RevistaEconomică
Journal of economic-financial theory and practice
Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

DETERMINANTS OF ACCESSING A TERM DEPOSIT IN A MARKETING CAMPAIGN - ANALYSIS ON A MARKETING CAMPAIGN AT A PORTUGUESE BANK

Maria-Manuela NECHITA[1], Mircea ASANDULUI [2], Ciprian TURTUREAN[3]

[1]*Alexandru Ioan Cuza University of Iasi, 0009-0007-8474-5091*
[2] *Alexandru Ioan Cuza University of Iasi, 0000-0002-6182-7103*
[3] *Alexandru Ioan Cuza University of Iasi, 0000-0003-4353-4267*

Abstract: *The main objective of this research is to identify the determining factors in the decision to open a term deposit following a telemarketing campaign carried out by a Portuguese bank. To support this objective, three secondary objectives were defined: (1) analyzing the impact of holding a personal loan or mortgage on a person's decision to access term deposits, (2) examining differences in education levels among individuals who open a term deposit, and (3) assessing the influence of individuals' responses to previous bank campaigns on their decision.The study is based on data collected by a Portuguese bank during a telemarketing campaign promoting term deposits. The data was gathered between May 2008 and November 2010. To achieve these objectives, we will employ data mining techniques such as logistic regression and decision trees. The analysis classifies individuals into two categories: those who accepted and those who did not accept the bank's offer. Results indicate that call duration and the outcome of the previous marketing campaign are the most significant predictors of a customer's decision. These findings provide valuable insights for optimizing direct marketing strategies in the banking sector. Future research could enhance predictive accuracy by integrating additional variables such as financial behavior, economic conditions, and demographic factors.*

Keywords: *Bank deposit, Marketing campaign, Data mining, Decision trees, Customer behavior*

JEL classification: *C25, D14, G21, M31, M37*

1. Introduction

The concept of marketing has progressively become a widely discussed topic, primarily aimed at identifying potential customers and effectively promoting various products or services to them. The strategies employed in the marketing process are diverse and carefully tailored to meet the needs and interests of those executing them. Among the most recognized forms of marketing are advertising campaigns, designed to reach a broad audience.

According to Ling and Li (2010), due to the intense global competition in recent years, the effectiveness of *advertising campaigns* has declined significantly. More precisely, less than 1% of people who are exposed to an advertising campaign end up purchasing the promoted product or service.

An alternative to this form of advertising is *direct marketing*, which targets a well-defined market segment. It is widely recognized as being a much more *"invasive"* method compared to other marketing techniques. Due to this characteristic, direct marketing is not recommended for every population segment. However, *direct marketing* focuses on understanding individual behavior by identifying the characteristics of potential customers, utilizing techniques such as *Machine Learning*.

Direct marketing is a method successfully employed by *banking* and *insurance institutions*, where selected clients are contacted directly to be presented with the promoted financial service. This

[1] *nechita.manuela@yahoo.com*
[2] *mircea.asandului@uaic.ro*
[3]*ciprian.turturean@uaic.ro*

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

method can be considered a component of the promotional strategy for these companies, as it facilitates direct interactions with clients and, implicitly, allows the collection and exploitation of gathered data *to identify the profiles of potential clients for new services/products.*

For banking institutions, *deposits represent an important source of income*, which is why they are directly interested in attracting as many clients as possible to access this product. This paper aims to analyze the impact of direct marketing conducted by a banking institution in Portugal on current clients, and implicitly on potential clients *(telemarketing)*. *The objective of the campaign* was to promote term deposits through phone calls.

The main objective (MO) of this research is to identify the determining *factors in the decision to open a term deposit as a result of a of a telemarketing campaign carried out by a Portuguese bank.*

The secondary objectives (SO), which support the main objective, are:

SO_1: Analyzing the impact that holding a personal loan/mortgage has on a person's decision to access term deposits.

SO_2: Analyzing the differences in the education levels of individuals who open a term deposit.

SO_3: Analyzing the impact of individuals' responses to the bank's previous campaigns on their decision to open a term deposit.

In accordance with the research objectives, we will define the research hypotheses (RH):

RH_1: Individuals who hold a personal loan are less interested in opening a term deposit compared to those who do not hold a loan.

RH_2: Individuals with higher education show more interest in opening a term deposit compared to those with secondary education.

RH_3: Individuals who responded positively to previous campaigns by the bank show more interest in opening a term deposit.

RH_4: The majority of people who did not participate in the bank's previous campaigns or who participated but did not purchase a banking product, did not open a term deposit.

2. Literature review

The role of marketing in the development of a business is a major one, as it becomes an integral part of the activities upon which a promotional campaign for certain goods/services is based (Borowik et al., 2016).

Kotler (2004) defines marketing as a process that integrates both the social and managerial aspects, with the goal of exchanging goods/services for a certain value between two or more people. A more suggestive definition provided by the same author considers marketing to be *"the art and science of selling."*

In recent years, marketing campaigns have become an essential part of business development, which is why their number has increased. However, the impact of these campaigns on potential buyers has diminished, leading to a *phenomenon of buyer desensitization*. Moreover, economic pressures, including competition, have caused marketing campaign managers to focus on campaigns targeted at specific audience segments (Moro, 2011).

In the paper *"The influence of banking advertising on bank customers: an examination of Greek bank customers' choices,"* Mylonakis (2008) focused on studying the effect of advertising campaigns conducted by banking institutions in attracting potential clients. The analysis was conducted for banks in Greece and demonstrated *that advertising is not a determining factor in choosing a bank*. In other words, the selection of a banking institution by an individual is based on other considerations, such as *the products and services offered, their costs, or the conditions* that must be met in order to access the respective service/product.

Charles et al. (2007) also analyzed the determining factors that influence individuals in choosing banking institutions. In this research, they examined a range of *cultural and economic scenarios* for bank customers in the USA. The conclusions highlighted that *promoting a banking institution does not necessarily lead to an increase in potential clients*. Potential clients of a bank are particularly interested

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

in *the features of the products offered by that bank, being less motivated to access a service based on the marketing campaigns it conducts.*

Moro et al. (2014) emphasize the importance of using data in evaluating the impact of banking marketing campaigns, showing that analyzing *customer behavior* through *machine learning methods* can significantly optimize their effectiveness. They also argue that *banking marketing* can be improved by integrating *predictive algorithms*, which allow for the rapid identification of the determining factors that lead to the success or failure of a campaign.

Lu at al. (2016), with a focus on identifying the determining factors behind the decision of bank clients to opt for term deposits, argues that the main factors influencing individuals' decisions to access a term deposit include: *the interest rate, the bank's risk management, and the associated costs*.

Loshin and Reifer (2013), in their paper *Using Information to Develop a Culture of Customer Centricity: Customer Centricity, Analytics, and Information Utilization*, argue that for marketing strategies to be successful, regardless of their field of application, they must be customer-oriented. This involves segmenting customers based on socio-demographic characteristics, with the goal of implementing different marketing strategies that have a much higher probability of success.

Telemarketing is one of the most commonly used marketing techniques within banking institutions. Kotler (2011) argues that one of the key benefits of this method is that it allows potential clients to directly signal their intention to purchase a product/service or participate in a campaign. However, Roach (2009) claims that not all segments of the population may react favorably to telemarketing campaigns. In another study, Mylonakis (2008) states that this form of marketing, which he considers much more "aggressive" compared to other methods, is not suitable for every population segment. For certain categories of clients, this technique may reduce their interest in accessing more products in the future and, in some cases, may even lead to the loss of existing clients.

The concept of relationship marketing was first introduced by Berry (1983). This technique is used in direct response marketing campaigns, with the aim of customer loyalty rather than simply making a sale. The fundamental principle of this marketing technique is to maximize long-term benefits for both parties involved. Since relationship marketing is client-oriented, it focuses on *strengthening relationships with customers* and on the *adaptability/personalization of the offer.*

Vella & Caruana (2012) believe that for banking marketing, the most important factors in developing operational marketing strategies, and implicitly in the development of technological solutions for customer relationship management (CRM), are *utility* and *ease of access* as perceived by the customers (Filip et al., 2016).

Bhambri (2011) argues that, regarding CRM systems, data mining methods have become *an extremely useful technique for banking institutions* in marketing campaigns. This is because they allow for the *development of highly personalized marketing strategies* that can be tailored to the specific demands and needs of customers, as well as *real-time detection of banking fraud.*

In conclusion, *direct marketing strategies target standardized profiles* that dominate each *defined market segment*, based on socio-economic characteristics. This approach will lead to an increased impact of the targeted effect through the campaign, whether it involves selling products/services, enhancing brand image, or strengthening the relationship between the client and the company/institution.

3. Data and methodology

3.1 Presentation of the dataset

The dataset used aims to analyze the process of selling term deposits within a telemarketing campaign conducted by a banking institution in Portugal. During this campaign, agents make phone calls to current clients to sell the deposit (outbound) or, if the client calls the call center for any other reason, they are asked to subscribe to such a deposit (inbound). The outcome is recorded through a binary variable with the values success/failure. Thus, the collected data is based on variables that capture

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

characteristics of the telemarketing campaign process (day, month, call duration), as well as items that address social and economic influence characteristics (person's occupation, their age).

The data was retrieved from https://archive.ics.uci.edu/dataset/222/bank+marketing and represents data collected from May 2008 to November 2010, as a result of calls made by agents of a financial institution in Portugal. In order toverify the probability of success of the campaign, i.e., whether the contacted person would be interested in participating in the campaign and eventually open a term deposit, it was necessary for a large portion of the potential interested individuals to be contacted multiple times.

The database used in this study initially contains 45.211 records and 15 variables, of which 14 are predictor variables and one is the outcome/target variable, indicating whether the client opened a term deposit or not (Table 1).

In the table below, I have grouped the variables according to their typology: binary categorical variables, categorical variables, and numerical variables. This categorization allows us to apply certain treatments for situations where we identify missing values, outliers, or intend to reduce the complexity of the variables and, implicitly, of the analysis process through predictive methods.

Tabel 1: Database structure

Variable	Variable type	Variable description
Y (target variable)	*binary categorical*	**Did the client open a term deposit**? Yes/No
default	*binary categorical*	**Is the client a defaulter?** No/Yes
housing		**Does the client have a mortgage loan?** No/Yes
loan		**Does the client have a personal loan?** Yes/No
marital	*categorical*	**Marital status:** divorced/ widowed, married, single,
Job		**Job:** Blue-collar, Entrepreneur, Other, Pink-collar, Self-employed, Technician, White-collar
education		**Education level:** unknown (missing value), primary, secondary, tertiary
day_of_week		**Day of the week when the client was last contacted::** 'mon', 'tue', 'wed', 'thu', 'fri'
Month		**Month when the client was last contacted**: 'jan', 'feb', 'mar', .. , 'dec'
poutcome		**Result of previous marketing campaign:** failure, success, not participate
Age	*numeric*	**Client's age**
balance		**Average annual balance (in euros)**
duration		**Duration of the last contact (in seconds)**
campaign		**Number of calls made during the campaign to the client (including the last call)**
previous		**Number of calls made during previous campaigns**

Source: Results obtained from data processing in R Studio by the authors

3.2 Methodology

The methodology adopted in this paper involves three stages:

1. *Preliminary Analysis*, which includes the following steps:
   a) Logical analysis of the data
   b) Analysis of missing data
   c) Analysis of outliers
   d) Analysis of data balancing
2. *Exploratory Analysis*, which involves the following steps:
   a) Visualization and tabulation of the data
   b) Data discretization
   c) Descriptive statistics

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

The typology of the data, whether *categorical* or *numerical*, requires the use of different methods, which is why the analyses will be grouped into *categorical* and *numerical* data analyses.

3. *Data Analysis*, which is the stage that will generate results that respond to the objectives set in this paper. It involves the use of two analytical methods:
   a) *Decision tree*-based data analysis
   b) *Logistic regression*-based data analysis

The first two stages ensure the quality of the inputs for the methods used in the third stage, a fundamental condition for obtaining high-quality results.

3.2.1 Preliminary data analysis

a. Logical data analysis

Logical data analysis is a preliminary step in preparing a dataset for statistical processing. At this stage, we aimed to identify any illogical data associations that could distort the analysis. For example, identifying an age that does not correspond to a completed level of education, such as 18 years old with tertiary education as the highest level completed. This type of logical analysis is typically conducted for demo-socio-economic variables, which are logically strongly correlated with other variables in the database. These analyses rely on examining bivariate tables (crosstabs) or scatterplot graphs.

After data verification, the only issue identified was the presence of missing values for certain variables, which were addressed in the *Missing Data Analysis and Treatment* stage.

b. Analysis and treatment of missing data

Missing values can significantly affect the quality and reliability of the analysis results, directly proportional to their proportion in the dataset. Once missing data is identified, we have two options: *elimination* or *substitution* with logically generated data or using certain numerical imputation methods. *Eliminating records* with missing data can cause major changes in the *structure of the database* and in determining the *importance of explanatory variables in modeling*. Therefore, in most cases, the preferred approach is to use methods for *completing the missing data while retaining the records*.

The only variable that presents unknown values is the *education level* variable.

Therefore, I opted for the missing data imputation method based on *decision trees*. For this, I considered the *education level* as the target variable and the following explanatory variables: *Job*, *Marital*, *Age*, and *Deposit*. *In the first stage*, I split the dataset into two subsets: *the first contained the records for which the education level is known* (43.354), and *the second subset contained the records for which the education level is unknown* (1.857). *In the second stage*, the *decision tree will be trained on the first subset*, which includes the first three levels of the education variable (*primary*, *secondary*, and *tertiary*), predicting the unknown values from the second subset. These predicted values will then be used to replace the missing data.

Table 2 presents the frequency distribution of the *Education* variable, grouped by the *Job* variable, *before* and *after* replacing the missing data using decision trees.

Table 2: Frequency distribution of the Education variable grouped by the Job variable before and after replacing missing data using decision trees

Occupation	Before replacing missing data				After replacing missing data		
	Education level			Mising Data	Education level		
	Primary	Secondary	Tertiary		Primary	Secondary	Tertiary
Blue-collar	3758	5371	149	454	3758	5825	149
Entrepreneur	183	542	686	76	183	542	762
Other	1147	2291	917	438	1147	2729	917
Pink-collar	972	3852	375	195	972	4047	375
Self-employed	130	577	833	39	130	577	872
Technician	158	5229	1968	242	158	5471	1968

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

White-collar	503	5340	8373	413	503	5340	8786

Source: Results obtained from data processing in R Studio by the authors

From the table, we observe that all 454 individuals working in the production sector (blue-collar) with an unknown education level have *secondary education* as their highest level completed. Similarly, all 76 individuals in the entrepreneur category with an unknown education level have *tertiary education* as their highest level completed. For the pink-collar category, 195 individuals, along with 438 individuals categorized as "other" and 242 individuals from the technician category, have *secondary education* as their highest level completed. The last two categories, self-employed (39 individuals) and white-collar (413 individuals), have *tertiary education* as their highest level completed.

The impact of replacing missing values for the *Education level* variable on the distribution of the target variable, *Deposit*, is presented in Table 3.

It can be observed that out of the 252 individuals who opened a term deposit following the campaign, 174 out of the 1,857 individuals, for whom the education level was unknown, have completed secondary education, and only 78 have completed tertiary education. Among the 1.605 individuals who *did not open a term deposit following the campaign*, for whom the education level was unknown, 1.157 have secondary education and only 450 have tertiary education.

Table 3: Frequency distribution of the deposit variable based on the Education variable before and after replacing missing data using decision trees

Deposit	**Before replacing missing data**				**After replacing missing data**		
	Education level				**Education level**		
	Primary	**Secondary**	**Tertiary**	**Mising data**	**Primary**	**Secondary**	**Tertiary**
No	6260	20752	11305	1605	6260	21907	11755
Yes	591	2450	1996	252	591	2624	2074

Source: Results obtained from data processing in R Studio by the authors

c. Outlier analysis of the data

While outlier analysis is meaningless for categorical variables, it is necessary for numerical variables. Identifying and replacing outliers for numerical variables reduces the degree of dispersion of the variables, which has positive effects on the modeling process.

To identify outliers for the variables age, balance, duration, previous, and campaign, we will use the band method. Values that fall outside the interval $[\bar{x} \pm 2.5s]$ are considered outliers.

Table 4: Identification of outliers for numerical variables

	Min	**Max**	**Mean**	**Std. Dev**	**L_{inf} =($\bar{x}$ - 2.5s)**	**No. $Outl_{inf}<L_{inf}$**	**L_{sup} =($\bar{x}$ + 2.5s)**	**No. $L_{sup}<Outl_{sup}$**	**% Outlin Total**
Age	18	95	40.94	10.61	14.42	0	67.47	634	**1.4%**
Balance	-8019	102127	1362	3044.76	-6249.90	2	8973.90	997	**2.2%**
Duration	0	4918	258.2	257.52	-385.60	0	902.00	1412	**3.1%**
Campaign	1	63	2.76	3.09	-4.97	0	10.49	1196	**2.6%**
Previous	0	275	0.58	2.3	-5.17	0	6.33	787	**1.4%**

Source: Results obtained from data processing in R Studio by the authors

The outlier values will be replaced as follows: *inferior outliers* (<Linf) and *superior outliers* (>Lsup) will be replaced with the *lower limit*, Linf = ($\bar{x}$ - 2.5s), and the upper limit, Lsup = ($\bar{x}$ + 2.5s),

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

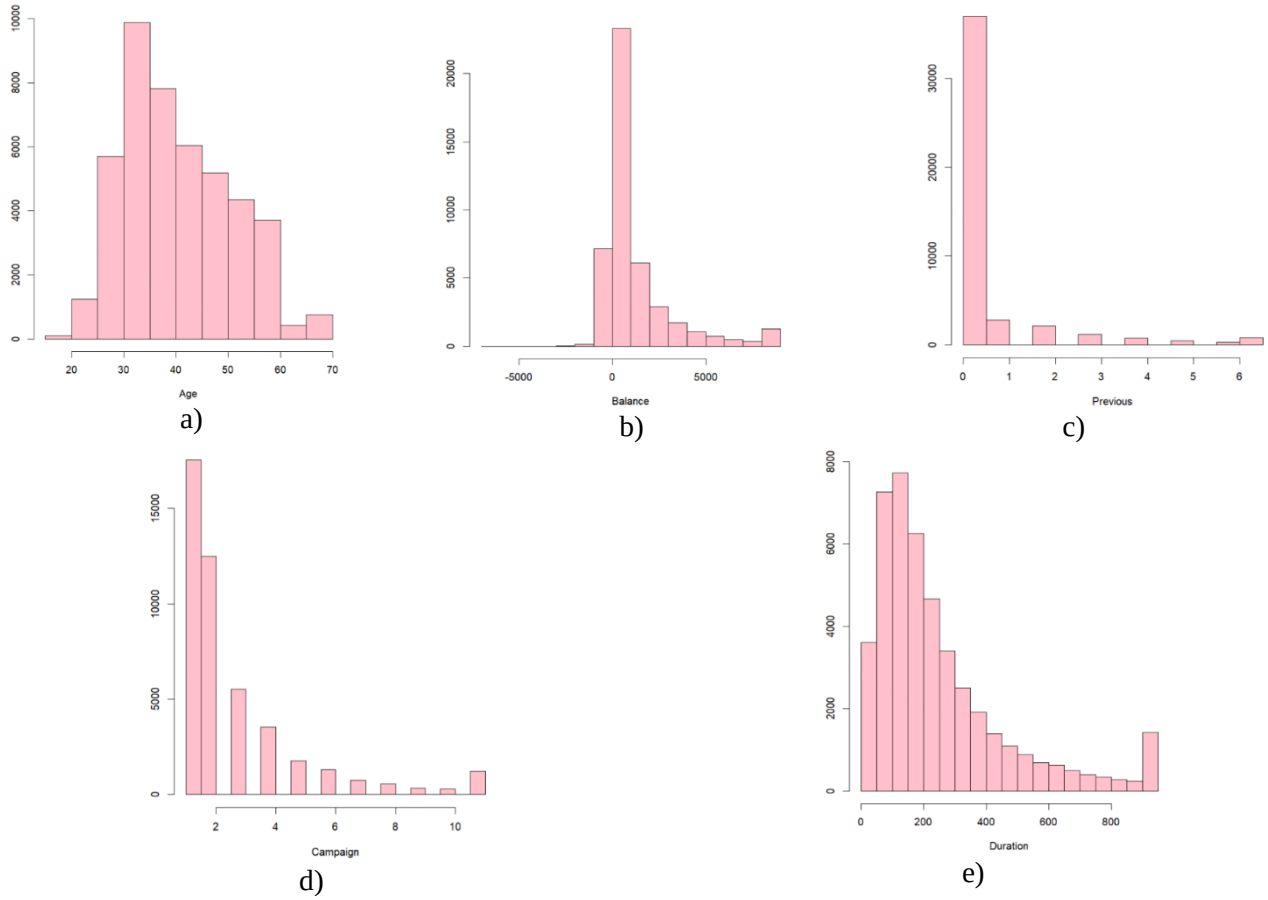
of the normal value range [$\bar{x} \pm 2.5s$]. The distributions obtained after replacing the outliers are presented in Figure 1.

The variable "Age" is a *continuous numerical* variable which, after replacing the outliers according to the band principle, has shifted from an age range of [18; 95] years to [18; 67.47] years. The frequency distribution of the variable "Age" is shown in *Figure 1.a*. These results align with the banking institutions' policy of targeting a demographic segment represented by the economically active population, aged 18-65, who are more likely to open term deposits. The "Age" variable presents 634 outliers, which account for 1.4% of the total records.

The variable Balance (annual average balance) is a *continuous numerical* variable which, after replacing the outliers, changed its range from [-8019; 102127] to [-6249.90; 8973.90], thus increasing its homogeneity level. The *Balance* variable presents 997 outliers, representing 2.2% of the total records. Of the 999 outliers, 849 belong to the *majority class*, represented by subjects who were not affected by the bank's campaign and did not open a term deposit. The frequency distribution of the *Balance* variable is shown in *Figure 1.b*.

The variable *Duration (call duration)* is a positive continuous numerical variable which, after replacing the 1347 outliers, changed its range from [0;4918] to [-385.60; 902]. Given the positive nature of the variable, the resulting range after replacing the outliers becomes [0; 902], meaning that the new variable will have a much lower level of homogeneity. Of the 1412 outliers, representing 3.1% of the total records, 538 belong to the majority class of the deposit variable, represented by clients for whom the bank's campaign had no effect. It is worth mentioning that this variable has the highest percentage of outliers, 3.1%, but still *falls below the maximum outlier threshold of 5%*. The frequency distribution of the *Balance* variable is shown in *Figure 1.c*.

Figure 1: Frequency distributions of the variables Age (a), Balance (b), Duration (c), Campaign (d), and Previous (e) after identifying and replacing the outliers



a) b) c) d) e)

Source: Results obtained from data processing in R Studio by the authors

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

The variable *Campaign (the number of calls made in the campaign to a given client)* is a *positive discrete numerical variable*, initially ranging from [1;63], which, after replacing the outliers, and considering its discrete and positive nature, becomes [0;11]. The variable with the outliers replaced is characterized by a much lower level of homogeneity compared to the initial variable. For the Campaign variable, we identified 1196 outliers, representing 2.6% of the total records, of which 1109 belong to the majority class of the deposit variable, represented by clients who did not open a term deposit after the bank's campaign. The frequency distribution of the *Campaign* variable is shown in *Figure 1.d.*

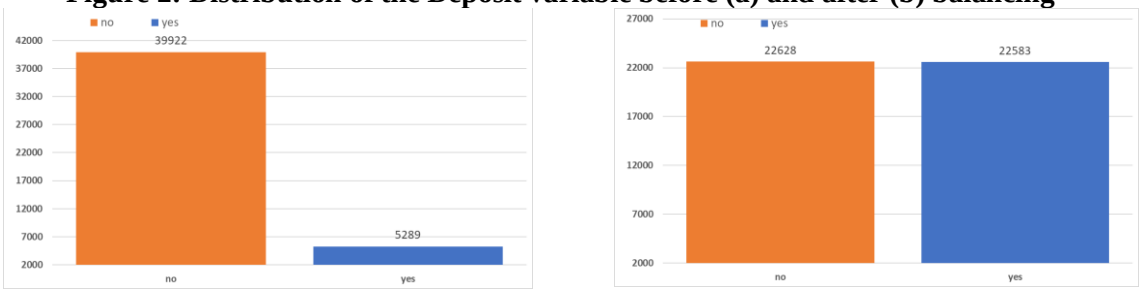
The variable *Previous (the number of calls made in previous campaigns)* is a positive *discrete numerical variable*, initially ranging from [0;275], which, after identifying and replacing the outliers, was reduced to [0;7]. For the Previous variable, we identified 787 outliers, representing 1.4% of the total records, of which 514 belong to the majority class of the deposit variable, represented by clients who did not open a term deposit after the bank's campaign. The frequency distribution of the *Previous* variable is shown in *Figure 1.e.*

d. Data balancing

Data balancing refers to adjusting the distribution of the *target variable*, which is usually a binary variable, so that approximately equal proportions of its two values are obtained. This is an essential condition to prevent influencing the classification/modeling process, where the number of misclassified records is of major importance. If the target variable is *highly imbalanced*, exceeding a ratio of 2:1 (67% to 33%), it can lead to naive classification by the algorithm, which, based on the confusion matrix, may default to *an accuracy of 66%*, classifying all records into the majority category.

Data balancing seeks to correct this imbalance by modifying the dataset, resulting in obtaining approximately equal weights for the classes of the target variable. In this case, we chose to use the data balancing algorithm implemented through the *ROSE (Random Over-Sampling Examples)* function, which generates synthetic cases by interpolating records (instances) from the minority class. It employs the Kernel Density Estimation (KDE) method, which estimates probability distributions used in generating synthetic cases. ROSE can also be successfully used to reduce overfitting, a situation that occurs when the same instances are replicated.

Figure 2: Distribution of the Deposit variable before (a) and after (b) balancing



a) b)

Source: Results obtained from data processing in R Studio by the authors

Figure 2.a. shows that the initial dataset was highly imbalanced, with 39,922 customers not opting for a term deposit, while 5.289 customers chose to open a term deposit as a result of the bank's campaign, resulting in a ratio of approximately 7:1. *Figure 2.b.* illustrates the effect of balancing through the ROSE algorithm, which reduced the imbalance to a 1:1 ratio (22,628 customers who did not opt for a term deposit and 22,583 customers who did), while maintaining the original data volume of 45.211 records. Therefore, we can conclude that, to achieve this result, the minority class was populated with synthetic records, while the majority class was depopulated by the random removal of records.

3.2.2 Exploratory data analysis

3.2.2.1 Exploratory analysis of categorical data after balancing

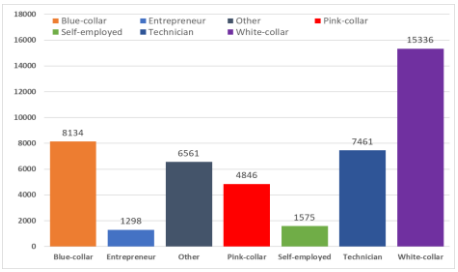
Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

a. Visualization of categorical data after balancing

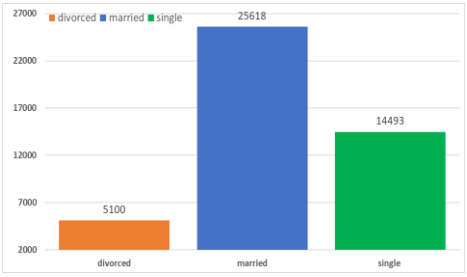
We will use *data visualization* to characterize the categorical variables. Based on this, we can state that most of the bank's customers who participated in the campaign:

- have a *job* that falls into the *white-collar* category (Figure 3.a.);
- *are married* (Figure 3.b.);
- have completed *secondary* education (Figure 3.c.);
- are *good payers* (Figure 3.d.);
- have a *mortgage loan* (Figure 3.e.);
- do not have a *personal loan* (Figure 3.f.);
- were contacted by phone during the campaign on *Wednesdays* (Figure 3.g.);
- were contacted in *May* (Figure 3.h.);
- *did not participate in previous bank campaigns* (Figure 3.i).

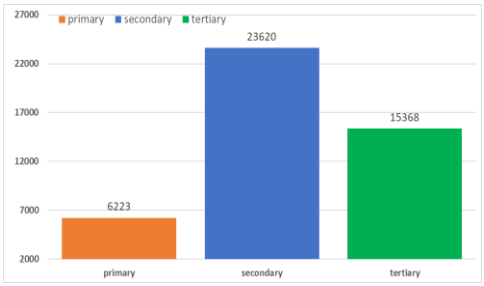
Figure 3: Frequency distributions of the variables Job (a), Marital Status (b), Education Level (c), Default (d), Housing (e), Loan (f), Day of Week (g), Month (h), Poutcome (i), and Deposit (j), target variable



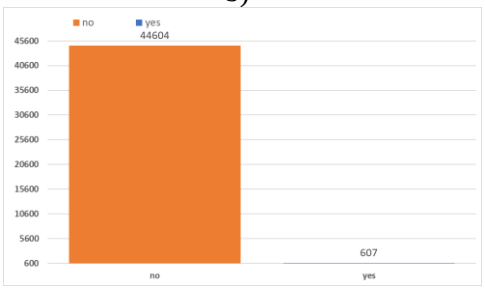
a)



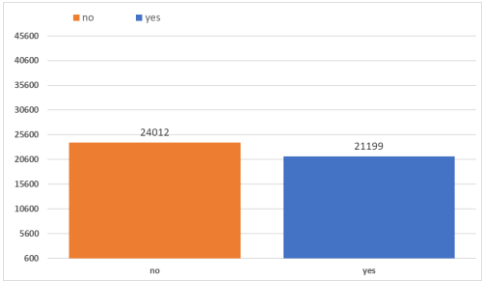
b)



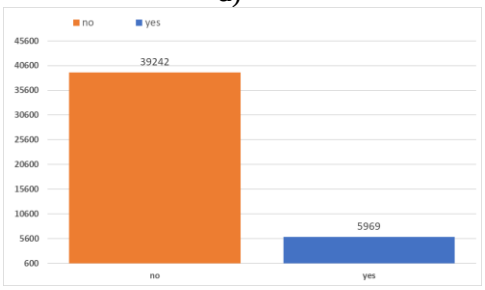
c)



d)

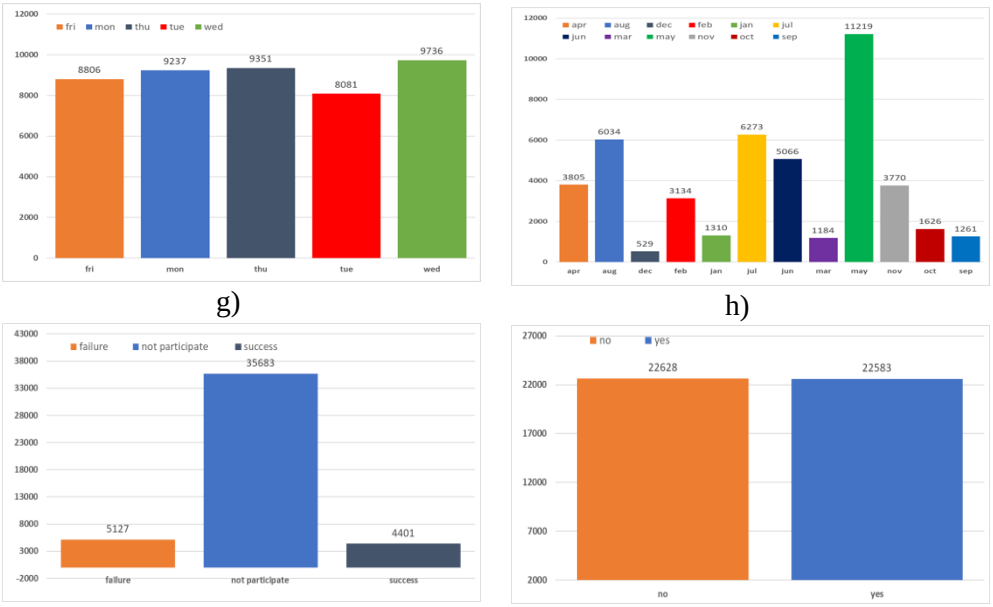


e)



f)

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034



g) h)

i) j)

Source: Results obtained from data processing in R Studio by the authors

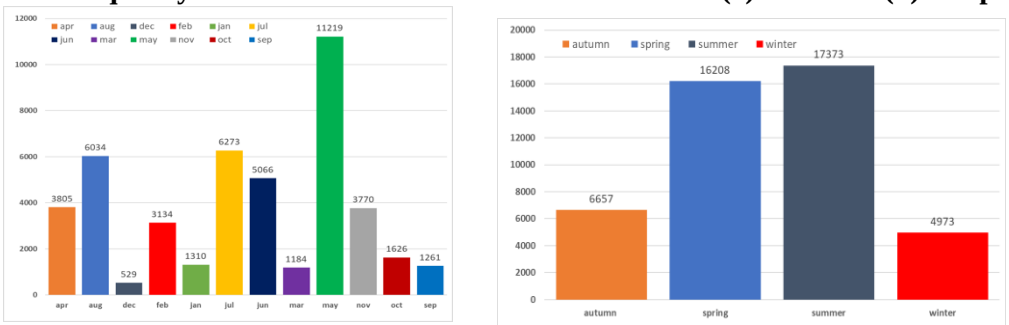
b. Discretization of categorical data after balancing

Discretization is a *grouping process specific to continuous numerical variables*, performed based on certain criteria aimed at *preserving relationships with other variables* in the database. This process reduces the complexity of both the analysis and the resulting outputs.

Although primarily used for *continuous numerical variables*, discretization can also be applied to numerical or categorical variables with a large number of value classes. The immediate effect is the compression of value classes, which decreases the complexity of the modeling process by significantly reducing the number of operations performed and, consequently, the processing time required. Discretization also impacts the transparency and accessibility of the results.

Due to the fact that the *Month* variable has 12 value classes, as well as significant differences in the size of each class (Figure 4.a), compressing the classes into seasons will reduce the complexity of the variable and, implicitly, of the modeling process in which it is involved. Additionally, it can be observed that the differences in frequencies between seasons are smaller in the compressed form, suggesting an increase in the homogeneity of the variable (Figure 4.b).

Figure 4. Frequency distribution of the Month variable before (a) and after (b) compression



Source: Results obtained from data processing in R Studio by the authors

c. Descriptive statistics for categorical data after balancing

The only descriptive statistical indicators, depending on the typology of the categorical variable (nominal or ordinal), are represented by frequencies, modal values, medians, and quantiles.

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

All categorical variables in the database are *nominal*, except for the *Education Level* variable. Therefore, all descriptive statistics can be accessed through the graphical representations presented above.

3.2.2.2 Exploratory analysis of quantitative data after balancing

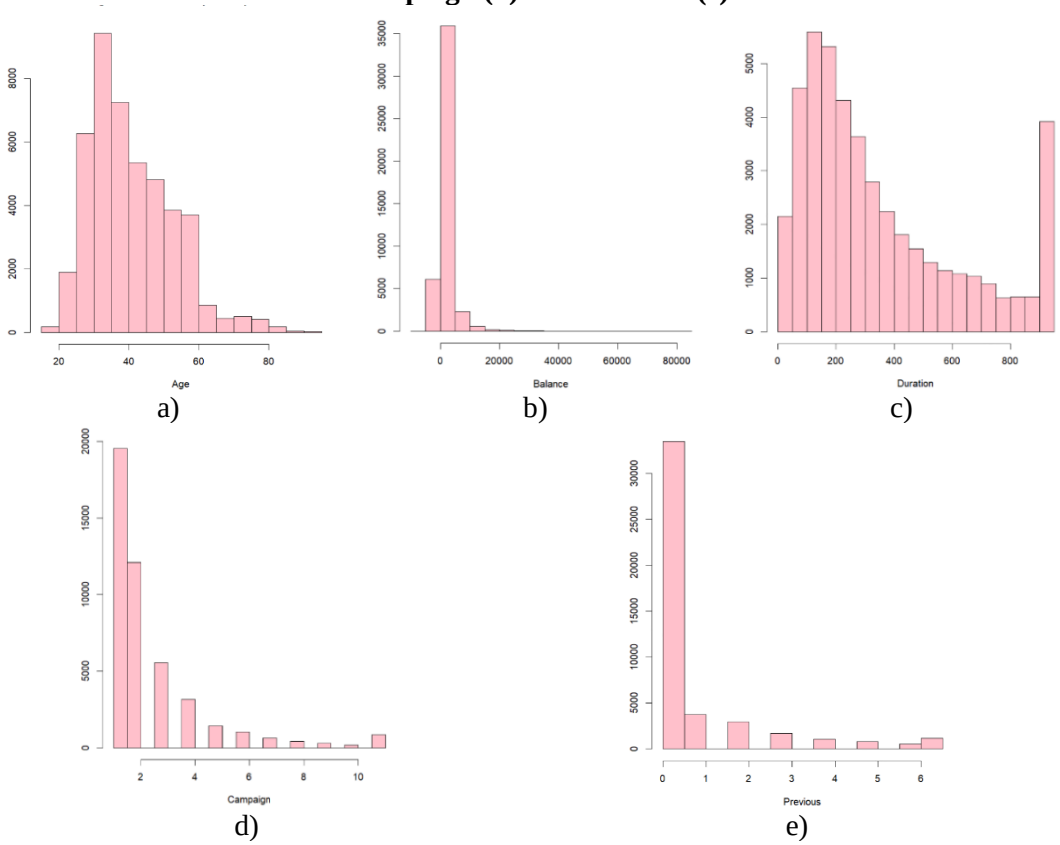
a.Visualization of quantitative data after balancing

By comparing the initial distributions of numerical variables (Figure 1) with the distributions after balancing (Figure 5), we can observe the influence of balancing using the ROSE algorithm.

As a general observation, we can state that the use of the ROSE algorithm for balancing did not significantly alter the initial distributions.

However, it is noteworthy that the ROSE algorithm has a positive effect on the variables *Age*, *Balance*, *Campaign*, and *Previous*, minimizing the accumulation of values in the tails of the distributions caused by the replacement of outliers. In contrast, for the variable *Duration*, the effect is reversed, with an accentuation of the right tail of the distribution.

Figure 5: Frequency distribution of the variables: Age (a), Balance (b), Duration (c), Campaign (d) and Previous (e)



a) b) c) d) e)

Source: Results obtained from data processing in R Studio by the authors

Additionally, from the analysis of the distributions in Figure 5, we can observe the following:

- for the variable *Age*, most individuals are aged between 30 and 39 years;
- for the variable *Balance* (average annual balance), a large number of individuals have an average annual balance close to 0;
- for the variable *Duration* (call duration), most calls last between 100 and 200 seconds;
- for the variable *Campaign* (number of calls made during the campaign to the respective client), most individuals received between 1 and 2 calls during the campaign;
- for the variable *Previous* (number of calls made during previous campaigns), the majority of the bank's clients are participating in the bank's campaign for the first time.

b. Discretization of numerical variables

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

Discretization, in the case of numerical variables, involves grouping data into homogeneous intervals based on specific criteria such as the frequency distributions of the variables, relationships with other values in the database, and the relationship with the target variable, among others. Ideally, after discretization, the variable's distribution and its relationships with other variables in the database should be preserved.

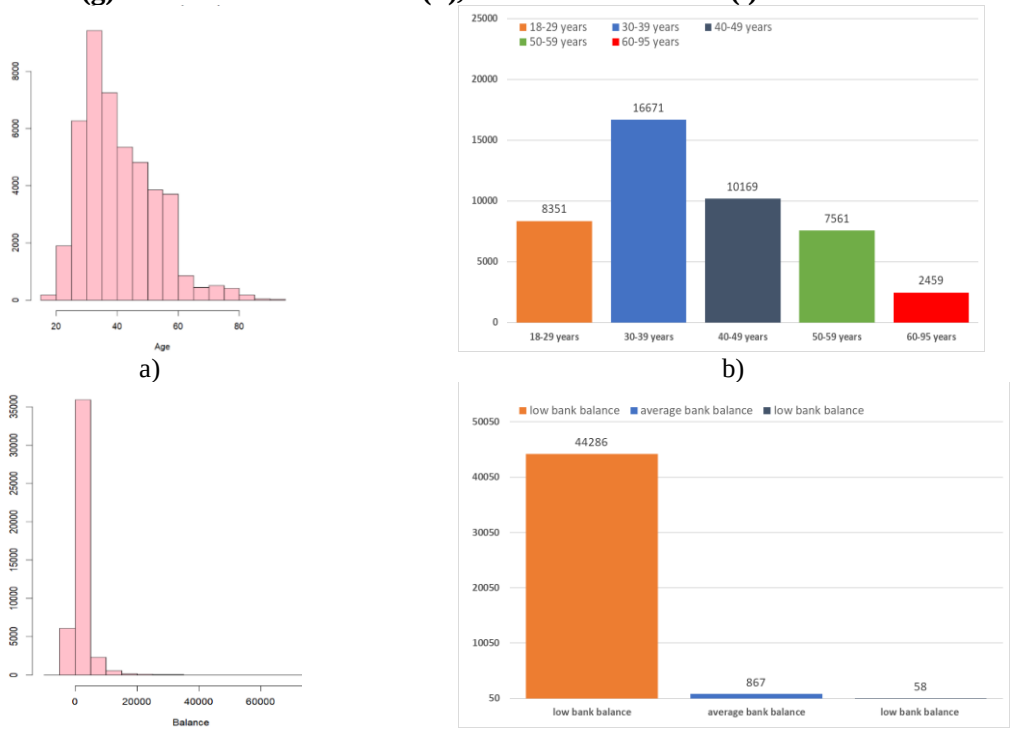
The main effects of discretization include reducing the complexity of variables, models, and calculations. Additionally, a secondary effect of discretization is the mitigation of the impact of errors on the analysis and modeling process, as erroneous values are less likely to influence the intervals in which they are placed through discretization. Once variables are divided into intervals, the model can more easily identify general trends or relationships between variables.

The discretization of numerical variables will be performed using the supervised MDLP (Minimum Description Length Principle) method. The MDLP principle is based on the idea of identifying discretizations that minimize the description of the data, so that they are represented in a compact format without significant information loss. Essentially, MDLP seeks to find a discretization that compresses the data to the maximum while reducing the complexity of the model used to describe it. In practice, the MDLP method divides continuous variables into discrete intervals using criteria that optimize data compression.

In the discretization process with MDLP, the optimal cut points are identified to transform a continuous variable into discrete intervals. In other words, this method involves finding an optimal discretization that minimizes the amount of information needed to describe the continuous variable in relation to the target variable.

The frequency distributions of the numerical variables before and after discretization with the MDLP method are shown in Figure 6.

Figure 6: Frequency distributions of the variable Age before (a) and after discretization (b), Balance before (c) and after discretization (d), Duration before (e) and after discretization (f), Campaign before (g) and after discretization (h), and Previous before (i) and after discretization (j)



a) b)

c) d)

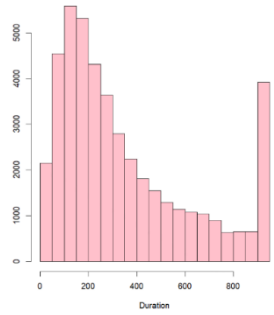
Source: Results obtained from data processing in R Studio by the authors

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

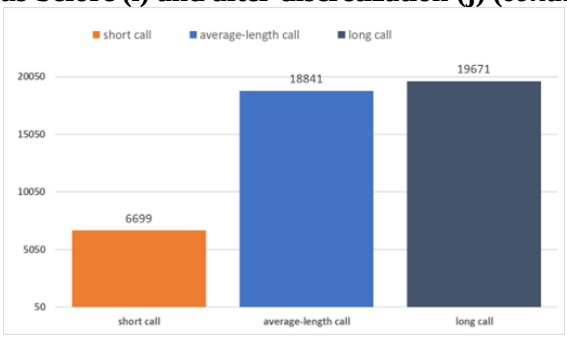
From the analysis of the distributions of the variables before and after discretization using the MDLP method, we can observe the following:

- For the variable *Age*, after discretization, the variable's distribution does not change, with most people being between 30 and 39 years old.
- For the variable *Balance*, the majority of clients have an average balance close to 0, and the distribution of this variable remains asymmetric even after discretization.
- For the variable *Duration*, most people contacted have an average call duration between 100 and 200 seconds (approximately 2-4 minutes), which can also be observed before discretization.
- For the variable *Campaign*, the number of calls made during the current campaign shows an asymmetric distribution both before and after discretization.
- For the variable *Previous*, the number of calls made during previous campaigns also shows an asymmetric distribution both before and after discretization, with most people receiving at most 5 calls.

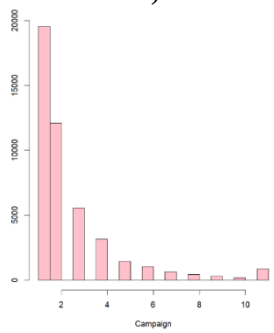
Figure 6: Frequency distributions of the variable Age before (a) and after discretization (b), Balance before (c) and after discretization (d), Duration before (e) and after discretization (f), Campaign before (g) and after discretization (h), and Previous before (i) and after discretization (j) (*continuation*)



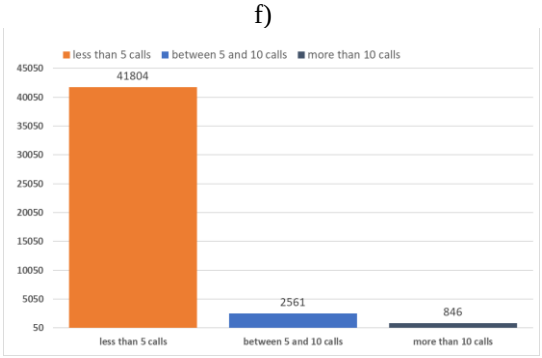
e)



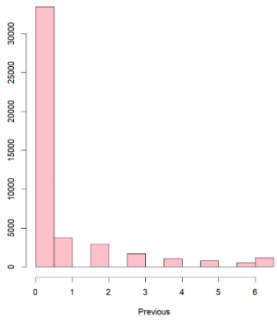
f)



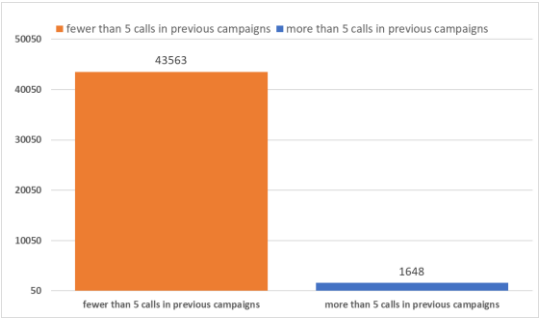
g)



h)



i)



j)

Source: Results obtained from data processing in R Studio by the authors

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

c. Descriptive statistics of quantitative data

For quantitative variables, descriptive statistics provide an accurate description of the characteristics of the variables. *Univariate descriptive statistics*, *shown in Table 5*, although representing the simplest way to characterize the *mean level, degree of dispersion, skewness*, and *kurtosis* of a series, offer important insights regarding the health of a variable.

In addition to *univariate descriptive statistics*, the correlation/association of the *explanatory variables* with the *objective variable*, shown in Table 6, is of particular importance.

Table 5: Descriptive analysis of numerical variables after replacing outliers

	Age	Balance	Duration	Campaign	Previous
Median	39.00	555.00	260.00	2.00	0.0
Min	18.00	-3697.77	0.00	1.00	0.0
Max	95.00	6118.00	772.33	8.04	3.71
1st Qu.	32.00	121.00	142.00	1.00	0.0
3rd Qu	49.00	1758.00	500.00	3.00	1.00
Mean	41.21	1273.02	334.28	2.32	0.60
Skwness	0.85	1.60	0.68	1.75	1.78
Kurtosis	0.62	1.71	-0.84	2.64	1.72
Std. Dev	12.03	1710.24	238.98	1.77	1.16

Source: Results obtained from data processing in R Studio by the authors

As a general observation, we can state that in most cases, the series exhibited upper outliers, and replacing these with the upper tolerance band values, $\bar{x} + 2.5s$ (Table 4), led to an increase in the thickness of the right tails of the distributions, thereby inducing right skewness.

Except for the Age variable, all other variables exhibit a high level of variation, which could negatively affect the modeling process. For the *Age*, *Balance*, *Campaign*, and *Previous* variables, the level of kurtosis is slightly above average, while for the *Duration* variable, we observe a slightly flattened distribution.

Table 6: Association Coefficients of the Target Variable with the Explanatory Variables

	Cramér's V/Eta coefficient	p-value		Cramér's V/Eta coefficient	p-value
Job	0.19	0.00	*Season*	0.15	0.00
Marital	0.11	0.00	*Poutcome*	0.61	0.00
Default	0.04	0.00	*Education*	0.11	0.00
Housing	0.12	0.00	*Age*	0.04	0.00
Loan	0.05	0.00	*Balance*	0.02	0.00
Day_of_week	0.29	0.00	*Campaign*	0.01	0.00
Duration	0.23	0.00	*Previous*	0.003	0.00

Source: Results obtained from data processing in R Studio by the authors

The association coefficients between the *target variable* and the *explanatory variables* are all significant and have low values (Table 6). The strongest associations with the *target variable* are observed for the variables *Poutcome* (0.61), followed at a considerable distance by *Day of week* (0.29), *Duration* (0.23), and *Loan* (0.05). Partially, these findings are confirmed by the ranking of the explanatory variables' importance using the *Random Forest* algorithm (Figure 12).

4. Research results

The objective of the research is to identify the *determinants in a telemarketing campaign* that *influence the decision* of clients of a Portuguese bank to open a term deposit at the end of the campaign.

Methodologically, this is reduced to studying the interaction between the independent variables: *default, housing, loan, marital, job, education, day_of_week, month, poutcome, age, balance, duration,*

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

campaign, *previous*, and the dependent variable, which records whether *the client opened a term deposit at the end of the campaign.*

The two methods used are *logistic regression* and *classification trees*. Logistic regression allows us to predict the outcome of a binary variable, which has only two categories (yes/no), based on the influence of independent variables. Classification trees are used to group records (instances) into classes based on their characteristics. Based on the results obtained, the hypotheses will be confirmed or disproven based on the target variable, whether or not the person accepted to open a term deposit as a result of the campaign.

To achieve the modeling objectives, the dataset was split into two subsets: *the training set*, representing 70% of the total available data, and *the testing set*, representing the remaining 30% of the data.

4.1 Logistic regression

Since the goal of the analysis is to classify clients into two categories—*those who accepted the bank's offer to open a term deposit* and *those who did not*—one of the most suitable methods for this type of analysis is *logistic regression*. This data mining technique is an appropriate method for addressing problems with binary outcome variables (yes/no). Before applying logistic regression, certain variables underwent some modifications. The "no" and "yes" levels of the variables *default, housing, loan*, and *deposit* were coded with the values *0* and *1*. Following the application of logistic regression, the data in Table 7 were obtained.

The estimated equation of the logistic model is:

$$g(x) = -1.30 - 0.43\cdot age_{30\text{-}39} - 0.49\cdot age_{40\text{-}49} - 0.51\cdot age_{50\text{-}59} + 0.65\cdot age_{60\text{-}95} - 0.22\cdot marit_{marr} + 0.16\cdot marit_{sing} + \\ + 0.30\cdot educ_{sec} + 0.58\cdot educ_{tert} - 0.34\cdot default_1 - 0.74\cdot housing_1 - 0.51\cdot loan_1 + 1.91\cdot durat_{aver\ length\text{-}call} + \\ + 4.03\cdot durat_{long\ call} - 0.52\cdot poutc_{not\ part} + 2.28\cdot poutc_{succ} - 0.50\cdot seas_{spring} - 0.9\cdot seas_{summer} - 0.32\cdot seas_{winter}$$

Table 7: Results of the logistic regression model

Variable	Beta	Std. Err.	Wald stat.	p-val.	Confid. Int.	Odds Ratio
(Intercept)	-1.30	0.09	-13.25	0.00	[0.22, 0.33]	0.27
$age_{30\text{-}39}$	-0.43	0.03	-11.63	0.00	[0.60, 0.70]	0.65
$age_{40\text{-}49}$	-0.49	0.04	-11.35	0.00	[0.56, 0.66]	0.61
$age_{50\text{-}59}$	-0.51	0.04	-10.75	0.00	[0.55, 0.66]	0.60
$age_{60\text{-}95}$	0.65	0.07	8.52	0.00	[1.65, 2.23]	1.92
$marit_{marr}$	-0.22	0.04	-5.48	0.00	[0.74, 0.86]	0.80
$marit_{sing}$	0.16	0.05	3.56	0.00	[1.07, 1.29]	1.17
$educ_{sec}$	0.30	0.03	7.75	0.00	[1.25, 1.45]	1.35
$educ_{tert}$	0.58	0.04	14.01	0.00	[1.65, 1.94]	1.79
$default_1$	-0.34	0.10	-3.24	0.00	[0.57, 0.87]	0.70
$housing_1$	-0.74	0.02	-25.93	0.00	[0.44, 0.50]	0.47
$loan_1$	-0.51	0.04	-13.73	0.00	[0.55, 0.64]	0.60
$durat_{aver\ length\text{-}call}$	1.91	0.05	34.35	0.00	[6.09, 7.58]	6.78
$durat_{\ long\ call}$	4.03	0.05	70.08	0.00	[49.12, 61.46]	54.88
$poutc_{not\ part}$	-0.52	0.04	-13.48	0.00	[0.54, 0.63]	0.58
$poutc_{succ}$	2.28	0.07	30.71	0.00	[8.51, 11.39]	9.83
$seas_{spring}$	-0.50	0.04	-11.86	0.00	[8.73, 13.37]	10.76
$seas_{summer}$	-0.90	0.04	-21.68	0.00	[0.57, 0.69]	0.62
$seas_{winter}$	-0.32	0.05	-9.12	0.00	[0.39, 0.48]	0.43

Source: Results obtained from data processing in R Studio by the authors

Based on the results obtained above, we will validate or invalidate the research hypotheses established at the beginning of the analysis as follows:

RH_1: Individuals who have a personal loan are less interested in opening a term deposit compared to those who do not have a loan.

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

The odds ratio associated with those who have a personal loan to open a bank deposit at the end of the telemarketing campaign decreases by 0.60 ($e^{-0.51}$=0.60) compared to those who do not have a loan. In other words, this hypothesis is confirmed.

RH_2: Individuals with higher education show a greater interest in opening a term deposit compared to those with secondary education.

We observe that the odds ratio for individuals with higher education (equivalent to a university degree) to open a term deposit increases by 1.79 ($e^{0.58}$=1.79) compared to those with secondary education. The hypothesis is confirmed.

RH_3: Individuals who responded positively to previous bank campaigns show a greater interest in opening a term deposit.

From Table 7, we observe that all variables have a p-value smaller than 5% for the Wald test statistic, which means their influence is statistically significant and the odds ratio for *a person who purchased banking products* in previous campaigns increases by 9.78 ($e^{2.28}$= *9.78*) compared to those who did not. The hypothesis is confirmed. The importance of the independent variables can be visualized in Table 8.

Table 8: Importance of variables - logistic regression

Variable	Value	Variable	Value	Variable	Value
$age_{30\text{-}39}$	11.63	$housing_1$	25.92	$educ_{sec}$	7.75
$age_{40\text{-}49}$	11.35	$poutcom_{not\ part}$	13.47	$educ_{tert}$	14.00
$age_{50\text{-}59}$	10.03	$poutcom_{succ}$	30.70	$default_1$	3.24
$age_{60\text{-}95}$	8.52	$seas_{spring}$	10.68	$loan_1$	13.72
$marit_{marr}$	5.47	$seas_{summer}$	9.12	$durat_{aver\ length\text{-}call}$	34.33
$marit_{sing}$	3.55	$seas_{winter}$	17.28	$durat_{long\ call}$	70.08

Source: Results obtained from data processing in R Studio by the authors

The two most important variables that influence the dependent variable, a person's decision to open a term deposit, are the duration of the call and the variable representing the outcome of previous marketing campaigns. To a lesser extent, the dependent variable is also influenced by two other variables: the level of education and whether the person has personal loans.

The *omnibus test* is a statistical test used to globally assess whether at least one independent variable has a significant effect on the dependent variable within a model. The hypotheses of the Omnibus test are: *H_0: There is no significant association between the independent variables and the dependent variable. H_1: At least one independent variable has a significant association with the dependent variable.*

Figure 7: Omnibus Test

Likelihood Ratio Test

Model 1:
Data_2$deposit ~ age + marital + education + default + housing + loan + duration + poutcome + season

Model 2:
deposit ~ 1

#Df LogLik Df Chisq Pr(>Chisq)

27 -20040

1 -31338 -26 22595 < 2.2e-16

Source: Results obtained from data processing in R Studio by the authors

We observe that the p-value = 0 < 0.05, so H_0 is rejected. The current model outperforms the null model in explaining the dependent variable.

The Hosmer-Lemeshow test is a statistical method used to assess how well a logistic regression model fits the observed data. It evaluates the difference between the observed and expected frequencies

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

in different population groups based on the probabilities predicted by the logistic regression model. The hypotheses of the Hosmer-Lemeshow test are*: H_0: There are no significant differences between the observed and predicted data (the model is a good fit). H_1: There are significant differences between the observed and predicted data (the model is not a good fit).*

Figure 8: Hosmer-Lemeshow Test

Hosmer and Lemeshow goodness of fit (GOF) test

data: bank.complete$deposit, fitted(model2)
p-value = 0.39

Source: Results obtained from data processing in R Studio by the authors

Since the p-value = 0.39 > 0.05, we do not reject H_0. There are no significant differences between the observed and predicted values, so we can conclude that the estimates fit the data at an acceptable level. Therefore, the model is adequate.

The performance of the model will be quantified using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) presented in Table 9. The ROC curve displays the true positive rate and false positive rate corresponding to different rankings of the records. An AUC value close to 1 indicates a very good model, while a value close to 0 indicates a poor model.

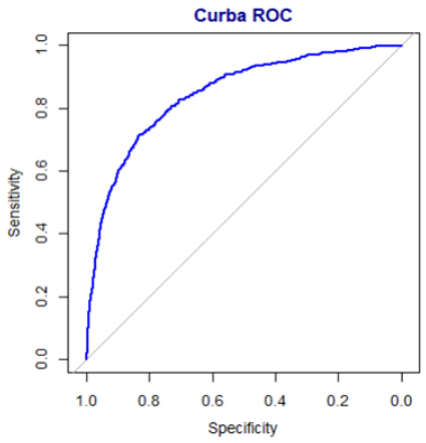
Table 9: AUC for the logistic regression model

	Training set	Testing set
Area unde curve (AUC)	0.8733	0.8743

Source: Results obtained from data processing in R Studio by the authors

Since the area under the ROC curve (which actually provides a measure of the quality of a classification model in a binary problem) has values greater than 0.8, both for the training set (0.8733) and the testing set (0.8743), we can state that the model has very good discrimination (the ability of the models to clearly and accurately distinguish between different categories or classes of data). Confusion matrix: We will make the prediction of the data by creating the confusion matrix.

Figure 9: ROC Curve (a) and Confusion Matrix (b) resulting from the application of logistic regression



a)

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 17597  4256
         1  5031 18327

               Accuracy : 0.7946
                 95% CI : (0.7908, 0.7983)
    No Information Rate : 0.5005
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5892

 Mcnemar's Test P-Value : 9.619e-16

            Sensitivity : 0.7777
            Specificity : 0.8115
         Pos Pred Value : 0.8052
         Neg Pred Value : 0.7846
             Prevalence : 0.5005
         Detection Rate : 0.3892
   Detection Prevalence : 0.4834
      Balanced Accuracy : 0.7946

       'Positive' Class : 0
```

b)

Source: Results obtained from data processing in R Studio by the authors

The resulting model has an accuracy of 79.46%, with a classification error rate of 20.54%. From the confusion matrix, we can see that out of the total 22,628 cases in category 0 (the person did not accept the bank's offer to open a term deposit), 17,597 were correctly classified, while 5,031 were

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

misclassified. For category 1, out of the total 22,583 cases, 18,327 were correctly classified, and 4,256 were misclassified.

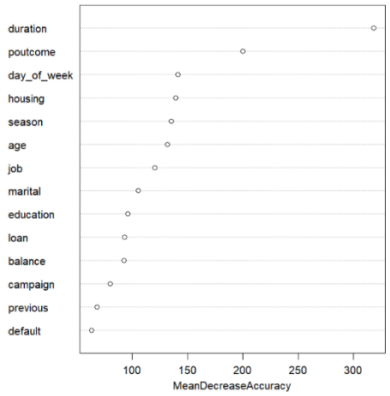
4.2 Decision trees

Decision trees, like logistic regression, are a technique used for classification problems, capable of capturing non-linear relationships between exploratory variables and the target variable. For constructing the decision tree, we will use the *CART* (Classification and Regression Trees) algorithm, a machine learning method used for both classification and regression problems.

Since the dataset is large, with 15 variables, and including all of them could result in a tree that is extremely complex and difficult to interpret, we will select the variables to include in the tree based on their importance using *the Random Forest algorithm*. The results obtained are presented in Figure 10. The variables are displayed in descending order based on their accuracy index value.

We observe that the most important variable is *duration*, followed by *poutcome* and *day*. Based on these results, we decide to include *the top 7 most important variables* in the tree.

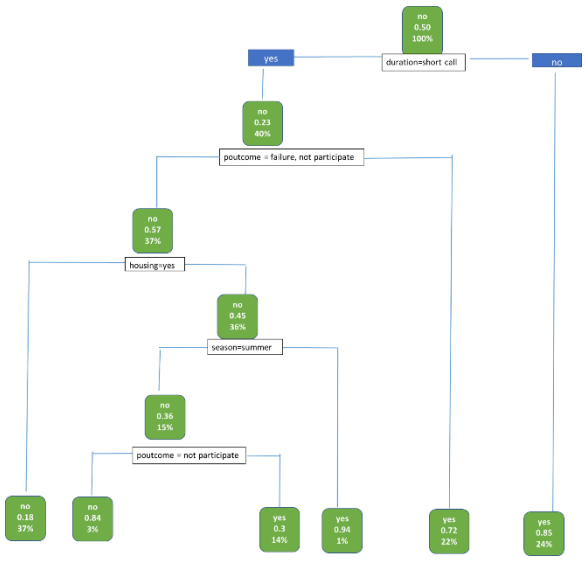
Figure 10: Variable Importance Using Random Forest



Source: Results obtained from data processing in R Studio by the authors

The results obtained from applying the CART algorithm are presented in Figure 11.

Figure 11: Decision Tree (a) built using the CART algorithm and the corresponding Confusion Matrix



a)

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  18084  4939
       yes  4544 17644

               Accuracy : 0.7903
                 95% CI : (0.7865, 0.794)
    No Information Rate : 0.5005
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5805

 Mcnemar's Test P-Value : 5.211e-05

            Sensitivity : 0.7992
            Specificity : 0.7813
         Pos Pred Value : 0.7855
         Neg Pred Value : 0.7952
             Prevalence : 0.5005
         Detection Rate : 0.4000
   Detection Prevalence : 0.5092
      Balanced Accuracy : 0.7902

       'Positive' Class : no
```

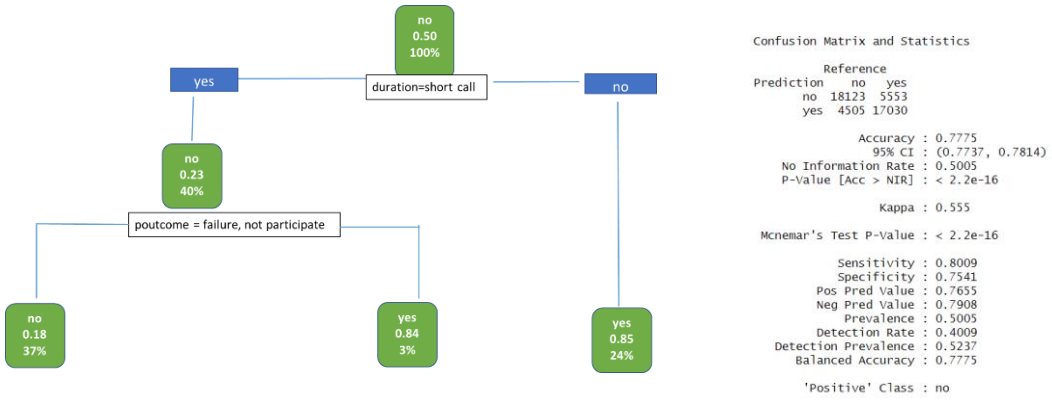
b)

Source: Results obtained from data processing in R Studio by the authors

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

From the confusion matrix (Figure 11.b), we observe that out of the total 22,628 people who responded "no," 18,084 were correctly classified and 4,544 were misclassified. Out of the total 22,583 people who accepted the bank's offer, 17,644 were correctly classified, and 4,939 were misclassified. The accuracy of the tree is 79.03%. Since we want to see if we can use a much simpler decision tree with similar performance, we decide to perform pruning on the tree. After pruning, the tree is presented in Figure 12.

Figure 12: Pruned Decision Tree (a) built using the CART algorithm and the corresponding Confusion Matrix



a) b)

Source: Results obtained from data processing in R Studio by the authors

The pruned tree (Figure 12.b.), although it has a lower accuracy (77.75% < 79.03%) compared to the initial one, is much simpler. The loss in accuracy is much smaller compared to the advantages brought by using a simpler decision tree. The algorithm determined that the other variables did not contribute significantly to the predictive power of the model. From Figure 12.a., we observe that the most important indicator is the duration of the call, and the second indicator is *poutcome* (the result of previous marketing campaigns). From the confusion matrix (Figure 12.b.), we observe that out of the total 22,628 people who responded "no," 18,123 were correctly classified and 4,505 were misclassified. Out of the total 22,583 people who accepted the bank's offer, 17,030 were correctly classified, and 5,553 were misclassified.

RH_4: The majority of people who did not participate in the bank's previous campaigns or who participated but did not purchase a banking product, did not open a term deposit.

According to the resulting classification tree, we observe that 37% of the people who participated for the first time in a bank campaign or who had participated previously but did not access any financial service, did not open a term deposit. The hypothesis is confirmed.

4. Conclusions

The main objective of this study was to identify the factors influencing an individual's decision to open a term deposit. The input data included both marketing attributes, such as the number of calls made, the outcome of the previous campaign, the call duration, and characteristics describing the subject, such as education level, marital status, and occupation. Since the analysis is one of classification, we aim to classify individuals into two categories: those who accepted and those who did not accept the bank's offer, using data mining methods, logistic regression, and decision trees. Secondary objectives were also set to provide a more realistic picture of the analysis conducted.

Based on the results obtained, all four research hypotheses were validated. After applying data mining methods, it was observed that the two most important variables influencing the dependent

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

variable, an individual's decision to open a term deposit, are the call duration and the outcome of the previous marketing campaign.

Regarding the performance of the two classification methods used, it was found that logistic regression had the highest accuracy value, 79.46%. After applying logistic regression, it was observed that individuals with a personal loan were less interested in opening a term deposit compared to those without a loan. Individuals with higher education showed more interest in opening a term deposit compared to those with secondary education, and those who responded positively to previous bank campaigns showed more interest in opening a term deposit. According to the confusion matrix, out of the total of 22,628 cases in category 0 (individuals who did not accept the bank's offer to open a term deposit), 17,597 were correctly classified, and 5,031 were incorrectly classified. For category 1, out of the total of 22,583 cases, 18,327 were correctly classified, and 4,256 were incorrectly classified.

The decision tree created indicated an accuracy of 77.75% after pruning. Since the accuracy value is greater than 50%, we can say that the tree performs well, correctly classifying individuals in most cases. As with logistic regression, the call duration was the most important variable, having the greatest influence on the target variable. From the confusion matrix, we can see that out of the total of 22,628 individuals who responded negatively, 18,123 were correctly classified, and 4,505 were incorrectly classified. Out of the total of 22,583 individuals who accepted the bank's offer, 17,030 were correctly classified, and 5,553 were incorrectly classified.

The main limitations observed in this research relate to the small number of variables. Therefore, in the future, this analysis can be extended by adding additional variables, such as factors related to financial behavior, economic conditions, and the demographic profile of clients, in order to obtain a more realistic picture of how an individual's decision to purchase a bank deposit is influenced.

The results obtained from the application of data mining techniques suggest that there are significant differences in customers' financial behavior, and certain factors may increase or decrease the probability of accepting bank offers.

References

- Ling, X., & Li, C. (2010). Data mining for direct marketing: Problems and solutions. *Proceedings of the 4th KDD Conference*. AAAI Press, https://cdn.aaai.org/KDD/1998/KDD98-011.pdf.
- Borowik, B., Suchacka, G., & Grzonka, D. (2016). Application of selected supervised classification methods to bank marketing campaign. Institute of Computer Science, Cracow University of Technology, https://www.researchgate.net/publication/303857692_Application_of_Selected_Supervised_Classification_Methods_to_Bank_Marketing_Campaign
- Kotler, P. (2004). Marketing de la A la Z, https://www.parteneriatefinantari.ro/wp-content/uploads/2023/12/10.-Philip-Kotler-Principles-Of-Marketing.pdf.
- Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the CRISP-DM methodology. European Simulation and Modeling Conference - ESM'2011, 117-121, https://www.researchgate.net/publication/236231158_Using_Data_Mining_for_Bank_Direct_Marketing_An_Application_of_the_CRISP-DM_Methodology
- Mylonakis, J. (2008). The influence of banking advertising on bank customers: An examination of Greek bank customers' choices. Banks and Bank Systems, 3(4), 44-49, https://www.researchgate.net/publication/291176120_The_influence_of_banking_advertising_on_bank_customers_An_examination_of_Greek_bank_customers'_choices
- Charles, B., Cheng, J., & Nancy, S. (2007). Determinants of bank selection in USA, Taiwan and Ghana. International Journal of Bank Marketing, 25(7), 469-489, https://www.researchgate.net/publication/235291454_Determinants_of_banks_selection_in_USA_Taiwan_and_Ghana.

Vol. 76, issue 4 Year 2024
DOI: 10.56043/reveco-2024-0034

- Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, https://www.sciencedirect.com/science/article/abs/pii/S016792361400061X.
- Lu, X.-Y., Chu, X.-Q., Chen, M.-H., Chang, P.-C., & Chen, S.-H. (2016). Artificial immune network with feature selection for bank term deposit recommendation. Journal of Intelligent Information Systems, 47(2), 267-285, https://www.researchgate.net/publication/299482174_Artificial_immune_network_with_feature_selection_for_bank_term_deposit_recommendation.
- Lewis, G. (2009). Consumer perceptions of mobile phone marketing: A direct marketing innovation. Direct Marketing: An International Journal, 3(2), 124–138, https://www.researchgate.net/publication/235278073_Consumer_perceptions_of_mobile_phone_marketing_A_direct_marketing_innovation.
- Berry, L. (1983). Relationship marketing. Emerging Perspectives in Services Marketing. American Marketing Association, https://link.springer.com/article/10.1177/009207039502300402.
- Vella, J., & Caruana, A. (2012). Encouraging CRM systems usage: A study among bank managers. Management Research Review, 35(2), 121-133, https://www.researchgate.net/publication/235317509_Encouraging_CRM_systems_usage_A_study_among_bank_managers
- Filip, A., Vrânceanu, D. M., Georgescu, B., & Marinescu, D. E. (2016). Relationship marketing stage of development in Romanian banking industry. Amfiteatru Economic, 18(41), 113-129, https://www.amfiteatrueconomic.ro/temp/Article_2499.pdf.
- Bhambri, V. (2011). Application of data mining in banking sector. International Journal of Computer Science and Technology, 2(2), 199-202, https://www.ijarcs.info/index.php/Ijarcs/article/download/2861/2844/5693.