12-2022

Identifying Prospective Clients for Long-Term Bank Deposit

Mohamed Al Hammadi

Identifying Prospective Clients for Long-Term Bank Deposit

by

Mohamed Al Hammadi

A Capstone Submitted in Partial Fulfilment of the Requirements

for the Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

December 2022

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Mohamed Al Hammadi**

Graduate Capstone Title**: Identifying Prospective Clients for Long-Term Bank Deposit**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr. Ehsan Warriach **Date:**

Member of committee

Acknowledgments

Words cannot express my gratitude to my mentor Dr. Ehsan Warriach and to my chair of committee Dr. Sanjay Modak for the valuable feedback. I also could not have undertaken or complete this journey without my defense committee, who generously provided expertise and knowledge. Furthermore, this endeavor would not have been possible without the support from the RIT management. I am also thankful to my cohort members, especially my classmates in every course, for their correcting help, late-night support, and moral support. Lastly, I would be remiss in not mentioning my family members, especially my parents, my spouse, and my son. Their belief in me has kept my motivation and spirits very high during this journey.

Abstract

The numerous characteristics of customers are often kept in bank databases, which are utilized to understand who they are. But it has been found in recent years that utilizing different Data Mining and Feature Selection (PCA) methods, customer traits and other factors connected to bank services have a big influence on consumers' decisions. Business analytics is an approach to conducting business that uses transactional data from an organization to acquire knowledge of how business operations can be enhanced by employing data mining methods to determine existing patterns that a firm can incorporate to generate significant data-driven choices to choose significant variables. In this project, we apply data mining techniques for the prediction of long-term bank deposits employing a well-known bank data collection. From PCA it is seen that customers' income level, pout come, p days, and previous (first PC) in general, may seem to have a higher impact on prospective clients, but this is indeed not the real. Also, the Banks' prior campaign and the social elements (Age, Marital Status, Education, Campaign, Duration) of the clients are primarily essential compared to other variables. Again k-means clustering is employed with reduced data by PCA to determine groups of potential customers which gives 87.76% accuracy scores.

Keywords: *Bank data, Deposit, Data Mining, Principal component Analysis, K-Nearest Neighbors Prediction*

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter-1

1.1 Background Information

A term deposit is an investment with a defined period that entails funding an account with a financial institution, for instance, a bank. Bank's term deposit usually lasts for more than 12 months which in some cases may be as long as 10 years and the longer store money, the better the interest rate we'll get (Proença, 2011).

One of two approaches is often used to implement bank marketing initiatives. These are Mass Marketing Campaigns and Targeted Marketing Campaigns. One is through general population-focused mass marketing initiatives, while the other is through specialized ads aimed at a select set of individuals. In comparison to focused marketing initiatives, mass marketing campaigns contains a lower positive response rate to service or product subscription purchases. As a result, large advertising efforts squander a lot of money even when they promote the product. However, the main goal is to sell the good or service, which may be done successfully with direct marketing (DM). For instance, with indirect telemarketing, a salesperson contacts a consumer via phone or mobile to promote a product. But it might be challenging to locate potential clients from a certain demographic. Marketing managers are now using statistical techniques to find potential customers for a product as a result of the rise of data-driven choices in recent years (Vajiramedhin & Suebsing, 2014). So, we apply data mining techniques (PCA and Cluster) for the prediction of long-term deposits using a well-known bank dataset.

1.2 Statement of the Problem

Bank profitability depends on long-term deposits. Targeted marketing tactics that let customers interact with banks directly are the main focus of bank marketing managers right now. This article discusses how a data mining approach was put into practice to anticipate long-term deposits using open-source Portuguese Telemarketing data.

1.3 Project Definition and Goals

Banks are required to sell more long-term bank deposits to enhance their cash reserves due to the contemporary economic situation in several nations. Marketing executives are therefore under pressure to persuade the general public to buy long-term deposits. To increase the positive response rate, marketing managers should build better use of their limited resources by making fewer calls to customers while closing more sales. Because they already contain data from prior campaigns to examine, managers may utilize multivariate data classification techniques to determine clients in near future. Data mining may help businesses acquire a competitive edge, obtain a better understanding of their customers, have more control over everyday operations, boost client acquisition rates, and find new business opportunities. To anticipate long-term deposit consumers, a data mining project based on the Data Mining approach was implemented, as described in this article.

Chapter-2

2.1 Literature Review

A literature review is a crucial, impartial overview of the body of published scientific literature that is pertinent to the area of study being considered. Its goal is to familiarize readers with the most recent theories and studies on a specific subject, and it might also provide justification for further investigation into a topic that was either ignored or underexplored in the past. The state of research on long-term bank deposits centered on the prospective of the client is provided beneath.

According to Amran et al (2017), market demands, shifting client prospective, and factors for the potential expansion of a bank all influence the desire for quality enhancement of banking services. The advancement of a commercial bank's level of customer-based strategy Palacio and Pérez (2018) ought to be, the goal of improving the responsibility scheme in its structural divisions, as this is conveyed in boosting the ethics of good belief and transparency in interactions with customers of banking products (Mallin, 2014).

Long-term bank deposits are among the investment services made available through DM, which enables financial entities like banks and credit cooperatives to concentrate on customers who have a high possibility of enrolling in their services. Businesses engage in outsourcing through marketing initiatives to increase their financial performance and gain an edge over their competitors (Apampa, 2016). Financial industries use DM to market their goods and services to customers who are especially in need of them. An effective DM program could (i) foster connections with new customers, (ii) give existing customers persuasive material they can distribute to prospective customers, and (iii) increase sales. Remote client centralization in a contact center makes it easier to manage programs operational. These businesses interact with

customers via a variety of platforms, with the phone (mobile or fixed) constituting the one that banks use the foremost to market services like long-term bank deposits. Due to the character of distance, DM can be operationally defined through a contact center termed telemarketing (Parlar & Acaravci, 2017).

As per Elsalamony (2014), the technique of locating prospective clients for services and marketing those things to this targeted client mass is known as DM. Recently, businesses have placed significant emphasis on DM strategies that focus on a particular group of clients because of the failure of mass marketing initiatives aimed at the general public. In particular, DM techniques work better in the banking sector since there is larger stress and rivalry there than in other sectors. By figuring out the variables that influence these initiatives, data mining techniques are utilized to make DM campaigns more successful. As a result, these techniques help to focus existing assets and produce a realistic and accurate pool of possible clients.

The paper by Mogaji (2020), examined the conceptual underpinnings of marketing analytics, a vast discipline that emerged from operations research, statistics, computer science, and marketing. One of the difficulties in doing a DM analysis, they claimed, is forecasting consumer behavior. They also covered customer relationship management (CRM), multidimensional scaling, correspondence analysis, and latent Dirichlet allocation as big data visualization techniques for the marketing sector. They discussed the relative value of geographic visualization for retail location research and the overall trade-off between its customary methods and art. They also expanded on discriminant analysis as a method for marketing forecasting. Techniques including ensemble learning, feature reduction, and extraction are used in discriminant analysis. These methods address issues with customer lifetime value, buying behavior, review ratings, brand awareness, customer loyalty, and sales (Cvijović et al, 2017).

CRM is "the business approach and style of functioning implemented to preserve and expand connections with successful clients and to control the expense of performing business with least profitable clients," according to a large retail bank (Giannakis-Bompolis & Boutsouki, 2014). CRM is defined by one of the biggest consulting companies for customer relationship marketing, Gartner Group, as an "IT-enabled business plan, the result of which optimize profit, income, and client satisfaction by arranging all over client portions, fostering client satisfaction actions, and incorporating a client-centric conduct". Additionally, they highlight in their description the role of IT in organizing CRM's efficient operation (Yeboah, 2012).

As per Sharma et al (2015), few research has looked at the use of different machine learning (ML) approaches to forecasting the effectiveness of bank telemarketing for the client prospective of long-term deposits. The goal of data mining techniques is to create a forecasting model that classifies information into a preset group (for instance, "yes" or "no"). Every bank's marketing techniques rely on the analysis of vast amounts of client electronic information (Raghunandan et al, 2018). A human investigator is incapable of extracting useful information from a large volume of data. The widely used data mining methods Naive Bayes (NB), logistic regression (LR), decision trees (DT), and support vector machines (SVM) are used by most of the investigators. The goal is to increase campaign efficiency by figuring out the key factors that influence performance. The degree of successful client participation closely relates to the performance of such programs (Ledhem, 2021). By using data mining techniques, most efforts, particularly those run by banks, may have a higher success rate.

In Hosseini, (2021) research, researchers offer a decision support system (DSS) that, depending on real-world statistics and a strengthened BN technique, forecasts the outcome of a bank telemarketing approach to selling long-term bank deposits. Through the implementation of enhanced NB, the architecture of BN is acquired. Using past data to target prospective

customers who are interested in subscribing to their service or product, such as DSS gives management insights they may use to optimize their marketing campaign.

Through CRM, authors Raju (2018), examined consumer behavior trends. They utilized the same data set used in the present study to apply NB, J48, and multi-layer perceptron NNs. Additionally, they used metrics of sensitivity, accuracy, and specificity to evaluate the effectiveness of their model. Their approach entails comprehending the subject and the data, creating an evaluation model, and then visualizing the results. The J48 classifier surpassed the competition with an accuracy of 89.40, according to the visualization of their findings.

Additionally, the same data set was used by Palaniappan (2017), for additional consumer profiling functions. Data mining techniques were used to propose a paradigm in this research. The database, which contains 41,188 occurrences and 21 characteristics, was collected from the UCI ML repository. On the expanded version of the data set analyzed in the current work, 3 algorithms NB, random forests, and DTs were employed. Before assessing the classifiers, preprocessing and normalization were carried out. For the sake of the trials and assessment processes, Rapid Miner was employed. Using a prior normalization approach, they demonstrated how each classifier's parameters might be adjusted. They also demonstrated how these parameter values affected recall, accuracy, and precision. DTs are the best classifier for consumer profile and behavior prediction providing a higher level of accuracy, according to their findings.

Four ML algorithms the multilayer perceptron NN, random forest, DT (C4.5), as well as was LR used by Asare-Frempong and Jayabalan (2017), to predict a framework. To encourage long-term deposits in banks among prospective clients of the bank, this research focuses on the necessity for telemarketing strategies. By utilizing several ML techniques, the research investigated the market for long-term deposits in the banks. The dataset, which includes 45147

cases with 17 variables, is acquired from the (UCI) ML Repository. Higher accuracy is provided by the random forest at 86.8%. The research findings also offer banks valuable knowledge for opting for telemarketing practices that would maximize the efficiency of long-term bank deposits to both current and prospective bank clients.

The research of Jiang (2018) looked at applying data mining techniques to forecast the performance of bank telemarketing. Among the most popular marketing strategies, today is telephone marketing, which might help consultancy firms find prospective clients. Data mining has indeed been included to accurately handle huge amounts of data in the Big Data age, according to prior studies. This research aims to identify the most suitable customer group for bank telemarketing by forecasting its performance. By using an LR model, a link involving performance and other variables can be seen. The (UCI) ML Repository was used to obtain the database. The data collection consists of 4119 occurrences and 21 characteristics. They employed DT, NNs, NB, LR, and SVMs. They found that of these 5 strategies, LR had the highest accuracy. This model had an accuracy rate of 92.03%.

A model employing ML techniques was developed by (Ilham et al, 2019). Prospective clients are frequently identified with the long-term bank deposit product through bank telemarketing techniques. This study has a key emphasis on how to increase client values via telemarketing tactics. Thus, a classification model for prospective clients was necessary to potentially boost business profits. Multiple methodologies were employed, including LR, K-Nearest Neighbor, NB, DT, SVM, NN, and Random Forest. A prepared database from the UCI ML repository was used without any preprocessing of the dataset's attributes. The assessment metrics for these models demonstrate that they are 97.07% accurate on average when utilizing the SVM. Comparing SVM to other categorization algorithms, it can be claimed that

SVM is the finest option for identifying prospective clients who may be keen on time deposit services that are provided via landline or mobile.

In contrast, Barraza (2019) developed, an LR model for predicting consumer behavior using the expanded data set. On top of certain feature selection methods, this model is constructed. The performance over false-positive findings is enhanced by mutual information (MI) and data-based sensitivity analysis (DSA). They minimized the number of feature sets affecting this marketing sector's success. In the situation of a minimal false-positive ratio with the nine criteria they chose, they discovered that DSA is superior. When 13 chosen characteristics from a large number of features have moderately high false-positive values, MI is marginally improved.

In addition, Moro (2014) proposed an approach of 3 feature selection methodologies to identify unique characteristics that directly impact data quality, which has a substantial impact on decision-making. The techniques involve contextual feature identification and historical feature assessment. To simplify the feature selection search space, an issue is split into smaller ones that can each be solved independently. The additional dataset utilized in the present study was put to the test using their approach. In their marketing activities, they wanted to focus on the top consumers. DSA was used to identify the candidacy of the highest correlated hidden characteristics. With the help of a subject-matter expert, the approach required creating new features for historical events.

In recent years, as scientific computing has grown more accessible, data-driven strategies for discovering new clients have proliferated. In previous research on bank telemarketing, dataset researchers relied only on class predictions using different classifying methods (Abu-Srhan, 2019) .

Only a few industries presently use data mining methods and resources to enhance their processes: healthcare, finance, retail, intelligence, and telecommunications. For example, data

mining and different multivariate data analysis techniques are being employed in the banking sector to develop models for risk assessment (Hormozi, 2004), Client Relationships, DM, and prevention of credit card scams. These data analysis approaches have acquired appeal in the healthcare and insurance industries for decreasing the con medical insurance and misuse and also forecasting patterns in the behavior of patients and their wellness.

Following these avenues, it would appear relevant to analyze the bank marketing data collection. This research aims to study the linkages and interactions between multiple elements of intellectual economy and company success in the Portuguese banking sector. Stakeholder orientation components were added to the idea of relational capital. PLS was used to create a model and test hypotheses on a sample of 253 respondents from 53 businesses. (Cabrita, 2008).

Data-driven ways of locating new clients have grown in popularity in contemporary years as scientific computing has grown highly accessible.

- Merely class predictions utilizing different categorizing algorithms were employed in the scholars' earlier studies on bank telemarketing data collection.
- None of the researchers who investigated this data collection took any steps to fix the dataset’s (Class) imbalanced design. They might use a class balancer or SMOTE to balance class at first and then apply a machine learning algorithm for better classification accuracy (Camacho, 2022).
- Besides no researcher applied PCA and Cluster at the same time on this dataset. Following these research gaps, an analysis of the bank marketing data collection seems to be relevant to the research where we will apply PCA, and k means clustering at a time to find an important factor (PCA) for response prediction and grouping with similar patterns (Cluster).

2.2 Summary

Several important substantial gaps are highlighted by the assessment of the literature review.

- First off, while the current data analysis techniques have indeed been utilized to produce a prediction model with respectable accuracy, they do not offer managers of financial firms any profound managerial knowledge. For instance, the causal linkages between the important factors that influence effective bank telemarketing (such as age, day, and housing condition) are not investigated.
- When implemented in the telemarketing issue, ML techniques like LR and NNs can only handle the connection or correlation between the major telemarketing parameters and the possibility that the customer would purchase products (output variable). On the other hand, the origin of factors is not investigated.
- More notably, financial firms like banks are keen on knowing the ideal value of each operator that maximizes the probability of marketing their goods to prospective clients so that they can concentrate on a certain segment of clients. This is in addition to their significance in estimating the probability that their marketing campaign will be successful.

Chapter-3

3.1 Methodology

The data analysis methodology's such as business understanding, data understanding, data preparation, modeling, and assessment processes are used in this project. To divide a customer list into those who wish to subscribe to a long-term bank deposit and those who do not, this project attempts to determine essential features and develop a prediction model. PCA is a method that lowers a dataset's dimensionality while keeping all of the data's information and enhancing interpretability. To increase variance, it will gradually introduce new uncorrelated variables. Which clients are more likely to subscribe can be determined using the rules that have been identified. This might be a result of their age, degree of education, marital status, gender, or other characteristics. This entails determining if bank employees' contributions to the campaign process affect how many clients subscribe to this service. For instance, the number of calls made to each client, the success of the last campaign, etc. Researchers studied these traits to look for trends that can enhance marketing initiatives and lead to improved services. The crisp-DM technique, which involves taking the measures listed beneath, is the optimal strategy to utilize to get the finest outcomes. The business Objective for this project is to identify the potential reasons that lead to Identifying Prospective Clients for Long-Term Bank Deposits: Data Mining approach. Data mining aims to concentrate on these groups utilizing a project plan once the target group has been identified. A preliminary data collection, a description of the data, an investigation of the data, and its verification are all included in the data comprehension phase. Age, sex, housing type, and other information will be collected as part of the data collection, which will aid in the development of a thorough grasp of the data type, like numerical or categorical data. The next stage is data preparation. Following the gathering phase, the data

will not be able to be put into an ML system and must go through additional phases. Clean the data, for instance, by eliminating null values, identifying outliers, and handling them as needed. In the modeling phase, we will be using R to create models to predict y based on the given dataset using a PCA and K means cluster ML algorithm. Additionally, we will be using excel as a tool in some preprocessing steps. The final step is an evaluation after training the model with the train set of data, the model needs to be evaluated with the test set to calculate the accuracy of the model. For a k-means model, a classification matrix will help to compare the actual results with the predicted results.

Apply k-means cluster to identify the clusters of the two categories of customers and compare them to the original dataset to determine how well the technique can distinguish the type of consumers after using PCA to reduce the dimensionality of the data.

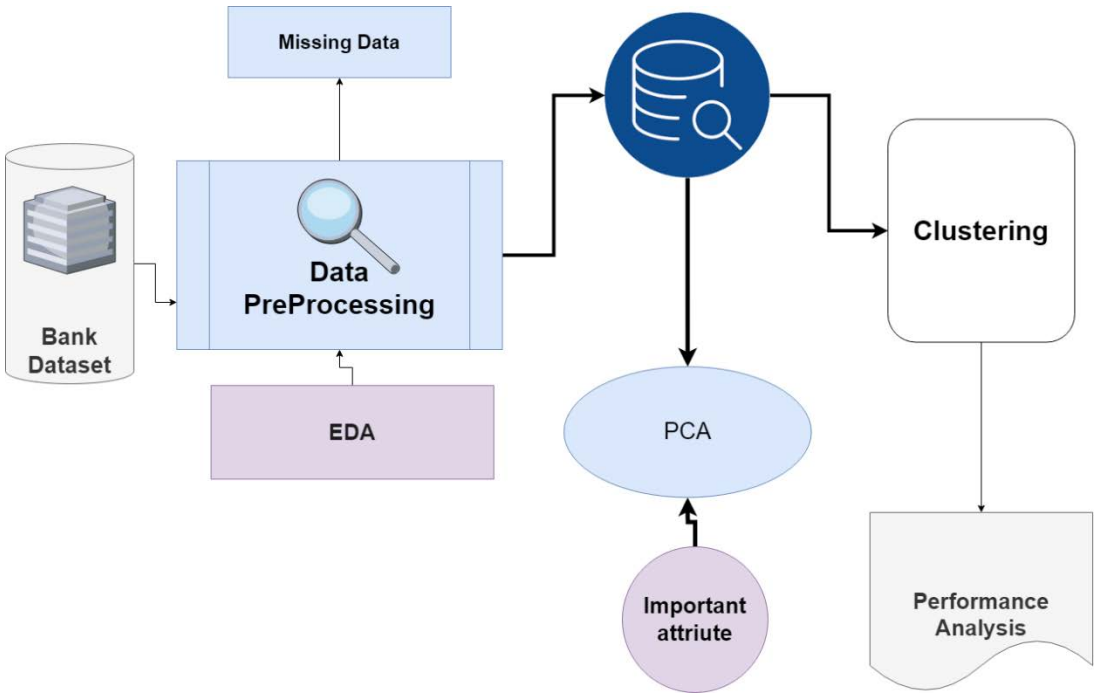

Figure 1 Proposed Methodology

3.2 Data Types

Gathering and analyzing data regarding various attributes to be investigated and evaluated as part of the qualitative and quantitative approaches to research. The various types of data encountered while conducting research are classified into the two categories mentioned below:

1. Categorical variables: If a variable only contains a few unique, often non-overlapping values, then it is categorical. Ordinal, nominal, and binary are the three distinct measurement scales that make up a categorical data type. An ordinal data type may be sorted according to precedence order and has an order of occurrence. A nominal data form is a list of categories with no particular order of occurrence. y is an example, as is yes or no. I have 11 categorical variables in the dataset.

2. Numerical variables: Numerical results contain both continuous and discrete variables. Discrete variables have a finite set of potential values, but continuous variables don't have any finite values. When a continuous variable is understood using categories, the continuous variable is categorized or discretized. The dataset has 10 numerical variables.

The categorical variables must be transformed into numerical variables before they can be used. Before principal component analysis, which will be addressed later, some of the categorical variables were omitted.

3.3 Dataset Preprocessing

To make a good data quality and good analysis with this data, the initial dataset was preprocessed. Since principal component analysis can only deal with continuous variables, we have transformed some of the variables into continuous ones and excluded the remaining variables except the variable y.

3.4 Principal Component Analysis

Principal component analysis is a technique for reducing the dimensionality of such datasets and minimizing information loss (Beattie, 2021). Using this analysis as a dimension-reduction approach, the key elements of a data set comprising bank marketing data are identified. Principal components are then applied in this study to forecast whether a consumer enrolled for a term deposit or not. This data collection includes details about a campaign run by a Portuguese financial institution to encourage its clients to sign up for term deposits.

3.5 K Means Clustering

In k-means we may use clustering, a type of unsupervised learning when you have unlabeled data (Ran, 2021). The K variable acts as a stand-in for the number of groups, and the goal of this method is to find groups in the data. By grouping or splitting data into groups of related items into clusters, clustering aids in our understanding of our data especially. K-means clustering reduces variations within clusters and tends to identify groups with similar geographical sizes.

3.6 Performance Matrix

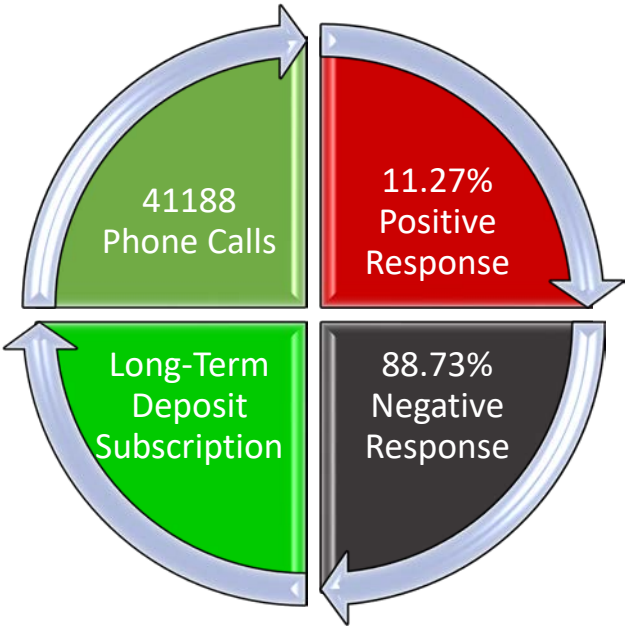

Figure 2 Performance Matrix of the Analysis

3.6 Exploratory Data Analysis

Exploratory data analysis (EDA) evaluates and examines bank data sets and summarizes their key characteristics, often using data visualization techniques. It can also help determine if the statistical methods that we're thinking of using for data analysis are suitable or not.

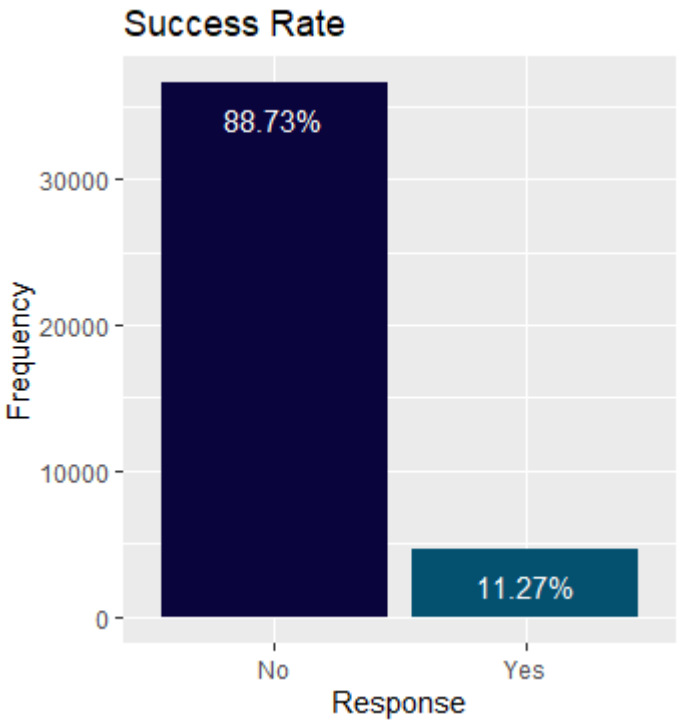

Figure 3 Success Rate

Only 11.27 percent of the 41188 phone calls to clients overall, as shown in Figure 3, had a good response. The success rate is not very high and fluctuates depending on several factors. To discover the pattern of how clients' decisions to subscribe to a long-term deposit alter, we must do more investigation. Since we won't be able to use some of the categorical variables, even after transforming them into numeric variables, let's find out whether the exclusion of the variables may have any impact or lose any valuable information or not.

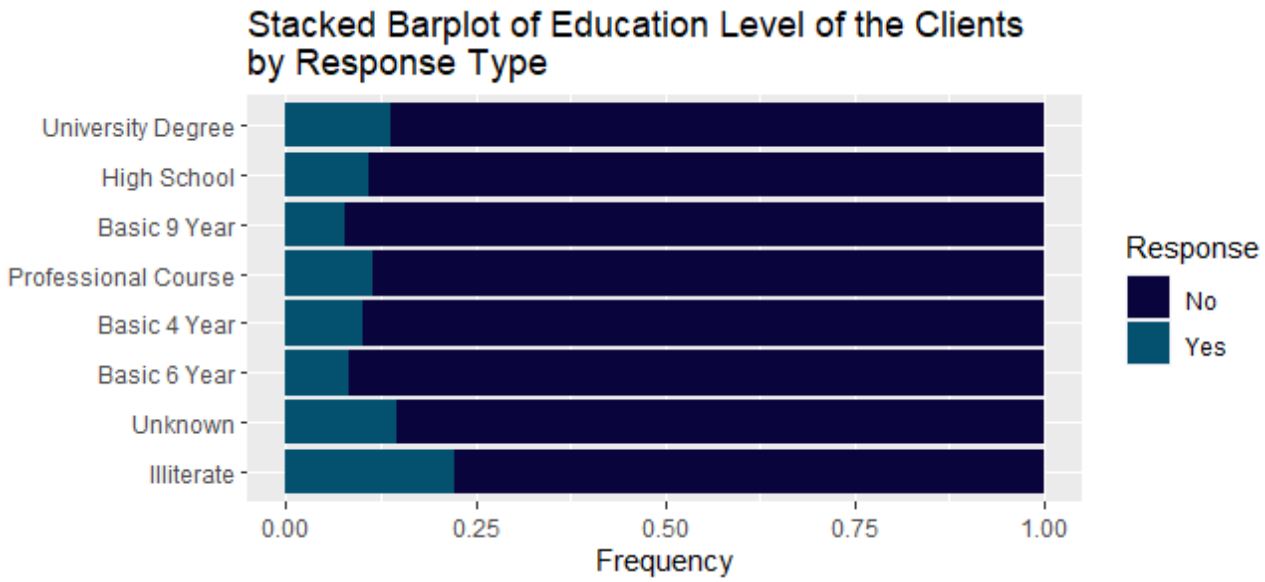

Figure 4 Bar plot that shows the proportion of responses by the education level

In figure 4 the education level of the responders shows some pattern in their responses. The illiterate peoples have the most positive response compared to others. The least positive response is seen among the customers who have completed 4, 6, and 9 years of basic schooling. The exclusion of this categorical variable might cause the loss of some valuable information from the analysis. So, we will force this categorical variable to a numeric variable in the sense that the years of schooling or education level could be transformed into some score, for example, 0 for illiterate or 16 for a university degree, etc.

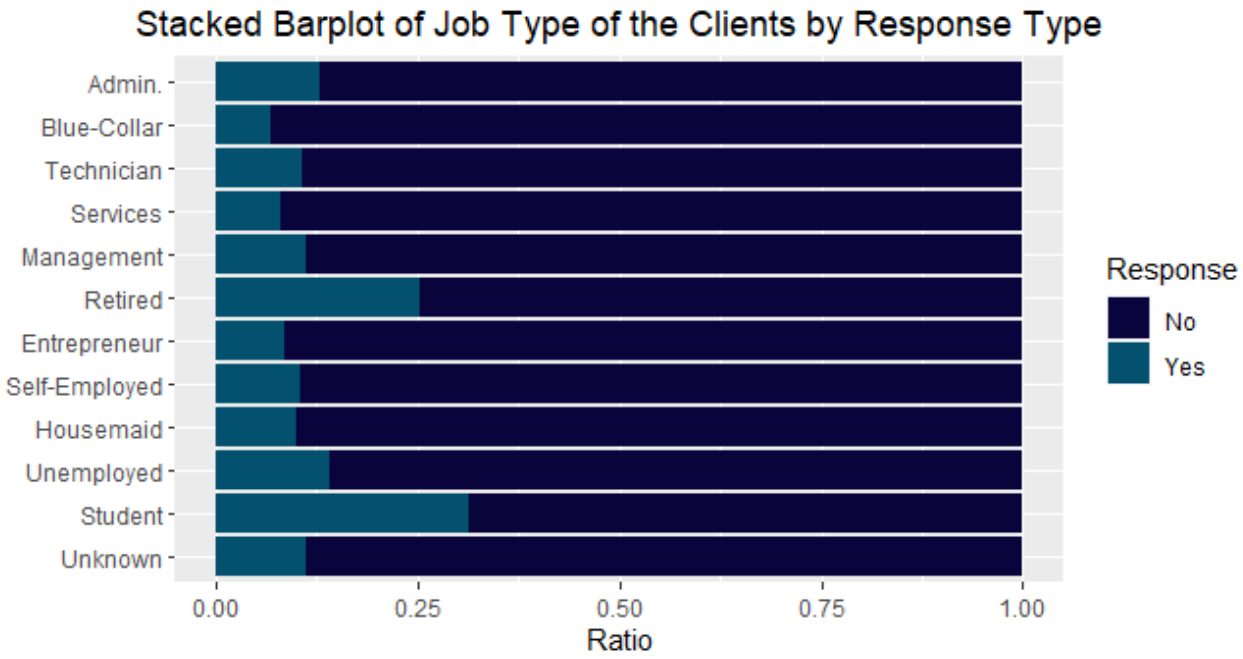

Figure 5 A stacked bar plot proportion of response by responders' job type

In figure 5 Like the customers' education level, the job type can also be analyzed to find some patterns. Retired students seem to have more positive responses compared to others. Like education level, job type cannot be transformed into any numeric variable because it will not make sense. So, we have no other choice than to drop this variable from the processed dataset for principal component analysis. And we can expect some information loss for this reason. The variable named default have only three responses that have the label yes. Since it doesn't show any meaningful association with the outcome y, we may also exclude this from the final analysis.

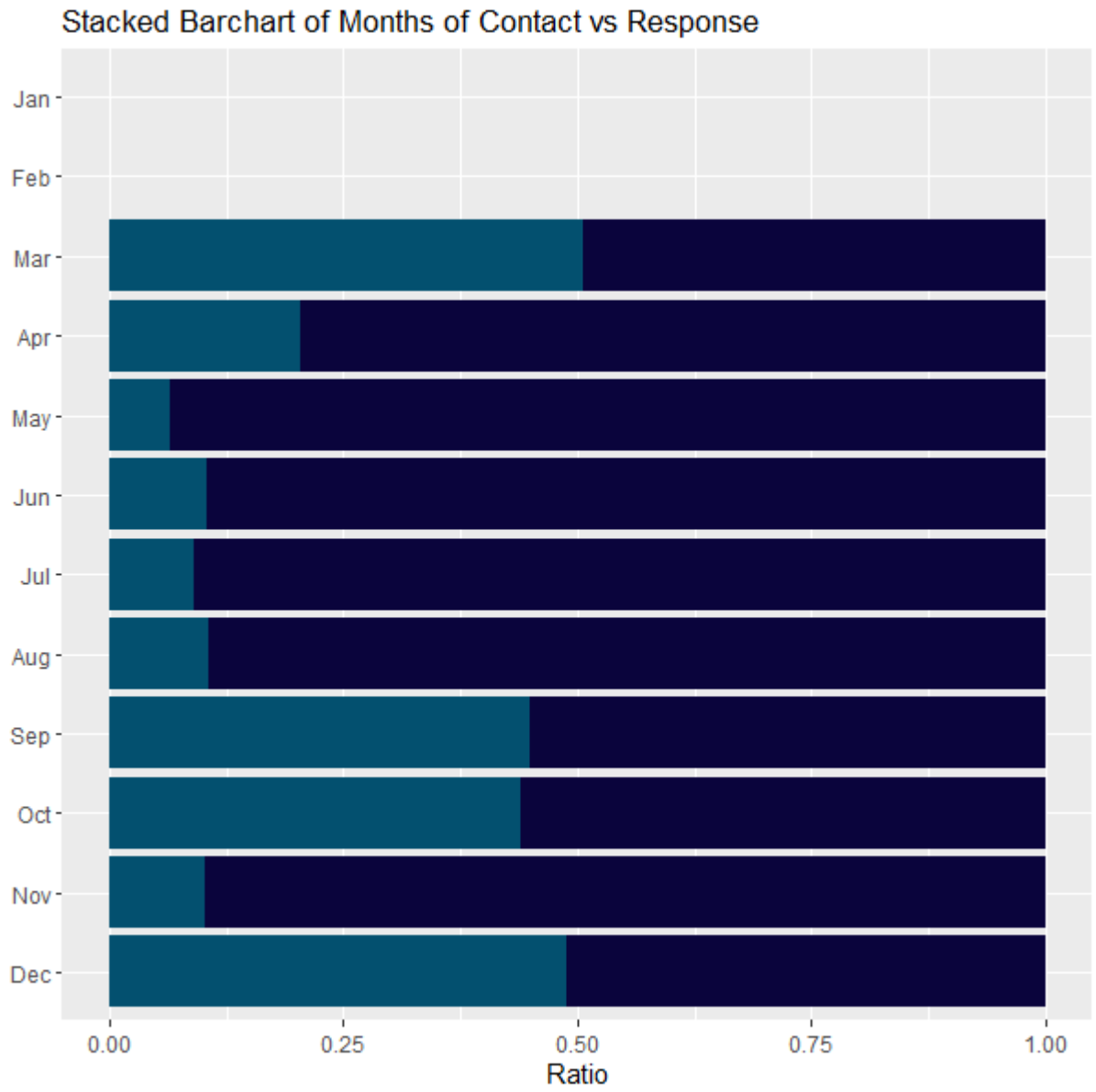

Figure 6 A stacked bar plot that shows the proportion of response by different month

In this figure 6, we can see the response rate concerning a different month. It is found that in the months of March and December the ratio reaches 0.5, which is the highest. On the other hand, in the month of May the ratio below 0.1, which is the lowest.

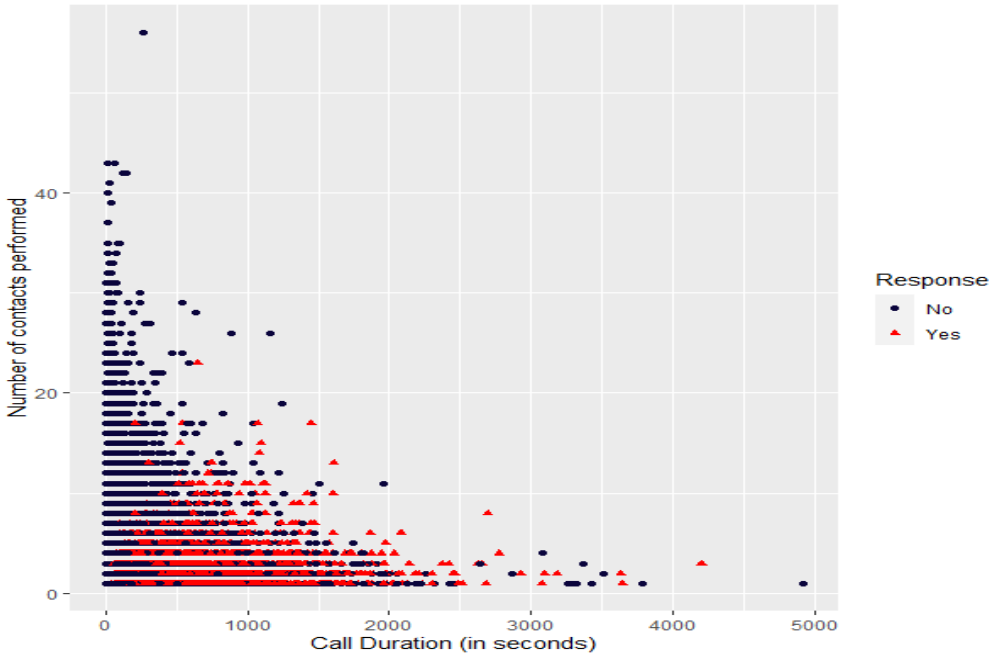

Figure 7 Response by number of contracted performed with call duration

In figure 7 we can see the response by some contracts performed with call duration. It is found that as the duration of call is increasing the possibility of number of contracts performed is decreasing.

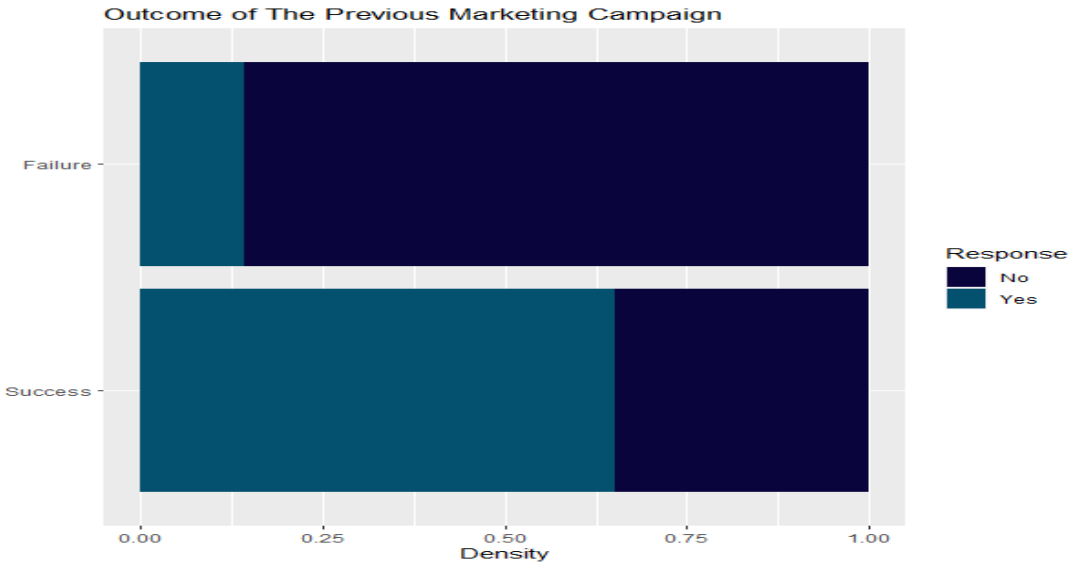

Figure 8 Response to the outcome of the previous marketing campaign

In figure 8 we can see the response to the outcome of a previous marketing campaign. It is noticed that the density of positive response from the marketing campaign is found with higher degree of ratio in the comparison to failed one.

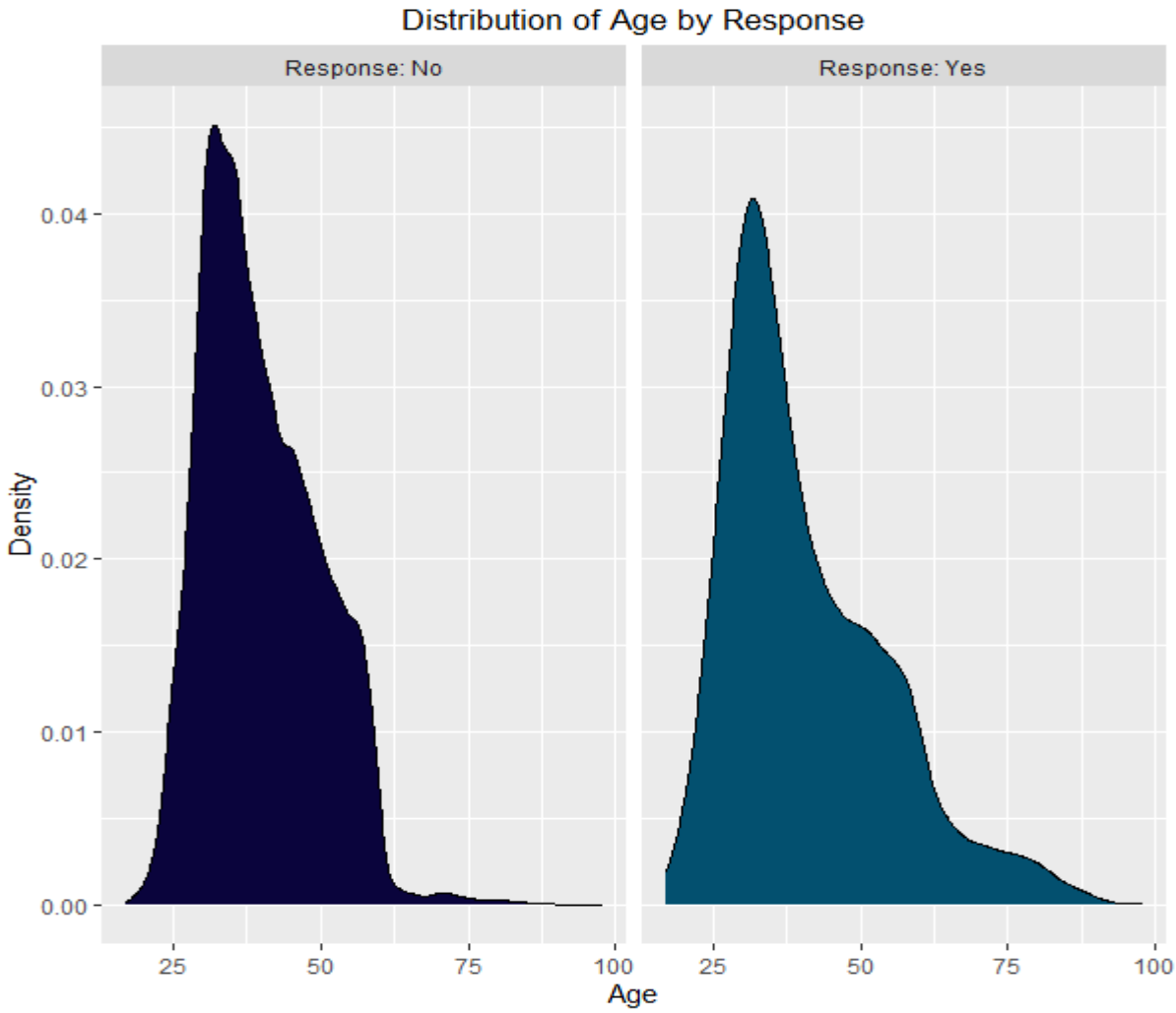

Figure 9 Ages vs Response

Figure 9 is all about age by response rate. It is found that the number of customers belonging to the age group 25 to 50 are found higher with negative response category and while the age range in case of positive response is around 25 to 75 years.

Chapter-4

Result and Discussions

4.1 Dataset Overview

Using the UCI Machine Learning Repository, actual data on bank telemarketing is gathered for this study (Moro, (2014)). Data were obtained from a Portuguese bank that used its contact center to undertake marketing campaigns to persuade and pull customers to its term deposit program to increase its customer base. The collection contains 17 campaigns that ran between May 2008 and November 2010. During the campaign, a user-friendly online application for long-term deposits with competitive interest rates was made available. According to the findings, the success rate is 11.27 percent.

Table 1: Data Overview

Number	Name
1	age
2	job
3	marital
4	education
5	default
6	housing
7	loan
8	contact
9	month
10	Day of week
11	duration

12	campaign
13	days the client was last contacted
14	previous
15	the outcome of the previous marketing campaign
16	Employment variation rate
17	Consumer price index
18	Consumer confidence index
19	Euribor 3-month rate
20	Employee number
21	**Y (Dependent Variable)**

4.2 Data Analysis

After the necessary data cleaning and modifications, we utilized principal component analysis. First, we'll take a look at the scree plot. The principal components are plotted against their respective eigenvalues, or the percentage of variances in data explained by the components, in a scree plot. The percentage of variations explained is displayed from biggest to smallest in figure 10's scree plot.

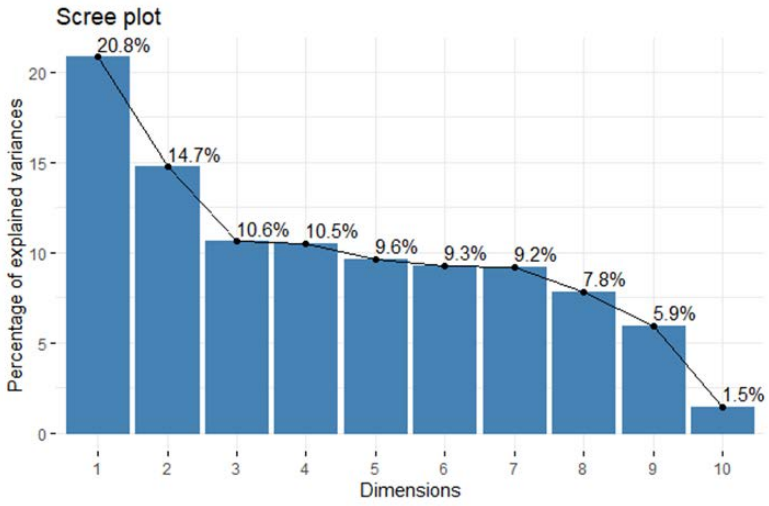

Figure 10 A scree plot that the number of variations explained by each

The above Scree plot shows that the 1st 4 principal components can explain 53.16% of the variations of the Portuguese bank data set, and the first 7 principal components can explain 84.81% of the variations of the Portuguese bank data set. The other components seem not so important. As a result, if some prediction analysis is intended for this data set, it can conclude that taking about 7 principal components seems to be best. Using principal component analysis, the attempt to represent all of the variables as linear combinations of a small number of eigenvectors. By doing so, it will be able to determine which variables contribute more to explaining variances than others.

The loading plot shows the coefficients of every variable for the pairs of principal components. Usually, a loading plot is used here to see how strongly each variable influences a principal component. Besides it can see how the vectors are pinned at the origin of principal components in each plot for 3 combinations (PC1 and PC2, PC2 and PC3, and PC3 and PC4). The coordinates values indicate how much weight they have on the PC.

In the first plot in figure-11, there are p-outcome, p-days, and previous; these 3 attributes strongly influence the 1st PC. The 1st PC explains 20.8% of the variations, which is the highest of all. So, it may say that the variables p-outcome, p-days, and previous are the most important variables in the bank dataset. And PC1 may represent the customers' decision according to the status of their previous calling outcome.

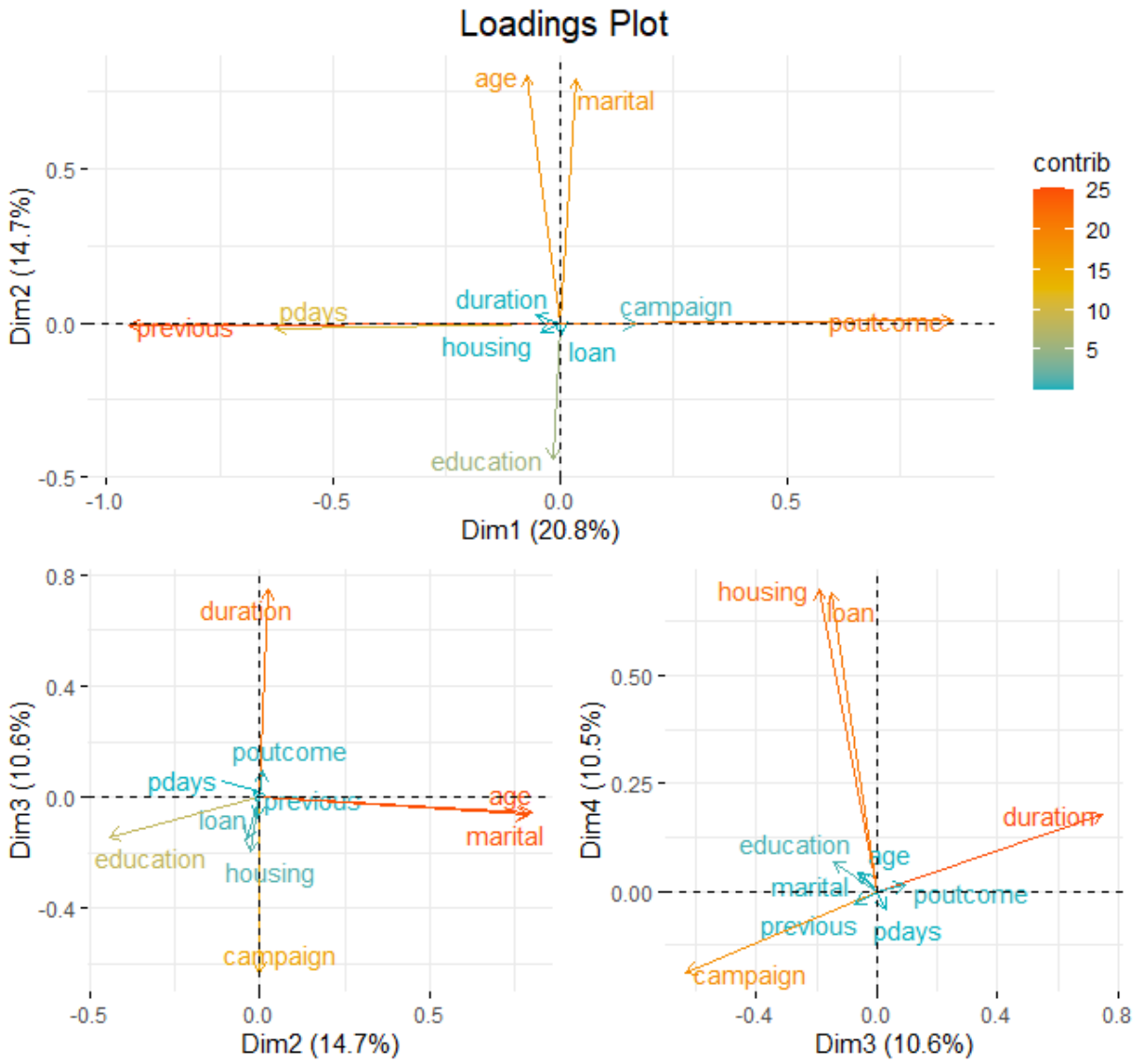

Figure 11 Loading Plots of the Variables

In 2nd PC, which accounts for 14.7% of the variances in the dataset, marital status, age, and education are the most important variables. This section can thus be used to group consumers according to their socioeconomic and age factors. Only 35.58 percent of deviations can be explained by the first two components in total, which is a relatively low percentage.

It may have a significant influence on the third PC's duration and campaign. Additionally, the component explains 10.6% of the variation on its own. Based on judgments made by bank managers and the effectiveness of call agents during prior campaigns, this section separates consumers into categories.

Table 2: Loadings' Table

	PrinComp1	PrinComp2	PrinComp3	PrinComp4	PrinComp5	PrinComp6	PrinComp7
age	-0.05	0.66	-0.05	0.04	-0.03	-0.13	0.16
marital	0.02	0.65	-0.06	0.04	-0.02	-0.13	0.24
education	-0.01	-0.36	-0.13	0.06	-0.11	-0.52	0.74
housing	-0.02	-0.02	-0.18	0.68	-0.57	-0.21	-0.34
loan	0.00	-0.03	-0.14	0.67	0.63	0.27	0.18
duration	-0.03	0.01	0.72	0.17	0.27	-0.52	-0.21
campaign	0.11	-0.0009	-0.61	-0.18	0.39	-0.49	-0.39
P days	-0.43	-0.01	0.02	-0.04	0.12	-0.19	-0.03
previous	-0.65	-0.007	-0.07	-0.02	0.02	0.00	-0.01
Pout come	0.60	0.005	0.09	0.01	0.02	-0.09	0.01

When looking at the fourth primary component, which explains 10.48 percent of the variations, it is clear that housing and loans are important in this situation. The financial activity or capability of a customer may thus be described in this area. The data reveals that this is not the case, even though these two characteristics should theoretically have a greater influence on distinguishing separate groups.

Through the guidance of the aforementioned processes, it can be determined that employing 7 parts for clustering algorithms would produce adequate results, with the parts accounting for the variances of 84.81%. It only contains a small number of variables, though, and clusters may be formed using any of the numeric variables. We did not create a biplot to display both observations and variables in one since there are so many observations in the dataset. Instead, there were two distinct stories planned. The biplot in figure 12 allows it to identify certain groupings. We can observe that the clusters contain a mixture of customer response labels.

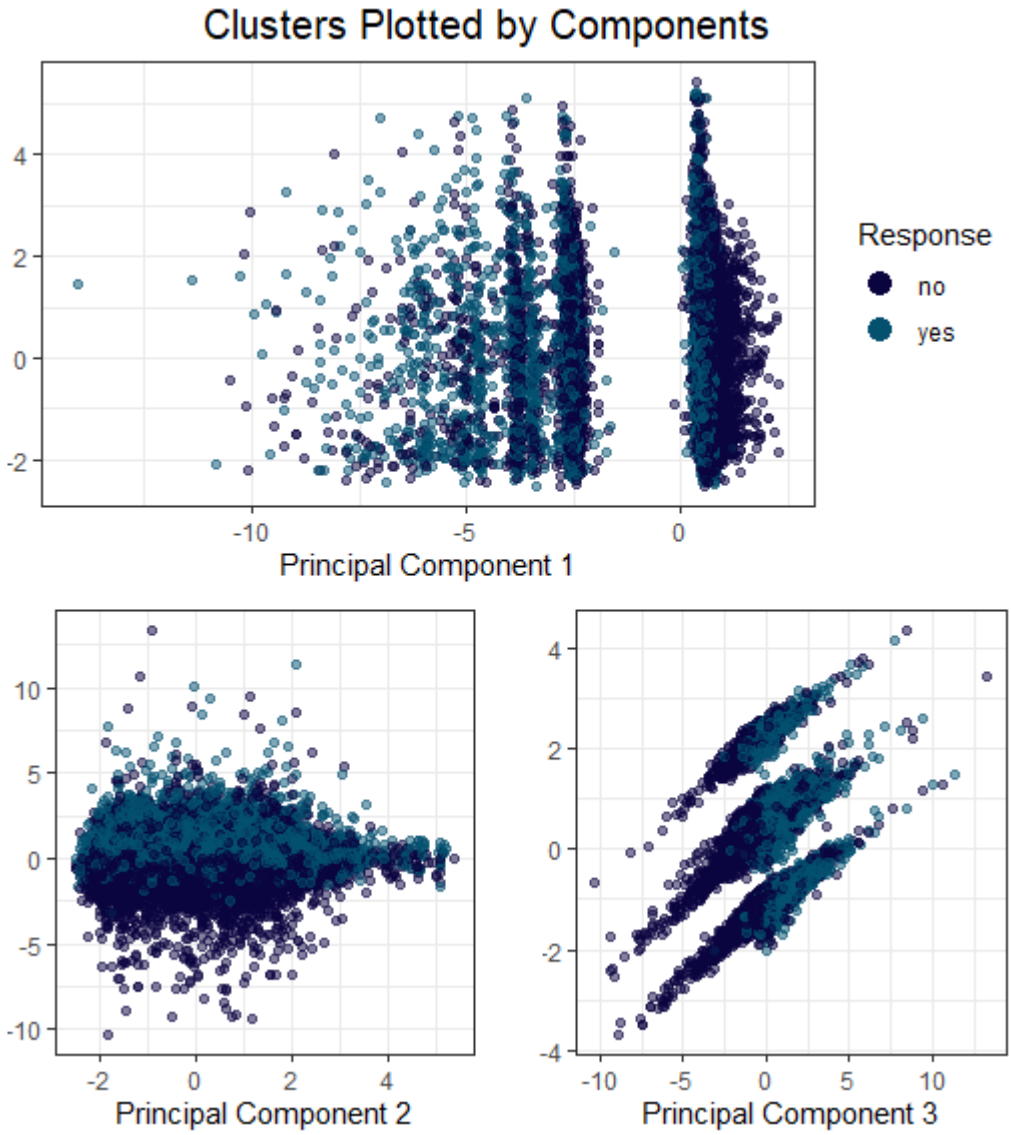

Figure 12 Biplot of the components

Now after identifying the factors that have the most influence on the dataset, we may go on to identify the two customer clusters and evaluate how well they perform in comparison to the original dataset.

As we know clustering is a set of techniques that finds subgroups of dataset observations where the observations in the same groups are similarly based on some factors. Since no response variable is present in cluster analysis, this is unsupervised. It means the data set that we have will be used to find which customers are alike and eventually categorize them into groups. K-means cluster evaluation has been employed as researchers are looking for certain distinct

categories, like yes and no replies. The k value is 2, reflecting our search for 2 clusters. So, we are using the k-means clustering algorithm. The algorithm finds the partition of k using a local optimum sum of squares inside the clusters, having the further restriction that no single data point migration from one cluster to the next could decrease the sum of squares inside the cluster. Using the r statistical software, the first values for the k means were chosen at random (RSTUDIO).

Table 3:K-Means Cluster Performance Analysis on Training Data

		Primary Response	
		No	Yes
Cluster Response	No	21226	2415
	Yes	604	755

We discovered an accuracy rate of 87.92 percent after running the k-means evaluation on the data training, after dividing it from the primary data in a ratio of 80/20. The contingency table reveals, however, that the clusters misclassified the majority of the affirmative replies. The classification of negative replies using the k-means cluster analysis was successful. This could be the case given that just 11.27 percent of the observations yield a positive reaction, compared to 88.73 percent of the observations that do. This misclassification issue in the affirmative response may be caused by a dearth of imbalanced (Class) data. The affirmative response has 87.76 percent misclassified outcomes.

Table 4: K-Means Cluster Performance Model on Test Data

		Primary Response	
		No	Yes
Cluster Response	No	4666	542
	Yes	133	147

We then performed the k-means cluster analysis on the test data to see how consistent the model performs. The outcome showed 87.7% accuracy, which is almost the same as before. But again, the positive responses are misclassified in a large portion, and the percentage of misclassification in positive responses is about 78.6% which is slightly lower than before but not a satisfactory result.

Thus, the current study collected data from the bank DM campaign and developed a performance matrix from the prospective of the client on a long-term bank deposit subscription plan through a telemarketing approach. The performance matrix evaluation model indicates that the telemarketing approach success rate is not high, and it fluctuated since out of 41188 phone calls to the client only 11.27% exhibited positive responses and 88.73% exhibited negative responses. Also, by employing the K-mean cluster algorithm it is validated that the results of the classification of negative responses are successful and the misclassification results on positive responses of client calls are slightly lower (78.6 %) and not a satisfactory outcome.

Chapter-5

5.1 Conclusion

In this post, we've looked at data from a bank's DM campaign using a variety of data analysis approaches. The data has been cleansed for use in the principal component analysis and k-means analysis after the descriptive analysis (EDA). We learned via PCA which variables are most effective in classifying the examples into distinct categories. An intriguing result of the PCA is that, despite the popular perception that consumers' income levels have a greater influence on possible customer groups, this is not the case. The social elements of the clients and the past marketing aspects of the banks are more significant than other variables. The effectiveness of the strategy in identifying groups of possible clients is tested using k-means clustering. The greater percentage of misclassification mistakes, however, suggests that while having a strong accuracy score, this may not be dependable for identifying new clients that banks may approach market term deposits.

5.2 Potential Obstacles

Since there were so many categorical variables in the data, we were unable to use some of the variables that had shown a general association with the response. Furthermore, transforming certain categorical variables into numerical variables may have been error-prone or misleading. Many of the observations have been dropped before the analysis for missing values and did not have any useful information. The main issue that we have seen after the k-means cluster analysis is the presence of so much misclassification, and it is because of having so much difference in the proportion of responses. If I could override these issues, maybe we could deliver more accurate results.

5.3 Recommendation and Future Works

Since the outcome of KNN method is found less efficient, due to which it is recommended to utilize better classification methods for predicting a customer's decision. The k-means clustering algorithm is an unsupervised learning method, the clusters may have represented something other than customers' decisions, or the number of k may seem misleading in contrast to this study. For increasing the competency of the work in future, it will be better to use of random forest and LR for finding the customer groups and respective efficiency might be augmented by including real-time primary dataset as a train and test data.

Bibliography (APA format)

1. Abu-Srhan, A., & Al-Sayyed, R. (2019). Visualization and Analysis in Bank Direct Marketing Prediction. *International Journal of Advanced Computer Science and Applications*, *10*(7).

   https://doi.org/10.14569/IJACSA.2019.0100785

2. Amran, A., Fauzi, H., Purwanto, Y., Darus, F., Yusoff, H., Zain, M. M., ... & Nejati, M. (2017). Social responsibility disclosure in Islamic banks: a comparative study of Indonesia and Malaysia. *Journal of Financial Reporting and Accounting*.

   http://dx.doi.org/10.1108/JFRA-01-2015-0016

3. Apampa, O. (2016). Evaluation of classification and ensemble algorithms for bank customer marketing response prediction. *Journal of International Technology and Information Management*, *25*(4), 6.

   https://scholarworks.lib.csusb.edu/jitim/vol25/iss4/6

4. Asare-Frempong, J., & Jayabalan, M. (2017). Predicting customer response to bank direct telemarketing campaign. In *2017 International Conference on Engineering Technology and Technopreneurship (ICE2T)* (pp. 1-4). IEEE.

   10.1109/ICE2T.2017.8215961

5. Barraza, N., Moro, S., Ferreyra, M., & de la Peña, A. (2019). Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study. *Journal of Information Science*, *45*(1), 53-67.

   https://doi.org/10.1177/0165551518770967

6. Beattie, J. R., & Esmonde-White, F. W. (2021). Exploration of principal component analysis: deriving principal component analysis visually using spectra. *Applied Spectroscopy*, *75*(4), 361-375.

https://doi.org/10.1177/0003702820987847

7. Cabrita, M. D. R., & Bontis, N. (2008). Intellectual capital and business performance in the Portuguese banking industry. *International Journal of Technology Management*, *43*(1-3), 212-237.

https://doi.org/ 10.1504/IJTM.2008.019416

8. Camacho, L., Douzas, G., & Bacao, F. (2022). Geometric SMOTE for regression. *Expert Systems with Applications*, 116387.

https://doi.org/10.1016/j.eswa.2021.116387

9. Cvijović, J., Kostić-Stanković, M., & Reljić, M. (2017). Customer relationship management in the banking industry: Modern approach. *Industrija*, *45*(3).

https://doi.org/10.5937/industrija45-15975

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

https://doi.org/ 10.5120/14852-3218

11. Giannakis-Bompolis, C., & Boutsouki, C. (2014). Customer relationship management in the era of the social web and social customer: an investigation of customer engagement in the Greek retail banking sector. *Procedia-Social and Behavioral Sciences*, *148*, 67-78.

https://doi.org/10.1016/j.sbspro.2014.07.018

12. Hormozi, A. M., & Giles, S. (2004). Data mining: A competitive weapon for banking and retail industries. *Information systems management*, *21*(2), 62-71.

https://doi.org/10.1201/1078/44118.21.2.20040301/80423.9

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

https://doi.org/10.1080/23270012.2021.1897956

14. Ilham, A., Khikmah, L., & Iswara, I. B. A. I. (2019). Long-term deposits prediction: a comparative framework of classification model for predicting the success of bank telemarketing. In *Journal of Physics: Conference Series* (Vol. 1175, No. 1, p. 012035). IOP Publishing.

https://doi.Org/ 10.1088/1742-6596/1175/1/012035

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

https://doi.org/ 10.11648/J.IJDST.20180401.15

16. Ledhem, M. A. (2021). Data mining techniques for predicting the financial performance of Islamic banking in Indonesia. *Journal of Modelling in Management*.

https://doi.org/10.1108/JM2-10-2020-0286

17. Mallin, C., Farag, H., & Ow-Yong, K. (2014). Corporate social responsibility and financial performance in Islamic banks. *Journal of Economic Behavior & Organization*, *103*, S21-S38.

https://doi.org/10.1016/j.jebo.2014.03.001

18. Mogaji, E., Soetan, T. O., & Kieu, T. A. (2020). The implications of artificial intelligence on the digital marketing of financial services to vulnerable customers. *Australasian Marketing Journal*, j-ausmj.

https://doi.org/10.1016/j.ausmj.2020.05.003

19. Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22-31.

https://doi.org/10.1016/j.dss.2014.03.001

20. Palacio, JRS, & Pérez, SR (2018). Corporate social responsibility in banking: Its application to the case of cooperative banking. *REVESCO: journal of cooperative studies*, (127), 204-227.

http://dx.doi.org/10.5209/REVE.59771

21. Palaniappan, S., Mustapha, A., Foozy, C. F. M., & Atan, R. (2017). Customer profiling using classification approach for bank telemarketing. *JOIV: International Journal on Informatics Visualization*, *1*(4-2), 214-217.

http://dx.doi.org/10.30630/joiv.1.4-2.68

22. 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), 692-696.

https://dergipark.org.tr/en/download/article-file/365990

23. Proença, J. F., & Rodrigues, M. A. (2011). A comparison of users and non-users of banking self-service technology in Portugal. *Managing Service Quality: An International Journal*.

https://doi.org/10.1108/09604521111113465

24. Raghunandan, G., Lavina, G. S., & Jose, S. (2018). Electronic customer relationship management is an effective tool in the banking sector. *Asian Journal of Management*, *9*(2), 913-919.

https://doi.org/10.5958/2321-5763.2018.00144.0

25. Raju, S. S., & Dhandayudam, P. (2018). Prediction of customer behaviour analysis using classification algorithms. In *AIP conference proceedings* (Vol. 1952, No. 1, p. 020098). AIP Publishing LLC.

https://doi.org/10.1063/1.5032060

26. Ran, X., Zhou, X., Lei, M., Tepsan, W., & Deng, W. (2021). A novel k-means clustering algorithm with a noise algorithm for capturing urban hotspots. *Applied Sciences*, *11*(23), 11202.

https://doi.org/10.3390/app112311202

27. Sharma, N., Kaur, A., Gandotra, S., & Sharma, B. (2015). Evaluation and comparison of data mining techniques over bank direct marketing. *International Journal of Innovative Research in Science, Engineering and Technology*, *4*(8), 7141-7147.

http://www.ijirset.com/upload/2015/august/60_Evaluation.pdf

28. Vajiramedhin, C., & Suebsing, A. (2014). Feature selection with data balancing for prediction of bank telemarketing. *Applied Mathematical Sciences*, *8*(114), 5667-5672.

https://doi.org/ 10.12988/AMS.2014.47222

29. Yeboah, E. (2012). Using customer relationship management as a strategic tool in financial institutions: a case study of the National Investment Bank, Ghana.

https://urn.fi/URN:NBN:fi:amk-2012060511589

The End