance across different levels of education? This question is grounded in the idea that education level may correlate with financial literacy, income, and overall financial stability. Higher educational attainment may be associated with higher-paying jobs and better money management, potentially leading to greater bank balances. To test this, we will use a one-way Analysis of Variance (ANOVA), which is suitable for comparing the means of a continuous variable (in this case, account balance) across more than two independent groups (education levels). ANOVA allows us to determine whether the differences in mean balance between the education categories—such as primary, secondary, and tertiary—are statistically significant.

Our second research question is: Is there an association between marital status and job type? This question explores potential demographic and occupational relationships. For example, certain job types may be more common among married individuals, while others may be prevalent among singles due to lifestyle choices or economic factors. Understanding these patterns can inform both customer profiling and targeted marketing strategies. To evaluate this question, we will use a chi-square test of independence. This statistical test is appropriate for determining whether two categorical variables—marital status and job type—are associated or independent from one another. It helps identify whether the observed distribution of job types varies significantly across different marital status groups.

The third research question investigates: Does call duration significantly predict whether a client will subscribe to a term deposit? This question is rooted in the hypothesis that longer calls may indicate greater client interest, engagement, or persuasion success. As call duration increases, we may expect the likelihood of a positive response to the marketing effort (i.e., a subscription) to also increase. To test this, we will apply a logistic regression model, where the binary outcome variable is whether or not the client subscribed (y), and the predictor variable is call duration. Logistic regression is the appropriate method when the goal is to model the probability of a binary outcome based on one or more predictor variables. It will allow us to quantify the relationship between call length and subscription likelihood, and assess whether this relationship is statistically significant.

4 Methods and Results