Predicting Term Deposit Subscriptions Using Statistical Techniques

STATISTICS FOR BUSINESS

**BY**

SHALINI PRIYA ELAVAZHAGAN

STUDENT ID:40459377

1. Introduction And Background

In the competitive world of banking, understanding customer behaviour is crucial for improving marketing strategies and upgrading customer care. Many financial institutions focus on promoting one key area called term deposits, an investment product offers secure way for the customers to grow their savings. Recently, bank had launched telephone marketing campaign aimed to encourage customers to subscribe term deposits. Research shows that the bank specific factors on deposits of banks in Ghana from 2008 to 2017 using a Random effects model increases their profitability (Yang, 2023). Here, I had used data from previous marketing campaign contains 40641 observations and 29 variables.

In this process, explains how factors like customer gender, occupation, salary, customers recently contacted months during campaign and previous marketing campaign. The aim is to find most effective factors related to subscription of term deposits.

In the financial services industry, marketing campaigns are critical tools used to generate new business and maintain customer loyalty. A term deposit is an interest-bearing deposit made by a customer with a fixed term, and typically, it offers a higher rate of return compared to a regular savings account. [2] states that gender can influence the likelihood of subscribing to products like term deposits, [3] shows an individual's job type can impact their financial risk tolerance and investment choices, including term deposits, [4] higher income is often associated with an increased likelihood of investing in savings products such as term deposits.[5] explains campaigns held during specific months can influence customer decisions, including those to subscribe to term deposits and [6] past marketing campaign outcomes enhances future customer decisions, including term deposit subscriptions.

Through analysis, I had considered **gender**, **Occupation**, **salary**, **month**(Marketing campaign last contacted month of year) and **previous marketing campaign outcome** will help in identifying the primary drivers behind term deposit subscriptions, with the aim of refining its marketing strategies and improving the customer experience.

In this study utilized statistical and machine learning techniques like correlation analysis, chi-square test and logistic regression model to determine the factors associated with term deposits subscriptions.

These insights gained will play a vital role in the banks strategic decision-making, improving future marketing campaigns and the overall customer experience, ultimately driving greater subscription rates to term deposits.

2.0 Methodology:

2.1 Descriptive statistics:

1. Raw data summary:

* Checked working directory using **getwd()**, then set working directory to locate the dataset by **setwd()**.
* Utilized RStudio packages were given below.

1. readxl→ to read data in excel format.
2. Dplyr → for data cleaning & manipulation.
3. Caret → for model evaluation.
4. ggplot2 → for data visualization.
5. tidyr → to structure data properly for analysis.
6. pROC → for curve analysis of binary classification model.
7. Vcd → for analyse and visualize the categorical data.

* Loaded the term deposit data in R and assigned it to variable data.
* Then summarized raw data using **Summary().** R code was listed in appendix A.1.

1. Data Cleaning and formatting:

Data cleaning and formatting is an important step in data analysis, in which invalid or unwanted data can be addressed to get an effective predictive model. Steps followed was listed below:

1. Checked NA values - used **is.na(data)** to check NA values in the dataset.
2. Replaced missing categorical by Unknown – used data[is.na(data)] <- "unknown" to replace NA values in a dataset with string “unknown”.

is.na() checks each element in data and returns a logical matrix, replaces the NA values to unknown.

1. Converting target variable(subscribed) to binary – check subscribed column in data, compares each value in subscribed column to string "yes". ifelse() conditional function works elementwise, if TRUE assigns value 1 or false assigns 0. as.factor() convert this numeric vector into a **factor** (categorical) variable.
2. Convert categorical variable to factor - categorical columns the names of the columns, lapply() to list vector, data[categorical columns], and the function as.factor convert selected columns into a factor. ensure that variables are treated as categorical data. R code can view in appendix A.2.
3. Data normalizing:

Data normalization is the process of scaling values of numerical features in a dataset. It improves the performance of machine learning algorithms. normalization can help improve convergence speeds during training in gradient-based methods like logistic regression or neural networks[7].

* At first, replaced ‘m’ to ‘male’ in gender column using **pipe operator** (%>%) from the dplyr package, allows multiple operations , mutate() to modify or create new variable in data and case\_when() for conditional transformation.
* Next removed rows contain value “unknown” in occupation column using pipe operator for multiple operation, mutate() to modify data, filter() to extract the selected rows and print() to print the results.
* Then ordered all the months in ascending in month column using pipe operator for multiple operation, mutate() to modify data, tolower() convert values, gsub() performs pattern replacement, factor() converts month to factor, arrange() to sort the data by month and print() to print results. R code listed in appendix A.3.

1. Descriptive Analysis:

statistical method to summarize data and makes data understandable and provides insights into trends and distributions without making predictions. Descriptive analysis provides a summary of key data attributes, such as central tendency and variability[8]. After cleaning the data used summary() to get descriptive analysis. R code viewed in appendix A.4.

1. Frequency table for categorical variables

Created frequency table to display count of unique values for selected variables gender, Occupation, salary, month and previous marketing campaign outcome and subscribed using table() Computes frequency of each unique value in the specified column.

Visualization:

It helps to view the extreme values in data, according to [9] effective data visualization transforms complex data into clear and impactful visual representations that enhance understanding and decision-making. In this ggplot package is used for visualization.

1. Bar plot for subscriptions- ggplot(): Initializes the plot using the data.

aes(x = subscribed): Maps the subscribed variable to the x-axis. geom\_bar() creates bar chart, fill to set color and labs() to add title.

1. Box plot for gender by subscription- to visualize the distribution of the gender variable with respect to the subscription status. aes(x = as.factor(subscribed)) maps the subscribed variable, , fill to set color and labs() to add title.
2. Bar plot for occupation by subscription- to visualize the distribution of occupation by subscription status used ggplot() for plotting, aes(x = subscribed) to map, geom\_bar() creates bar chart, fill to set color and labs() to add title.
3. Bar plot for salary by subscription- to visualize the distribution of salary by subscription status used ggplot() for plotting, aes(x = subscribed) to map, geom\_bar() creates bar chart, fill to set color and labs() to add title.
4. Bar plot for month by subscription- to visualize the distribution of months by subscription status used ggplot() for plotting, aes(x = subscribed) to map, geom\_bar() creates bar chart, fill to set color and labs() to add title.
5. Bar plot for previous campaign outcomes by subscription- to visualize the distribution of previous campaign outcomes by subscription status used ggplot() for plotting, aes(x = subscribed) to map, geom\_bar() creates bar chart, fill to set color and labs() to add title. R code view in Appendix A.5.

2.2 Measures of correlation:

1. Correlation for numeric: dplyr packages is used, correlation matrix shows relationships between numeric variables in the dataset. **data %>% select\_if(is.numeric)** uses the **select\_if()** function from the **dplyr** package to select only the numeric columns from the data frame. **cor(numeric\_data)** calculates **correlation matrix** of numeric data.
2. chi-square test for categorical: determine whether significant association between the two categorical variables. Performs for gender and subscribed **table()** generates **contingency table** from two categorical variables gender and subscribed. **chisq.test()**performs the **Chi-square test of independence.**

3-way Contingency table for for gender, salary, subscribed: used to perform **Chi-square tests of independence** for two pairs of categorical variables gender vs subscribed and salary vs subscribed. A contingency table displays frequency counts of the three categorical variables. table(data$gender, data$salary, data$subscribed) creates three-way table. R code listed in A.6.

2.3 Split data into train & test data:

to predict logistic regression model, splits data into **training data** → subset of dataset to train model, **test data** → to evaluate the model. Using this approach, divided 70%→ train data, 30%→ test data.

1. set .seed to ensure reproducibility, student id#40459377 → can use any integer to define seed.
2. Then used caret package, for training, tuning the model and it was excellent choice for creating predictive models.
3. Created train index→ 70% for training and test data→ 30% for evaluating model. R code listed in A.7.
   1. Creating a model:

Created logistic regression model used to analyse the categorical data. Logistic regression works by modelling the relationship between a dependent binary variable and one or more independent variables. The dependent variable represents the outcome of interest (e.g., subscription to a term deposit)[10]. Used selected variables gender, occupation, salary, previous campaign, month and used summary() to summarise the data.

1. Predict and evaluate model – in these generated predictions using trained logistic regression model on a data set.
2. Convert probabilities to binary predictions- performed to convert predicted probabilities into binary predictions based on specified threshold 0.5.
3. Confusion matrix- created to generate a confusion matrix for evaluating the performance of a binary classification model.
4. Calculate accuracy- created to calculate the **accuracy** of a binary classification model by comparing the predicted values (test\_data$predicted) with the actual values (test\_data$subscribed).
5. calculate and Plot the AUC-ROC Curve – used to calculate and plot AUC-ROC curve, which is commonly used for evaluating the performance of a binary classification model. R code in A.8.

3.0 Result and Discussion:

3.1 Descriptive analysis:

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

Table1. descriptive statistics

* performed descriptive analysis of 40641 obs. of 29 variables. The last contact duration was 999 was not reliable. However, all the variables are categorical, analysed further.

3.2 Visualizations:

1. Bar plot for subscriptions:

A blue rectangular object with black text

Description automatically generated

* This graph represents customer subscription distribution, 0 – no subscription, 1 – yes, only 5000 customers subscribed term deposits so far and more than 35000 customers were not subscribed yet.

1. Box plot for gender by subscription:

A white rectangular object with black lines

Description automatically generated

* It shows female subscribed(1) and non-subscribed(0) for deposits are comparatively lower than male subscribed(1) and non-subscribed(0) for deposits. It results that males are eager in deposits than female.

1. Bar plot for occupation by subscription:

A graph with red and blue bars

Description automatically generated

* In this retired people scores more in term deposits than other fields. Secondly administrative, thirdly blue-collar & technician people subscribed for term deposits.

1. Bar plot for salary by subscription:

A graph of a number of red and blue rectangular bars

Description automatically generated

* Customers has high pay salary shows interests for subscription is high compared to medium & low income. Hence in future, marketing can be focused on those people.

1. Bar plot for month by subscription:

A graph of a number of bars

Description automatically generated

* Based on previous data, March, September, October month scores more subscription compared to other months.

1. Bar plot for previous campaign outcomes by subscription:

A graph of a campaign distribution

Description automatically generated with medium confidence

* Analysing previous campaign outcomes non-existent criteria scores more than the others.

3.3 measures of correlation:

A screenshot of a computer

Description automatically generated

In this except ID, age, contact duration, previous contacts all other criteria was in negative.

* chi-square test for categorical variables

1. for gender and subscribed

A white background with black text

Description automatically generated

1. for gender, salary, subscribed

A screenshot of a computer code

Description automatically generated

3.4 logistic regression model:

A screenshot of a computer

Description automatically generated

A black text on a white background

Description automatically generated

1. Confusion matrix

A screenshot of a computer

Description automatically generated

1. Calculate accuracy

A screenshot of a computer

Description automatically generated

1. calculate and Plot the AUC-ROC Curve



A graph with a blue line

Description automatically generated

4.0 Conclusion:

This project aimed to identify key factors influencing customers' decisions to subscribe to term deposit products through a comprehensive analysis of a large banking dataset. Using statistical and machine learning techniques such as descriptive statistics, chi-square tests, correlation analysis, and logistic regression several insights were uncovered.

The analysis indicated that variables like occupation, salary level, month of contact, and previous campaign outcomes show a stronger association with term deposit subscriptions than gender alone. Notably, retired individuals and higher-income earners were more likely to subscribe, suggesting that targeted campaigns focusing on these segments could yield higher conversion rates. Additionally, months such as March, September, and October saw increased subscription activity, offering potential for seasonally optimized marketing strategies.

Although the logistic regression model provided some predictive capability, the overall accuracy and significance of results suggest there is room for improvement. Incorporating additional variables such as customer age, contact duration, or external economic indicators may enhance model performance and deliver more robust insights.

In conclusion, this analysis underscores the importance of data-driven marketing in the financial sector and highlights the potential to refine campaign strategies by leveraging customer demographics and historical behaviour.

References

[1] Yang, Y., 2023. *Variables that Influence Customers' Choice to Purchase Term Deposits.* [Online]   
Available at: https://www.researchgate.net/publication/373896931\_Variables\_that\_Influence\_Customers'\_Choice\_to\_Purchase\_Term\_Deposits  
[Accessed 13 sep 2023].

[2] Agarwal, S., Driscoll, J. C., Gabaix, X. and Laibson, D. (2009) 'The Age of Reason: Financial Decisions over the Life Cycle and Implications for Regulation', *Brookings Papers on Economic Activity*, 2009(2), pp. 51–117.

[3] Kumar, P., Mishra, R. and Tripathi, S. (2018) 'Impact of Occupation on Investment Behavior of Investors', *Journal of Financial Planning and Analysis*, 8(2), pp. 29-41.

[4] Guiso, L., Jappelli, T. and Terlizzese, D. (1996) 'Income Risk, Borrowing Constraints, and Portfolio Choice', *The American Economic Review*, 86(1), pp. 158-172.

[5] Chandon, P., Wansink, B. and Laurent, G. (2000) 'A Benefit Congruency Framework of Sales Promotion Effectiveness', *Journal of Marketing*, 64(4), pp. 65-81.

[6] Blake, A. and Crosbie, P. (2016) 'Predicting Consumer Behavior in Financial Products Using Historical Data: A Study of Past Marketing Campaigns', *Marketing Science*, 35(4), pp. 562-579.

[7] Kuhn, M. (2021). *caret: Classification and Regression Training*. R package version 6.0-88. Available at: https://CRAN.R-project.org/package=caret.

[8] Mann, P.S., 2021. **Introductory Statistics**. 10th ed. Hoboken: Wiley.

[9] Kirk, A., 2016. *Data Visualisation: A Handbook for Data Driven Design*. London: Sage Publications.

[10] Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.