

BUAN 6356.006 – Business Analytics with R – Fall’23

PREDICTING BANK TERM DEPOSIT SUBSCRIPTION

***Submitted By:***

Sai Lahari Korsipati

Sri Sravya Jamallamudi

Yash Taneja

Lishaan Parinkala

Bhargav Reddy Munnangi

(Group 15)

***Presentation Link:***

<https://cometmail-my.sharepoint.com/:v:/g/personal/sxk230101_utdallas_edu/EVzxojR200FJvPw-yg81Q-0BqtcSOsjQw4s2q92u_Bjugg?referrer=Teams.TEAMS-ELECTRON&referrerScenario=MeetingChicletGetLink.view.view>

# Acknowledgement

To begin with, we would like to express our gratitude to Prof. Zhe (James) Zhang and Teaching Assistant Arjun Singh Pathania for giving us the opportunity for completing a project and helping us understanding the subject and concepts throughout the semester and coursework. Their enthusiasm, patience, insightful comments, helpful information, practical advice, and unceasing ideas have always helped us tremendously in our analysis and writing of this project. Their immense knowledge, profound experience and professional expertise in Business Analytics has enabled us to complete this project successfully. Without their support and guidance, this project would not have been possible.

We also wish to express my gratitude to all the developers who helped and played a part in making the dataset in Kaggle.com because of whom the project base line has been formed.

Thank You

# Executive Summary

There has been a revenue decline in the Portuguese Bank and their customers are not investing enough in long-term deposits. So, the bank would like to identify existing customers that have a higher chance to subscribe for a long-term deposit and focus marketing efforts on such customers. Thus, the project’s motive and objective is to predict the term insurance subscription with the bank dataset using machine learning models.

The existing data shows there is 12.6% term subscription rate. Since 12.6% is a small percentage, we must accurately predict this rate, as the bank is more interested in finding this group and targeting them for getting furthermore subscriptions.

We have used 5 machine learning models to predict the subscription rate such as K-Nearest Neighbors, Logistic Regression, SVM, Random Forest, and Decision Tree. During the analysis Random Forest models showed the highest accuracy with 93.63% and a CI @95% of 93.15% to 94.09%.

# Key Words

Term Deposit Subscription, Banking Sector, Regression, Correlation, Logistic Regression, Decision Tree, Random Forest, K Near Neighbor, Support Vector Machine.

# Table of Contents

|  |  |  |
| --- | --- | --- |
|  | **Title** | **Page No.** |
| **Chapter 1** | Introduction | 5 |
| Objectives of Study | 5 |
| **Chapter 2** | Methodology | 5 |
| Data Set | 5 |
| Data Description | 5 |
| Data Cleaning | 6 |
| **Chapter 3** | Demographic Analysis – Histograms | 7 |
| Demographic Analysis – Bar Plots | 8 |
| Modelling Techniques | 10 |
| Logistic Regression | 11 |
| Decision Tree | 11 |
| Random Forest | 11 |
| K Nearest Neighbor | 11 |
| Support Vector Machine | 12 |
| **Chapter 4** | Best Technique | 12 |
| Conclusion | 13 |
| **Misc.** | References | 13 |

List of Figures

|  |  |  |
| --- | --- | --- |
| **Figure** | **Figure Title** | **Page No** |
| Figure 1 | Histograms | 7 |
| Figure 2 | Bar Plots | 8 |
| Figure 3 | Modelling Techniques | 10 |
| Figure 4 | Confusion Matrix - Logistic Regression | 11 |
| Figure 5 | Confusion Matrix - Decision Tree | 11 |
| Figure 6 | Confusion Matrix - Random Forest | 11 |
| Figure 7 | Confusion Matrix - K Nearest Neighbor | 11 |
| Figure 8 | Confusion Matrix - Support Vector Machine | 12 |
| Figure 9 | ROC – Random Forest | 13 |

# Introduction

In today's competitive banking landscape, financial institutions are constantly seeking strategies to enhance customer engagement and boost profitability. One such strategy involves predicting customer behavior, particularly in terms of their propensity to subscribe to term deposits. Term deposits, also known as time deposits, offer customers a higher interest rate in exchange for committing their funds for a predetermined period. Accurately predicting term deposit subscription rates enables banks to effectively target potential customers and optimize their marketing campaigns, leading to increased revenue and overall business growth. This report delves into the development of a predictive model for term deposit subscription using a comprehensive dataset of customer information and relevant financial factors. The model aims to identify the key drivers influencing customer decisions regarding term deposit subscriptions, enabling banks to tailor their marketing strategies accordingly and maximize the likelihood of customer uptake.

# Objectives of Study

The objective of the project is to use BI techniques like clustering and association to predict if the client will subscribe (yes/no) a term deposit (variable y). The various steps that will be followed in the project will be:

* + - Identify and visualize which factors which are contributing to the subscription for bank term deposit.
    - Build a prediction model to predict the subscription of bank term deposit.
    - Based on model performance, providing necessary recommendations to the bank, which will increase the future term deposit subscriptions.

# Methodology

To determine the characteristics that influence the subscription for a bank term deposit, we will use five different techniques: logistic regression, K-nearest neighbor, support vector machine, decision tree, and random forest. The model that yields the greatest accuracy will be chosen.

# Dataset

We have taken the Portuguese bank dataset from Kaggle. The multivariate dataset has 32950 instances and 15 attributes (5 numerical and 10 categorical). The target variable is a binary feature (Yes/No).

# Data Description

The attributes for the given dataset are mentioned below along with their respective categories

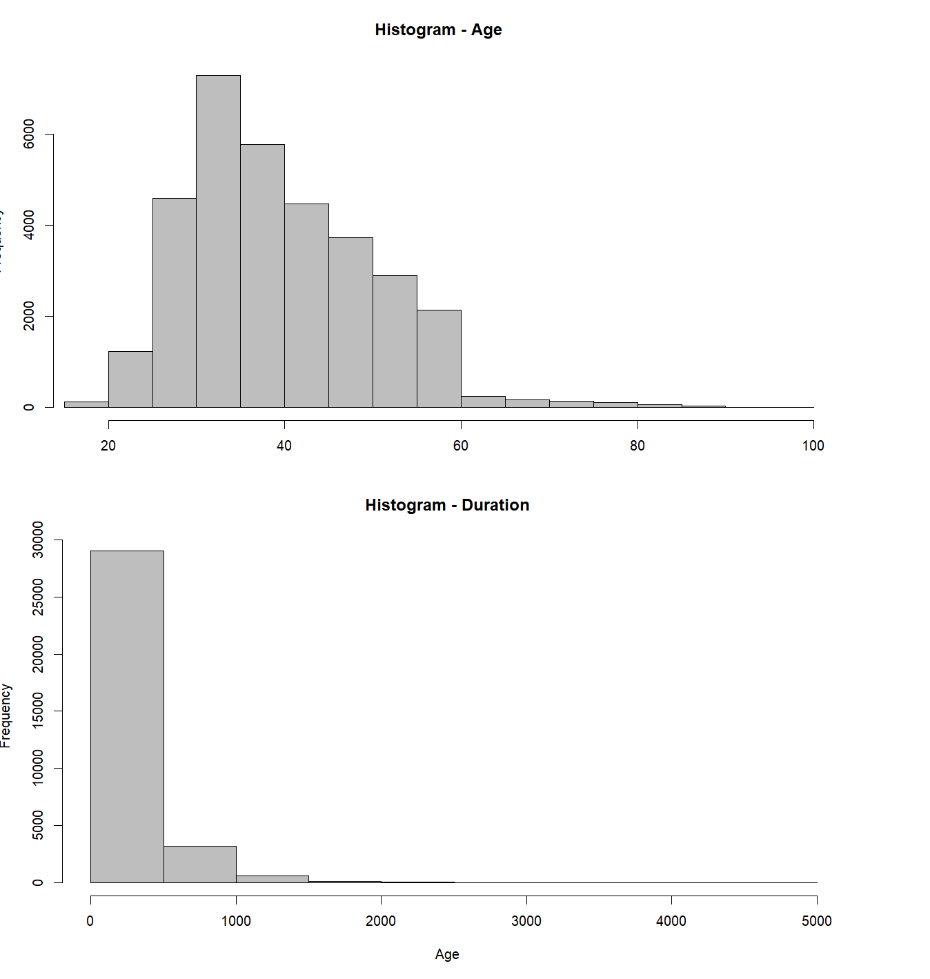
1. age: age of a person (numeric)
2. job: type of job ('admin.', 'bluecollar', 'entrepreneur', 'housemaid', 'management', 'retired', 'selfemployed', 'services', 'student', 'technician', 'unemployed', 'unknown') (Categorical, nominal)
3. marital: marital status ('divorced','married','single','unknown'; note: 'divorced' means divorced or widowed (categorical,nominal)
4. education:('basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown') (categorical,nominal)
5. default: has credit in default? ('no','yes','unknown') (categorical,nominal)
6. housing: has housing loan? ('no','yes','unknown') (categorical, nominal)
7. loan: has personal loan? ('no','yes','unknown') (categorical, nominal)
8. contact: contact communication type ('cellular','telephone') (categorical, nominal)
9. month: last contact month of year ('jan', 'feb', 'mar', …, 'nov', 'dec') (categorical, ordinal)
10. day\_of\_week: last contact day of the week ('mon', 'tue', 'wed', 'thu', 'fri') (categorical, ordinal)
11. duration: last contact duration, in seconds. Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no') (numeric)
12. campaign: number of contacts performed during this campaign and for this client (includes last contact) (numeric)
13. pdays: number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted) (numeric)
14. previous: number of contacts performed before this campaign and for this client (numeric)
15. poutcome: outcome of the previous marketing campaign ('failure','nonexistent','success') (categorical,nominal)
16. y: has the client subscribed to a term deposit? ('yes','no') (binary)

# Data Cleaning

The dataset has 32950 records and there are no null values. There are 8 duplicates in the dataset and have been removed. The data has been further checked for unknown values and there are 8643 records which consist of single or multiple unknowns, they have been removed as they will not be useful for building the model and prediction. After removing the unknowns there are a total of 24299 records.

# Demographic Analysis – Histograms

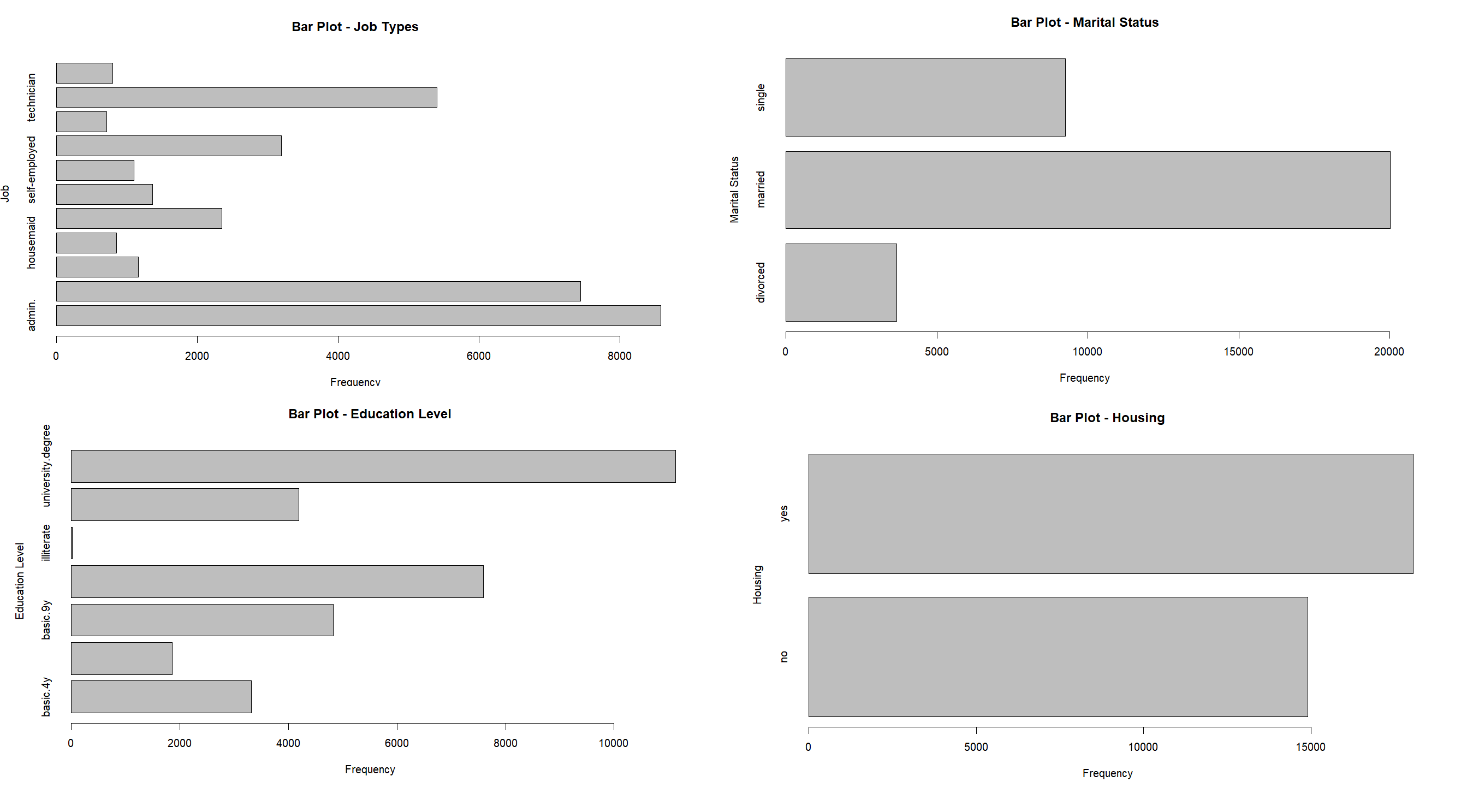
The histograms and boxplots are used to depict different numerical variables. Both Age and Duration are rightly skewed as shown below.

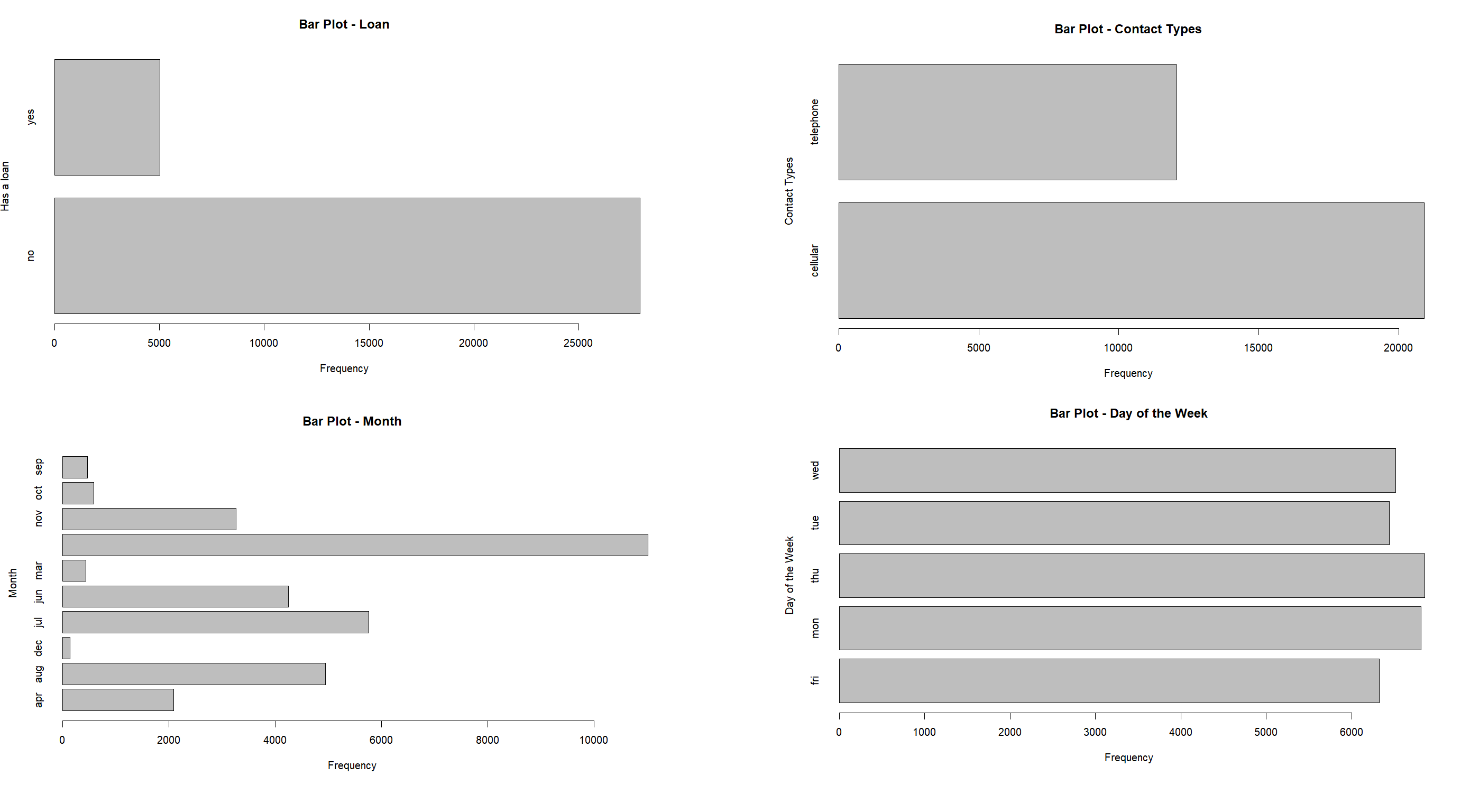


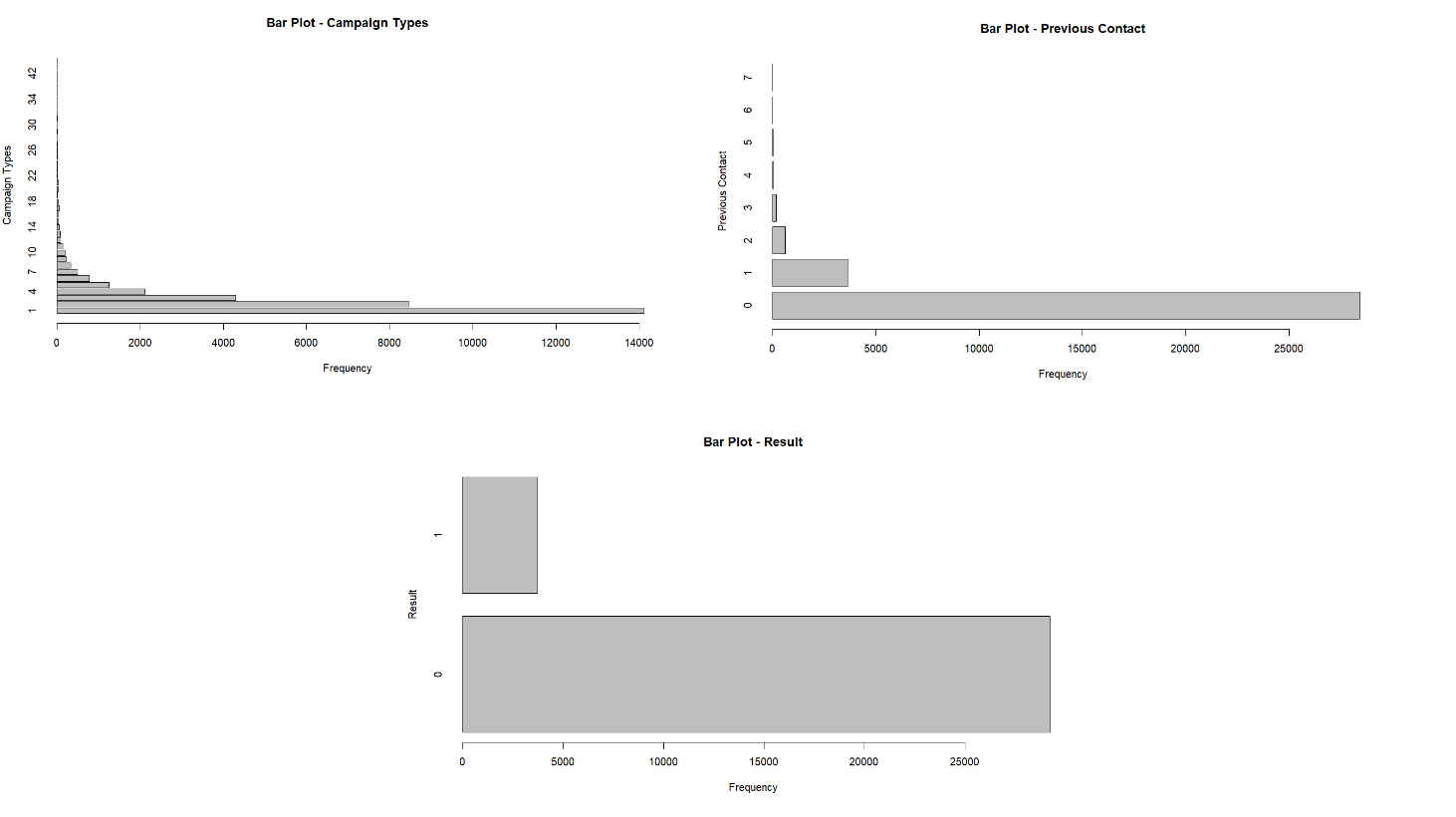
*Figure 1 Histogram*

# Demographic Analysis – Bar Plots

Bar plots are used to depict different categorical variables.







*Figure 2 Bar Plots*

# Modelling Techniques Used

Approximately 12.6% of the bank's customers have subscribed for Term Insurance. Therefore, the basic model may project a 12~13% subscription rate. We will be using Client subscription (Y/N) as a dependent variable and run the 5 techniques to classify. We will be taking the best model which has the highest accuracy.



SVM

Decision

Tree

Modelling

Techniques

KNN

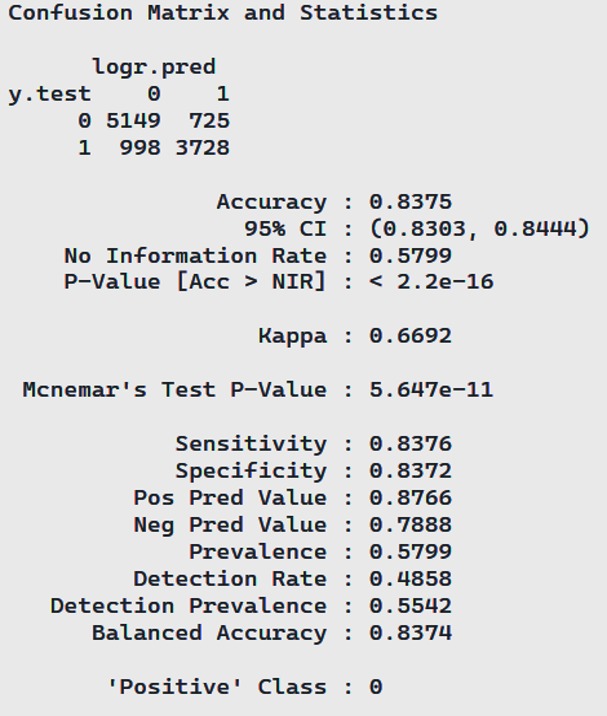
Random

Forest

Logistic Regression

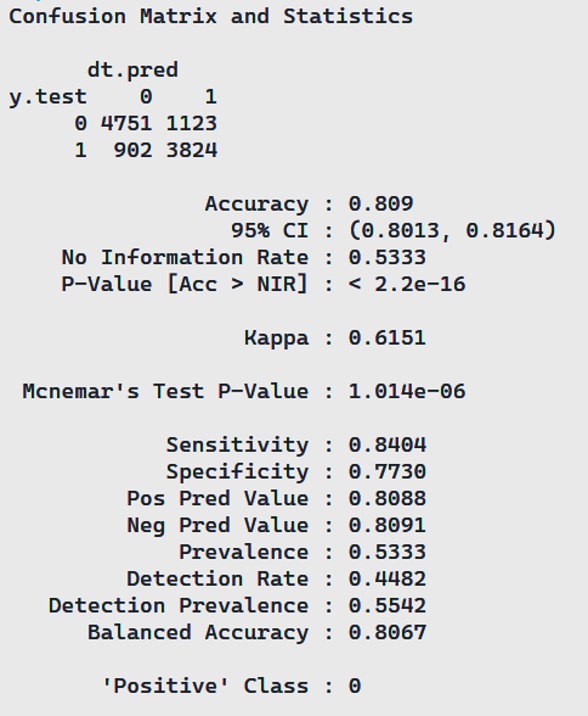
*Figure 3 Modelling Techniques*

# Logistic Regression



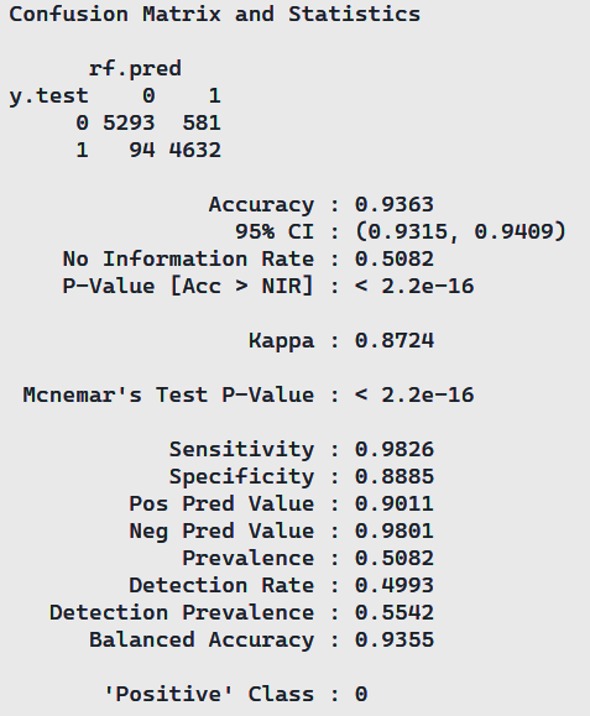
*Figure 4 Confusion Matrix and Statistics - Logistic Regression*

# Decision Tree



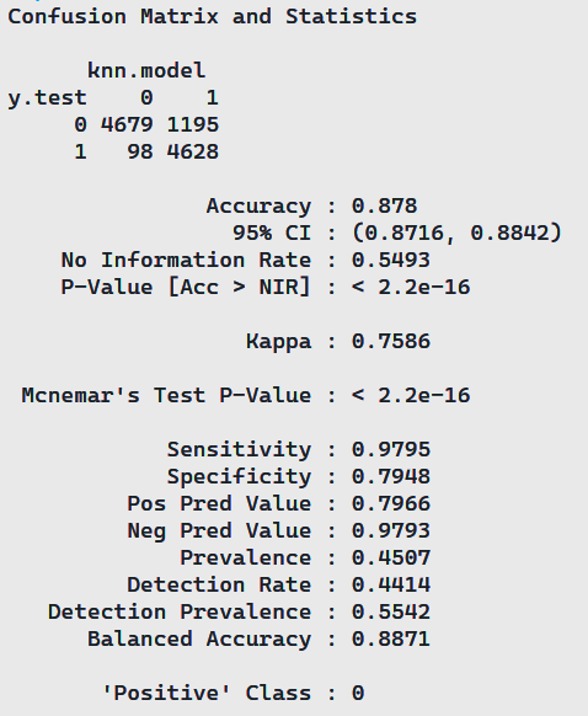
*Figure 5 Confusion Matrix and Statistics - Decision Tree*

# Random Forest



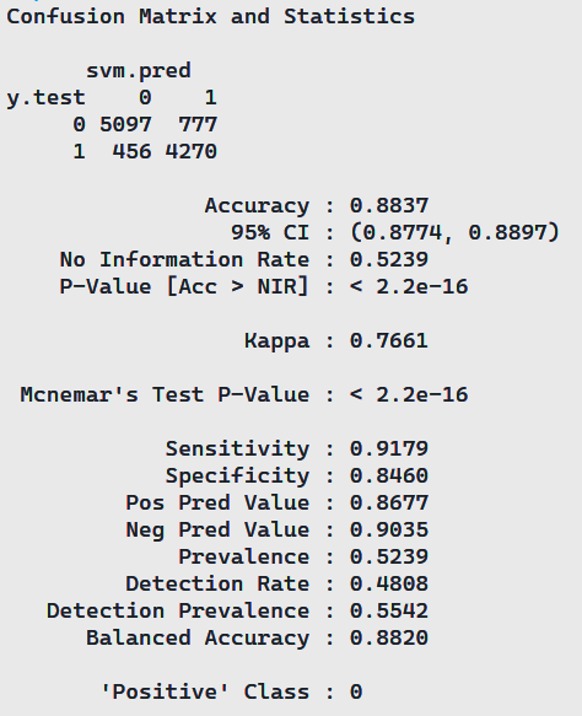
*Figure 6 Confusion Matrix and Statistics - Random Forest*

# K Nearest Neighbor



*Figure 7 Confusion Matrix and Statistics - KNN*

# Support Vector Machine

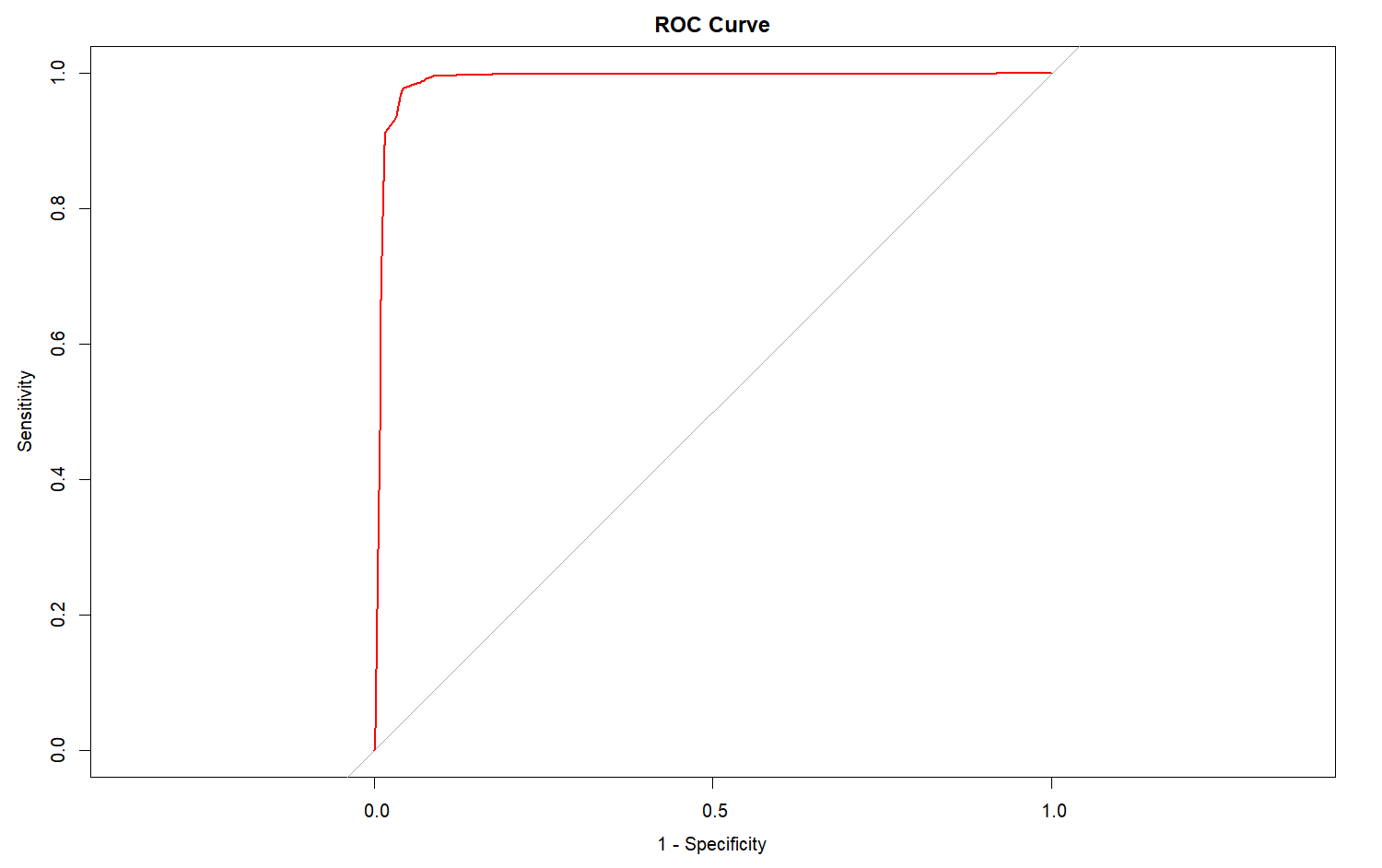


*Figure 8 Confusion Matrix and Statistics - SVM*

# Best Technique

Since the accuracy of the Random Forest is highest. Hence, we will select that model for the future predictions.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Logistic Regression** | **Decision Tree** | **Random Forest** | **KNN** | **SVM** |
| Accuracy | 0.8375 | 0.809 | 0.9363 | 0.878 | 0.8837 |
| Sensitivity | 0.8376 | 0.8404 | 0.9826 | 0.9795 | 0.9179 |
| Specificity | 0.8372 | 0.773 | 0.8885 | 0.7948 | 0.846 |
| 95% CI | (0.8303, 0.8444) | (0.8013, 0.8164) | (0.9315, 0.9409) | (0.8716, 0.8842) | (0.8774, 0.8897) |



*Figure 9 ROC Curve for Random Forest*

# Conclusion

In conclusion, this project addresses the imperative need for the Portuguese Bank to reverse its revenue decline by strategically targeting customers for long-term deposits. Faced with a modest 12.6% term subscription rate, our comprehensive analysis utilizing five machine learning models reveals the potential for accurate prediction. Among these models, the Random Forest approach emerges as the most promising, boasting a remarkable 93.63% accuracy and a 95% confidence interval between 93.15% and 94.09%. By leveraging such predictive insights, the bank can efficiently channel its marketing efforts towards the identified customer segment, fostering increased long-term deposit subscriptions and revitalizing its financial landscape. It suggests that the model can effectively identify customers who are likely to subscribe to a term deposit, which can be valuable for Portuguese bank in several ways:

**Targeted marketing:** The Random Forest model can be used to identify high-potential customers and target them with personalized marketing campaigns for term deposits. This can lead to increased subscription rates and improved returns on marketing investments.

**Customer segmentation:** This model can help banks segment their customer base based on their likelihood of subscribing to term deposits. This allows banks to tailor their product offerings and communication strategies to each segment, leading to a more relevant and engaging customer experience.

**Resource allocation:** Portuguese bank can even use the model to predict demand for term deposits and allocate resources accordingly. This can help optimize staffing, branch operations, and marketing efforts, leading to greater efficiency and cost savings.

**Reduced churn:** By identifying customers who are less likely to subscribe to term deposits, banks can proactively address their concerns and offer them alternative products or services. This can help reduce customer churn and strengthen customer relationships.

Overall, a 93% accuracy with a random forest model is a strong indication that the model can be a valuable tool for predicting bank term deposit subscriptions.

# Bibliography

* + - <https://www.kaggle.com/>
    - <https://www.ibisworld.com/>
    - <https://www.statista.com/>
    - <https://www.investopedia.com/>
    - <https://www.kdnuggets.com/>
    - <https://www.optimove.com/>