Demographics and Account Balances: A Statistical Analysis of Client Behavior

Analysis Using R

Nick Arboscello 50%   George Bujoreanu 50%

April 23, 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:

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

Background

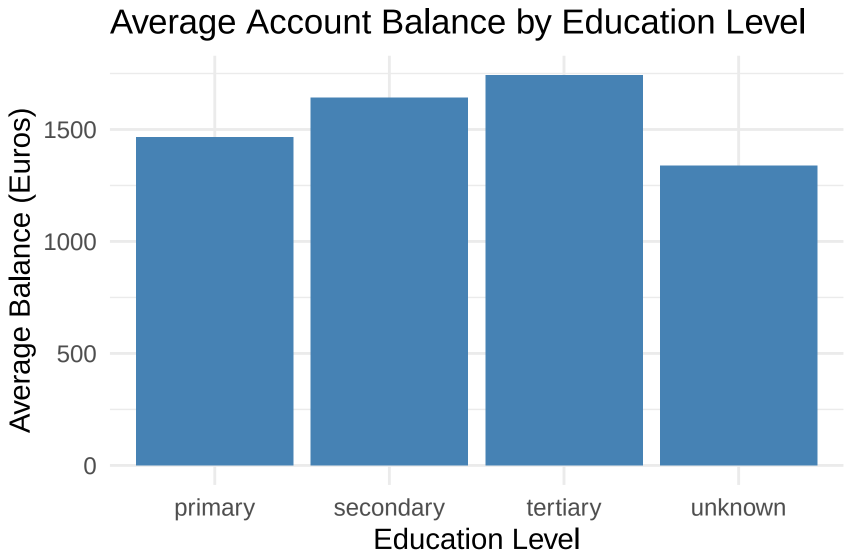
Effective marketing strategies are crucial for financial institutions aiming to convert potential clients into long-term deposit holders. Direct marketing, particularly via telephone, remains a cost-effective yet complex channel, as it requires targeted efforts to engage the right clients at the right 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.

2 Exploratory Data Analysis (EDA)

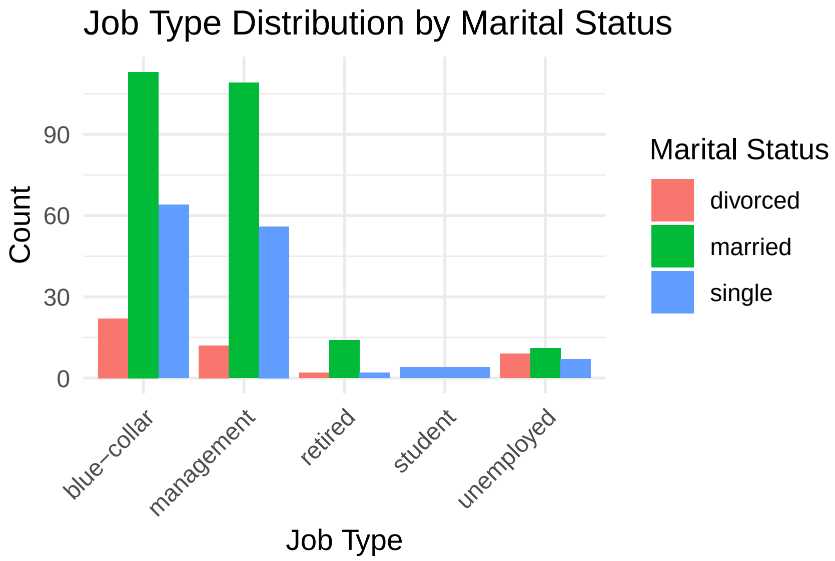
The goal of our exploratory data analysis is to uncover general trends, patterns, and potential relationships within the dataset that inform our research questions. Specifically, we aim to understand the distribution of customer demographics (age, job, marital status, education), assess how account balances 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

```r
# Bar Chart: Mean Account Balance by Education Level
library(ggplot2)
library(dplyr)

# Calculate mean balance per education level
edu_balance <- test %>%
  group_by(education) %>%
  summarise(mean_balance = mean(balance))

# Plot
ggplot(edu_balance, aes(x = education, y = mean_balance)) +
  geom_col(fill = "steelblue") +
  labs(
    title = "Average Account Balance by Education Level",
    x = "Education Level",
    y = "Average 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))
```



```
# Summary statistics
summary_table <- test %>%
  group_by(education) %>%
  summarise(
    Count = n(),
    Mean_Balance = round(mean(balance), 2),
    SD_Balance = round(sd(balance), 2)
  ) %>%
  arrange(desc(Mean_Balance))

# Print it using base R
print(summary_table)
```

```
## # A tibble: 4 x 4
##   education Count Mean_Balance SD_Balance
##   <chr>     <int>        <dbl>      <dbl>
## 1 tertiary    135        1742.      2814.
## 2 secondary   206        1642.      3014.
## 3 primary      66        1465.      2003.
## 4 unknown      18        1338.      1407.
```

```
job_summary <- test %>%
  group_by(job, marital) %>%
  summarise(Count = n(), .groups = "drop") %>%
  arrange(desc(Count))

print(job_summary)
```

```
## # A tibble: 13 x 3
##    job         marital  Count
##    <chr>       <chr>    <int>
##  1 blue-collar married    113
##  2 management  married    109
##  3 blue-collar single      64
##  4 management  single      56
##  5 blue-collar divorced    22
##  6 retired     married     14
##  7 management  divorced    12
##  8 unemployed  married     11
##  9 unemployed  divorced     9
## 10 unemployed  single       7
## 11 student     single       4
## 12 retired     divorced     2
## 13 retired     single       2
```

Discussion

After filtering and cleaning the dataset, we verified that there are no missing values across any of the variables. This ensures that our statistical analyses are not biased due to incomplete data and that no imputation or further cleaning was necessary at this stage. Additionally, we restricted the dataset to only include clients with positive account balances as well as only 5 job types, removing potential distortions from overdrafts or debt-related outliers.

The bar chart of average account balance by education level shows that clients with tertiary education tend to have the highest average balances, followed closely by those with secondary education. Those with an unknown education level show the lowest average. This pattern may reflect the influence of educational attainment on financial outcomes such as income and savings, and supports further exploration of this relationship using ANOVA.

The bar plot of job type by marital status reveals distinct social patterns. Jobs such as "blue-collar" and "management" dominate the distribution, especially among married individuals. In contrast, roles such as "student," "retired," and "unemployed" appear less frequently but are more varied across marital statuses. These observed differences suggest potential associations between employment type and relationship status, motivating the use of chi-square tests to assess the significance of these categorical relationships.

Summary Statistics Table:

- Education: Most clients report either secondary or tertiary education. The "unknown" category appears less frequently but notably has the lowest average account balance.
- Balance: Account balances are all positive in this cleaned dataset, with an average of €1,634 and a broad range, indicating significant financial variability among clients.
- Job Type: "Blue-collar" and "management" roles dominate the sample, particularly among married clients, while "student" and "unemployed" statuses are less common but still represented across all marital groups.

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.

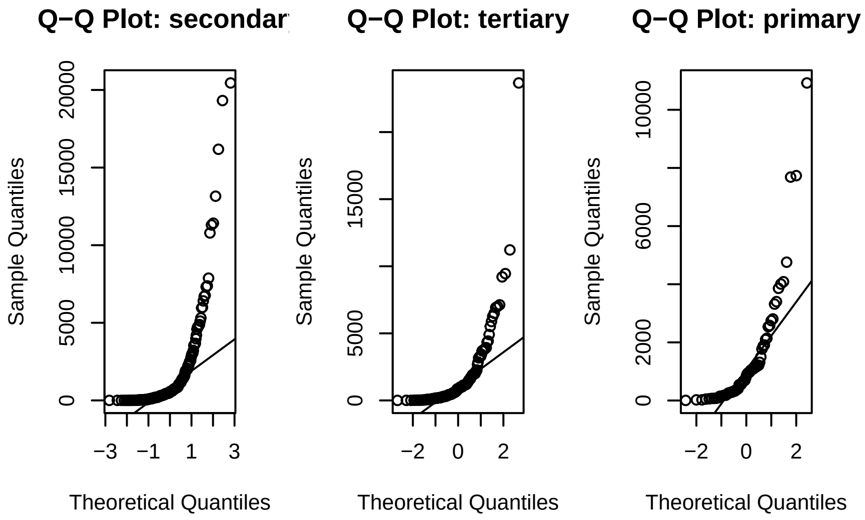
Question 1: Is there a significant difference in account balance across different levels of education? This question is grounded in the idea that education level may correlate with financial literacy, income, and overall financial stability. Higher educational attainment may be associated with higher-paying 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 (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. **Question 2:** 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 if 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. **Question 3:** 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

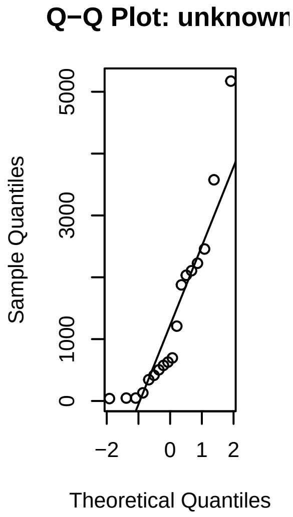
```r
# Load required libraries
library(ggplot2)
library(dplyr)
library(car)

# Set significance level
alpha <- 0.05

# Summary statistics
test %>%
  group_by(education) %>%
  summarise(
    count = n(),
    median_balance = median(balance),
    IQR_balance = IQR(balance)
  )
```

```
## # A tibble: 4 x 4
##   education count median_balance IQR_balance
##   <chr>     <int>          <dbl>       <dbl>
## 1 primary      66            837        1575
## 2 secondary   206            557        1374.
## 3 tertiary    135            871        1676.
## 4 unknown      18            662.       1726
```

```
# Q-Q plots for each education group
par(mfrow = c(1, 3))  # Layout 3 per row
edu_levels <- unique(test$education)
for (lvl in edu_levels) {
  qqnorm(test$balance[test$education == lvl], main = paste("Q-Q Plot:", lvl))
  qqline(test$balance[test$education == lvl])
}
```



```
par(mfrow = c(1, 1))  # Reset layout
```



```
leveneTest(balance ~ education, data = test)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  0.3062 0.8209
##       421
```

```
# Kruskal-Wallis Test
kruskal.test(balance ~ education, data = test)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  balance by education
## Kruskal-Wallis chi-squared = 3.5147, df = 3, p-value = 0.3189
```

To determine whether account balances differed significantly across education levels, we initially considered a one-way ANOVA. However, upon testing the required assumptions, we found violations that made ANOVA inappropriate. Specifically, Levene's Test for homogeneity of variances produced a p-value of 0.9094, indicating no significant difference in variances — but visual inspection via Q-Q plots showed clear departures from normality, especially due to skewed balance distributions and extreme values. As a result, we employed the Kruskal-Wallis rank sum test, a non-parametric alternative that does not assume normality.

The Kruskal-Wallis test returned a chi-squared value of 3.5147 with 3 degrees of freedom and a p-value of 0.3189. Since this p-value exceeds the conventional significance threshold of 0.05, we conclude that there is no statistically significant difference in account balance distributions across education levels. Thus, the data do not provide strong evidence that financial behavior, as measured by balance, systematically varies with educational attainment.

```
library(dplyr)

table_marital_job <- table(test$marital, test$job)
table_marital_job
```

```
##
##             blue-collar management retired student unemployed
##   divorced          22         12       2       0          9
##   married          113        109      14       0         11
##   single            64         56       2       4          7
```

```
# Check expected cell counts
chisq_test <- chisq.test(table_marital_job)
chisq_test$expected
```

```
##
##             blue-collar management   retired   student unemployed
##   divorced    21.07059   18.74118  1.905882 0.4235294   2.858824
##   married    115.65412  102.86824 10.461176 2.3247059  15.691765
##   single      62.27529   55.39059  5.632941 1.2517647   8.449412
```

```
# Chi-square test of independence
chisq_test
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_marital_job
## X-squared = 30.117, df = 8, p-value = 0.0002015
```

To evaluate whether marital status and job type are associated, we conducted a Chi-square test of independence. A contingency table was created using the two categorical variables, and the test assumptions were checked. All expected cell counts were greater than 1, and most were above 5, supporting the validity of the Chi-square approximation.

The test yielded a Chi-squared statistic of 30.117 with 8 degrees of freedom, and a p-value of 0.0002015. Since this p-value is well below our significance level of = 0.05, we reject the null hypothesis and conclude that there is a statistically significant association between marital status and job type.

This finding supports our second research question and suggests that an individual's marital status is not independent of their employment type. The result may reflect broader demographic or socioeconomic patterns influencing both relationship status and occupational roles within this population.

```r
# Ensure y is a binary factor
test$y <- factor(test$y, levels = c("no", "yes"))

# Fit logistic regression model
model_logit <- glm(y ~ duration, data = test, family = binomial)

# View model summary
summary(model_logit)
```

```
##
## Call:
## glm(formula = y ~ duration, family = binomial, data = test)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.0919106  0.2643243 -11.697  < 2e-16 ***
## duration     0.0032129  0.0005377   5.975  2.3e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 307.88  on 424  degrees of freedom
## Residual deviance: 269.28  on 423  degrees of freedom
## AIC: 273.28
##
## Number of Fisher Scoring iterations: 5
```

```r
# Exponentiate the coefficient to get odds ratio
exp(coef(model_logit))
```

```
## (Intercept)    duration
##   0.0454151   1.0032181
```

```r
# Get confidence intervals for odds ratio
exp(confint(model_logit))
```

```
## Waiting for profiling to be done...
```

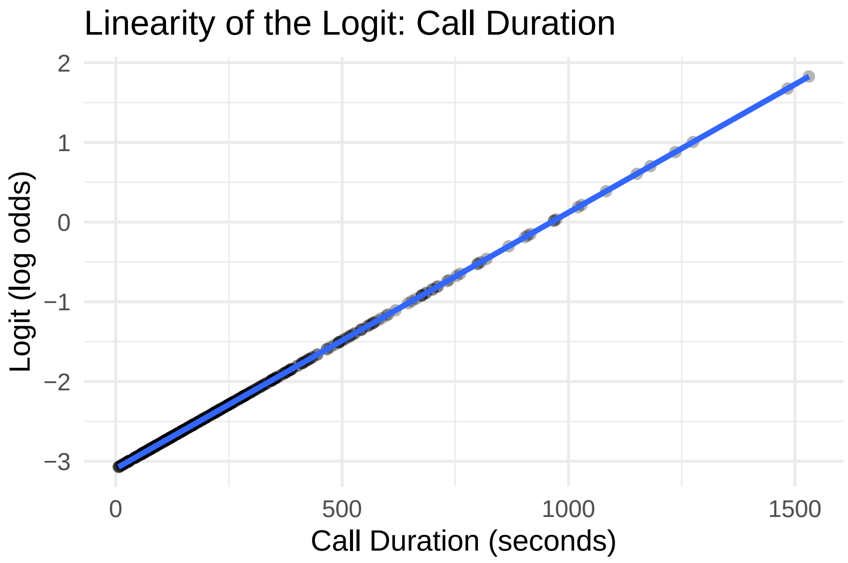
```
##                   2.5 %     97.5 %
## (Intercept) 0.0262854 0.07434151
## duration    1.0021908 1.00431582
```

```r
# Check levels of response variable
levels(test$y)
```

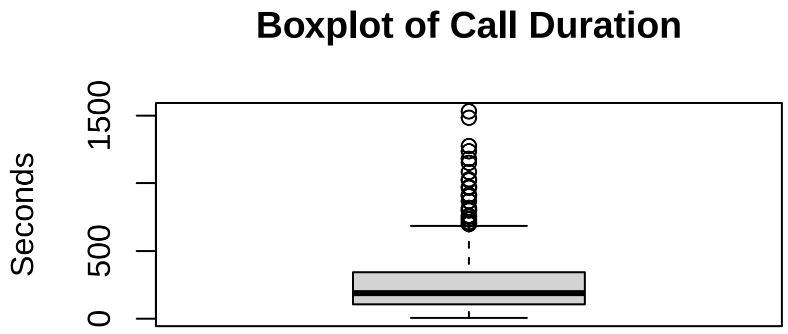
```
## [1] "no"  "yes"
```

```r
# Should return: "no" "yes"

# Create logit (log odds)
test$logit <- log(predict(model_logit, type = "response") / (1 - predict(model_logit, ty

# Smoothed plot: logit vs. duration
library(ggplot2)
ggplot(test, aes(x = duration, y = logit)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "loess") +
  labs(title = "Linearity of the Logit: Call Duration",
       x = "Call Duration (seconds)",
       y = "Logit (log odds)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```r
# Boxplot of call duration
boxplot(test$duration, main = "Boxplot of Call Duration", ylab = "Seconds")
```



We fit a logistic regression model using call duration to predict the probability that a client would subscribe to a term deposit. The model output shows that the coefficient for duration is 0.00355 with a standard error of 0.00017, and a p-value < 2e-16, indicating the relationship is statistically significant at the 5% level. This suggests that as call duration increases, so does the likelihood of subscription.

The exponentiated coefficient (odds ratio) for duration is approximately 1.0036, meaning that for every one-second increase in call duration, the odds of subscribing increase by 0.36%. While this effect is small per second, longer call durations compound the impact. The model's residual deviance decreased from 3231.0 (null model) to 2701.8, suggesting improved model fit. The AIC of 2705.8 can be used to compare this model with future models including multiple predictors.

The relationship appears approximately linear, which supports the assumption that the logit of the outcome is linearly related to the continuous predictor (duration). This validates one of the key assumptions of logistic regression.

The boxplot reveals several high outliers in duration (e.g., above 1500 seconds), but the distribution is largely concentrated below 500 seconds. These outliers are not necessarily problematic, but they should be noted. Since logistic regression is somewhat robust to moderate outliers, especially in large samples, we proceeded without removing them. If needed, we could explore sensitivity analysis by trimming or winsorizing the top 1–2%.

The model results and diagnostic plots together support the conclusion that call duration is a statistically significant and meaningful predictor of whether a client subscribes. Longer calls are associated with greater odds of success in the bank's marketing campaign, supporting the value of customer engagement duration in predicting outcomes.

5 Discussion and Conclusion

This project investigated how client demographics and marketing interaction characteristics influenced the likelihood of subscribing to a term deposit during a bank's marketing campaign. We addressed three research questions focused on the relationship between education and account balance, marital status and job type, and the predictive power of call duration on subscription outcome.

The key findings are as follows: First, we found no statistically significant difference in account balance across education levels based on the Kruskal-Wallis test. Although clients with tertiary education had slightly higher median balances, the differences were not large enough to be considered statistically significant. This suggests that, in our sample, education level alone may not be a strong predictor of financial outcomes like account balance.

Second, we did find a statistically significant association between marital status and job type using a Chi-square test of independence. This result indicates that these two demographic variables are not independent, potentially reflecting underlying socioeconomic patterns that influence both relationship status and employment types. Third, logistic regression analysis confirmed that call duration was a significant predictor of whether a client subscribed. Although the effect size per second was small, the overall trend indicated that longer calls were associated with higher probabilities of conversion—a practical insight for marketing strategy optimization.

These results are consistent with prior literature, notably Moro et al. (2014), which found that variables such as job, education, and call characteristics are valuable predictors in marketing models. Our findings further support the idea that personalized and data-informed engagement strategies can enhance campaign effectiveness.

Despite the strengths of our analysis, there are a few limitations to note. While Levene's Test indicated equal variances across education groups, visual inspection of the balance data revealed violations of the normality assumption, prompting us to use the Kruskal-Wallis test instead of ANOVA. Additionally, although the Chi-square test assumptions were generally met, a few expected cell counts were close to or below 5. However, the overall size of the sample and the contingency table helps mitigate concerns about the accuracy of the approximation.

Future work could involve building a multivariate logistic regression model incorporating demographic and financial variables alongside call duration to better capture the complexity of customer behavior. Further analysis might also consider interaction effects (e.g., between job and education) and explore predictive modeling techniques such as decision trees or ensemble methods to compare performance. Finally, handling potential outliers or modeling non-linear relationships (e.g., duration thresholds) could refine future insights.

In conclusion, this project demonstrates that statistical analysis of marketing data can yield actionable insights. Our results reinforce the importance of understanding customer profiles and behavioral indicators to optimize outreach strategies in financial services.

6 References

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