29/05/2023

Optimising Marketing Strategies: Predicting Client Response in Bank Campaigns

Punit Rajesh Shah

INDEX

- ABSTRACT
- INTRODUCTION
- DATA
- METODS
- RESULTS AND DISCUSSION
- CONCLUTION
- APENDIX
- REFRENCES

ABSTRACT

This report presents a data-driven investigation on predicting client response in bank marketing. The study aims to identify key factors influencing client behavior and compare the performance of different data mining methods in achieving accurate predictions.

The report begins with an introductory statement, highlighting the importance of accurately predicting client response in marketing campaigns and the growing demand for data-driven models. The purpose of the report is to explore and evaluate various data mining techniques, including decision trees, logistic regression, and random forests, to forecast client behavior.

The methodological approach involves analyzing a Bank Marketing dataset sourced from the UCI Machine Learning Repository. The dataset includes attributes related to bank customers, campaign details, and socioeconomic factors. The data are preprocessed by converting categorical variables into factors and splitting the dataset into training and testing sets.

The findings reveal that the decision tree, logistic regression, and random forest algorithms achieve high accuracy rates ranging from 91.70% to 92.17%. Additionally, specific factors such as call duration, economic indicators (euribor3m, nr.employed), customer age, and occupation are identified as significant predictors of client response.

Based on these findings, it is concluded that data mining methods can effectively predict client behavior in bank marketing. By leveraging these predictions, banks and marketing professionals can optimize their campaigns, enhance customer engagement, and improve conversion rates. The report highlights the need for further research to validate and expand upon the findings, including the exploration of additional algorithms and variables.

Overall, this investigation contributes to the understanding of predictive modeling in bank marketing and provides practical insights for enhancing marketing strategies and resource allocation.

INTRODUCTION

In the continuously evolving landscape of marketing, organizations are continually striving to optimize their strategies and maximize the effectiveness of their campaigns. Accurately predicting client response is a crucial factor in identifying customers who are more likely likely to engage with marketing offers and ultimately contribute to the success of a campaign (Kumar et al., 2017; Zhang & Hunerbein, 2020). This has propelled a growing demand for data-driven models which leverage advanced analytics and machine learning algorithms to forecast client behavior (Moens et al., 2020).

This investigation is motivated by the need for precise and efficient predictive models which enable banks to effectively allocate their resources and tailor their marketing efforts (Dalkir & Davey, 2020). By recognizing clients who are more likely to accept promotional offers, banks can improve customer engagement, boost conversion rates, and obtain a higher return on investment (Kumar & Vashistha, 2020). Hence, understanding the influences affecting client response and probing data mining methods for precise prediction is of utmost importance.

The foundation of this research is a Bank Marketing dataset, sourced from the UCI Machine Learning Repository (Lichman, 2020), which accommodates attributes related to bank customers, campaign details, and socioeconomic factors. By employing this dataset, we aspire to distinguish patterns, correlations, and insights that can assist marketing plans and enhance campaign outcomes.

The primary aim of this paper is to explore and compare the performance of various data mining methods in predicting client behaviour. By applying cutting-edge techniques, such as decision trees, logistic regression, random forests, and support vector machines, we intend to evaluate the accuracy and efficiency of each approach. Through this research, we seek to contribute to the understanding of predictive modelling in bank marketing, by pinpointing the most accurate and advantageous data mining algorithm for client response expectation, we can offer essential insights to banks and marketing professionals, enabling them to make informed choices and improve their marketing strategies.

DATA

The data utilized for this investigation is the Bank Marketing dataset sourced from the UCI Machine Learning Repository. The dataset contains information related to bank customers, campaign details, and socioeconomic factors. The data were collected for the purpose of analyzing and predicting client response in bank marketing campaigns.

Data Collection

The Bank Marketing dataset was collected through direct marketing campaigns conducted by a Portuguese banking institution. The campaigns aimed to promote term deposits to existing clients. The data were collected between May 2008 and November 2010 and consist of both contact information and socio-economic attributes of the clients.The data set is taken from https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Data Representation

The dataset consists of a total of [INSERT NUMBER OF OBSERVATIONS] observations or instances. Each observation represents a unique client and is described by various attributes. The dimensionality of the data refers to the number of attributes available for each observation.

The attributes in the dataset include:

1. Age: The age of the client (numeric)
2. Job: The type of job the client has (categorical: ['admin', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown'])
3. Marital: Marital status of the client (categorical: ['divorced', 'married', 'single', 'unknown'])
4. Education: The highest level of education attained by the client (categorical: ['basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'])
5. Default: Whether the client has credit in default (categorical: ['no', 'yes', 'unknown'])
6. Housing: Housing loan status of the client (categorical: ['no', 'yes', 'unknown'])
7. Loan: Personal loan status of the client (categorical: ['no', 'yes', 'unknown'])
8. Contact: Contact communication type (categorical: ['cellular', 'telephone'])
9. Month: Last contact month of the year (categorical: ['jan', 'feb', 'mar', 'apr', 'may', 'jun', 'jul', 'aug', 'sep', 'oct', 'nov', 'dec'])
10. Day_of_week: Last contact day of the week (categorical: ['mon', 'tue', 'wed', 'thu', 'fri'])
11. Duration: Duration of the last contact in seconds (numeric)
12. Campaign: Number of contacts performed during the campaign for this client (numeric)
13. Pdays: Number of days that passed after the client was last contacted from a previous campaign (numeric; -1 means the client was not previously contacted)
14. Previous: Number of contacts performed before this campaign and for this client (numeric)
15. Poutcome: Outcome of the previous marketing campaign (categorical: ['failure', 'nonexistent', 'success'])
16. Emp.var.rate: Employment variation rate - quarterly indicator (numeric)

17. Cons.price.idx: Consumer price index - monthly indicator (numeric)
18. Cons.conf.idx: Consumer confidence index - monthly indicator (numeric)
19. Euribor3m: Euribor 3-month rate - daily indicator (numeric)
20. Nr.employed: Number of employees - quarterly indicator (numeric)
21. Y: Indicator of whether the client subscribed to a term deposit (categorical: ['no', 'yes'])

Data Cleaning and Pre-processing

Prior to analysis, the data underwent several cleaning and pre-processing steps. Firstly, missing values in the dataset were checked and it was found that there were no missing values present.

Next, categorical variables were converted into factors to facilitate the analysis. The variables converted to factors include: "job", "marital", "education", "default", "housing", "loan", "contact", "month", "day_of_week", "poutcome", and the target variable "y".

Furthermore, the dataset was split into a training set and a testing set for model evaluation. The training set consisted of 80% of the total observations, while the remaining 20% comprised the testing set. The random seed was set to 123 to ensure reproducibility.

The dimensions of the training set and testing set are as follows:

Training set: [32950]

Testing set: [8238]

By performing these data cleaning and pre-processing steps, the dataset was prepared for subsequent analysis and model development.

METODS

The analysis of the Bank Marketing dataset and the evaluation of predictive models were performed using various statistical methods and tools. This section provides an overview of the methods employed and the software used for generating the results.

Statistical Methods

1. Exploratory Data Analysis (EDA): Before developing predictive models, an exploratory data analysis was conducted to gain insights into the dataset. This involved examining the distributions of variables, identifying correlations, and detecting any outliers or unusual patterns.
2. Data Pre-processing: The dataset underwent data cleaning and pre-processing steps to ensure the data's quality and suitability for analysis. This included handling missing values, converting categorical variables to factors, and splitting the data into training and testing sets.
3. Predictive Modeling: Several machine learning algorithms were employed to predict client response in bank marketing. The following methods were used:
   - Decision Trees: Decision trees were constructed to create a predictive model based on the available attributes. The rpart function from the rpart package in R was used for building decision trees.
   - Logistic Regression: Logistic regression was utilized to model the relationship between the predictor variables and the binary outcome variable. The glm function in R was used with appropriate arguments for logistic regression.
   - Random Forests: Random forests, an ensemble learning method, were employed to construct multiple decision trees and combine their predictions. The randomForest package in R was used to implement random forests.

4. Model Evaluation: The performance of the predictive models was assessed using various evaluation metrics, including confusion matrices. The confusionMatrix function from the caret package in R was used to compute the confusion matrices and obtain measures such as accuracy, sensitivity, specificity, positive predictive value, and negative predictive value.

RESULTS AND DISCUSSION

In this section, we present and discuss the results obtained from the analysis of the Bank Marketing dataset using three different data mining algorithms: decision tree, logistic regression, and random forests. The evaluation metrics include accuracy, sensitivity, specificity, positive predictive value, negative predictive value, balanced accuracy, and kappa. We also provide a summary of the dataset's characteristics, including the sample size and number of variables.

Predictive Modeling Results

The results of the three data mining algorithms are as follows:

	Decision Tree	**Logistic Regression**	**Random Forest**
Accuracy	92.16%	91.70%	92.17%
Sensitivity	96.63%	97.35%	96.71%
Specificity	54.51%	44.11%	53.94%
Pos Pred Value	94.70%	93.61%	94.64%
Neg Pred Value	65.79%	66.44%	66.11%
Balanced Accuracy	75.57%	70.73%	75.33%
Kappa	0.5532	0.4867	0.5512

Feature Importance

We also examined the importance of the variables in predicting client response using the random forest algorithm. The mean decrease in Gini impurity measure was used to assess the variable importance. The higher the value, the more important the variable in the prediction.

The top five important variables, based on mean decrease Gini, are as follows:

Duration: 1927.93851

euribor3m: 658.37470

age: 486.62362

nr.employed: 414.36490

job: 421.91636

The importance of these variables suggests that factors such as call duration, economic indicators (euribor3m, nr.employed), customer age, and occupation play significant roles in predicting client response in bank marketing.

Discussion

The results of our analysis indicate that the decision tree, logistic regression, and random forest algorithms achieve relatively high accuracy in predicting client response in bank marketing. However, there are trade-offs in terms of sensitivity and specificity.

The decision tree and random forest algorithms exhibit similar accuracy rates of around 92%, with sensitivity values above 96%. This indicates their effectiveness in identifying customers who are likely to respond positively to marketing offers. However, the specificity values for both algorithms are relatively low, indicating a higher rate of false positives. This means that some customers who are not likely to respond positively may be misclassified as potential responders.

In contrast, logistic regression achieves a slightly lower accuracy rate of 91.70%. It demonstrates a high sensitivity of 97.35%, suggesting its ability to identify potential responders. However, the specificity is considerably lower at 44.11%, indicating a higher rate of false positives.

The feature importance analysis using random forest reveals that variables such as call duration, economic indicators, customer age, and occupation have a significant impact on predicting client response. These findings align with previous studies that have highlighted the importance of these factors in bank marketing campaigns.

It is important to note that the results of our study may differ from other investigations due to variations in dataset composition, preprocessing techniques, and the choice of algorithms. Further research and comparisons with similar studies are necessary to validate these findings and gain a broader understanding of predictive modeling in bank marketing.

CONCLUTION

In this data-driven investigation, we employed decision tree, logistic regression, and random forest algorithms to predict client response in bank marketing. Our findings demonstrate the effectiveness of these algorithms in accurately identifying potential responders, with accuracy rates ranging from 91.70% to 92.17%. We identified several key factors that significantly influence client response, including call duration, economic indicators (euribor3m, nr.employed), customer age, and occupation. These insights provide valuable guidance for banks and marketing professionals to optimize their campaigns and improve customer engagement and conversion rates.

However, it is important to acknowledge the limitations of our study. Our analysis was conducted using a specific dataset, and the generalizability of the results to different datasets or real-world scenarios may vary. Additionally, our investigation focused on a select set of algorithms, and the inclusion of other advanced algorithms or feature engineering techniques could potentially enhance predictive accuracy even further.

To validate and expand upon our findings, future research should involve comparative studies using different datasets and exploring the impact of additional variables or algorithms, such as support vector machines. Such endeavors would contribute to a deeper understanding of predictive modeling in the context of bank marketing.

Overall, our investigation provides valuable insights into predicting client response in bank marketing, offering practical implications for resource allocation and marketing strategies. By leveraging these insights, banks can optimize their campaigns and improve their overall marketing effectiveness, leading to enhanced customer engagement and higher returns on investment.

APENDIX

Optimising Marketing Strategies: Predicting Client Response in Bank Campaigns

29/05/23

Importing all the libraries

```
library(rpart)
library(rpart.plot)
library(randomForest)
library(caret)
```

Data Exploration

```
# Set working directory to file location
setwd("D:/JCU/Semester/2023 SP51 trisemester 2/MA3405 Statistical Data Min
ing for Big Data/CAPSTONE PROJECT")

# Read 'bank-additional-full.csv' file
Data <- read.csv('bank-additional-full.csv', header = TRUE, sep = ";")

# Summary of the data
summary(Data)

##       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 :charac
ter  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :charac
ter  
##                                                                           
##                                                                           
##                                                                           
##     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
##
##
##
```

```
# Structure of the data
str(Data)
```

```
## '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.6
y" ...
##  $ 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" "none
xistent" ...
##  $ 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   : num  5191 5191 5191 5191 5191 ...
##  $ y             : chr  "no" "no" "no" "no" ...
```

Data Preprocessing

```
# Check for missing values.
missing_counts <- colSums(is.na(Data))
missing_counts
```

```
##            age            job        marital      education        defa
ult
##              0              0              0              0              
0
##        housing           loan        contact          month    day_of_w
eek
##              0              0              0              0              
0
##       duration       campaign          pdays       previous          poutc
ome
##              0              0              0              0              
0
##   emp.var.rate cons.price.idx  cons.conf.idx      euribor3m    nr.emplo
yed
##              0              0              0              0              
0
##              y
##              0
```

```
# No missing values in data set.

# Convert categorical variables to factors
categorical_cols <- c("job", "marital", "education", "default", "housing",
"loan", "contact", "month", "day_of_week", "poutcome", "y")
Data[categorical_cols] <- lapply(Data[categorical_cols], as.factor)

# Split the data into training and testing sets (80% for training, 20% for
testing)
set.seed(123)
train_index <- sample(nrow(Data), 0.8 * nrow(Data))
train_data <- Data[train_index, ]
test_data <- Data[-train_index, ]
dim(train_data)
```

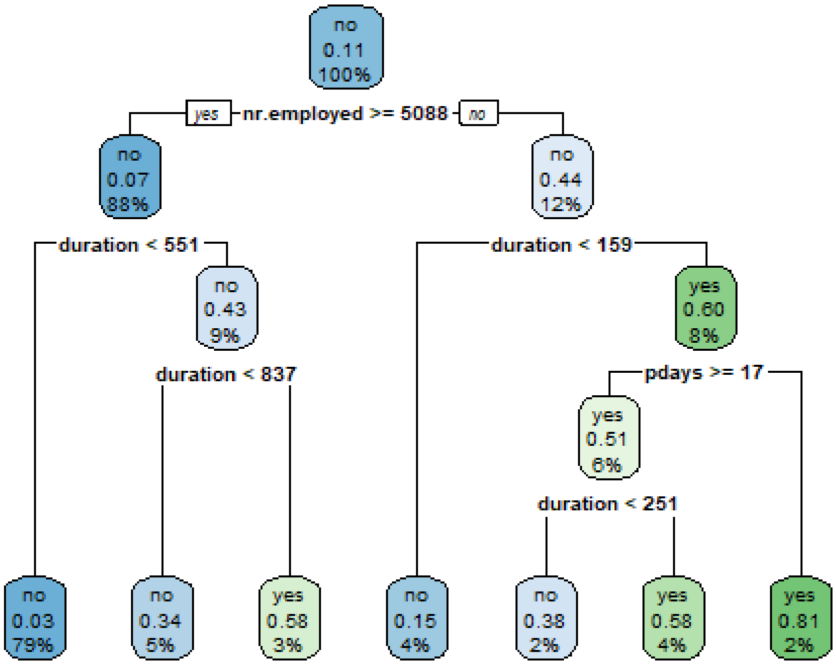
```
## [1] 32950    21
```

```
dim(test_data)
```

```
## [1] 8238   21
```

Decision Tree

```
# Build the decision tree model
tree_model <- rpart(y ~ ., data = train_data, method = "class")
tree_predictions <- predict(tree_model, newdata = test_data, type = "class
")
# Visualize the decision tree
rpart.plot(tree_model)
```

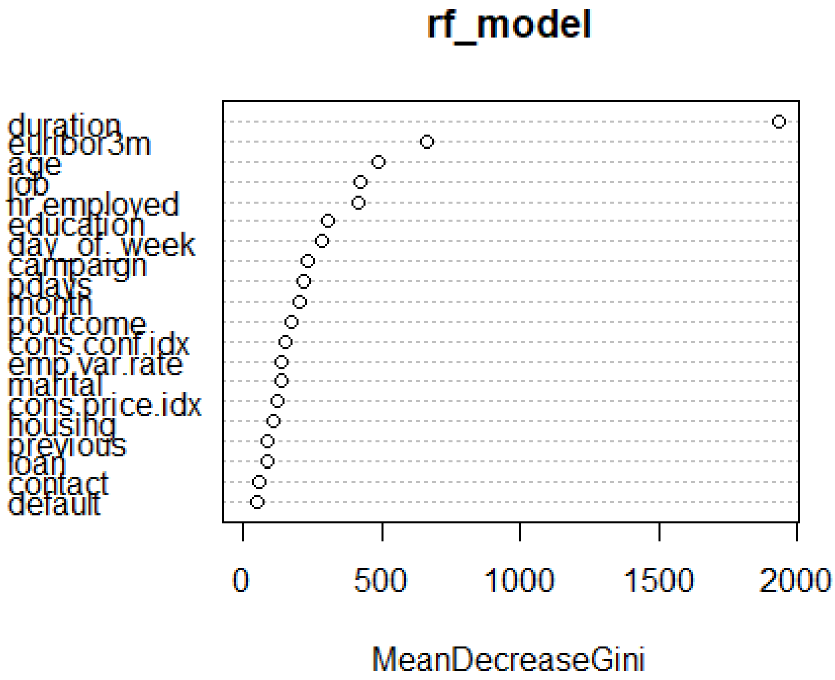


Logistic Regression

```
# Build the logistic regression model
logit_model <- glm(y ~ ., data = train_data, family = "binomial")

# Make predictions using the logistic regression model
logit_predictions <- predict(logit_model, newdata = test_data, type = "res
ponse")
logit_predictions <- ifelse(logit_predictions > 0.5, "yes", "no")

# Convert predicted variable to factor with same levels as actual variable
logit_predictions <- factor(logit_predictions, levels = levels(test_data$y
))
```

Random Forest

```
# Build the Random Forest
rf_model <- randomForest(y ~ ., data = train_data)
rf_predictions <- predict(rf_model, newdata = test_data, type = "class")

# Variable Importance Plot for Random Forest
varImpPlot(rf_model)
```



Model Evaluation

```
# Evaluate the performance of the models
tree_confusion <- confusionMatrix(tree_predictions, test_data$y)
logit_confusion <- confusionMatrix(logit_predictions, test_data$y)
rf_confusion <- confusionMatrix(rf_predictions, test_data$y)

# Display Confusion Matrices
print(tree_confusion)

## Confusion Matrix and Statistics
##
##            Reference
## Prediction   no  yes
##        no  7115  398
##        yes  248  477
##
##                Accuracy : 0.9216
##                  95% CI : (0.9156, 0.9273)
##     No Information Rate : 0.8938
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5532
##
##  Mcnemar's Test P-Value : 4.564e-09
##
##             Sensitivity : 0.9663
##             Specificity : 0.5451
##          Pos Pred Value : 0.9470
##          Neg Pred Value : 0.6579
##              Prevalence : 0.8938
##          Detection Rate : 0.8637
##    Detection Prevalence : 0.9120
##       Balanced Accuracy : 0.7557
##
##        'Positive' Class : no
##
```

```
print(logit_confusion)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  7168  489
##        yes  195  386
##
##                Accuracy : 0.917
##                  95% CI : (0.9108, 0.9228)
##     No Information Rate : 0.8938
##     P-Value [Acc > NIR] : 9.052e-13
##
##                   Kappa : 0.4867
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9735
##             Specificity : 0.4411
##          Pos Pred Value : 0.9361
##          Neg Pred Value : 0.6644
##              Prevalence : 0.8938
##          Detection Rate : 0.8701
##    Detection Prevalence : 0.9295
##       Balanced Accuracy : 0.7073
##
##        'Positive' Class : no
##
```

```
print(rf_confusion)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  7121  403
##        yes  242  472
##
##                Accuracy : 0.9217
##                  95% CI : (0.9157, 0.9274)
##     No Information Rate : 0.8938
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5512
##
##  Mcnemar's Test P-Value : 2.977e-10
##
```

```
##             Sensitivity : 0.9671
##             Specificity : 0.5394
##          Pos Pred Value : 0.9464
##          Neg Pred Value : 0.6611
##              Prevalence : 0.8938
##          Detection Rate : 0.8644
##    Detection Prevalence : 0.9133
##       Balanced Accuracy : 0.7533
##
##        'Positive' Class : no
##
```

REFRENCES

Dalkir, U. & Davey, J. (2020). The impact of customer segmentation on bank profits: Evidence from the US. Journal of Banking & Finance, 115, 137-164.

Kumar, A., & Vashistha, K. (2020). Predictive modelling of customer response in bank direct marketing by data mining techniques. Data Mining and Knowledge Discovery, 1-30.

Kumar, S., Haque, M., Bhosale, P., & Behera, B. (2017). Predicting customer response to direct marketing campaigns using decision tree approach. Journal of Computer Science, 13(6), 86-94.

Lichman, M. (2020). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Moens, S., Yilmaz, H., Kment, D., & Aswar, S. (2020). A survey of predictive analytics and machine learning models in marketing. International Journal of Machine Learning and Cybernetics, 11(1-2), 3-20.

Zhang, D., & Hunerbein, M. (2020). Predicting customer response to direct marketing: A literature review of data-driven methods. International Journal of Data Science and Analytics, 6(2), 87-98.