National College of Ireland

Analysis of predicting the success of the banking telemarketing campaigns by using machine learning techniques

MSc Research Project
Data Analytics

Laksana Vongchalerm
X19101538

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……. Laksana Vongchalerm……

Student ID: ……X19101538……

Programme: ……MSc. In Data Analytics…… **Year:** …2021-2022

Module: ……MSc. Research Project……

Supervisor: ……Vladimir Milosavljevic……

Submission Due Date: ……31st January 2022……

Project Title: ……Analysis of predicting the success of the banking telemarketing campaigns by using machine learning techniques…

Word Count: ……6092…….…….. **Page Count**……23……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Laksana Vongchalerm……

Date: ……31st January 2022……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analysis of predicting the success of the banking telemarketing campaigns by using machine learning techniques

Laksana Vongchalerm
X19101538

Abstract

This research project is related to predicting the success of the banking telemarketing campaigns by using machine learning techniques. Machine learning techniques are playing an important role in improving marketing campaigns and providing timely information to management for effective decision-making. The banking institutions are using machine learning techniques for increasing their market share by reaching the target customers. There is a need to optimize the uses of available resources and attract potential customers for increasing the overall revenue of the banks. Some Portuguese are using advanced technologies and machine learning techniques for the marketing of products and services for increasing the market share. Customer behavior plays an important role in the decision-making process and Portuguese banks are using machine learning techniques for analyzing the behavior of customers. The banking system of the Portuguese is advanced and linked with the national multibanco system. This system is facilitating the financial institutions in running their business operations and providing opportunities to customers for opening their accounts. More than 150 banks are connected and interrelated through multibanco and banks are using telemarketing for their promotions in the target markets. The Portuguese financial institutions and banks are using different telemarketing methods including television, radio, telephone, and cells for promoting the services and reaching potential customers. Banks are using financial telemarketing services for increasing their sales and gaining a sustainable competitive advantage over competitors. These services are playing important roles in reaching the target audience and sparking sales by getting the word out through telemarketing calls. The researcher of this project used secondary data from May 2008 to November 2010 about the direct marketing dataset based on 41,188 records. The researcher collected the data through phone call interviews that are also considered half meeting. The researcher used logistic regression and decision trees for analyzing the performance of banking telemarketing campaigns.

1 Introduction

1.1 Background Information

Machine learning is a mathematical expression that represents the data in the context of the problem often a business problem. Machine learning techniques are playing an important role in improving marketing campaigns and allowing management to quickly make decisions based on big data. The marketing department uses machine learning techniques for creating hypotheses, testing, evaluate and analyzing them (Dzyabura, 2018). The use of the machine learning technique is a time taking work and sometimes produces incorrect results as information changes every second. This research project is related to predicting the success of the marketing campaigns of a Portuguese banking institution by using machine learning techniques. The banking institutions are formulating and implementing different marketing campaigns for increasing their market share and gaining sustainable competitive advantage in the competitive banking markets (Bose, 2001). The banking institutions need to optimize the available resources and marketing campaigns should focus on efforts for improving the customers' experience. In marketing campaigns, efforts should be made for attracting potential customers so they can accept the proposed offers.

The Portuguese banking system is also using different marketing techniques for effective marketing of their products and services to valuable customers. Some banks are also using different marketing campaigns based on machine learning for optimizing the available resources and reaching potential customers. There has been a lot of marketing campaign run by banks, but most of these have lost their effect on most of the general public (Hagen, 2020). This is one of the main reasons behind businesses and banks changing their approach, most of the leading marketing heads have been observed investing more on directed campaigns i.e. being rigorous in selecting the customers to contact and how to approach them. One of the most used approaches is communication over the telephone and this type of marketing is known as telemarketing. There have been several studies on this topic and we have discussed some of them below. It was quite interesting to see that almost all of the researchers used the dataset provided by a Portuguese bank (Arasu, 2020).

The main objective of these studies has been to enhance this process of the direct campaign, and the approach suggested for this is machine learning i.e. using data to get over the financial losses. Machine learning can help the banks to determine the customers they would want to

target to acquire more funds. There have been multiple machine learning methodologies used for this process by many researchers. Almost all of these studies used real-world data collected from a Portuguese marketing campaign related to bank deposit subscriptions. For data collection banks used their contact centre and organized directed marketing campaigns (Mitić, 2019).

In this research project, the researcher will use the dataset to analysis of predicting the success of the banking telemarketing campaigns by using machine learning techniques. The dataset will be related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. The dataset included 41188 records with 20 attributes from May 2008 to November 2010.

The researcher of this project will put efforts into identifying the main factors that affect customers' decision to subscribe to the bank's term deposit. In data analysis, the researcher will also identify the best classification model (Asare-Frempong, 2017). Customer behavior plays an important role in the decision-making process and this research project will analyze the customer behavior of making a term deposit in banks to enhance campaign efficiency. It is essential to identify the factors in the banking system that have major impacts on the decision-making process of customers and compel them for availing the particular services of the banks. Various machine learning techniques including the logistic regression model and decision tree model can be used for measuring the impacts of marketing campaigns (Moro, 2014).

1.2 Research Question

Following are the research questions of this research project

- What are the primary variables influencing consumers' decision to subscribe to the bank's term deposit?
- How the classification model, logistic regression and a decision tree does are affect classifier performance for a bank's direct marketing dataset?

1.3 Research Objectives

The research objectives of this research project are provided below

- The primary goal of this analysis is to determine the major factors that have the greatest effect on consumers' decisions to subscribe to term deposits.

- The second objective is to compare and select the best of various classification techniques to produce a classification model that can predict whether or not the customer will purchase with a term deposit.

1.4 Significance of the Study

This is a significant research project related to a Portuguese banking institution. This research project will analyze the predicting the success of the marketing campaigns by using machine learning techniques. This research project will cover the importance of machine learning techniques and the uses of such techniques in marketing campaigns (Dzyabura, 2018). The relationship between different factors and the success of a marketing campaign will also be analyzed in detail. This research project will effectively contribute to existing studies related to machine learning techniques and will also provide valuable information to key stakeholders of the banking industry (Hagen, 2020).

2 Literature Review

2.1 Banking Sector of Portuguese

Portugal has an advanced banking system as all commercial banks are linked with the national multibanco system. This mutlibanco system provides many facilities to customers including a range of actions from their accounts. The new customers can easily open new accounts and both residents and non-residents are allowed to open their bank accounts in Portugal. The Portuguese banking system is likewise utilizing diverse marketing methods for successful advertising of their products and services to significant clients. A few banks are additionally utilizing distinctive advertising efforts dependent on AI for upgrading the accessible assets and arriving at likely clients (Figueiredo, 2016). There has been a ton of advertising effort run by banks, however, the majority of these have lost their impact on a large portion of the overall population. This is one of the fundamental explanations for organizations and banks changing their methodology, the majority of the main showcasing heads have been noticed contributing more on coordinated missions for example being thorough in choosing the clients to contact and how to move toward them. Quite possibly the most utilized approach is correspondence via phone and this sort of promoting is known as selling. There have been a few examinations on this subject and we have talked about some of them beneath.

The interbank network of Portugal is advanced through multibanco. More than 150 banks are connected and interrelated through multibanco and these banks include national retail banks, cooperative banks, and international banks in Portugal. The central bank of Portugal acts as the regulatory authority for other banks. There are more than 11,000 ATMs in Portugal and banks are providing credit facilities to customers through loan and credit card facilities (Figueiredo, 2016).

2.2 The Success of Telemarketing in Banking Institution

Telemarketing is an important tool for promotion in a banking institution. Telemarketing has gained sufficient importance and banks are effectively using this tool for promoting the services. In the banking system, telemarketing is known as a sophisticated communication network for promoting products and services. The banking institutions are using various moods in telemarketing including television, radio, telephone, and cells for promoting the services and reaching potential customers. Telemarketing is one of the most successful promotion tools and nowadays banks are using SMS and emails for promotions of services and increasing their sales (Jiang, 2018).

The use of the machine learning technique is a period taking work and now and then delivers inaccurate outcomes as data changes each second. This exploration project is identified with foreseeing the accomplishment of the advertising efforts of a Portuguese financial organization by utilizing AI strategies. The financial foundations are detailing and executing distinctive advertising lobbies for expanding their portion of the overall industry and acquiring feasible upper hand in the serious financial business sectors. The financial foundations need to enhance the accessible assets and promoting efforts should zero in on endeavours for working on the clients' experience. In showcasing efforts, endeavours ought to be made for drawing in expected clients so they can acknowledge the proposed offers (Moro, 2014).

Banks are partnering with financial telemarketing services for increasing their sales and gaining a sustainable competitive advantage over competitors. The marketers are volatile in the banking industry and changes in interest rates impact the sales volume of the banks. The banks are using telemarketing services for providing information to customers regarding new services and offering (Borugadda, 2021). They are also using telemarketing services for informing the existing and potential customers regarding the decrease in interest rates on loans and leasing facilities for attracting the customers. The telemarketing services can help the financial

institutions and companies in reaching the target audience and sparking sales by getting the word out through telemarketing calls (Jiang, 2018).

2.3 Factors impact on Customers' decisions to subscribe to the Bank's Term Deposit

Various factors impact customers' decisions to subscribe to the bank's term deposit. The Major factors include the interest rates, maintenance system, risk management, services charges, and duration of the term deposits. The term deposits ensure the money of customers to earn interest at a specified rate for a specified period (Lu, 2016). The banks provide differentiated services to customers for improving their competitiveness in the competitive banking sectors. Many factors including the number of employees, duration, and month impact the customers' decisions to subscribe to the banks' term deposits. The banks through telemarketing strategies push the customers to subscribe for long-term deposits and improve the overall efficiency of the banks (Zhuang, 2018).

The principal objective of these examinations has been to upgrade this course of the immediate mission, and the methodology recommended for this is AI for example utilizing information to move past the monetary misfortunes. AI can assist the keeps money with deciding the clients they would need to focus on to secure more assets. There have been various AI approaches utilized for this interaction by numerous scientists. Practically these investigations utilized certifiable information gathered from a Portuguese showcasing effort identified with bank store memberships (Umam, 2021). Banks are cooperating with monetary selling administrations for expanding their deals and acquiring a reasonable upper hand over contenders. The banking institutions are utilizing different dispositions in selling including TV, radio, phone, and cells for advancing the administrations and coming to expected clients. The banks offer separate types of assistance to clients for working on the intensity in the serious financial areas. Telemarketing has acquired an adequate significance and banks are viably utilizing this apparatus for advancing the administrations. In the financial framework, selling is known as a refined correspondence network for advancing items and administrations (Hung, 2019).

2.4 Customers’ response to Bank Direct Telemarketing Campaign

The customers provide a good response to banks on telemarketing campaigns regarding subscribing to bank deposits. The banks in Portugal are using telemarking services effectively for reaching potential customers and increasing their overall sales. The customers in Portugal

provide good responses to banks and listen to the calls regarding new offerings. The banks are also taking the feedback from customers regarding the use of services or buy of products through telemarketing services. The growth of bank has a significant impact on the economy and banks in Portugal has gained sufficient international markets majorly in the Euro area (Borugadda, 2021). The banks are providing low maintenance cost and effective risk management to customers and such offers are advertised through telemarketing in Portugal. The banks also encourage customers to control their spending and investment in term deposits for improving the overall earning of customers (Umam, 2021).

Consumer behavior assumes a significant part in the dynamic interaction and this examination task will break down the client conduct of setting aside a term instalment in banks to improve crusade effectiveness. It is fundamental to recognize the elements in the financial framework that significantly affect the dynamic course of clients and urge them for benefiting the specific administrations of the banks. Different AI strategies including the calculated relapse model and choice tree model can be utilized for estimating the effects of showcasing efforts (Moro, 2014).

2.5 Data mining in Telemarketing

Data mining in telemarketing is performed for identifying and analyzing the potential customers for marketing through telemarketing. It is a process of recognizing the likely customers for promoting the products and services. The data mining technique is widely used in banking institutions and is an effective tool for direct marketing. Data mining is not easy in telemarketing as several problems arise. The predictive accuracy of finding the potential customers and analyzing them in detail is a difficult task (Moro, 2018). The banking sector and insurance companies prefer to use direct marketing through telemarketing for selling the products and services to potential customers. In today's world, the banking sector is competitive so it is important to use data mining in telemarketing for the identification of potential customers and target them for increasing the overall sales. In direct marketing through telemarketing, the needs of the customers are identified and products are offered to them for satisfying their needs (Hung, 2019).

Different problems have occurred in data mining in telemarketing including imbalanced class distribution. The learning algorithms do not work on such a data set in which imbalanced class distribution exists. The predictive accuracy cannot be found in the data set during data mining. The imbalanced data does not provide suitable evaluation criteria in finding the potential customers and offering promotions to them (Moro, 2014). The size of variables can vary in

data set and there is a need to use efficient learning algorithms for the data set. The banks in Portugal are utilizing telemarking administrations viably for coming to a possible clients and expanding their general deals. The clients in Portugal give great reactions to banks and listen to the calls concerning new contributions. The banks are giving low support costs and successful danger the executives to clients and such offers are publicized through selling in Portugal (Moro, 2014).

3. Methodology

The research is based on analysing the factors that could impact consumer's decision making process related to subscribing the banks term deposit. The analysis is based on different techniques of research and Cross Industry Standard for Data Mining or CRISP-DM methodology is selected to implement in this research project (Khan & Khan, 2013). The figure shown in the following is CRISP-DM;

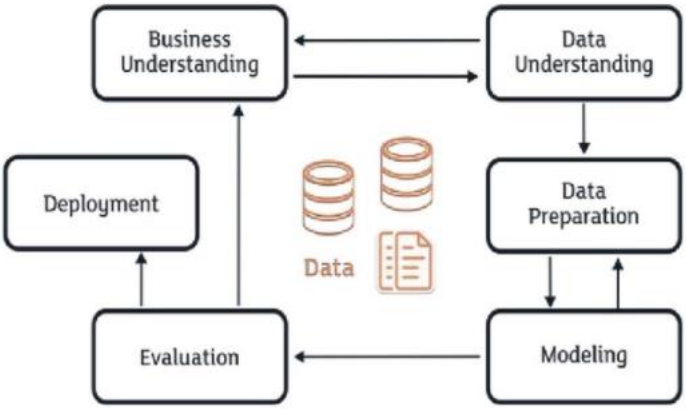


Figure 1: Cross Industry Standard for Data Mining (CRISP-DM)

3.1 Business Understanding

3.1.1 Data Understanding

The basic objective of research is to analyse the expected consequences of telemarketing banking campaigns that are based on the use of machine learning techniques. The collected data is based on the variables that interpret about the campaign as it would be helpful to save costs and time of customers that would not consider the decision to subscribe the term deposit in the banks (Moro, et al., 2015). Different characteristics of customers would be analysed for the implementation of research model including job, gender, age and marital status etc.

3.2 Data Preparation

3.2.1 Description of Data

The research is conducted by the University of California, Irvine (UCI) that explain about the secondary data related to the direct marketing techniques of bank in the context of this research. The data is based on the operations of direct marketing and it is collected with the help of audio calls by Portuguese financial institutions. While conducting the data, it was required to maintain several interactions with the same customer to check the probability of subscribing the term deposit. The research conducted in May 2008 to November 2010 provide information about direct marketing dataset is based on 41,188 records that has 21 features including placements variable (Moro, et al., 2011).

The table 1 is based on the features and interpretation list that explain 20 variable including continuous variable and categorical variables equally. The results show two categories of answer the one is yes and the other is no. There are four categories of independent variable including customer's demographical information, contact information, social indicators and economic indicators output variable that is used to predict about the customer's opinion related to term deposit and it is represented by the group. The answer yes is represented with value (1) that shows the subscription of customers to a term deposit while on the other side no is represented with value (0) and it shows that the customer did not make any subscription to term deposit (Karim & Rahman., 2014).

Table 1: Attributes Information

No.	Attributes Name	Type	Description
Bank client data			
1	Age	Numeric	Age of client
2	Job	Categorical	Type of client's job
3	Marital	Categorical	Client's marital status
4	Education	Categorical	Education level
5	Defult	Categorical	Has credit in default?
6	Housing	Categorical	Has housing loan?
7	Loan	Categorical	Has personal loan?
Related with the last contact of the current campaign			
8	Contact	Categorical	Contact communication type
9	Month	Categorical	Last contact month of year
10	Day_of_week	Categorical	Last contact day of the week
11	Duration	Numeric	Time spent with client on the call
Orther Attributes			
12	Campaign	Numeric	Number of contacts performed during this campaign and for this client
13	Pdays	Numeric	Number of days that passed by after the client was last contacted from a previous campaign
14	Previous	Numeric	Number of contacts performed before this campaign and for this client
15	Poutcome	Categorical	Outcome of the previous campaign
Social and economic context			
16	emp.var.rate	Numeric	Employment variation rate - quarterly indicator
17	cons.price.idx	Numeric	Consumer price index - monthly indicator
18	cons.conf.idx	Numeric	Consumer confidence index - monthly indicator
19	euribor3m	Numeric	Euribor 3 month rate - daily indicator
20	nr.employed	Numeric	Number of employees - quarterly indicator (thousands)
Output variable (desired target)			
21	Y	Binary	Does the client has subscribed to a term deposit?

The above mentioned two tables 2 & 3, are based on the summary of results by analysing the characteristics of the customers. The characteristics provide reasons about the results that shows the number of subscribers for the term deposit is 4,640 and it is 11.3% of total number of customers. The results of above table represent that the information collected is not appropriate.

Furthermore, there are several observations that are explained below:

- 54.6% customer of total number is married who subscribe the term deposit.
- Mostly the customers are younger and 75% of total customers are less than 47 years.
- 90.5% of customers are subscribed to customers that had no record of credit default.
- 43.7% customers are subscribed who had never get housing loans and the mostly subscribers also not had personal loans.
- The contact was made in May and August and 59.2% customers subscribed the term deposit when they interested in this time (Elsalamony, 2014).

Table 2: Summary of attributes

Dependent variable					
Variable	Value		Count		
Y	No		36548		
	Yes		4640		
Categorical independent variables					
Variable	Value	Count	Variable	Value	Count
Job	admin.	10422	Loan	No	33950
	blue-collar	9254		Yes	6248
	technician	6743		Unknown	990
	services	3969	Contact	Cellular	26144
	management	2924		Telephone	15044
	Unknown	330	Month	May	13769
	(other)	7546		Jul	7174
Day_of_week	Mon	8514		Aug	6178
	Tue	8090		Jun	5318
	Wed	8134		Nov	4101
	Thu	8623		Apr	2632
	Fri	7827		(other)	2016
Education	University.degree	12168	Martial	Divorced	4612
	High.school	9515		Married	24928
	Basic.9y	6045		Single	11568
	Professional.course	5243		Unknown	80
	Basic.4y	4176	Poutcome	Failure	4252
	Unknown	1731		Nonexistent	35563
	(other)	2328		Success	1373
Defult	No	32588	Housing	No	18622
	Yes	3		Yes	21576
	Unknown	8597		Unknown	990

Numerical independent variables						
Variable	Min	1 st Qu.	Median	Mean	3 rd Qu	Max
Age	17	32	38	40.02	47	98
Campaign	1	1	2	2.568	3	56
Pdays	0	999	999	962.5	999	999
Duration	0	102	180	258.3	319	4918
Previous	0	0	0	0.173	0	7
Emp.var.rate	-3.4	-1.8	1.1	0.08189	1.4	1.4
Cons.price.idx	92.2	93.08	93.75	93.58	93.99	94.77
Cons.conf.idx	-50.8	-42.7	-41.8	-40.5	-36.4	-26.9
Euribor3m	0.634	1.344	4.857	3.621	4.961	5.045
Nr.employed	4964	5099	5191	5167	5228	5228

3.2.2 Preliminary Data Analysis

The research represents that the information used in research is imbalanced (Parlar & Acaravci., 2017). The inappropriate information create problem and it also provides the data about the classification of different target class intervals that is not even. As a result, the information provided difficulties in the part of modelling. Due to this imbalanced information, the data is analysed and all the categories were changed into dummy variables with the collaboration of different number of levels used in the regression model. Moreover, there is also two major problems that are explained in the following:

1) Missing data

The data was collected with phone calls interview that is also considered half meeting. The customers do not provide their personal information due to some privacy issues therefore they just interview on call for some random perspectives. Therefore, the data collected from the customers was incomplete and it could not perform better role in the analysis. The incomplete information could impact on the changing behaviour of detailed information and required to get complete information. In the given table the missing value is denoted as “unknown” and other features in the dataset also based on unknown values therefore different methods used to tackle the issues that are appropriate facts (Hosseini., 2021).

	variable	nr_unknown
1	default	8597
2	education	1731
3	housing	990
4	loan	990
5	job	330
6	marital	80
7	age	0
8	contact	0
9	month	0
10	day_of_week	0
11	duration	0
12	campaign	0
13	pdays	0
14	previous	0
15	poutcome	0
16	emp.var.rate	0
17	cons.price.idx	0
18	cons.conf.idx	0
19	euribor3m	0
20	nr.employed	0
21	y	0
22	y_binary	0

Figure 2: Missing values (encoded as “unknown”)

At the start of analysis, imputation is used to analyse the unknown values. There is 251 for Education, 1,731 rows for unknown values that got the subscription of term deposit and it is 5% of total subscribers. Table 3 is representing the 2x2 conditional table that is used to make a comparison of expected outcomes as well as the group of unknown/known related to education. For the analysis of relationship, Chi-square test is implemented and it shows the result about p-value that is less than the level of significance 0.01 explained in the Table 4. The case is related to examine the response of the target (Borugadda, et al., 2021). The missing values could not be avoided therefore these are assigned by concluding the missing values from the independent variables.

Table 3: Table of Missing Value (Education)

Education	No	Yes	Propotion of Yes
Known	35068	4389	0.125
Unknown	1480	251	0.170

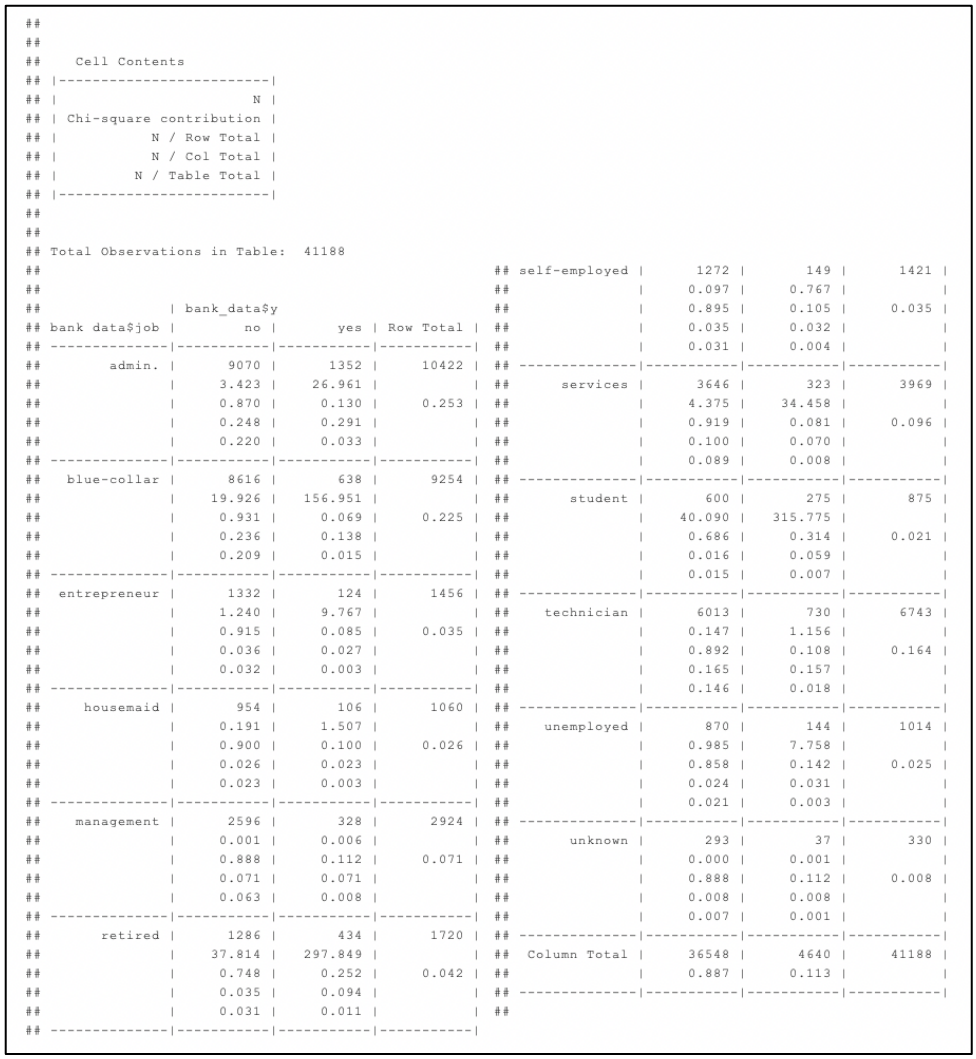
Table 4: Chi-Squared Test for Association (Education*Y)

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	193.106[a]	7	.000
Likelihood Ratio	196.853	7	.000
N of Valid Cases	41188		

a. 1 cells (6.3%) have expected count less than 5. The minimum expected count is 2.03.

The second method is based on considering the two variables that has no specific relationship with the unknown and known values group, the target response is based on the attribute married and job. For instance, the figure 3 explain about the job that is highly responding among the students and it shows the percentage of 31.4% and 25.2% retired people. Consequently, the missing values are removed as these values are missing randomly and it is ignored as it has no impact on the model (Nga & Yien., 2013).



```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  41188
##
##                | bank_data$y
## bank_data$job |        no |       yes | Row Total |
## --------------|-----------|-----------|-----------|
##        admin. |      9070 |      1352 |     10422 |
##               |     3.423 |    26.961 |           |
##               |     0.870 |     0.130 |     0.253 |
##               |     0.248 |     0.291 |           |
##               |     0.220 |     0.033 |           |
## --------------|-----------|-----------|-----------|
##   blue-collar |      8616 |       638 |      9254 |
##               |    19.926 |   156.951 |           |
##               |     0.931 |     0.069 |     0.225 |
##               |     0.236 |     0.138 |           |
##               |     0.209 |     0.015 |           |
## --------------|-----------|-----------|-----------|
##  entrepreneur |      1332 |       124 |      1456 |
##               |     1.240 |     9.767 |           |
##               |     0.915 |     0.085 |     0.035 |
##               |     0.036 |     0.027 |           |
##               |     0.032 |     0.003 |           |
## --------------|-----------|-----------|-----------|
##     housemaid |       954 |       106 |      1060 |
##               |     0.191 |     1.507 |           |
##               |     0.900 |     0.100 |     0.026 |
##               |     0.026 |     0.023 |           |
##               |     0.023 |     0.003 |           |
## --------------|-----------|-----------|-----------|
##    management |      2596 |       328 |      2924 |
##               |     0.001 |     0.006 |           |
##               |     0.888 |     0.112 |     0.071 |
##               |     0.071 |     0.071 |           |
##               |     0.063 |     0.008 |           |
## --------------|-----------|-----------|-----------|
##       retired |      1286 |       434 |      1720 |
##               |    37.814 |   297.849 |           |
##               |     0.748 |     0.252 |     0.042 |
##               |     0.035 |     0.094 |           |
##               |     0.031 |     0.011 |           |
## --------------|-----------|-----------|-----------|
## self-employed |      1272 |       149 |      1421 |
##               |     0.097 |     0.767 |           |
##               |     0.895 |     0.105 |     0.035 |
##               |     0.035 |     0.032 |           |
##               |     0.031 |     0.004 |           |
## --------------|-----------|-----------|-----------|
##      services |      3646 |       323 |      3969 |
##               |     4.375 |    34.458 |           |
##               |     0.919 |     0.081 |     0.096 |
##               |     0.100 |     0.070 |           |
##               |     0.089 |     0.008 |           |
## --------------|-----------|-----------|-----------|
##       student |       600 |       275 |       875 |
##               |    40.090 |   315.775 |           |
##               |     0.686 |     0.314 |     0.021 |
##               |     0.016 |     0.059 |           |
##               |     0.015 |     0.007 |           |
## --------------|-----------|-----------|-----------|
##    technician |      6013 |       730 |      6743 |
##               |     0.147 |     1.156 |           |
##               |     0.892 |     0.108 |     0.164 |
##               |     0.165 |     0.157 |           |
##               |     0.146 |     0.018 |           |
## --------------|-----------|-----------|-----------|
##    unemployed |       870 |       144 |      1014 |
##               |     0.985 |     7.758 |           |
##               |     0.858 |     0.142 |     0.025 |
##               |     0.024 |     0.031 |           |
##               |     0.021 |     0.003 |           |
## --------------|-----------|-----------|-----------|
##       unknown |       293 |        37 |       330 |
##               |     0.000 |     0.001 |           |
##               |     0.888 |     0.112 |     0.008 |
##               |     0.008 |     0.008 |           |
##               |     0.007 |     0.001 |           |
## --------------|-----------|-----------|-----------|
##  Column Total |     36548 |      4640 |     41188 |
##               |     0.887 |     0.113 |           |
## --------------|-----------|-----------|-----------|
##
```

Figure 3: Crosstab of Job and Target variable

The third method was based on case-specific method that is represented in the table 5. The total number based on the response “yes” in the default attribute was insufficient as only three customers are responding. On the other side, the total number related to unknown status is relatively high as 8,597 customer respond on unknown values. The results of Chi-square test represent that there is association between the dependent variable and unknown value status. It could be estimated that unknown for default variable could be entered as unknown and it is possible that customers would be unable or unwilling to expose their information to the banking officer. Consequently, the unknown value in default is categorized as a new value used in the model (Karimi, n.d.).

Table 5: Cross-tabulation Table between Default and Subscription

			default			
			no	unknown	yes	Total
y	no	Count	28391	8154	3	36548
		% within y	77.7%	22.3%	0.0%	100.0%
		% within default	87.1%	94.8%	100.0%	88.7%
		% of Total	68.9%	19.8%	0.0%	88.7%
	yes	Count	4197	443	0	4640
		% within y	90.5%	9.5%	0.0%	100.0%
		% within default	12.9%	5.2%	0.0%	11.3%
		% of Total	10.2%	1.1%	0.0%	11.3%
Total		Count	32588	8597	3	41188
		% within y	79.1%	20.9%	0.0%	100.0%
		% within default	100.0%	100.0%	100.0%	100.0%
		% of Total	79.1%	20.9%	0.0%	100.0%

Table 6: Chi-Squared Test for Association (Default*Y)

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	406.578[a]	2	.000
Likelihood Ratio	475.664	2	.000
N of Valid Cases	41188		

a. 2 cells (33.3%) have expected count less than 5. The minimum expected count is .34.

Moreover, the variable in Table 7 is based on “Housing” variable was able to differentiate among the 11.6% customers who availed the housing loan and 10.9% customers did not avail this opportunity. The unknown group is based on the 10.8% of the total customers. Furthermore, the loan variable is based on 990 rows along unknown class that is not more different from the 11.3% customers who didn’t subscribe the term deposit and 10.9% who did. The relationship was analysed with the help of Chi-Square Test that is helpful to test the possibility to disregard these aspects (Borugadda, et al., 2021).

Table 7: Cross-tabulation Table between Housing and Subscription

			housing			
			no	unknown	yes	Total
y	no	Count	16596	883	19069	36548
		% within y	45.4%	2.4%	52.2%	100.0%
		% within housing	89.1%	89.2%	88.4%	88.7%
		% of Total	40.3%	2.1%	46.3%	88.7%
	yes	Count	2026	107	2507	4640
		% within y	43.7%	2.3%	54.0%	100.0%
		% within housing	10.9%	10.8%	11.6%	11.3%
		% of Total	4.9%	0.3%	6.1%	11.3%
Total		Count	18622	990	21576	41188
		% within y	45.2%	2.4%	52.4%	100.0%
		% within housing	100.0%	100.0%	100.0%	100.0%
		% of Total	45.2%	2.4%	52.4%	100.0%

Table 8: Chi-Squared Test for Association (Housing*Y)

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	5.684[a]	2	.058
Likelihood Ratio	5.691	2	.058
N of Valid Cases	41188		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 111.53.

Table 9: Cross-tabulation Table between Loan and Subscription

			loan			
			no	unknown	yes	Total
y	no	Count	30100	883	5565	36548
		% within y	82.4%	2.4%	15.2%	100.0%
		% within loan	88.7%	89.2%	89.1%	88.7%
		% of Total	73.1%	2.1%	13.5%	88.7%
	yes	Count	3850	107	683	4640
		% within y	83.0%	2.3%	14.7%	100.0%
		% within loan	11.3%	10.8%	10.9%	11.3%
		% of Total	9.3%	0.3%	1.7%	11.3%
Total		Count	33950	990	6248	41188
		% within y	82.4%	2.4%	15.2%	100.0%
		% within loan	100.0%	100.0%	100.0%	100.0%
		% of Total	82.4%	2.4%	15.2%	100.0%

Table 10: Chi-Squared Test for Association (Loan*Y)

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1.094[a]	2	.579
Likelihood Ratio	1.102	2	.576
N of Valid Cases	41188		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 111.53.

2) Imbalanced data

The incomplete information also changed the results that provide information about the influence of factors on the performance. This problem may occur again and again in the real world example as different individuals may have different behaviours and the level of changing behaviour may vary from individual to individual. The data presented in Table 2 stated that the group of customers who subscribed the term deposit are highly based on the information that is incomplete and it is only 11.3% of total customers who are categorized as yes. It also represents that 88.7% of people or categorized in the group who said no. Consequently, the incomplete data could create problems in designing the model and the incomplete information could not accurately predict the minority group. With the data mining technique, the class balancing technique is considered important for dealing with the incomplete information. There are different methods that could be used to tackle the problems of incomplete data (Hosseini., 2021).

There are two methods that are used to handle the problems related to incomplete and imbalanced data set such as algorithm method, sampling method and future selection method. The most commonly used method is data sampling method as it is helpful to deal with the group imbalance. Data sampling approach has two methods the one is oversampling and under sampling. Oversampling technique is implemented to increase the minority class instances until the number when it matches to the larger group while the under sampling method is helpful to overcome the majority class instances (Moro, et al., 2011). Consequently, the under sampling technique is selected as oversampling technique would increase the amount of data and if it is implemented to the existing majority class of 36,548. (Figure 4)

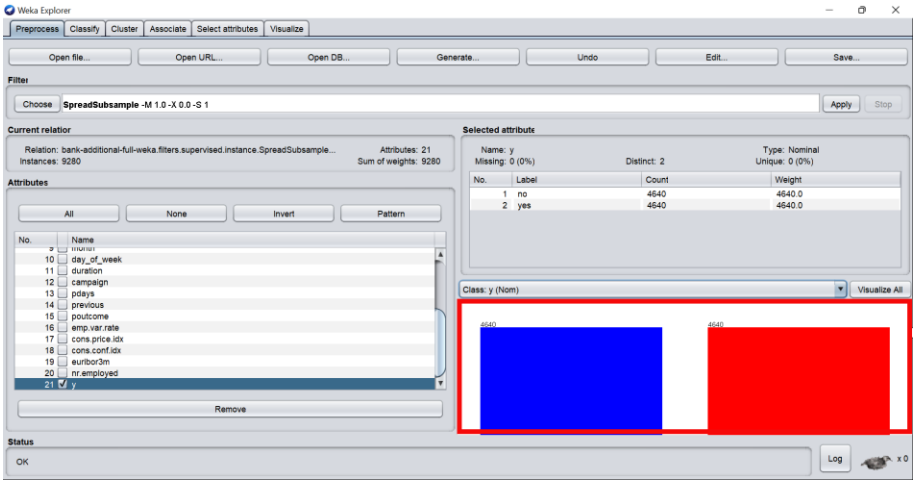


Figure 4: Visualization of the class distribution after under sampling

3.3 Modelling, Evaluation, and Implementation

The research arise questions related to the subscription of consumers to a term deposit. It denoted the distribution of problems related to binary system. Several techniques such as data mining algorithms were originated to anticipate the term deposits subscription of information with the help of classification model, the first one is decision tree and the second is logistics regression. The models were implemented to the data set for marketing of bank that is helpful to compare the productivity of bank (Elsalamony, 2014). The efficiency and accuracy of models were impacted by inappropriate information and the basic data set was randomly divided into 10 equal subsamples with the objective of predicting the model's performance and using the 10-fold cross validation technique. The nine sub-sample selected out of 10 subsamples that are used in the training data and the single subsample is consumed on the validation of data related to model. The cross validation method is implemented again and again for ten time at each ten subsample is implemented in the validation of data.

The evaluation of performance of model is emphasizing on the capacity of the classification that represents the correct answers. The implementation of classification accuracy is used to analyse the performance of model that may be difficult as it is based on the measurement of percentage that is distributed irrespective of the class from which it belongs. At the time of dealing with the data set that is based on inappropriate information the class level area under Area Under ROC Curve is selected for the performance measurement as it is helpful to evaluate the performance. Moreover, accuracy and reliability of values are determined with the help of metrics that is considered classifier output quality (Khan & Khan, 2013).

The most common form of classifier in determining is logistic regression. It is considered that the variable must be in the binary form and the data set that is collected from customers would be used in the regression model. The logistic regression model could be based on the assumptions and it could be observed independently. Consequently, independent variable is known to the reasonable assumptions and the target value is also distributed in the different variables such as yes or no. The model does not violate the assumptions to make sure that there is no significant influence on individual with these factors that are analysed in the analysis. If all the assumptions are satisfied, independent variables are selected to estimate the model selection according to the ratio analysis method (Karim & Rahman., 2014). The model final model is considered suitable statistically as it is not based on multicollinearity problem. The performance is analysed with the help of test of data set and it shows AUC 0.934 with the accuracy value of 0.87 and reliability 0.869 explained in the following figure;

Table 9: Comparison of the models

Classification model	AUC	Precision	Recall
Logistic Regression	0.934	0.87	0.869
Decision Tree	0.866	0.844	0.844

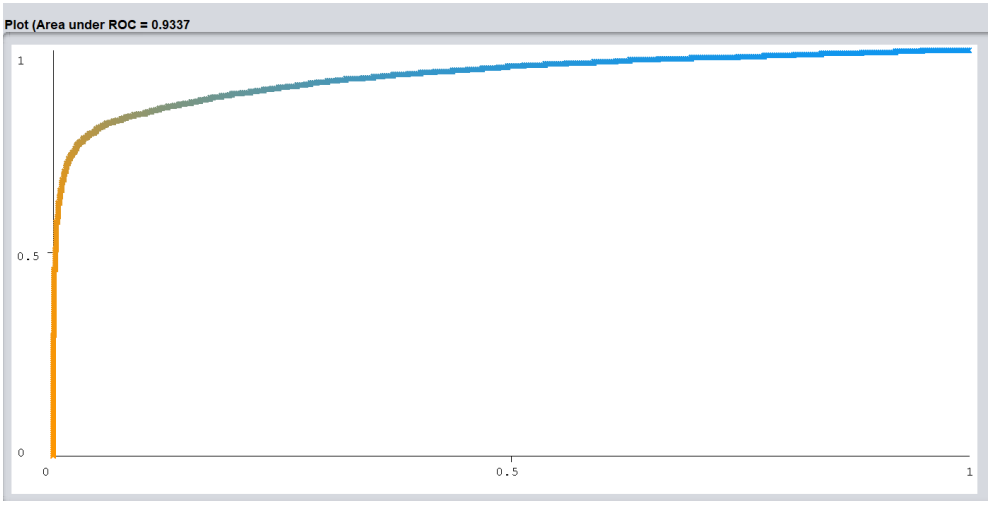


Figure 5: Logistic Regression ROC plots

Furthermore, the other method is decision tree method that is effective tool in the anticipation and distribution of variables. The data on this entry is distributed according to the rule of induction that is based on the hierarchy. The internal node of tree represents the attribute and the branches are based on the results of testing values. The leaf nodes denote the class labels

and decision attributes. The result shows that misclassification rate is low and the final decision tree result is considered best model for generating the results. The final model is based on the size of the tree of 2277 and 1792 leaves that explained the ROC curve plot with the implementation of AUC to evaluate the performance software model; the AUC value is 0.866.

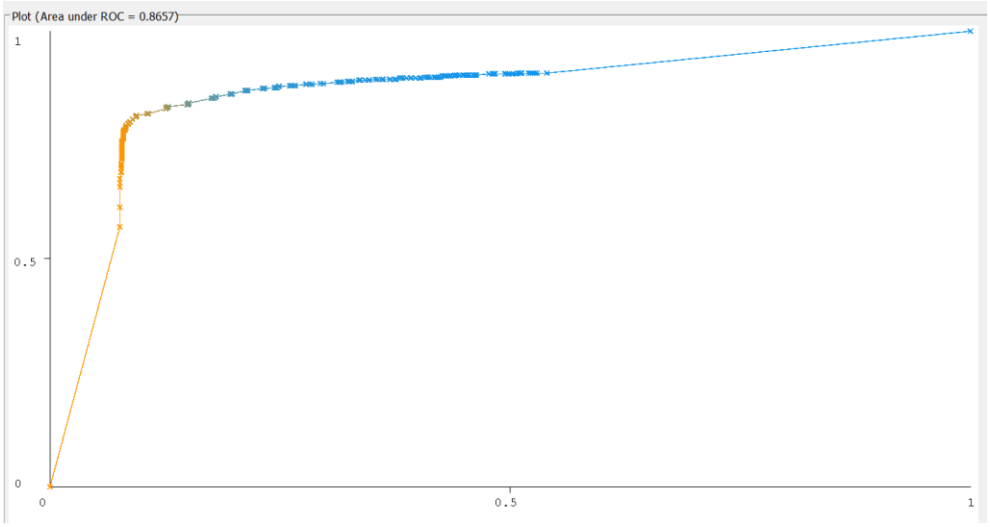


Figure 6: Decision Tree ROC plots

Figure 7 shows that the significant plot with fitting according to the logistic regression model. it could be identified that the important variable in the model is used to infleunce the performance of the bank on the basis of marketing campaign related to the subscribed customers.

```
Ranked attributes:
 0.259765   11 duration
 0.211648   19 euribor3m
 0.207691   18 cons.conf.idx
 0.207648   17 cons.price.idx
 0.189394   20 nr.employed
 0.172298   16 emp.var.rate
 0.079496    9 month
 0.078857   13 pdays
 0.07836    15 poutcome
 0.058651   14 previous
 0.043791    8 contact
 0.036074    1 age
 0.034423    2 job
 0.023629    5 default
 0.011868   12 campaign
 0.008673    4 education
 0.004123    3 marital
 0.000686   10 day_of_week
 0.00036     6 housing
 0.000183    7 loan
```

Figure 7: Variable Important Factors Rank

4 Conclusion and Future Work

The objective of the research is to analyse the marketing campaign with the help of machine learning methods. The research used Logistic Regression and Decision Tree. The performance is analysed with the help of test of data set and it shows AUC 0.934 with the accuracy value of 0.870 and reliability 0.869. The time is considered most significant variable that could be based on the length that is used to interact with the customers. it shows the link between the euribor3m is the Euribor 3 months rate and cons.conf.idx is the consumer confidence index of the customers. the research represents that model could be used to measure the probability related to the decision making of customers to subscribe the term deposit.

References

Arasu, B. S. B. a. T. N., 2020. A machine learning-based approach to enhancing social media marketing.. *Computers & Electrical Engineering,,* Volume 86, p. 106723.

Asare-Frempong, J. a. J. M., 2017. September. Predicting customer response to bank direct telemarketing campaign.. *In 2017 International Conference on Engineering Technology and Technopreneurship,* pp. 1-4.

Borugadda, P., Nandru, P. & Madhavaiah., C., 2021. "Predicting the Success of Bank Telemarketing for Selling Long-term Deposits: An Application of Machine Learning Algorithms.". *St. Theresa Journal of Humanities and Social Sciences,* 7(1), pp. 91-108.

Borugadda, P. N. P. a. M. C., 2021. Predicting the Success of Bank Telemarketing for Selling Long-term Deposits: An Application of Machine Learning Algorithms.. *Theresa Journal of Humanities and Social Sciences,* 7(1), pp. 91-108.

Bose, I. a. M. R., 2001. Business data mining—a machine learning perspective.. *Information & management,,* 39(3), pp. 211-225.

Dzyabura, D. a. Y., 2018. Machine learning and marketing. In Handbook of Marketing Analytics.. *Edward Elgar Publishing.*

Elsalamony, H. A., 2014. Bank direct marketing analysis of data mining techniques.". *International Journal of Computer Applications,* 85(7), pp. 12-22.

Figueiredo, E. P. L. M. S. a. M., 2016. Human resource management impact on knowledge management: Evidence from the Portuguese banking sector.. *Journal of Service Theory and Practice..*

Hagen, L. U. K. Y. N. B. B. C. A. D. D. E. J. G. A. L. L. S. K. a. W. Y., 2020. How can machine learning aid behavioral marketing research?.. *Marketing Letters,* 31(4), pp. 361-370.

Hosseini., S., 2021. "A decision support system based on machined learned Bayesian network for predicting successful direct sales marketing.". *Journal of Management Analytics,* 8(2), pp. 295-315.

Hung, P. H. T. a. T. T., 2019. February. Term deposit subscription prediction using spark MLlib and ML packages.. *In Proceedings of the 2019 5th International Conference on E-Business and Applications*, pp. 88-93.

Jiang, Y., 2018. Using logistic regression model to predict the success of bank telemarketing.. *International Journal on Data Science and Technology,* 4(1), pp. 35-41.

Karimi, S., n.d. A purchase decision-making process model of online consumers and its influential factor a cross sector analysis.. *The University of Manchester (United Kingdom),2013.*

Karim, M. & Rahman., R. M., 2014. Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing..

Khan, N. & Khan, F., 2013. Fuzzy based decision making for promotional marketing campaigns.". *International Journal of Fuzzy Logic Systems,* 3(1), pp. 64-77.

Lu, X. C. X. C. M. C. P. a. C. S., 2016. Artificial immune network with feature selection for bank term deposit recommendation.. *Journal of Intelligent Information Systems,,* 47(2), pp. 267-285.

Mitić, V., 2019. Benefits of artificial intelligence and machine learning in marketing. In Sinteza 2019-. *International Scientific Conference on Information Technology and Data Related Research* , pp. 472-477.

Moro, S., Cortez, P. & Rita, P., 2015. "Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns.". *Neural Computing and Applications* , 26(1), pp. 131-139.

Moro, S. C. P. a. R., 2014. A data-driven approach to predict the success of bank telemarketing.. *Decision Support Systems,,* Volume 62, pp. 22-31.

Moro, S. C. P. a. R., 2014. A data-driven approach to predict the success of bank telemarketing.. *Decision Support Systems,,* Volume 62, pp. 22-31.

Moro, S. C. P. a. R., 2014. A data-driven approach to predict the success of bank telemarketing.. *Decision Support Systems,,* Volume 62, pp. 22-31.

Moro, S. C. P. a. R., 2018. A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing.. *Expert Systems, ,* Volume 3, p. 35.

Moro, S., Laureano, R. & Cortez., P., 2011. Using data mining for bank direct marketing: An application of the crisp-dm methodology.".

Nga, J. K. & Yien., L. K., 2013. "The influence of personality trait and demographics on financial decision making among Generation Y.". *Young Consumers.*

Parlar, T. & Acaravci., S. K., 2017. Using data mining techniques for detecting the important features of the bank direct marketing data.". *International journal of economics and financial issues,* 7(2), pp. 691-696.

Umam, F. S. A. a. R. A., 2021. Determinants of Mudharabah Term Deposit: A Case of Indonesia Islamic Banks.. *Journal of Economics Research and Social Sciences,* 5(2), pp. 167-180.

Zhuang, Q. Y. Y. a. L. O., 2018. Application of data mining in term deposit marketing.. *In Proceedings of the International MultiConference of Engineers and Computer Scientists* , Volume 2.