Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Foundation of Business Analytics - Group Assignment

By

Chenyu Wang

Geunju Park

Panagiotis Georgiadis

Shanshan Tan

Xiaoxue Ji

PROGRAMME NAME

MSc Business Analytics 2024/25

BU7154 - Foundations of Business Analytics

Module Leader:

Baidyanath Biswas

Trinity Business School

TRINITY COLLEGE

UNIVERSITY OF DUBLIN

November 2024

Declaration

I declare that this work has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

Generative AI Declaration

Please choose A or B with regards to your use of ChatGPT & other generative AI tools in this project:

☐ **A. Nothing to declare.** I did not use ChatGPT or any other generative AI software. (see note)

■ **B. I used ChatGPT** or other generative AI software (see note)

NOTE:

- If you answer A and the corrector/supervisor, finds evidence that you have indeed used ChatGPT, this behaviour will be considered as unethical and you will be penalized accordingly with reference to the TCD policy on plagiarism.
- If you answer B, please clearly explain for which chapters or parts of your dissertation you used ChatGPT and how it helped you to improve your learning process within ethical guidelines. You may include your answer – 300 to 600 words approx.- in the appendix.

Signed: ID No.

1. Chenyu Wang	1. 24336364
2. Geunju Park	2. 24340216
3. Panagiotis Georgiadis	3. 24364488
4. Shanshan Tan	4. 24342976
5. Xiaoxue Ji	5. 24332507

We used ChatGPT for the parts below.

- Writing R code for visualizations related to research question 1.
- Crafting Tableau calculated field functions and generating the finding about contact fatigue for research question 2.
- Implementing R code, creating distribution diagrams, and refining statements for research question 3.

Date: 14-November-2024

Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Assignment Submission Cover Sheet

Programme Title:	**MSc Business Analytics 2024/25**
Module Code and Title:	**BU7154 Foundation of Business Analytics**
Assessment Title:	**Group Assignment – Project Report**
Group Number:	**13**

Student Name and Contribution	**%**		**%**
1.Chenyu Wang		**4.Shanshan Tan**	
2.Geunju Park		**5.Xiaoxue Ji**	
3.Panagiotis Georgiadis		**6.**	

For group work – individual % contributions need to be stated **only** where they **are not equal.**

Please read the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at: http://www.tcd.ie/calendar

Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', is located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write

Table of Contents

1. Executive Summary

This project analyses customer data from a Portuguese bank's phone marketing campaigns to generate insights for improving customer targeting and engagement. It is revealed that specific demographic and financial characteristics strongly influence a client's likelihood of subscribing to a term deposit. Management and technician roles show high subscription rates, suggesting these as priority segments, while blue-collar workers display lower engagement, indicating potential product relevance challenges. Retired individuals and students, though smaller groups, demonstrate untapped potential for growth.

Economic conditions influence customer behaviour, as many clients fall within the 0-1000 euro or negative balance ranges, showing financial strain. A post-2010 recovery points to potential for greater engagement as interest rates stabilized. Cellular outreach is the most effective contact method, while traditional phone contact shows limited success, suggesting the need for re-evaluation. Additionally, job type and education level are key factors in housing loan eligibility, helping to identify priority demographics for targeted loan offerings.

These insights suggest targeted marketing and outreach strategies for specific demographic and financial profiles to maximize engagement. Future actions involve expanding the dataset with more financial metrics and pilot-testing these recommendations to assess impact.

2. Project Goals

This project aims to develop a strategy to improve a Portuguese bank's performance through modelling and data visualisation, focusing on increasing the success rate of term deposit subscription calls and identifying potential housing loan customers.

In this project we will complete the following tasks:

- Analyse demographics to draw a customer segment strategy.
- Optimise phone call performance to increase subscription rates for term deposits.
- Use demographic data to determine which customer segments are more likely to have a housing loan.

The dataset contains 45,211 records and 17 variables, covering demographic information, account details, loan status, and marketing interaction metrics. We cleaned the data and added relevant variables to better support these project goals.

3. Literature Review

Telemarketing has become a widely used, cost-effective direct marketing method for banks, yet its effectiveness remains challenged. Baek and Morimoto (2012) emphasize the importance of leveraging consumer data to personalize product offerings, helping customers quickly identify relevant products and boosting efficiency. Similarly, Moro, Cortez, and Rita (2018) stress the need to analyze message relevance and identify potential buyers to increase

customer engagement and minimize waste. Yan, Li, and Liu (2020) further caution that poor data analysis can waste resources and harm customer relationships, highlighting the need for strategic, data-driven approaches.

In applying these insights to bank marketing strategies, several customer demographics were analysed for their impact on deposit subscriptions and housing loans. Kearns (2019) mentioned that age plays a significant role, as individuals in their thirties and forties are more likely to have mortgages, particularly if they have stable employment. Unemployment within this group can lead to higher arrears rates. However, for older workers, younger workers without mortgages, or secondary income earners in households already managing repayments comfortably, job loss may have minimal impact on arrears. Regarding employment type, Xie and Zhang (2023) note that entrepreneurs, the self-employed, and the unemployed tend to show lower willingness to deposit compared to administrators, who have the highest probability of subscribing to time deposit schemes. Marital status also influences financial behaviour, with married individuals often having greater economic and social responsibilities, making them more inclined to save for long-term goals like children's education. This insight can help banks target married customers with tailored savings products. Additionally, education level is closely linked to financial decision-making, with higher education typically resulting in higher income and better financial literacy. This can improve housing loan eligibility and the likelihood of informed financial choices (The Mortgage Reports, 2024).

4. Methodology

1) Research Question 1

We developed a Random Forest model for its robustness in handling diverse predictors, making it ideal for analysing subscription outcomes. Data preprocessing included converting key variables to factors and handling missing data. Feature engineering added ‘age_group’ and ‘balance_group’ columns for better segmentation. The data was split into training and testing. Model performance was evaluated using a confusion matrix and accuracy score. Visual analyses and summary statistics provided insights into the distributions of age, balance, contact type, and job by subscription status, revealing trends and patterns in customer behaviour.

2) Research Question 2

To assess the relationship between variables and prior campaign outcomes, we utilized Tableau to create visualizations that reveal correlations between each variable and campaign success metrics. This approach enabled a deeper analysis of demographic and behavioral trends associated with successful outcomes, guiding more targeted marketing strategies. Additionally, we incorporated the ‘Weekday’ variable to explore daily variations in campaign performance, providing a detailed view of timing effects on engagement. By examining these visualizations, we identified actionable insights for strategy optimization based on key patterns in campaign success.

3) Research Question 3

We constructed a decision tree model for its simplicity and intuitive nature, which makes it suitable for segmenting customer groups in this project. Its hierarchical structure provides a clear view of each feature's importance, enabling the bank to better target customer characteristics in its business operations. However, its limitation is relatively low accuracy, which may be due to the model using only basic customer demographic information.

5. Discussion of Findings

1) Analyse Demographics to Draw a Customer Segment Strategy

a) Key Factors Influencing Subscription Decisions

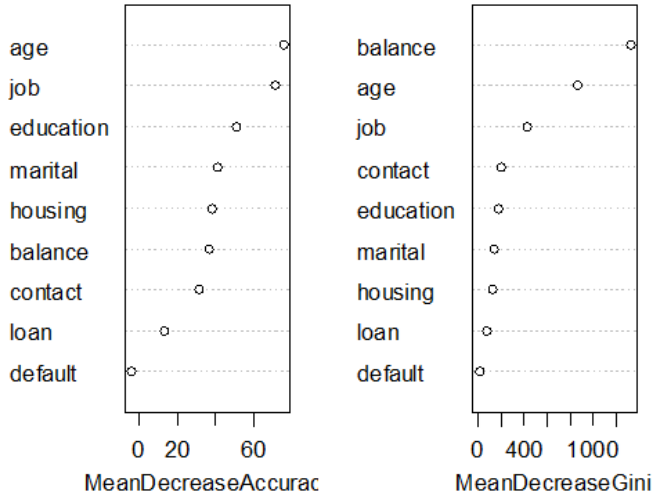


Figure 1. Results of Population Analysis

Based on the results of the population analysis, the primary factors influencing the decision to subscribe include age, account balance, and occupation type. Balance consistently ranks highest in both accuracy and Gini importance, indicating its strong influence on subscription decisions, likely due to its reflection of financial capability. Age and job also show high importance, as older customers and those in specific occupations tend to have higher subscription likelihood due to financial stability and professional status.

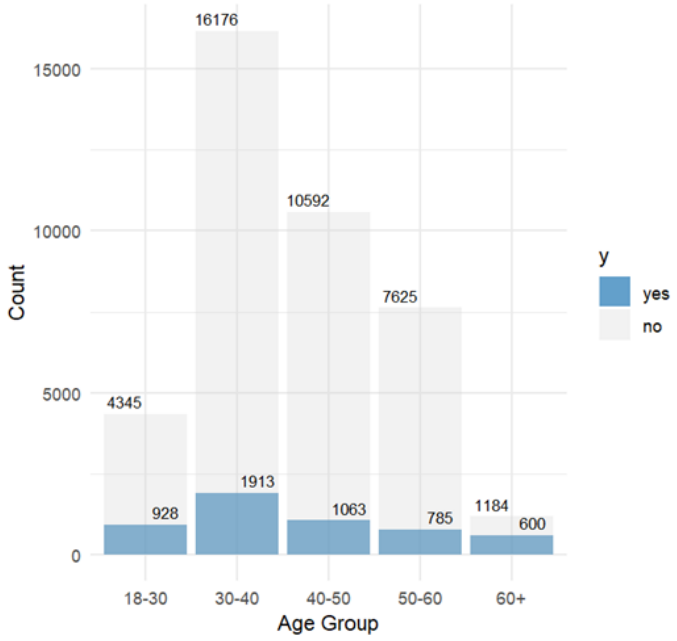


Figure 2. Age Group Distribution of Subscription Status

From figure 2, we observe that customers aged 60+ have a relatively high subscription rate, indicating strong potential for targeted high-conversion strategies. Meanwhile, the 30-60 age group represents the largest portion of subscribers, making it the primary market to focus on for maximizing subscription volume.

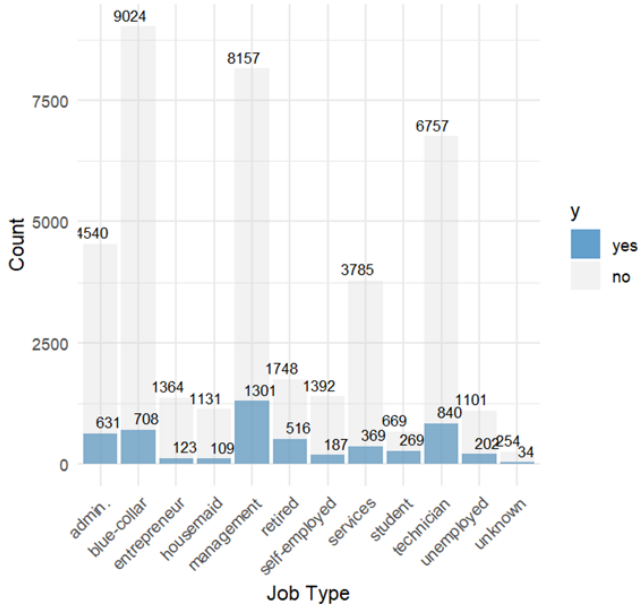


Figure 3. Job Type Distribution by Subscription Status

Figure 3 indicates that occupations such as admin, blue-collar, management, retired, and technician show relatively high subscription rates. Further analysis suggests an age-related distribution within these job types, with retired individuals primarily falling into the older age group, which aligns with their higher subscription potential.

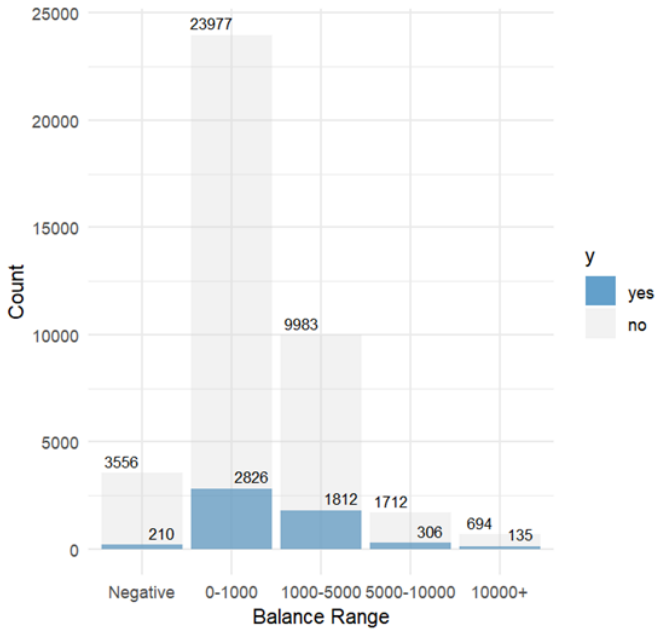


Figure 4. Balance Distribution by Subscription Status

During the 2008-2010 financial crisis, Portugal's low interest rates reduced returns on savings, straining consumer finances. In this context, customers with balances over 1000 demonstrated greater financial resilience, as shown by their higher subscription rates. Focusing on this stable segment enables targeted engagement with those likely to invest in additional financial products.

2) How can we maximise phone call success rates for term deposit subscriptions.

To solve the second question, we made following four analytics;

a) Effective Call Scheduling

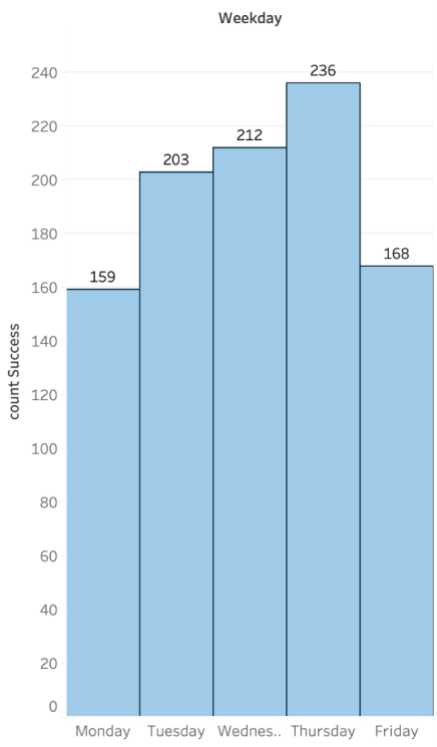


Figure 5. Count Success by Weekday

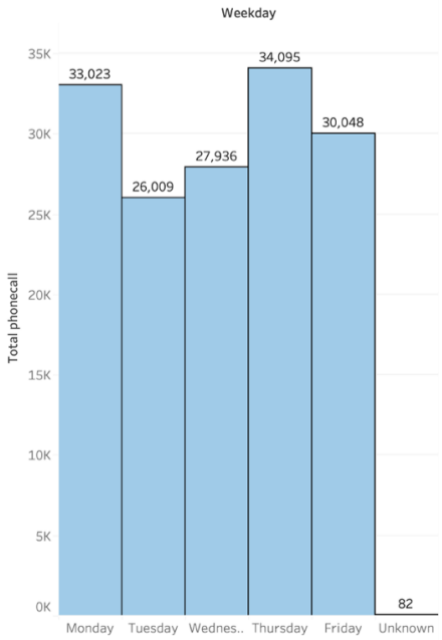


Figure 6. Total Count of Phone calls by Weekday

Data analysis shows that customer call success rates are highest from Tuesday to Thursday, with Monday calls yielding fewer successful outcomes despite higher volumes. This suggests that timing, rather than call volume, plays a key role in success. Customers may be more receptive mid-week due to their work or personal schedules, highlighting the need for optimized call scheduling.

b) Duration for Success

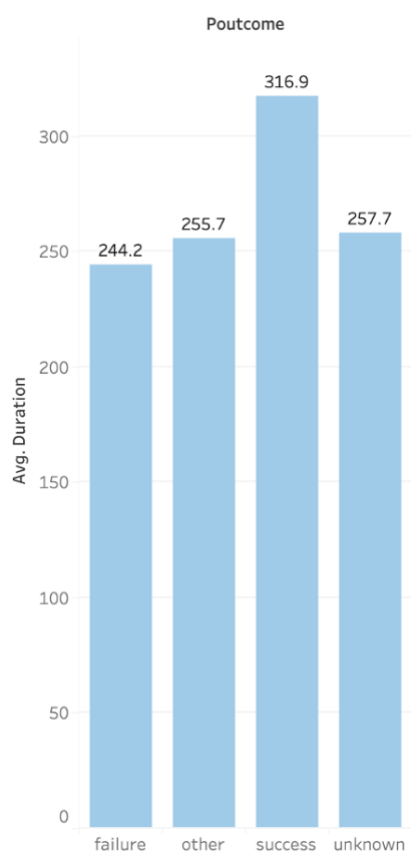


Figure 7. Average Phone Calls Duration Second from Previous Marketing Outcome

Given that the "most recent call" could either be the final call in the communication process or simply an ongoing step in a longer engagement, our insights and recommendations need to account for both possibilities. Here's a refined analysis with insights:

(i) 260 Seconds as a Foundational Threshold

Regardless of whether a call is the final or an ongoing one, reaching 260 seconds appears to be a crucial threshold. Calls that last beyond 260 seconds tend to be more successful, indicating effective customer engagement.

(ii) Longer Calls Indicate High Engagement

Calls lasting around 317 seconds often reflect deeper customer interest.

(iii) Different Stages, Different Durations

Calls in the "Other" and "Unknown" categories average around 260 seconds, suggesting that early-stage calls can effectively filter potential customers at this duration. For successful conversions, a duration closer to 316 seconds seems ideal, especially if the customer is nearing a decision point or the call is likely to be the final touchpoint.

c) Number of phone calls for success

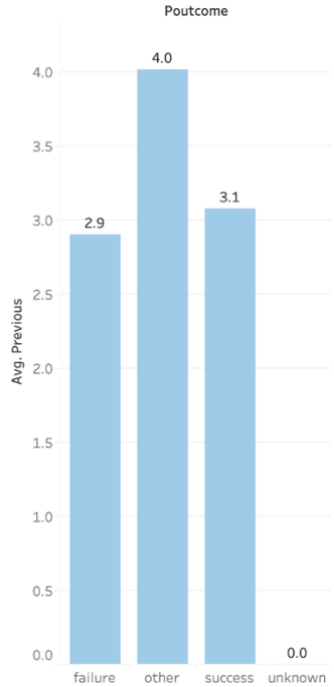


Figure 8. Average Phone calls Count based on Previous Marketing Outcome

The analysis shows that a minimum of three calls is typically necessary to secure customer commitment, with successful conversions often occurring after several interactions. Customers categorized as "Other" represent those who are still in consideration, indicating untapped conversion potential. This suggests that regular follow-ups can nurture customer interest, providing essential reminders and addressing evolving questions. Additionally, tracking customer responses to each call can identify ideal touchpoints and guide follow-up timing, especially for those marked as "Other."

d) Campaign Period Contact Strategy

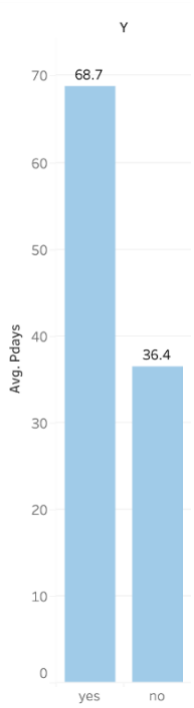


Figure 9. Average Pdays based on Deposit

Contrary to initial expectations, maintaining contact with customers between campaigns does not seem to enhance conversion rates. Continuous follow-ups may instead dilute impact, as customers might perceive frequent contact as intrusive or unnecessary. Focusing efforts within active campaigns may allow for more concentrated and purposeful engagement. It suggests that customer receptivity may peak during active campaign phases, with prolonged contact periods potentially leading to message fatigue.

3) How Can Personal Information Factors Predict Housing Loan Customers

a) Descriptive Analysis and Results of Visualization

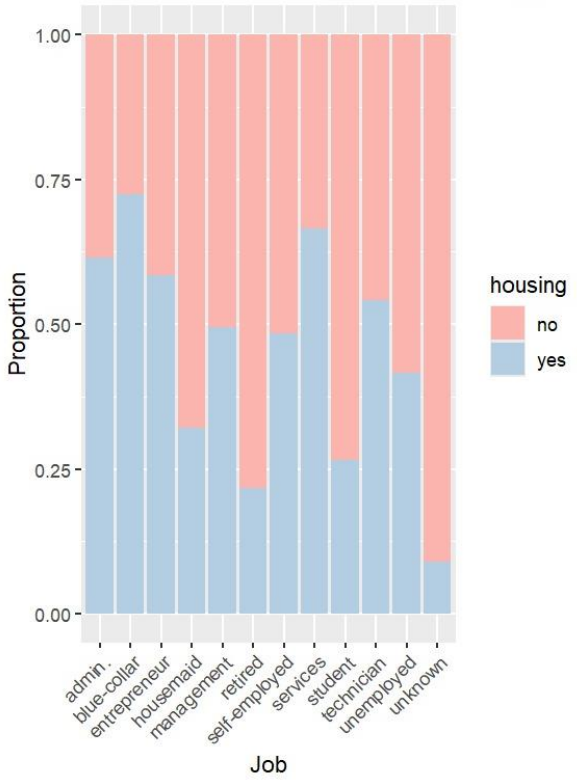


Figure 10. Proportion of Housing Loan by Job Type

Figure 10 shows the proportion of customers with and without a housing loan for each job category. Different job types have varying levels of housing loan ownership.

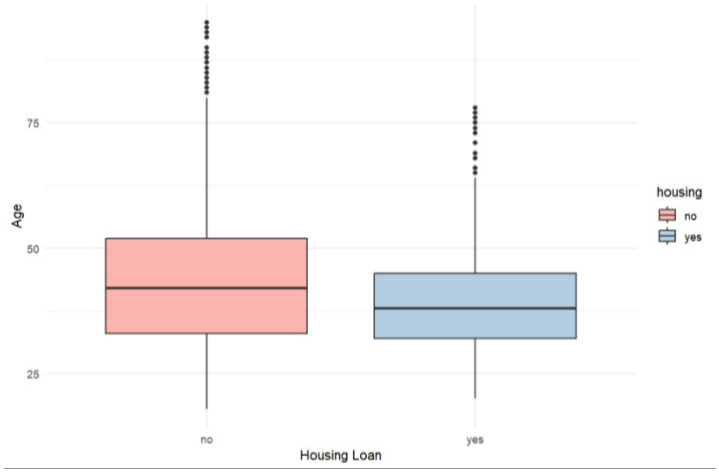


Figure 11. Distribution of Age by Housing Loan Status

Figure 11 compares the age distributions of customers with and without a housing loan. Customers with housing loans generally tend to be slightly younger than those without, as seen by the lower median age in the "yes" group.

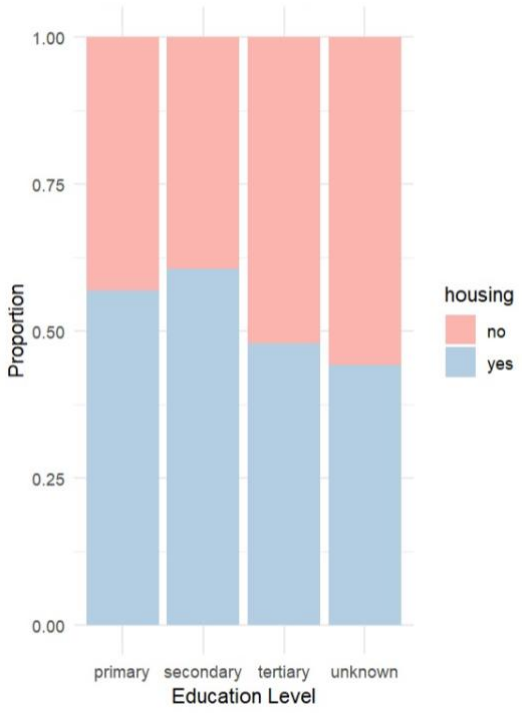


Figure 12. Proportion of Housing Loan by Education Level

Figure 12 shows the proportion of customers with and without a housing loan for different education levels. We can find that customers with primary and secondary education are more likely to have housing loans, especially those with secondary education.

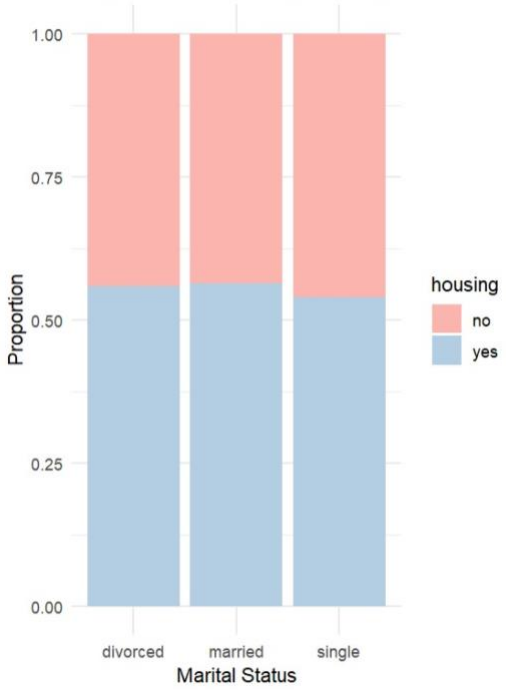


Figure 13. Proportion of Housing Loan by Marital Status

Figure 13 displays the proportion of customers with and without a housing loan based on marital status. The proportions across marital status categories are relatively similar, suggesting that marital status may not be a strong differentiator for housing loan ownership. However, we can still find that the proportion of single customers with housing loans is slightly lower.

b) Results of modelling indicators

		Actual 0	Actual 1
Predicted	0	True Negative (TN) = 2778	False Negative (FN)=1761
Predicted	1	False Positive (FP) = 3197	True Positive (TP) = 5828

Table 1. Confusion Matrix

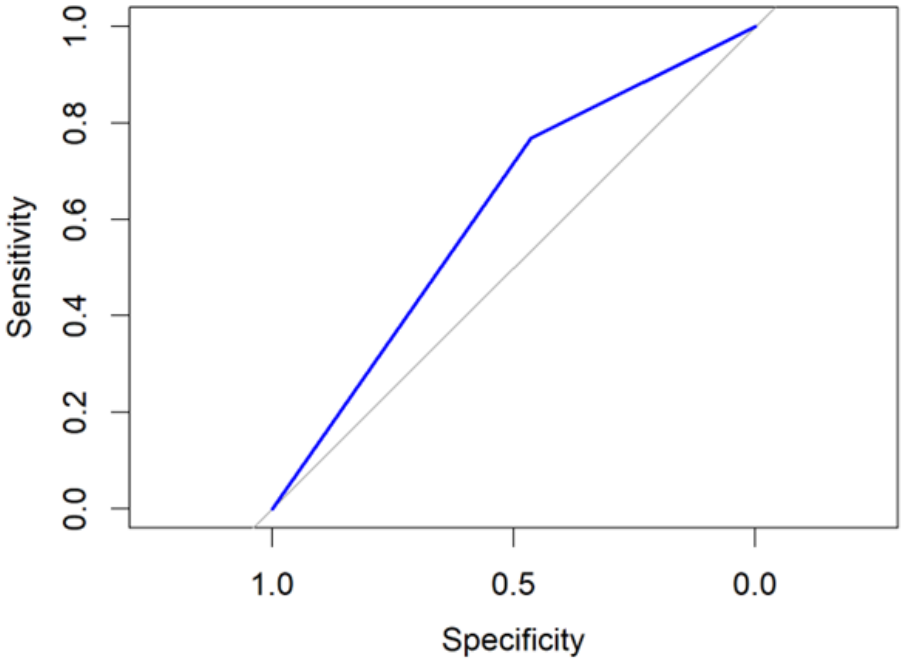


Figure 14. ROC Curve for House Loan Prediction

Overall, the model is generally effective, with an accuracy rate of 63.45% meaning that made correct predictions in about two-thirds of the cases, and a recall rate of 76.8% indicates that most customers having a housing loan were identified, which is crucial for the bank as it includes the majority of potential customers in its target group. The precision rate of 64.58% implies that although the model was able to identify a certain number of customers in need of a housing loan, it also has some misclassification, with non-need customers included, potentially causing inefficient marketing. AUC is not very high, reflecting the limited

predictive value. This may be due to the dataset's primary use in modelling customer subscriptions or to limited feature selection, selecting only basic personal information about the customer, which does not reflect the full picture of the customer very well.

c) Description of the results of the decision tree

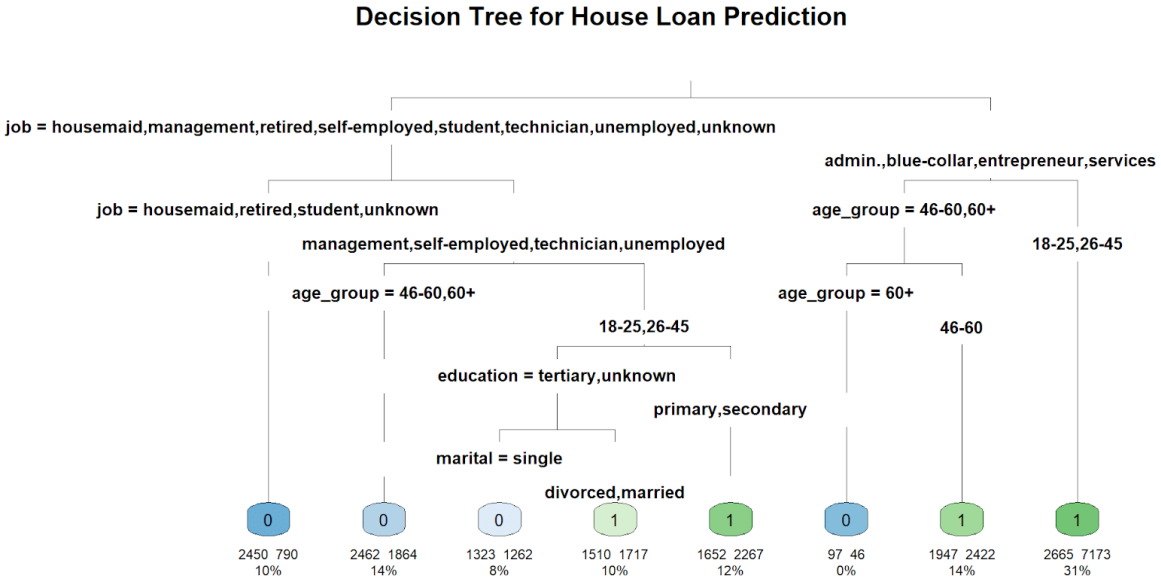


Figure 15. Decision Tree For House Loan Prediction

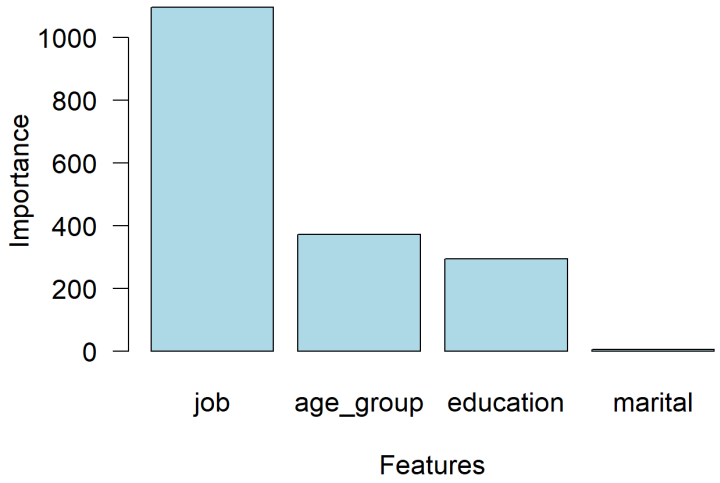


Figure 16. Feature Importance in Decision Tree

(i) Based on the figure 15 and 16, we can observe that Job has the greatest impact on housing loan demand, followed by age. For customers aged 18-60, especially those aged 18-45, those working in admin, blue-collar, entrepreneur, and services

sectors are more likely to have a housing loan. Customers working as housemaids, retired, students, or with an unknown job are generally less likely to have a housing loan.

(ii) Education and marital play an auxiliary role in predicting housing loan demand. For customers aged 18-45, working in management, self-employed, technician, and unemployed roles, and with primary or secondary education, are more likely to have a housing loan. But if their education level is tertiary or unknown, they are more likely to have a housing loan if they are married or divorced.

6. Recommendations

1) Recommendations for Subscription

a) Digital Marketing

(i) Set a Customer Segment for Better Targeting

Based on our analysis, customers can be segmented into three primary levels by age, balance, and occupation, each with distinct subscription potential:

Customer Level	Segment Name	Age Range	Balance Range	Occupation	Subscription Rate	Total Individuals Subscribed
Level 1	Senior Market	60+	>1000	Management, Retired	39.40%	268
Level 2	Working Professionals Market	30-60	0-5000	Blue-collar, Management, Technician, Admin	14.50%	2493
Level 3	General Market	All	Any	Any	9.20%	2528

Table 2. Customer Segmentation

Senior Market (Level 1): Customers aged 60+ with balances over 1000, particularly those in management or retired roles, show the highest subscription rate (39.4%) due to financial stability, making them ideal for targeted, high-conversion strategies.

Working Professionals Market (Level 2): Customers aged 30-60 with moderate balances (0-5000), including blue-collar, management, technician, and admin roles, offer high subscription volume (14.5%) and suit large-scale marketing.

General Market (Level 3): A diverse segment with lower subscription rates (9.2%) but significant population share, allowing for broader engagement opportunities.

(ii) KPI Setting

Tailor calling KPIs to specific days, prioritizing business development calls from Tuesday to Thursday. For Mondays and Fridays, allocate resources to administrative tasks or internal planning. Conduct further tests on time-of-day preferences within these days to fine-tune call scheduling for optimal results.

(iii) Phone Calls Practice Advice

(a) Duration

Call Duration	Purpose	Recommendation
260 seconds (Minimum Standard)	Baseline for all calls (introductory or final)	Set 260 seconds as the minimum call length to ensure adequate time for product information and customer interest assessment.
316 seconds (Benchmark for Final Calls)	Indicator of strong customer engagement	If a call reaches around 316 seconds, treat it as a strong engagement signal. Use this time to answer final questions and guide toward a decision.
Flexible Adjustment	Based on customer engagement level	Start with a 260-second baseline, but extend to 316+ seconds if the customer shows interest. For less engaged customers, aim to wrap up at 260 seconds and consider scheduling a follow-up.

Table 3. Call Duration

(b) Phone Calls Times Recommendations

Step	Objective	Details
First Call	Introduce key benefits	Schedule first call to highlight the main benefits of the campaign.

Second Call	Address specific questions	Answer any customer-specific questions and clarify details.
Third Call	Reinforce the offer or close if ready	Emphasise the offer and encourage closing if the customer is ready.
Follow-Ups for 'Other' Customers	Gradually share new information, testimonials, or limited-time offers	Avoid repetitive messaging; engage with tailored, non-intrusive information.

Table 4. Phone Calls Time Recommendations

(c) Utilisation of Phone Calls Fatigue

Limit customer contact to active campaign periods to avoid overexposure and maintain the effectiveness of messaging. Focus resources on maximizing interaction quality during campaigns, incorporating high-value updates or personalized offers to boost engagement.

Frequent follow-ups outside of active campaign periods may lead to customer fatigue. Building on the “at least three calls” suggestion, concentrating these calls within the campaign period can leverage peak customer interest and accelerate conversions. For “other” customers, intensive follow-ups during the campaign period may facilitate conversions, while follow-up frequency can be reduced outside campaign periods.

(d) Contact Method

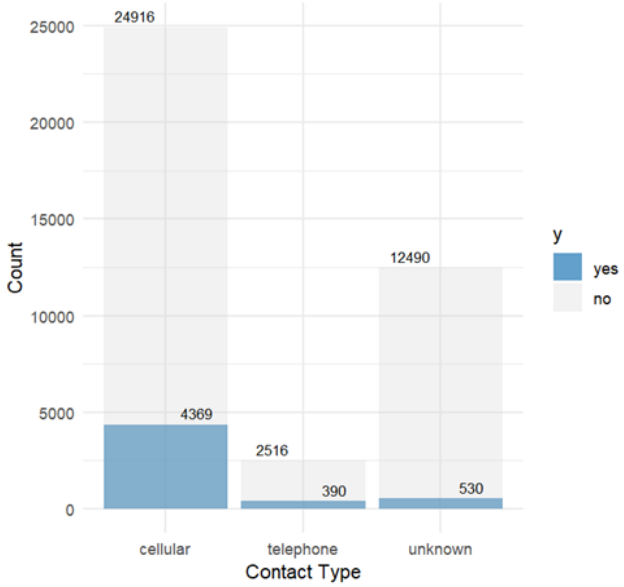


Figure 17. Contact Type by Subscription Status

Figure 17 indicates that cellular and telephone are the most effective contact methods for driving subscriptions, with cellular yielding the highest number of positive responses.

Therefore, prioritising cellular and telephone contact is recommended to maximise subscription outcomes.

b) E-commerce Team

Automate Reminders for Regular Follow-Ups: Implement an automated reminder system for each lead to ensure consistent follow-up without overloading any single customer. For example, set reminders to revisit “Other” customers after a set cooling period if they have not converted by the end of the current campaign, ensuring they remain engaged without feeling pressured.

2) Recommendations for Housing Loan

a) Bank Team

Since job is the most significant influencing factor, the bank can prioritize targeted marketing based on customers' jobs.

For customers working in admin, blue-collar, entrepreneur, and services sectors, the bank should increase efforts to promote housing loans. For the blue-collar and services groups, given the potential fluctuation in their income, the bank can offer more flexible loan repayment plans to increase their interest in housing loan products.

For customers working as housemaids, retired, students, or with unknown jobs, the bank should reduce the marketing resources allocated to housing loans, as these groups have relatively low demand for housing loans. Instead, the bank can recommend savings accounts, retirement plans, or other financial products to fully tap into the value of these customer segments.

For customers aged over 60, who generally have lower demand for housing loans, the bank can design products that meet the housing needs of their children, such as family housing loans, or promote products like retirement plans or time deposits.

Although marital ranks are lower in the importance of characteristics, married or divorced customers are more likely to have a house loan than single customers. Banks can pay special attention to married or divorced customers by launching house loans products to meet their family needs.

b) E-commerce Team

The bank can collaborate with e-commerce platforms related to furniture, home renovation, and appliances to offer joint promotional activities for customers with housing loan needs.

For example, providing furniture purchase discounts or home renovation loan benefits to eligible housing loan customers.

7. References

Baek, T.H. & Morimoto, M., 2012. Stay away from me: Examining the determinants of consumer avoidance of personalized advertising. *Journal of Advertising*, 41(1), pp.59–76.

Moro, S., Cortez, P. & Rita, P., 2018. A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. *Expert Systems*, 35(3), e12253.

Yan, C., Li, M. & Liu, W., 2020. Prediction of bank telephone marketing results based on improved whale algorithms optimizing S_Kohonen network. *Applied Soft Computing*, 92, Article 106259.

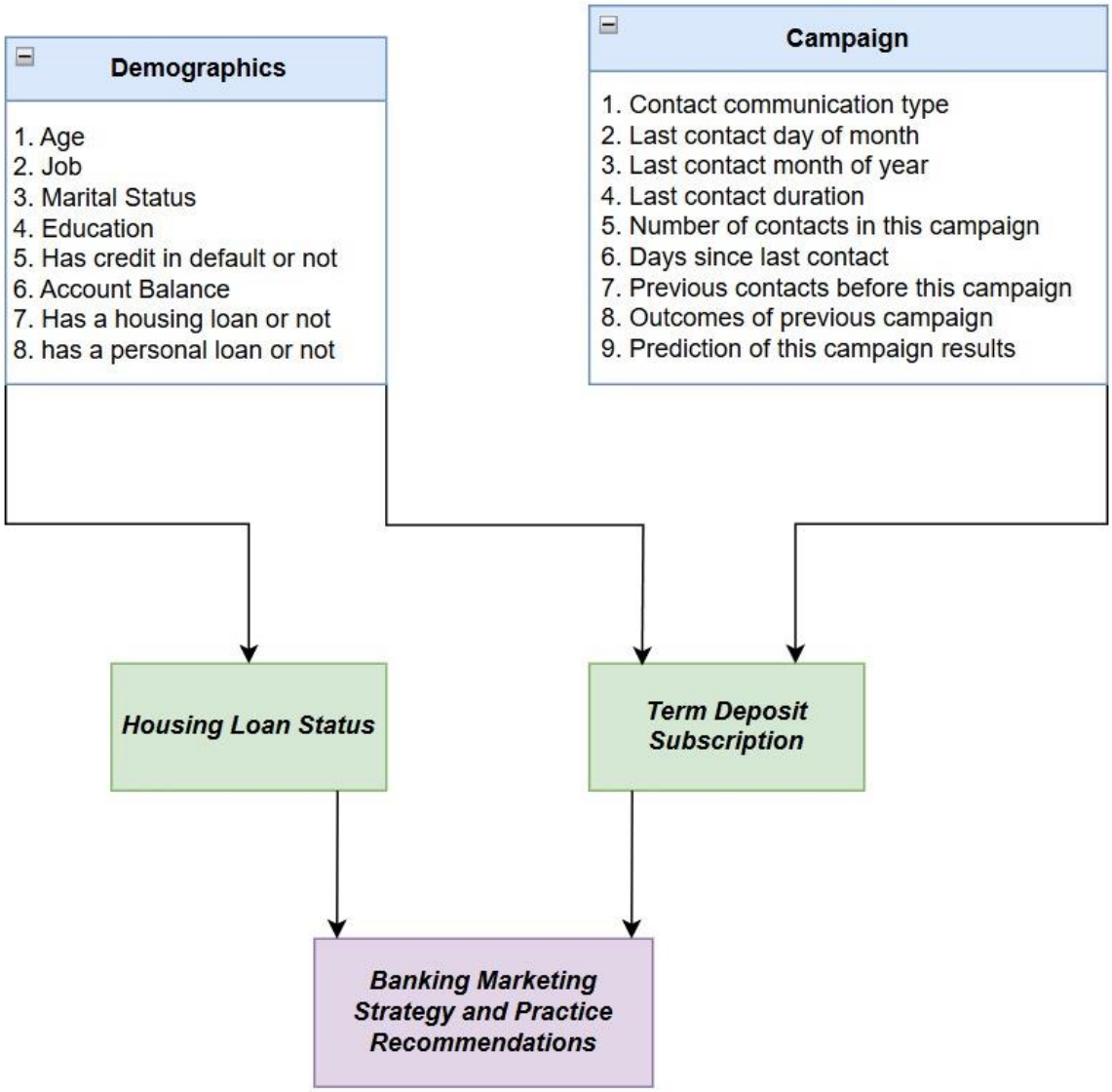
Kearns, J., 2019. *Understanding rising housing loan arrears*. Speech presented at Property Leaders' Summit, Canberra, 18 June 2019. Reserve Bank of Australia.

Quader, S.M., Shamsi, N.I. & Abdullah, M.N., 2020. Expansion and profitability of bank branches: A study on selected rural branches of Bangladesh. *Macroeconomic and Finance in Emerging Market Economies*, 13(3), pp.295–315.

Xie, C., Zhang, J.-L., 2023. How to improve the success of bank telemarketing? Prediction and interpretability analysis based on machine learning, *Computers & Industrial Engineering* 175, pp. 2, 7–9.

The Mortgage Reports, 2024. How much home can you buy with your education level. [online] Available at: https://themortgagereports.com/47354/how-much-home-can-you-buy-with-your-education-level [Accessed 11 November 2024].

*Modelling Guide



8. Appendices

The model for Research question 1:

```
library(tidyverse)
library(caret)
library(randomForest)
library(ggplot2)
library(dplyr)

data <- read.csv("bank-full.csv")

head(data)
summary(data)

data$job <- as.factor(data$job)
data$marital <- as.factor(data$marital)
data$education <- as.factor(data$education)
data$y <- as.factor(data$y)
data$default <- as.factor(data$default)
data$housing <- as.factor(data$housing)
data$loan <- as.factor(data$loan)
data$contact <- as.factor(data$contact)

data <- na.omit(data)



data$age_group <- cut(data$age,
                      breaks = c(18, 30, 40, 50, 60, 100),
                      labels = c("18-30", "30-40", "40-50", "50-60", "60+"),
                      right = FALSE)

data$balance_group <- cut(data$balance, breaks = c(-Inf, 0, 1000, 5000, 10000, Inf),
```

```
                labels = c("Negative", "0-1000", "1000-5000", "5000-10000", "10000+"),
                right = FALSE)

set.seed(123)
train_index <- createDataPartition(data$y, p = 0.7, list = FALSE)
train_data <- data[train_index, ]
test_data <- data[-train_index, ]

rf_model <- randomForest(y ~ age + job + marital + education + default + balance + housing + loan + contact,
                data = train_data, importance = TRUE)

importance_model = importance(rf_model)
print(importance_model)
varImpPlot(rf_model)
#Age, balance, contact, job.

rf_predictions <- predict(rf_model, test_data)

# Confusion Matrix
conf_matrix_rf <- confusionMatrix(rf_predictions, test_data$y)
print(conf_matrix_rf)

# Accuracy of Random Forest Model
rf_accuracy <- conf_matrix_rf$overall['Accuracy']
print(paste("Accuracy of Random Forest Model:", round(rf_accuracy, 2)))

# 1. Age
ggplot(data, aes(x = age_group, fill = y)) +
  geom_bar(data = subset(data, y == "yes"), alpha = 0.7, position = "dodge") +
  geom_bar(data = subset(data, y == "no"), alpha = 0.3, position = "dodge") +
  geom_text(
```

```
  stat = 'count',
  aes(label = ..count..),
  position = position_dodge(width = 0.9),
  vjust = -0.5,
  size = 3
 ) +
 scale_fill_manual(values = c("yes" = "#1f77b4", "no" = "#d3d3d3")) +  # Dark blue for 'yes', Light gray for 'no'
 labs(title = "Age Group Distribution by Subscription Status",
     x = "Age Group", y = "Count") +
 theme_minimal()

# 2. Balance Binning
ggplot(data, aes(x = balance_group, fill = y)) +
 geom_bar(data = subset(data, y == "yes"), alpha = 0.7, position = "dodge") +
 geom_bar(data = subset(data, y == "no"), alpha = 0.3, position = "dodge") +
 geom_text(
  stat = 'count',
  aes(label = ..count..),
  position = position_dodge(width = 0.9),
  vjust = -0.5,
  size = 3
 ) +
 scale_fill_manual(values = c("yes" = "#1f77b4", "no" = "#d3d3d3")) +  # Dark blue for 'yes', Light gray for 'no'
 labs(title = "Balance Distribution by Subscription Status (Binned)",
     x = "Balance Range", y = "Count") +
 theme_minimal()

# 3. Contact Type Distribution by Subscription Status
ggplot(data, aes(x = contact, fill = y)) +
 geom_bar(data = subset(data, y == "yes"), alpha = 0.7, position = "dodge") +
 geom_bar(data = subset(data, y == "no"), alpha = 0.3, position = "dodge") +
```

```
  geom_text(
    stat = 'count',
    aes(label = ..count..),
    position = position_dodge(width = 0.9),
    vjust = -0.5,
    size = 3
  ) +
  scale_fill_manual(values = c("yes" = "#1f77b4", "no" = "#d3d3d3")) +  # Dark blue for 'yes', Light gray for 'no'
  labs(title = "Contact Type by Subscription Status",
       x = "Contact Type", y = "Count") +
  theme_minimal()

# 4. Job Type Distribution by Subscription Status
ggplot(data, aes(x = job, fill = y)) +
  geom_bar(data = subset(data, y == "yes"), alpha = 0.7, position = "dodge") +
  geom_bar(data = subset(data, y == "no"), alpha = 0.3, position = "dodge") +
  geom_text(
    stat = 'count',
    aes(label = ..count..),
    position = position_dodge(width = 0.9),
    vjust = -0.5,
    size = 3
  ) +
  scale_fill_manual(values = c("yes" = "#1f77b4", "no" = "#d3d3d3")) +  # Dark blue for 'yes', Light gray for 'no'
  labs(title = "Job Type Distribution by Subscription Status",
       x = "Job Type", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))



age_summary <- data %>%
```

```
  group_by(age_group, y) %>%
  summarise(count = n()) %>%
  group_by(y) %>%
  mutate(percentage = count / sum(count) * 100) %>%
  arrange(y, desc(percentage))

print(age_summary)
# Pie chart for 'No' Subscription with percentages
no_data <- age_summary %>% filter(y == "no")
no_data <- no_data %>% mutate(percentage_label = paste0(round(percentage, 1), "%"))

ggplot(no_data, aes(x = "", y = count, fill = age_group)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  scale_fill_manual(values = c("30-40" = "#1f77b4", "40-50" = "#ff7f0e",
                               "50-60" = "#2ca02c", "18-30" = "#d62728",
                               "60+" = "#9467bd")) +
  geom_text(aes(label = percentage_label), position = position_stack(vjust = 0.5)) +
  labs(title = "Age Group Distribution for 'No' Subscription") +
  theme_void()

# Pie chart for 'Yes' Subscription with percentages
yes_data <- age_summary %>% filter(y == "yes")
yes_data <- yes_data %>% mutate(percentage_label = paste0(round(percentage, 1), "%"))

ggplot(yes_data, aes(x = "", y = count, fill = age_group)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  scale_fill_manual(values = c("30-40" = "#1f77b4", "40-50" = "#ff7f0e",
                               "50-60" = "#2ca02c", "18-30" = "#d62728",
                               "60+" = "#9467bd")) +
```

```
  geom_text(aes(label = percentage_label), position = position_stack(vjust = 0.5)) +
  labs(title = "Age Group Distribution for 'Yes' Subscription") +
  theme_void()



balance_summary <- data %>%
  group_by(balance_group, y) %>%
  summarise(count = n()) %>%
  group_by(y) %>%
  mutate(percentage = count / sum(count) * 100) %>%
  arrange(y, desc(percentage))

print(balance_summary)
# Pie chart for 'No' Subscription with percentages (Balance Groups)
no_balance_data <- balance_summary %>% filter(y == "no")
no_balance_data <- no_balance_data %>% mutate(percentage_label = paste0(round(percentage, 1), "%"))

ggplot(no_balance_data, aes(x = "", y = count, fill = balance_group)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  scale_fill_manual(values = c("0-1000" = "#1f77b4", "1000-5000" = "#ff7f0e",
                               "Negative" = "#2ca02c", "5000-10000" = "#d62728",
                               "10000+" = "#9467bd")) +
  geom_text(aes(label = percentage_label), position = position_stack(vjust = 0.5)) +
  labs(title = "Balance Group Distribution for 'No' Subscription") +
  theme_void()

# Pie chart for 'Yes' Subscription with percentages (Balance Groups)
yes_balance_data <- balance_summary %>% filter(y == "yes")
yes_balance_data <- yes_balance_data %>% mutate(percentage_label = paste0(round(percentage, 1), "%"))
```

```
ggplot(yes_balance_data, aes(x = "", y = count, fill = balance_group)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  scale_fill_manual(values = c("0-1000" = "#1f77b4", "1000-5000" = "#ff7f0e",
                               "Negative" = "#2ca02c", "5000-10000" = "#d62728",
                               "10000+" = "#9467bd")) +
  geom_text(aes(label = percentage_label), position = position_stack(vjust = 0.5)) +
  labs(title = "Balance Group Distribution for 'Yes' Subscription") +
  theme_void()



contact_summary <- data %>%
  group_by(contact, y) %>%
  summarise(count = n()) %>%
  group_by(y) %>%
  mutate(percentage = count / sum(count) * 100) %>%
  arrange(y, desc(percentage))

print(contact_summary)

# Pie chart for 'No' Subscription with percentages (Contact Types)
no_contact_data <- contact_summary %>% filter(y == "no")
no_contact_data <- no_contact_data %>% mutate(percentage_label = paste0(round(percentage, 1), "%"))

ggplot(no_contact_data, aes(x = "", y = count, fill = contact)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  scale_fill_manual(values = c("cellular" = "#1f77b4", "unknown" = "#ff7f0e", "telephone" = "#2ca02c")) +
  geom_text(aes(label = percentage_label), position = position_stack(vjust = 0.5)) +
  labs(title = "Contact Type Distribution for 'No' Subscription") +
  theme_void()
```

```
# Pie chart for 'Yes' Subscription with percentages (Contact Types)
yes_contact_data <- contact_summary %>% filter(y == "yes")
yes_contact_data <- yes_contact_data %>% mutate(percentage_label = paste0(round(percentage, 1), "%"))

ggplot(yes_contact_data, aes(x = "", y = count, fill = contact)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  scale_fill_manual(values = c("cellular" = "#1f77b4", "unknown" = "#ff7f0e", "telephone" = "#2ca02c")) +
  geom_text(aes(label = percentage_label), position = position_stack(vjust = 0.5)) +
  labs(title = "Contact Type Distribution for 'Yes' Subscription") +
  theme_void()

job_summary <- data %>%
  group_by(job, y) %>%
  summarise(count = n()) %>%
  group_by(y) %>%
  mutate(total_count = sum(count),
         percentage = count / total_count * 100) %>%
  arrange(y, desc(percentage))

print(job_summary)

# Bar plot for Job Distribution for 'No' Subscription with percentage labels
ggplot(job_summary %>% filter(y == "no"), aes(x = job, y = percentage, fill = y)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_dodge(width = 0.8), vjust = -0.5, size = 3) +  # Add percentage labels
  labs(title = "Job Distribution for 'No' Subscription",
       x = "Job Type", y = "Percentage") +
  scale_fill_manual(values = c("no" = "#d3d3d3")) +  # Custom color for 'No'
```

```
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +  # Rotate and reduce font size
  theme_minimal()

# Bar plot for Job Distribution for 'Yes' Subscription with percentage labels
ggplot(job_summary %>% filter(y == "yes"), aes(x = job, y = percentage, fill = y)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_dodge(width = 0.8), vjust = -0.5, size = 3) +  # Add percentage labels
  labs(title = "Job Distribution for 'Yes' Subscription",
       x = "Job Type", y = "Percentage") +
  scale_fill_manual(values = c("yes" = "#1f77b4")) +  # Custom color for 'Yes'
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +  # Rotate and reduce font size
  theme_minimal()
```

The model for Research question 3:

decision tree for predicting housing loan

```
library(rpart)
library(rpart.plot)

# read data
house_loan_data <- read.csv("house_loan1.csv")

# check the data
head(house_loan_data)

# Convert categorical variables to factors
house_loan_data$job <- as.factor(house_loan_data$job)
house_loan_data$marital <- as.factor(house_loan_data$marital)
```

```
house_loan_data$education <- as.factor(house_loan_data$education)


hist(house_loan_data$age,
    main = "Age Distribution",
    xlab = "Age",
    ylab = "Frequency",
    col = "lightblue",
    border = "black")


house_loan_data$age_group <- cut(house_loan_data$age,
                        breaks = c(0, 25, 45, 60, Inf),
                        labels = c("18-25", "26-45", "46-60", "60+"),
                        right = FALSE)


table(house_loan_data$age_group)


# remove the original age column
house_loan_data <- house_loan_data[,-which(names(house_loan_data) == "age")]


# Convert target variable (housing) to 0 and 1, where "yes" is 1 and "no" is 0
house_loan_data$housing <- ifelse(house_loan_data$housing == "yes", 1, 0)
house_loan_data$housing <- as.factor(house_loan_data$housing)


# Split data into training and test sets (70% training, 30% testing)
set.seed(123)
train_indices <- sample(1:nrow(house_loan_data), size = 0.7 * nrow(house_loan_data))
train_data <- house_loan_data[train_indices, ]
test_data <- house_loan_data[-train_indices, ]


# Build a decision tree model with housing as the target variable
tree_model <- rpart(housing ~ age_group + job + marital + education,
```

```
            data = train_data,
            method = "class",
            control = rpart.control(cp = 0.001, maxdepth = 5))

# Visualize the decision tree
rpart.plot(tree_model,
       type = 3,
       extra = 101,
       under = TRUE,
       fallen.leaves = TRUE,
       cex = 0.6,
       main = "Decision Tree for House Loan Prediction")

# Predict house loan class on the test set
test_data$predicted <- predict(tree_model, newdata = test_data, type = "class")

# Generate a confusion matrix to evaluate model performance
confusion_matrix <- table(Predicted = test_data$predicted, Actual = test_data$housing)
print(confusion_matrix)

# Calculate model accuracy
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(paste("Model accuracy on the test set:", round(accuracy * 100, 2), "%"))

# Extract TN, FP, FN, TP
TN <- confusion_matrix[1, 1]
FN <- confusion_matrix[1, 2]
FP <- confusion_matrix[2, 1]
TP <- confusion_matrix[2, 2]

# Calculate Recall
```

```
recall <- TP / (TP + FN)
print(paste("Recall:", round(recall * 100, 2), "%"))

# Calculate Precision
precision <- TP / (TP + FP)
print(paste("Precision:", round(precision * 100, 2), "%"))

# ROC
library(pROC)
roc_curve <- roc(test_data$housing, as.numeric(test_data$predicted))
plot(roc_curve, main = "ROC Curve for House Loan Prediction", col = "blue")
auc <- auc(roc_curve)
print(paste("AUC:", round(auc, 2)))

feature_importance <- tree_model$variable.importance
print(feature_importance)

# Plot feature importance
barplot(feature_importance, main = "Feature Importance in Decision Tree",
        xlab = "Features", ylab = "Importance", col = "lightblue", las = 1)
```