**Portfolio Project – Project Code and GitHub Link**

Prakash Bajracharya

MIS 581 – Predictive Analytics

Colorado State University-Global Campus

Dr. Justin Bateh

June 1, 2024

**Project Code and GitHub Link**

I used the R Markdown application to seamlessly integrate R code with the results, providing a comprehensive and reproducible report. The code is divided into two main sections: Data Exploration and Hypothesis Testing.

**Data Exploration:** This section delves into the descriptive and statistical data to provide a comprehensive overview of the dataset. Using R packages such as dplyr and ggplot2, I conducted data cleaning, transformation, and visualization. This process involved summarizing key statistics, identifying patterns, and visualizing data distributions through histograms, box plots, and scatter plots. These visualizations helped in understanding the central tendency, dispersion, and relationships between variables.

**Hypothesis Testing:** In this section, I tested the formulated hypotheses to determine if the observed patterns in the data were statistically significant. Using R’s glm function for logistic regression and chisq.test for chi-square tests, I evaluated hypotheses related to factors influencing term deposit subscriptions. The results, including p-values and confidence intervals, were presented alongside the code to ensure transparency and reproducibility. This approach not only confirmed the significant predictors but also provided insights into their effect sizes and directions.

By structuring the analysis into these sections and using R Markdown, I ensured that the entire workflow—from data exploration to hypothesis testing—was documented, reproducible, and easy to follow.

Below is the GitHub link and R code script:

GitHub Link: https://github.com/bajracharya1/MIS581-Portfolio-Project

# Data exploration

# Loading necessary library  
  
library(dplyr)  
library(ggplot2)  
  
#Dataset Statistical Summary  
  
## Loading banking data into data frame  
  
bank\_df <- read.csv("F:/LearningStuff/CSUGLOBAL/MIS-581/bank-additional-full.csv")  
  
## Viewing all data in the data frame  
View (bank\_df)  
  
## Displaying the structure  
str(bank\_df)

## 'data.frame': 41188 obs. of 21 variables:  
## $ age : int 56 57 37 40 56 45 59 41 24 25 ...  
## $ job : chr "housemaid" "services" "services" "admin." ...  
## $ marital : chr "married" "married" "married" "married" ...  
## $ education : chr "basic.4y" "high.school" "high.school" "basic.6y" ...  
## $ default : chr "no" "unknown" "no" "no" ...  
## $ housing : chr "no" "no" "yes" "no" ...  
## $ loan : chr "no" "no" "no" "no" ...  
## $ contact : chr "telephone" "telephone" "telephone" "telephone" ...  
## $ month : chr "may" "may" "may" "may" ...  
## $ day\_of\_week : chr "mon" "mon" "mon" "mon" ...  
## $ duration : int 261 149 226 151 307 198 139 217 380 50 ...  
## $ campaign : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ pdays : int 999 999 999 999 999 999 999 999 999 999 ...  
## $ previous : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutcome : chr "nonexistent" "nonexistent" "nonexistent" "nonexistent" ...  
## $ emp.var.rate : num 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...  
## $ cons.price.idx: num 94 94 94 94 94 ...  
## $ cons.conf.idx : num -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...  
## $ euribor3m : num 4.86 4.86 4.86 4.86 4.86 ...  
## $ nr.employed : int 5191 5191 5191 5191 5191 5191 5191 5191 5191 5191 ...  
## $ y : chr "no" "no" "no" "no" ...

## Statistical Summary for all the attributes in the data frame  
summary(bank\_df)

## age job marital education   
## Min. :17.00 Length:41188 Length:41188 Length:41188   
## 1st Qu.:32.00 Class :character Class :character Class :character   
## Median :38.00 Mode :character Mode :character Mode :character   
## Mean :40.02   
## 3rd Qu.:47.00   
## Max. :98.00   
## default housing loan contact   
## Length:41188 Length:41188 Length:41188 Length:41188   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## month day\_of\_week duration campaign   
## Length:41188 Length:41188 Min. : 0.0 Min. : 1.000   
## Class :character Class :character 1st Qu.: 102.0 1st Qu.: 1.000   
## Mode :character Mode :character Median : 180.0 Median : 2.000   
## Mean : 258.3 Mean : 2.568   
## 3rd Qu.: 319.0 3rd Qu.: 3.000   
## Max. :4918.0 Max. :56.000   
## pdays previous poutcome emp.var.rate   
## Min. : 0.0 Min. :0.000 Length:41188 Min. :-3.40000   
## 1st Qu.:999.0 1st Qu.:0.000 Class :character 1st Qu.:-1.80000   
## Median :999.0 Median :0.000 Mode :character Median : 1.10000   
## Mean :962.5 Mean :0.173 Mean : 0.08189   
## 3rd Qu.:999.0 3rd Qu.:0.000 3rd Qu.: 1.40000   
## Max. :999.0 Max. :7.000 Max. : 1.40000   
## cons.price.idx cons.conf.idx euribor3m nr.employed   
## Min. :92.20 Min. :-50.8 Min. :0.634 Min. :4964   
## 1st Qu.:93.08 1st Qu.:-42.7 1st Qu.:1.344 1st Qu.:5099   
## Median :93.75 Median :-41.8 Median :4.857 Median :5191   
## Mean :93.58 Mean :-40.5 Mean :3.621 Mean :5167   
## 3rd Qu.:93.99 3rd Qu.:-36.4 3rd Qu.:4.961 3rd Qu.:5228   
## Max. :94.77 Max. :-26.9 Max. :5.045 Max. :5228   
## y   
## Length:41188   
## Class :character   
## Mode :character   
##   
##   
##

## First 6 (default) rows of the data frame  
head(bank\_df)

## age job marital education default housing loan contact month  
## 1 56 housemaid married basic.4y no no no telephone may  
## 2 57 services married high.school unknown no no telephone may  
## 3 37 services married high.school no yes no telephone may  
## 4 40 admin. married basic.6y no no no telephone may  
## 5 56 services married high.school no no yes telephone may  
## 6 45 services married basic.9y unknown no no telephone may  
## day\_of\_week duration campaign pdays previous poutcome emp.var.rate  
## 1 mon 261 1 999 0 nonexistent 1.1  
## 2 mon 149 1 999 0 nonexistent 1.1  
## 3 mon 226 1 999 0 nonexistent 1.1  
## 4 mon 151 1 999 0 nonexistent 1.1  
## 5 mon 307 1 999 0 nonexistent 1.1  
## 6 mon 198 1 999 0 nonexistent 1.1  
## cons.price.idx cons.conf.idx euribor3m nr.employed y  
## 1 93.994 -36.4 4.857 5191 no  
## 2 93.994 -36.4 4.857 5191 no  
## 3 93.994 -36.4 4.857 5191 no  
## 4 93.994 -36.4 4.857 5191 no  
## 5 93.994 -36.4 4.857 5191 no  
## 6 93.994 -36.4 4.857 5191 no

## Getting dimension of the data frame  
dim(bank\_df)

## [1] 41188 21

## attributes name in data frame  
names(bank\_df)

## [1] "age" "job" "marital" "education"   
## [5] "default" "housing" "loan" "contact"   
## [9] "month" "day\_of\_week" "duration" "campaign"   
## [13] "pdays" "previous" "poutcome" "emp.var.rate"   
## [17] "cons.price.idx" "cons.conf.idx" "euribor3m" "nr.employed"   
## [21] "y"

## Calculate additional descriptive statistics for numerical data  
bank\_df %>%  
 summarise(across(where(is.numeric), list(mean = mean, sd = sd, median = median)))

## age\_mean age\_sd age\_median duration\_mean duration\_sd duration\_median  
## 1 40.02406 10.42125 38 258.285 259.2792 180  
## campaign\_mean campaign\_sd campaign\_median pdays\_mean pdays\_sd pdays\_median  
## 1 2.567593 2.770014 2 962.4755 186.9109 999  
## previous\_mean previous\_sd previous\_median emp.var.rate\_mean emp.var.rate\_sd  
## 1 0.172963 0.4949011 0 0.0818855 1.57096  
## emp.var.rate\_median cons.price.idx\_mean cons.price.idx\_sd  
## 1 1.1 93.57566 0.57884  
## cons.price.idx\_median cons.conf.idx\_mean cons.conf.idx\_sd  
## 1 93.749 -40.5026 4.628198  
## cons.conf.idx\_median euribor3m\_mean euribor3m\_sd euribor3m\_median  
## 1 -41.8 3.621291 1.734447 4.857  
## nr.employed\_mean nr.employed\_sd nr.employed\_median  
## 1 5167.019 72.17807 5191

### data frame with 0 columns and 1 row  
## Count frequencies of categorical data  
bank\_df %>%  
 summarise(across(where(is.factor), ~table(.)))

## data frame with 0 columns and 1 row

### data frame with 0 columns and 1 row

# Hypothesis Test

## Converting columns to factor for proper statistical analysis and visualization  
bank\_df$contact <- as.factor(bank\_df$contact)  
bank\_df$y <- as.factor(bank\_df$y)  
bank\_df$age <- as.numeric(bank\_df$age)  
  
## Hypothesis 1: Method of contact  
contact\_effect <- table(bank\_df$contact, bank\_df$y)  
contact\_test <- chisq.test(contact\_effect)  
contact\_test

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: contact\_effect  
## X-squared = 862.32, df = 1, p-value < 2.2e-16

## Plotting the results  
ggplot(bank\_df, aes(x = contact, fill = y)) +  
 geom\_bar(position = 'fill') +  
 labs(title = 'Subscription Rate by Contact Method', y = 'Proportion', x = 'Contact Method')

A graph of a number of colored squares

Description automatically generated with medium confidence

## Hypothesis 2: Age effect  
age\_groups <- bank\_df %>%  
 mutate(age\_group = cut(age, breaks = c(18, 35, 60, 100), labels = c('18-35', '36-60', '60+'))) %>%  
 group\_by(age\_group, y) %>%  
 summarise(count = n()) %>%  
 ungroup()  
  
age\_effect <- chisq.test(table(age\_groups$age\_group, age\_groups$y))  
age\_effect

##   
## Pearson's Chi-squared test  
##   
## data: table(age\_groups$age\_group, age\_groups$y)  
## X-squared = 0, df = 2, p-value = 1

## Plotting the results  
ggplot(age\_groups, aes(x = age\_group, y = count, fill = y)) +  
 geom\_bar(stat = 'identity', position = 'dodge') +  
 labs(title = 'Subscription Rate by Age Group', y = 'Count', x = 'Age Group')

A graph of a number of people

Description automatically generated

## Hypothesis 3: Previous engagement  
bank\_df$previous <- as.factor(ifelse(bank\_df$previous > 0, 'yes', 'no'))  
previous\_effect <- table(bank\_df$previous, bank\_df$y)  
previous\_test <- chisq.test(previous\_effect)  
previous\_test

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: previous\_effect  
## X-squared = 1540.5, df = 1, p-value < 2.2e-16

## Plotting the results  
ggplot(bank\_df, aes(x = previous, fill = y)) +  
 geom\_bar(position = 'fill') +  
 labs(title = 'Subscription Rate by Previous Engagement', y = 'Proportion', x = 'Previous Engagement')

A graph with red and blue squares

Description automatically generated