STAT 516 Midterm 3: Course Project

Analysis Using R

Nick Arboscello 50%   George Bujoreanu 50%

April 22, 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)

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.

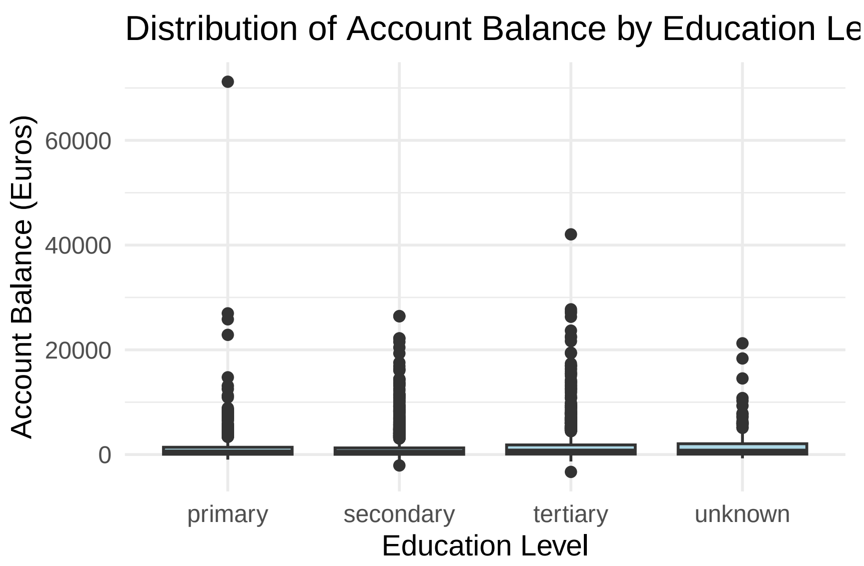
Reference: Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

2 Exploratory Data Analysis (EDA)

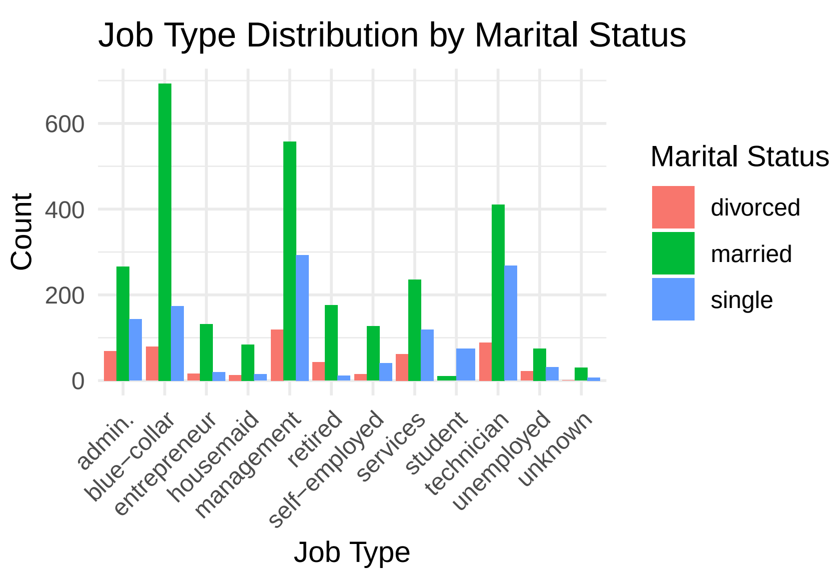
The goal of our exploratory data analysis is to uncover general trends, patterns, and potential relationships within the dataset that inform our research questions. Specifically, we aim to understand the distribution of customer demographics (age, job, marital status, education), assess how account balances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

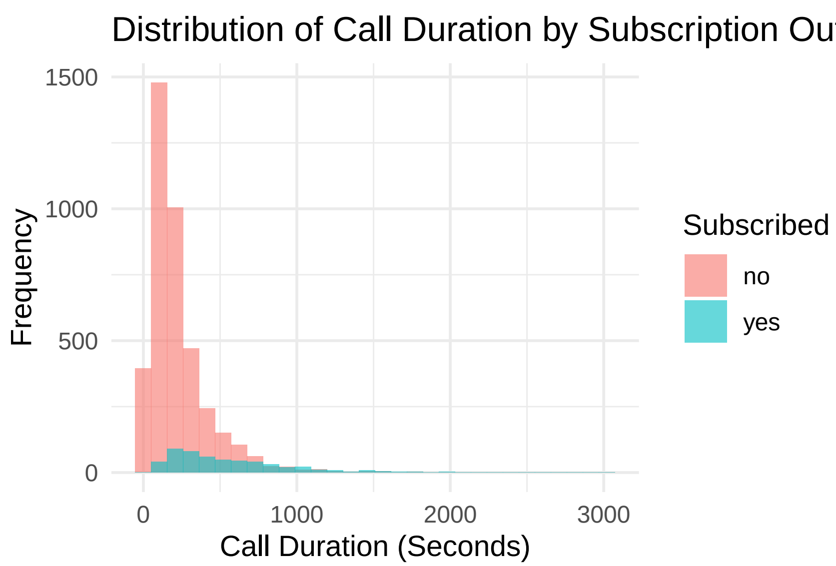
```r
# Boxplot: Account Balance by Education Level
library(ggplot2)

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()
```




```r
# 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 (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 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.

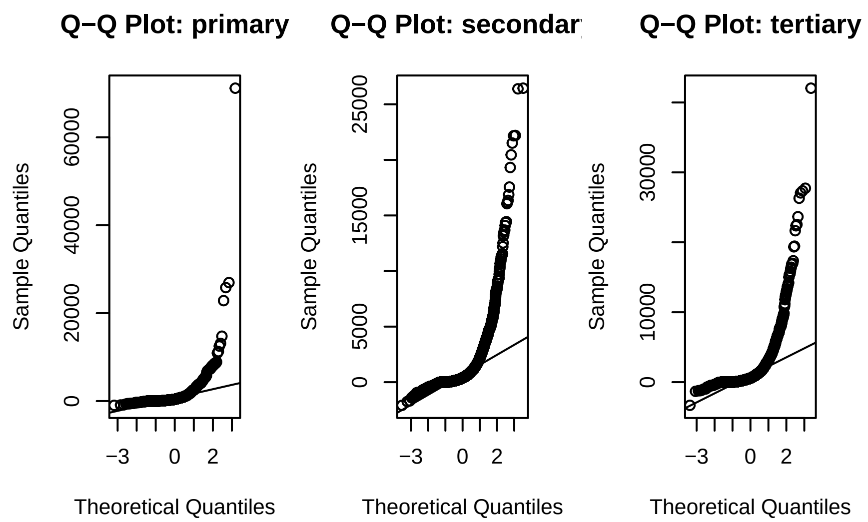
The third research question investigates: Does call duration significantly predict whether a client will subscribe to a term deposit? This question is rooted in the hypothesis that longer calls may indicate greater client interest, engagement, or persuasion success. As call duration increases, we may 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

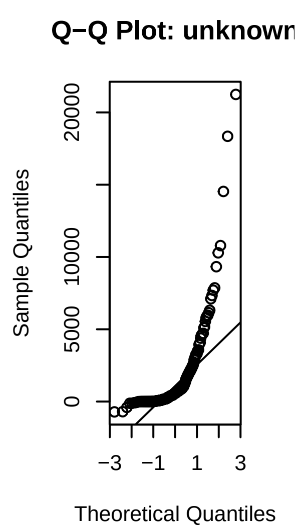
```r
# Load required libraries
library(ggplot2)
library(dplyr)
library(car)

# Set significance level
alpha <- 0.05

# Check summary of balance by education
test %>%
  group_by(education) %>%
  summarise(
    count = n(),
    mean_balance = mean(balance, na.rm = TRUE),
    sd_balance = sd(balance, na.rm = TRUE)
  )
```

```
## # A tibble: 4 x 4
##   education count mean_balance sd_balance
##   <chr>     <int>        <dbl>      <dbl>
## 1 primary     678        1412.      3714.
## 2 secondary  2306        1197.      2420.
## 3 tertiary   1350        1775.      3461.
## 4 unknown     187        1701.      2981.
```

```r
# 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])
}
```




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



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

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)    
## group    3  8.7238 9.094e-06 ***
##       4517                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Fit one-way ANOVA model
anova_model <- aov(balance ~ education, data = test)

# Summary of ANOVA model
summary(anova_model)
```

```
##               Df    Sum Sq   Mean Sq F value   Pr(>F)    
## education      3 3.002e+08 100071422   11.12 2.86e-07 ***
## Residuals   4517 4.064e+10   8997474                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Tukey's HSD to compare group means
TukeyHSD(anova_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = balance ~ education, data = test)
## 
## $education
##                          diff         lwr       upr     p adj
## secondary-primary  -214.72985 -551.509939  122.0502 0.3569521
## tertiary-primary    363.87946    1.015576  726.7433 0.0490540
## unknown-primary     289.70174 -347.040848  926.4443 0.6463248
## tertiary-secondary  578.60931  314.430452  842.7882 0.0000002
## unknown-secondary   504.43159  -81.709515 1090.5727 0.1201556
## unknown-tertiary    -74.17771 -675.684792  527.3294 0.9889855
```

To assess whether the assumption of equal variances was met for one-way ANOVA, we first conducted Levene's Test for homogeneity of variances. The test produced a p-value of 9.094e-06, which is well below our chosen significance level of $= 0.05$. Therefore, we reject the null hypothesis of equal variances and conclude that the variances in account balance differ significantly across education levels. Despite this violation, we proceeded with the ANOVA since it is relatively robust to deviations from this assumption, particularly with large sample sizes. **Ask about this**

The results of the one-way ANOVA showed a statistically significant effect of education level on account balance (F(3, 4517) = 11.12, p = 2.86e-07). This indicates that at least one group mean differs significantly from the others. To further explore these differences, we conducted a Tukey's HSD post-hoc test. The test revealed that the mean balance for clients with tertiary education was significantly higher than that of those with secondary education ($p < 0.001$). There was also a marginally significant difference between the "unknown" and "primary" groups ($p = 0.049$). All other pairwise comparisons were not statistically significant at the 5% level.

In conclusion, the analysis supports our research question by showing that account balance varies meaningfully with education level, reinforcing the idea that education may be a key factor in financial behavior.

```r
library(dplyr)

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

```
##
##             admin. blue-collar entrepreneur housemaid management retired
##   divorced      69          79           16        13        119      43
##   married      266         693          132        84        557     176
##   single       143         174           20        15        293      11
##
##             self-employed services student technician unemployed unknown
##   divorced             15       62       0         89         22       1
##   married             127      236      10        411         75      30
##   single               41      119      74        268         31       7
```

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

```
##
##                admin. blue-collar entrepreneur housemaid management   retired
##   divorced  55.82482    110.4818     19.62044  13.08029   113.1679  26.86131
##   married  295.72351    585.2603    103.93630  69.29086   599.4897 142.29374
##   single   126.45167    250.2579     44.44326  29.62884   256.3424  60.84495
##
##             self-employed  services   student technician unemployed   unknown
##   divorced       21.37226  48.70073  9.810219   89.69343   14.94891  4.437956
```

```
##   married      113.21632 257.98474 51.968149  475.13736   79.18956 23.509401
##   single        48.41141 110.31453 22.221632  203.16921   33.86154 10.052643
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  table_marital_job
## X-squared = 373.18, df = 22, p-value < 2.2e-16
```

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 examined. The expected cell counts showed that all values were greater than 1; however, some expected frequencies were below 5, triggering a warning that the chi-squared approximation may be inaccurate. Despite this, we proceeded with the test given the overall large sample size, which helps mitigate this concern.**Ask about counts under 5**

The test produced a chi-squared statistic of 373.18, with 22 degrees of freedom, and a p-value less than 2.2e-16. Since this p-value is far 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 job type in this dataset, potentially reflecting demographic or socioeconomic patterns that influence both employment and relationship status.

```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) -3.2559346  0.0845767  -38.50   <2e-16 ***
## duration     0.0035496  0.0001714   20.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3231.0  on 4520  degrees of freedom
## Residual deviance: 2701.8  on 4519  degrees of freedom
## AIC: 2705.8
##
## Number of Fisher Scoring iterations: 5
```

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

```
## (Intercept)    duration
##  0.03854478  1.00355586
```

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

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

```
##                 2.5 %     97.5 %
## (Intercept) 0.0325604 0.04536511
## duration    1.0032233 1.00389759
```

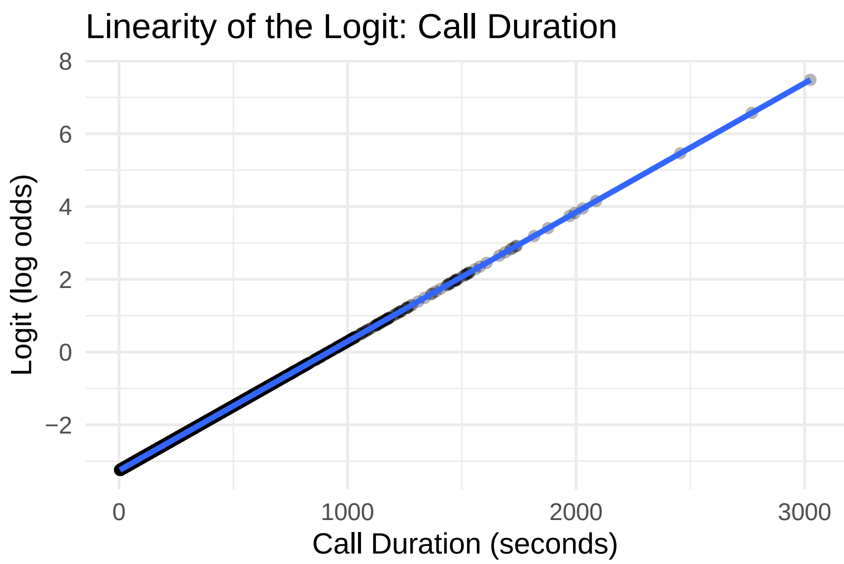
```r
# Check levels of response variable
levels(test$y)
```

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

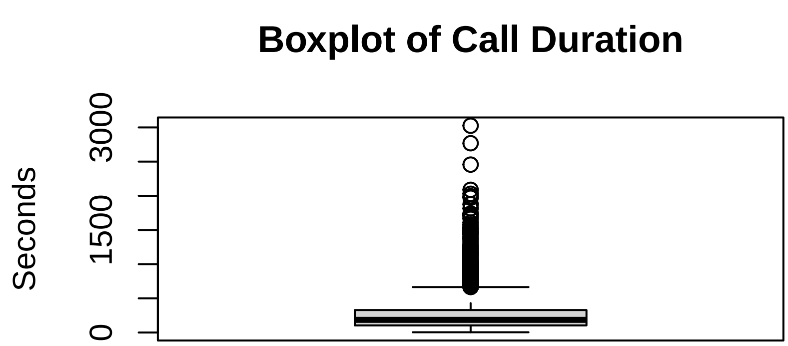
```r
# Should return: "no" "yes"

# 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) +
```

```r
  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 (y). The model output shows that the coefficient for duration is 0.00355 with a standard error of 0.00017, and a p-value $< 2e\text{-}16$, 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, there was a statistically significant difference in account balance across education levels, with clients holding tertiary education exhibiting significantly higher average balances compared to those with secondary education. This finding aligns with economic theories suggesting that higher educational attainment correlates with better financial outcomes. Second, we found a strong and statistically significant association between marital status and job type using a chi-square test, suggesting that these demographic variables are not independent and may reflect broader socioeconomic patterns. 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. The Levene's Test for ANOVA indicated unequal variances across education groups, which technically violates the homogeneity assumption, though ANOVA is fairly robust to this issue given our large sample size. The chi-square test triggered a warning due to some expected cell counts being below 5, which could affect the test's accuracy—though this concern is somewhat mitigated by the overall size of the contingency table.

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.

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