UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Universität Paderborn
Fakultät für Wirtschaftswissenschaften
Department Wirtschaftsinformatik
Lehrstuhl für Wirtschaftsinformatik, insb. Data Analytics

Individual Study Research

Data mining for direct Marketing in the banking sector

submitted to:

Mr. Prof. Dr. Oliver Müller

Due date:

26. April 2021

submitted by:

Sonia Djankou Nana	Matriculation number: 6832815
Fotso Tenku Joel Cedric	Matriculation number: 6810782

Table of Contents

List of figures IV

List of tables..VI

List of abbreviations VII

1 Introduction 2

2 Related works 3

3 Data describtion 5

3.1 Exploratory Data Analysis (EDA) 6

3.1.1 EDA process: Numerical Variables 7

3.1.2 Pearson correlation 8

3.1.3 Data visualization 9

3.1.3.1 Numerical variable: age 9

3.1.3.2 Numerical variable: pdays 10

3.1.3.3 Numerical variable: previous 11

3.1.3.4 Numerical variable: balance 12

3.1.3.5 Numerical variable: day 13

3.1.3.6 Numerical variable: duration 13

3.1.3.7 Numerical variable: campaign 14

3.2 Categorical Variables 15

3.2.1 Categorical variable: job 16

3.2.2 Categorical variable: marital 17

3.2.3 Categorical variable: education 18

3.2.4 Categorical variable: default 19

3.2.5 Categorical variable: housing 20

3.2.6 Categorical variable: loan 21

3.2.7 Categorical variable: poutcome 22

3.2.8 Summary of categorical variables 23

4 Data wrangling...24

4.1 Numerical variables 24

4.2 categorical variables 25

5 Experiments and results 26

5.1 Machine learning algorithm 26

5.2 Model training and results 26
5.2.1 Logistic regression (LR) 26
5.2.2 K Nearest Neighbor (KNN) 28
5.2.3 Support Vector Machine (SVM) 29
5.2.4 Decision Tree (DT) 30
5.2.5 Random Forest (RF) 31
5.2.6 Neural Network (NN) 32
5.2.7 Comparison of all models 33
6 Discussion and limitations 34
Conclusion 35
References V
Statutory declaration VI

List of figures

Figure 1: KDD Process........5
Figure 2: Pearson Correlation........8
Figure 3: density of age........9
Figure 4: Frequency of age toward good and bad clients........9
Figure 5: Good clients per group of age........10
Figure 6: Bad and clients for the pdays variables........11
Figure 7: number of days after contact for those who subscribed........11
Figure 8: Number of previous contacts toward good and bad clients. 12
Figure 9: percentage of subscription before the campaign........12
Figure 10: Average of yearly balance in euro........13
Figure 11: Day of the week the last contact was made........13
Figure 12: duration in second........14
Figure 13: last contact duration in second........14
Figure 14: Number of campaign contacts........15
Figure 15: professional occupation........16
Figure 16: professional occupation toward good and bad clients........17
Figure 17: clients by professional occupation and age........17
Figure 18: Marital situation chart........18
Figure 19: clients by marital situation and age........18
Figure 20: Education level........19
Figure 21: clients by education level and age........19
Figure 22: good and bad clients according to their default loan........20
Figure 23: clients by default loan and age........20
Figure 24: housing loan situation........21
Figure 25: clients by housing and age........21
Figure 26: personal loan toward good and bad clients........22
Figure 27: clients by loan and age........22
Figure 28: Outcome of the previous marketing campaign........23
Figure 29: clients by previous outcome and age........23
Figure 30: percentage of subscription for categorical variables........24
Figure 31: ROC curve........27
Figure 32: Prevision according to the LR model........27

Figure 33: ROC curve KNN 28
Figure 34: Matrix of predicted values KNN 29
Figure 35: ROC curve SVM 29
Figure 36: Matrix of predicted values SVM 30
Figure 37: ROC curve DT 30
Figure 38: Matrix of predicted values DT 31
Figure 39: ROC curve RF 31
Figure 40: Matrix of predicted values RF 32
Figure 41: ROC curve NN 33
Figure 42: Matrix of predicted values NN 33

List of tables

Table 1: Previous research 3
Table 2: Validation method and result 4
Table 3: statistic description of numerical variables 7
Table 4: Percentage of subscription per group of age 10
Table 5: statistic description of categoric variables 15
Table 6: P-value and Pearson correlation 25
Table 7: Pearson correlation and p-value.. 26
Table 8: summary classification models.. 34

List of abbreviations

EDA	Exploratory Data Analysis
KNN	K-Nearest Neighbour
LR	Logistic Regression
DT	Decision Tree
SVM	Support Vector Machine
RF	Random Forest
NN	Neural Network
AUC	Area Under the Curve
ROC curve	Receiver Operating Characteristic curve
CRM	Customer Relationship Management
KDD	Knowledge Discovery in Database

Abstract

The increasingly vast number of marketing campaigns of banking institutions over the time has reduced its effect on the public. Furthermore, economical pressure and competition has led with the advancement of the data science and machine learning to the fact that most of banks are adapting to a data-drive decision using direct marketing to generate an increasing interest. Despite of numerous studies that have provided important insights into the direct marketing, the understanding of this topic of growing interest and importance still remains deficient. Therefore, in this project we are going to use the EDA processing and data mining method to build a predictive model for a Portuguese bank to provide the necessary suggestion for marketing campaign team in order to target a specific category of customer which is more susceptible to subscribe to the new product as well the term deposit.

Keywords: Data Mining, Banking Sector, Marketing Campaign, Machine Learning. Direct Marketing

1 Introduction

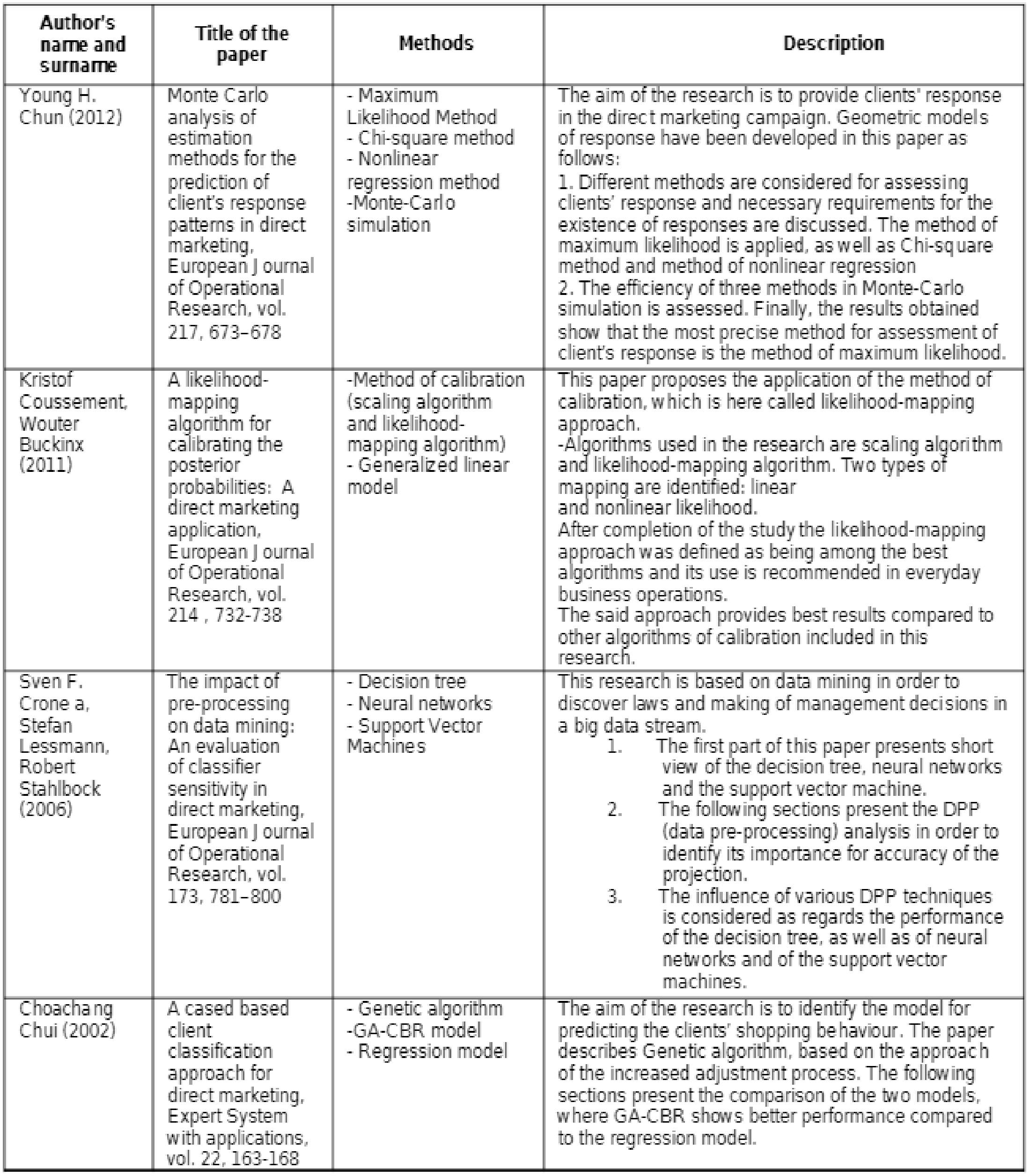
The global banking sector is rapidly changing with the developing in which innovations are applied widely. Especially, in recent years, redesigning of working manner and activity structure in the banking sector has become widespread in all around the world. The developments in behavior and preferences of consumers, competition from different sectors and continuously changing legislations have created serious pressure on banks. Today, customer satisfaction is more important key factor on being ahead of this highly competitive sector. To increase customer satisfaction, banks should improve with creative products and distribution channels to make differentiation in the powerful competition environment. Banks's store huge amount of information about their customers to offer them for several campaigns or products like term deposit. There are two main approaches for enterprise to promote products and or services, through mass campaign, targeting general indiscriminate public or direct marketing which consist of targeting a specific set of contacts. Direct marketing is very effective and widely used strategy of contacting customers or potential customers. It is the process of identifying potential buyers of certain products and promoting the products accordingly. In this case study, the data is related with direct marketing campaigns for bank term deposits of a Portuguese banking institution. The marketing campaigns were based on phone calls and often more than one contact to the same client was required, in order to know if the product (bank term deposit) would be "yes" or "no", that mean if customer will subscribe to the term deposit or not. In order to target the specific client that will subscribe or not we used data mining process as well the prediction using the machine learning algorithm to increase the success of marketing campaigns. The goal of this project is first to use the Exploratory data analysis (EDA) approach to analyze and understand the structured data then build a predictive model which can improve the efficiency of the marketing campaigns and helping the decision makers by reducing the number of features of the bank direct marketing data using some popular data mining techniques such as KNN, decision trees, support vector machines (SVM), logistic regression etc.…

The first part of the paper focus on descriptive analysis of dataset by visualizing all variables and their relationship with the target variable (yes or no). in overall using the EDA approach, and the second part aim to be about the modelling and the result.

2 Related works

There are so many works related to the data mining in banking sector. In the project of Dejana Pavlovic, Marija Reljic and Sonja Jacimovic which consist in the application of data mining in direct marketing in banking sector, they first condensed the previous research done in the direct marketing field in the following table.

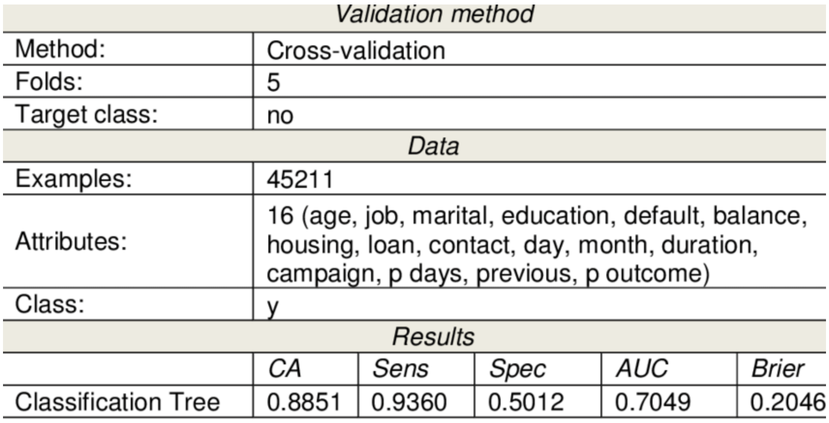
Table 1: Previous research



Author's name and surname	Title of the paper	Methods	Description
Young H. Chun (2012)	Monte Carlo analysis of estimation methods for the prediction of client's response patterns in direct marketing, European J ournal of Operational Research, vol. 217, 673–678	- Maximum Likelihood Method - Chi-square method - Nonlinear regression method -Monte-Carlo simulation	The aim of the research is to provide clients' response in the direct marketing campaign. Geometric models of response have been developed in this paper as follows: 1. Different methods are considered for assessing clients' response and necessary requirements for the existence of responses are discussed. The method of maximum likelihood is applied, as well as Chi-square method and method of nonlinear regression 2. The efficiency of three methods in Monte-Carlo simulation is assessed. Finally, the results obtained show that the most precise method for assessment of client's response is the method of maximum likelihood.
Kristof Coussement, Wouter Buckinx (2011)	A likelihood-mapping algorithm for calibrating the posterior probabilities: A direct marketing application, European J ournal of Operational Research, vol. 214 , 732-738	-Method of calibration (scaling algorithm and likelihood-mapping algorithm) - Generalized linear model	This paper proposes the application of the method of calibration, which is here called likelihood-mapping approach. -Algorithms used in the research are scaling algorithm and likelihood-mapping algorithm. Two types of mapping are identified: linear and nonlinear likelihood. After completion of the study the likelihood-mapping approach was defined as being among the best algorithms and its use is recommended in everyday business operations. The said approach provides best results compared to other algorithms of calibration included in this research.
Sven F. Crone a, Stefan Lessmann, Robert Stahlbock (2006)	The impact of pre-processing on data mining: An evaluation of classifier sensitivity in direct marketing, European J ournal of Operational Research, vol. 173, 781–800	- Decision tree - Neural networks - Support Vector Machines	This research is based on data mining in order to discover laws and making of management decisions in a big data stream. 1. The first part of this paper presents short view of the decision tree, neural networks and the support vector machine. 2. The following sections present the DPP (data pre-processing) analysis in order to identify its importance for accuracy of the projection. 3. The influence of various DPP techniques is considered as regards the performance of the decision tree, as well as of neural networks and of the support vector machines.
Choachang Chui (2002)	A cased based client classification approach for direct marketing, Expert System with applications, vol. 22, 163-168	- Genetic algorithm -GA-CBR model - Regression model	The aim of the research is to identify the model for predicting the clients' shopping behaviour. The paper describes Genetic algorithm, based on the approach of the increased adjustment process. The following sections present the comparison of the two models, where GA-CBR shows better performance compared to the regression model.

Dejana Pavlovic et al, say that the key to successful business operation lies in a good communication with clients, so their strategy was to use the customer relationship management (CRM) to analyze and understand the customer's behavior and characteristics, and to reach the necessary answers based on the implementation of the direct marketing campaign. They applied the data mining in CRM by using the decision tree model, which shows data mining based on the client's response. They used the database of the Bank of Portugal and the information was collected from the marketing campaigns conducted in the period May 2008 – November 2010. Client were contacted by telephone and offered an attractive interest rate for the long-term form of saving. In their work they analyzed the group of clients they selected form the database by using the classification method and the clustering method in order to find the best model that achieves to the high-performance predictions. As classification method they used the decision tree, and they represented the results and accuracy in the table below.

Table 2: Validation method and result



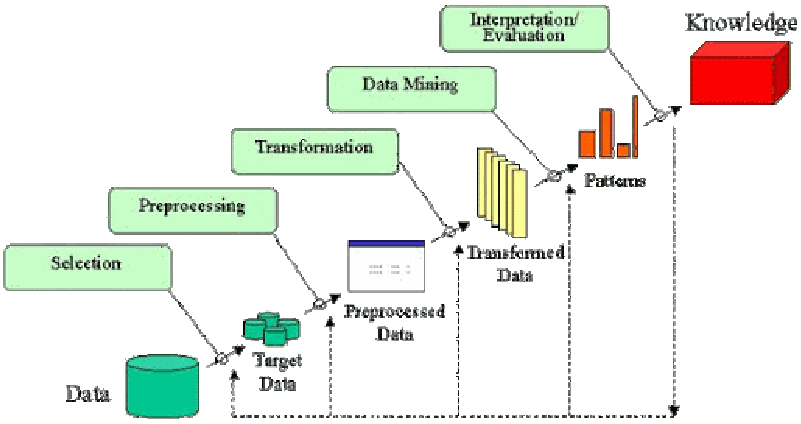
Validation method					
Method:	Cross-validation				
Folds:	5				
Target class:	no				
Data					
Examples:	45211				
Attributes:	16 (age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, p days, previous, p outcome)				
Class:	y				
Results					
	CA	*Sens*	*Spec*	*AUC*	*Brier*
Classification Tree	0.8851	0.9360	0.5012	0.7049	0.2046

Source: Author

The second method they used was the clustering and the purpose was to make marketing segmentation through valid classification of clients into certain clusters. However, the group was not pre-defined, and the grouping was performed based on the similarities found *(Sarcevic M. et al, 2010).* By this approach, they identified several segments that managers used to make decisions in the next marketing campaigns. While they focus their work on the application of the data mining in

CRM, Lilian Sing'oi and Jiayang Wang worked on the data mining framework for direct marketing. Due to competitive market environment, advanced technology and changing behavior of clients, direct marketing has generated an increasing interest among academics and practitioners. Their objective was to provide a comprehensive framework to guide research efforts focusing on direct marketing strategy. Indeed, as data mining process they used the Knowledge Discovery in Databases (KDD) which refers to the overall process of the discovering useful knowledge form data. The figure below displays all the step of this process.

Figure 1: KDD Process



Source: Author

The data mining step was crucial to extract potential useful patterns. In fact, they used the decision tree as classification method. In fact, they build the decision tree algorithm in c steps and called it C4.5 i.e., Choose attribute for root node, Create branch for each value of that attribute, Split cases according to branches and Repeat process for each branch until all cases in the branch have the same class. To evaluate their model, they used 10-fold Cross-validation and achieved 93.37% accuracy.

3 Data describtion

the purpose of our study is to build model that can predict with a high precision if the client will subscribe or not a term deposit so that the manager of the bank would be able to target a category of client. The dataset we are going to use in our work is coming from the UIC[1] archive Repository and is related to a direct marketing campaign of a Portuguese Banking Institution. Indeed, we used the bank-full.csv of the bank.zip folder. Those data were collected by telephone from May 2008 to November 2010 *(Moro et al., 2011).* The dataset contains 45211 observations and 17 variables i.e., 16 inputs variable and the target variable y (yes or no response toward the new term deposit) or output variable.

The input variables provide information about each client such as age, marital status, job and education level. Inputs are subdivided into qualitative and quantitative variables *(Moro et al., 2011).* As numerical variables there are:

- **Age**: the age of each client of the bank
- **Balance**: the average yearly balance in euro
- **Pdays**: the number of days since the client was contacted concerning the previous campaign.
- **Duration**: the contact duration of each client in seconds
- **Campaign**: the number of contacts performed during this campaign and for each client.
- **Previous**: the number of contacts performed before the campaign
- **Day**: the day of the week since the latest contact was made

and as qualitative variables:

- **Job**: as job there are admin, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services and unknown job recorded.
- **Marital**: the marital status (married, single, and divorced)
- **Education**: it is categorised in 4 classes namely tertiary, secondary, primary, and unknown class.
- **Loan**: if the client has a personal loan or no
- **Default**: it is about to know if the client has credit default or no.
- **Housing**: the client has a housing loan or no
- **Contact**: it refers to the way the clients were contacted (unknown, cellular, telephone)
- **Poutcome**: it is the outcome of the previous marketing campaign and is classified in 4 categories namely: Unknown, Failure, Other, Success
- **Month**: it refers to the month of year the latest contact was made.

3.1 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis, which is fundamentally a creative process, was developed by the American mathematician John Tukey in the 1970s. This process

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/00222/

is used to analyse, discover, and investigate data sets and summarize their main characteristics, often employing data visualization methods. it can also help detect outliers and find relation among the variables. For this work we did the EDA process for the numerical variables and the categorical variables as well.

3.1.1 EDA process: Numerical Variables

The table below displays about the important factors of each variables. Using the command describe(), we bring out for each numerical variable the count, the mean, the standard deviation, the minimum, the maximum, the variance, the 25%, 50% and 75%.

Table 3: statistic description of numerical variables

	age	**balance**	**days**	**duration**	**campaign**	**pdays**	**previous**
count	45211	45211	45211	45211	45211	45211	45211
mean	40.94	1.36 e+03	15.81	258.16	2.76	40.19	0.58
std	10.62	3.04 e+03	8.32	257.53	3.09	100.13	2.30
min	18	-8.02 e+03	1	0	1	-1	0
25%	33	7.2 e+01	8	103	1	-1	0
50%	39	4.48 e+02	16	180	2	-1	0
75%	48	1.43 e+03	21	319	3	-1	0
max	95	1.02 e+05	31	4918	63	871	275
variance	112.75	9.27e+06	69.26	66320.57	9.59	10025.77	5.30

Age: According to the results in the table, the youngest client is 18 years while the oldest is 95, with a median of 39 years whereas the average is 40.

Balance: Concerning the balance, the minimum is -8000€ while the maximum is about 102000€ with an average of 1360€.

Duration: The minimum duration of calls is zero, it may mean that the client could not be contacted at all, but for clients who were successfully called, the highest duration is about 4918 seconds with an average of 258 seconds per call.

Campaign: During the campaign, each client has been contacted at least one time, while the maximum of calls was 63.

Pdays: For the days of the previous campaign, the minimum is -1 and probably means the client was never contacted during the previous campaign. Therefore, 871 days max pasted after the latest contacted client of a previous campaign.

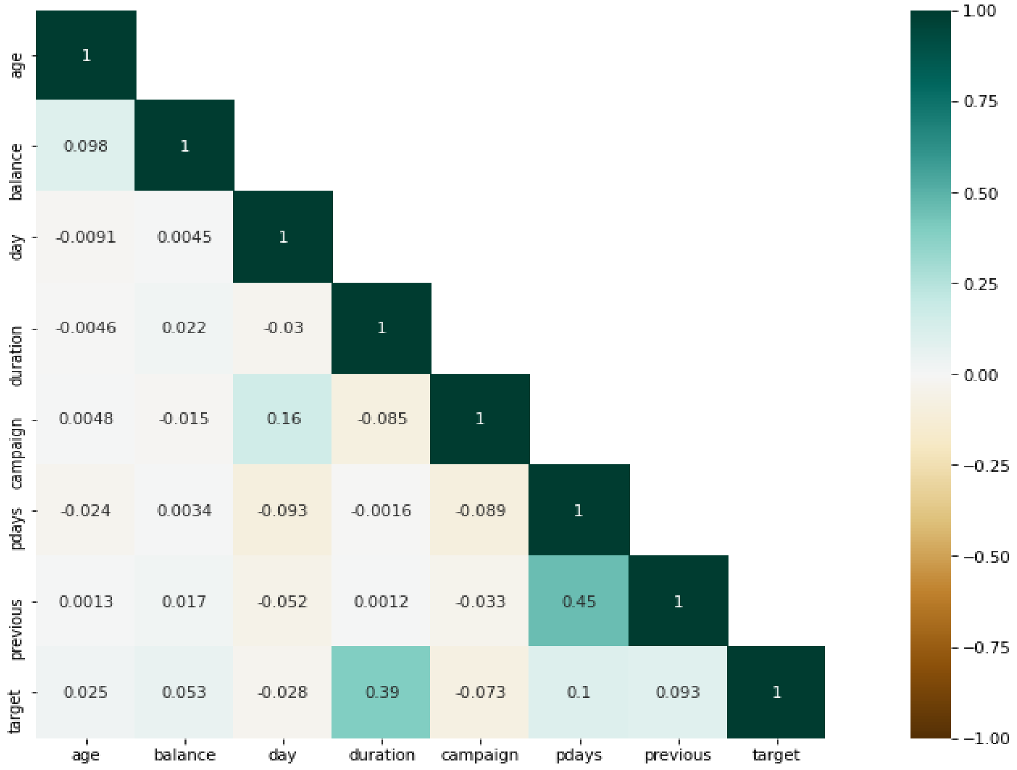
Previous: according to the table, most clients have never been contacted before the campaign.

3.1.2 Pearson correlation

The Pearson Correlation[2] Coefficient is a statistical measure that always must be between -1 and 1; it describes how variables are linearly correlated and how strong their statistical relationship is.

According to the figure 2, the duration is the most correlated with the target, but overall, the correlation between all numerical variables is not significant enough, so we will take each variable in account because there is no linear dependency between them.

Figure 2: Pearson Correlation



[2] https://www.spss-tutorials.com/pearson-correlation-coefficient/.

3.1.3 Data visualization

We will look for each numeric variable the percentage of subscription to the term deposit or not and as well as the good and bad customers.

3.1.3.1 Numerical variable: age

According to the figure 3, the density represents the frequency of age of each client and the occurrence curve (mean) shows that the recurrent age or the average is around 35.

Figure 3: density of age

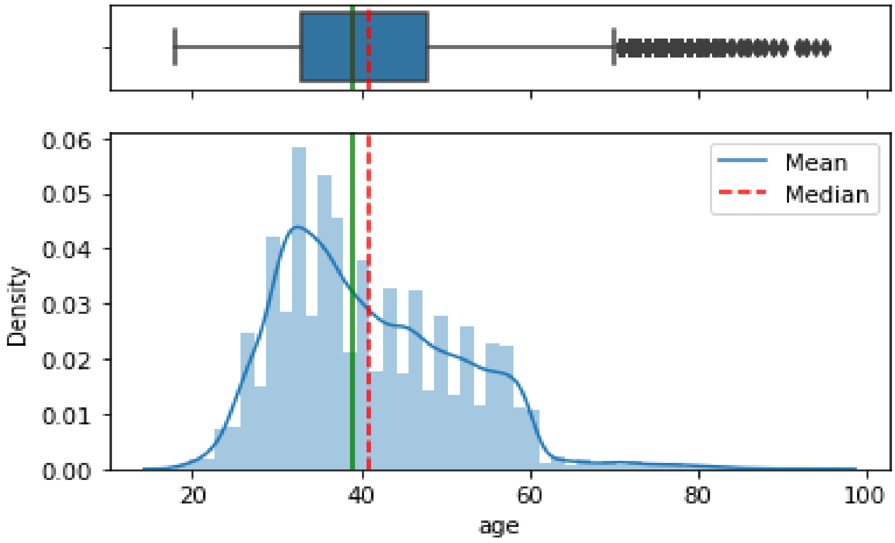
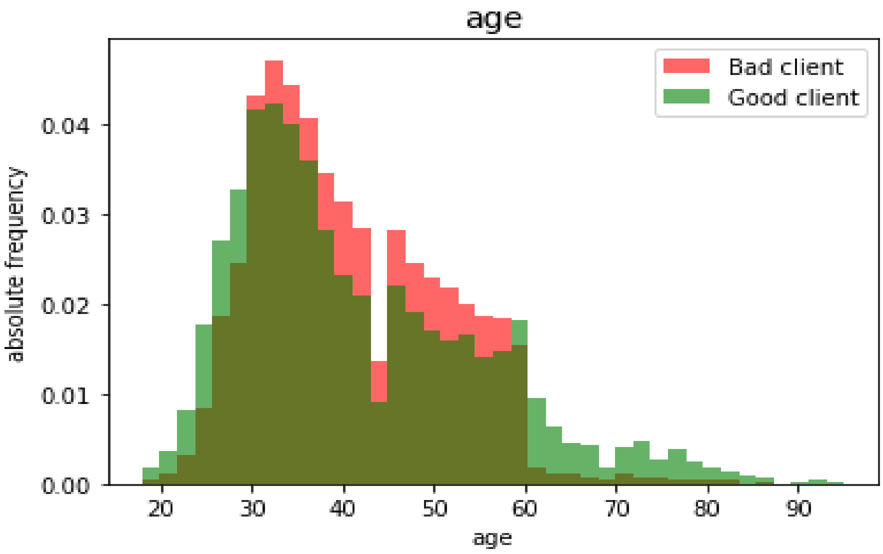


Figure 4: Frequency of age toward good and bad clients



The figure 6 displays about the frequency of age toward good clients who subscribed to the new term deposit and bad clients who did not subscribed, and it shows that there are more bad clients than good. However, to see which age group subscribes the most in order to target only the one with the highest percentage of subscriptions. Therefore, we classified the age in 3 categories as well:

- the senior: which represents the clients over 60 years old,
- the Adult: which is between 30 and 60 years old
- the Young: which is less than 30 years old.

Figure 5: Good clients per group of age

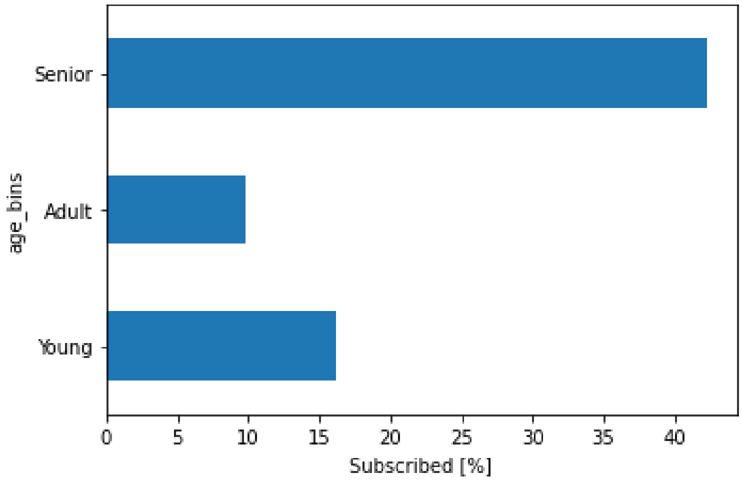


Table 4: Percentage of subscription per group of age

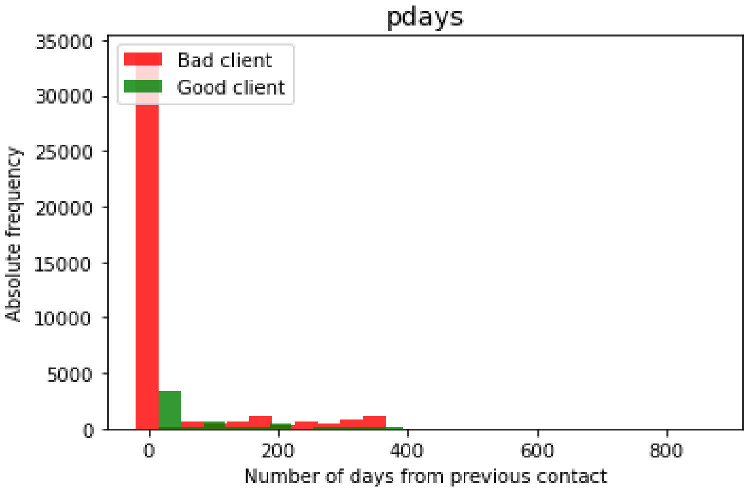
Category	Subscribed (%)
Young	16.215446
Adult	9.845106
Senior	42.255892

According to the table 4, the group of age with the highest percentage of subscription is the senior with about 43% of subscription, thus this group is the best one to target.

3.1.3.2 Numerical variable: pdays

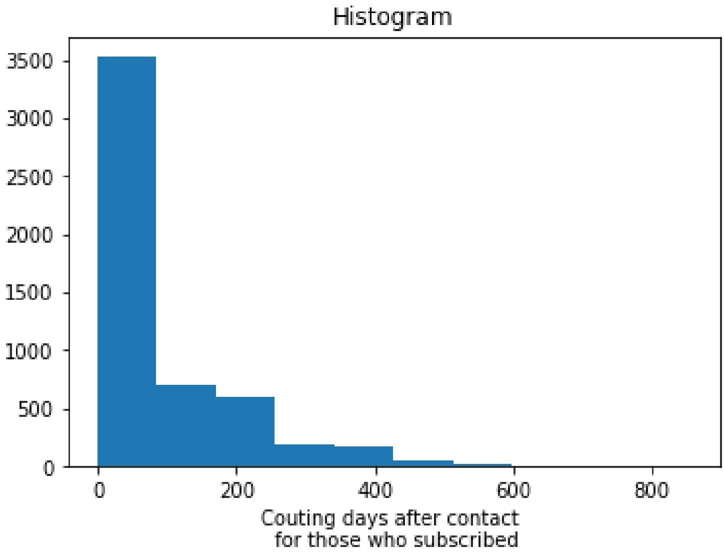
Concerning the pdays variable, it corresponds to the number of days after the latest contacted client, and this chart shows that they were more bad clients than good.

Figure 6: Bad and clients for the pdays variables



However, the following graph gives us a detailed overview of the number of days after contact for those who subscribed. So, we can see that they more days past, they smaller there are subscriptions to a term deposit.

Figure 7: number of days after contact for those who subscribed



3.1.3.3 Numerical variable: previous

According to the figure 8, clients who were contacted before the campaign agree more to the new term deposit. Concerning the bad clients, we started counting the number of previous contacts at zero, and did not take the before in account, because at zero no clients have been contacted.

Figure 8: Number of previous contacts toward good and bad clients.

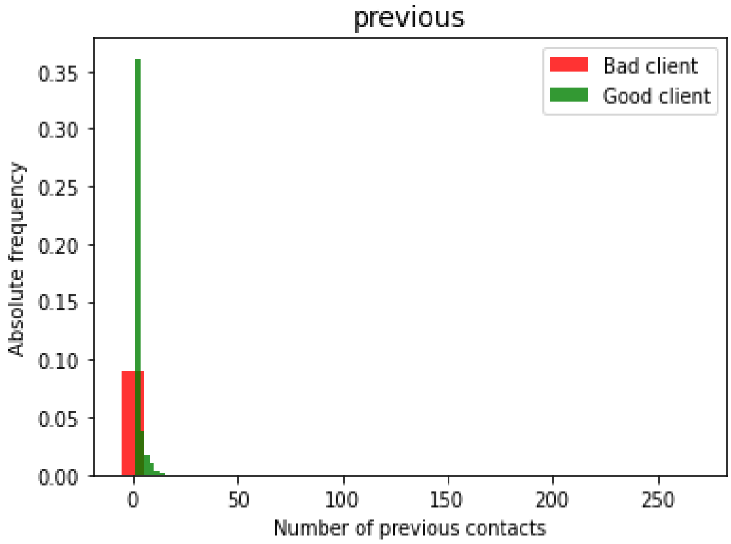
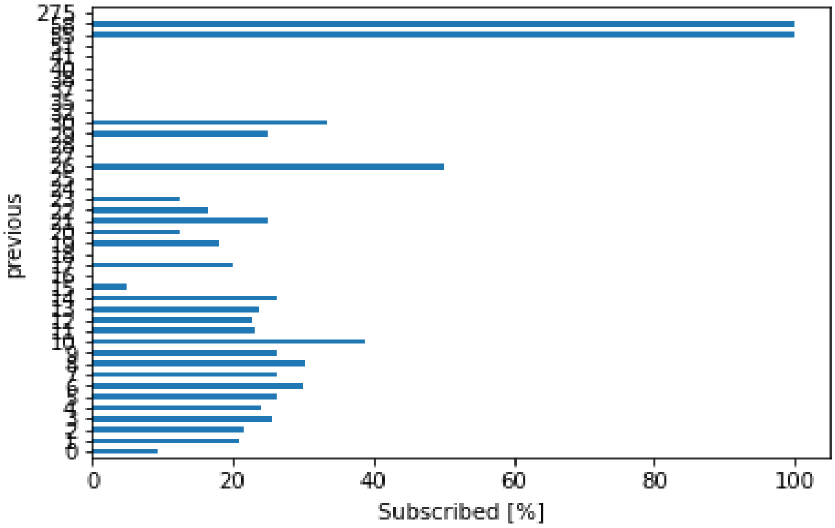


Figure 9: percentage of subscription before the campaign

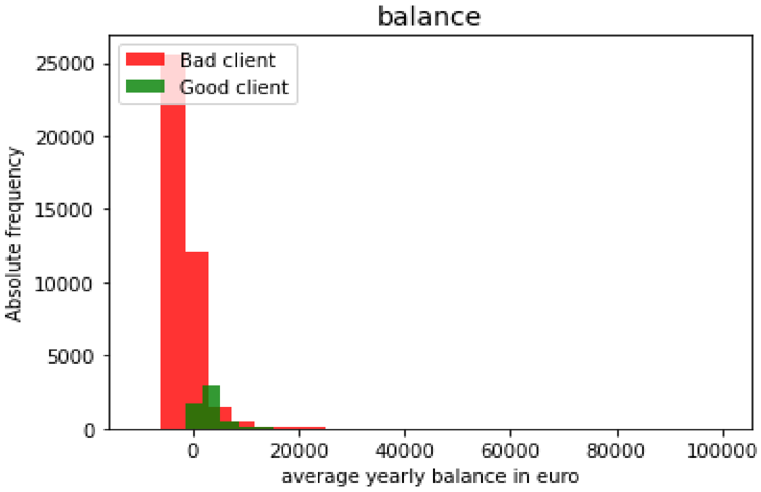


Clients who were previously contacted subscribed in a much higher rate to the term deposit, while for clients who have never been contacted only 10% subscribed to the deposit. For clients who were previously contacted more than twenty times the campaign successfully increases to 45%.

3.1.3.4 Numerical variable: balance

The chart below shows the yearly negative and positive balance of clients, and the majority of clients with negative yearly balance (≤ 0) refuse the new product, while a few numbers of clients with positive yearly balance agree.

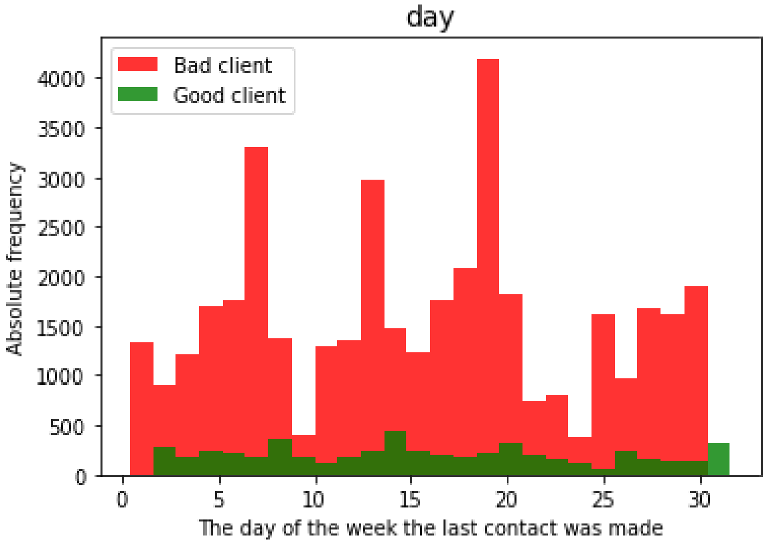
Figure 10: Average of yearly balance in euro



3.1.3.5 Numerical variable: day

For clients who have been contacted on different days, the most have not subscribed to the new term deposit as show in the graph below.

Figure 11: Day of the week the last contact was made



3.1.3.6 Numerical variable: duration

From the duration, the most of has not subscribed to the new term deposit, however, we notice according to the figure 13, that the more the duration of the call is long, the smaller the number of subscriptions to the new product. Thus, the subscription depends strongly to the duration of the call.

Figure 12: duration in second

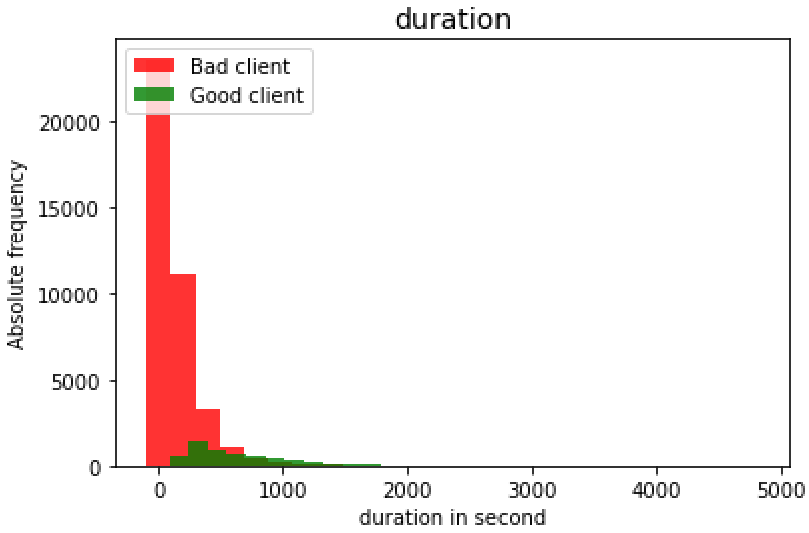
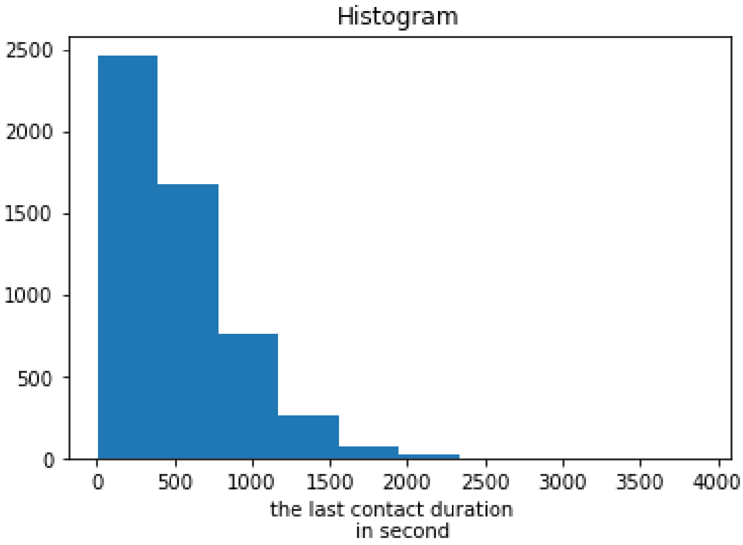


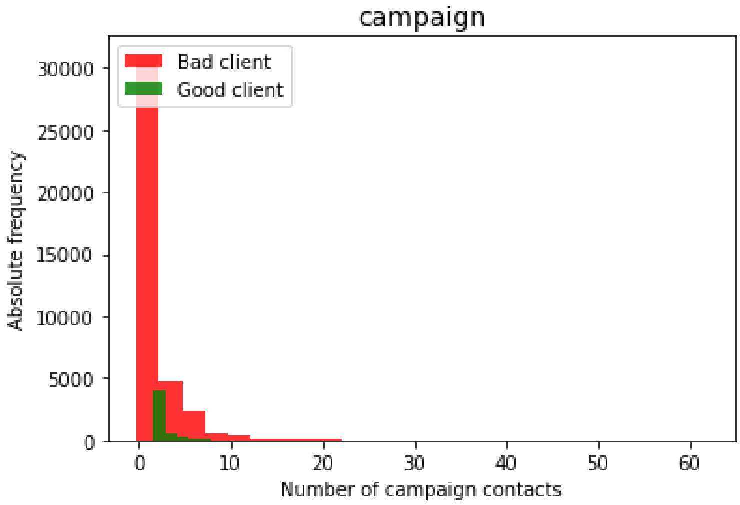
Figure 13: last contact duration in second



3.1.3.7 Numerical variable: campaign

The graph shows the number of contacts that has been performing during the campaign, and we can see that most of contacted clients during the campaign did not agree to the new term deposit.

Figure 14: Number of campaign contacts



3.2 Categorical Variables

As mentioned above, there are in total 9 categorical variables as well job, marital, education, default, housing, loan, poutcome, contact, and month. The table below shows the number of unique values of each variable, the count of each variable, the most representative variables (top).

Table 5: statistic description of categoric variables

	job	**marital**	**education**	**default**	**housing**	**loan**	**contact**	**month**	**poutcome**	**y**
Count	45211	45211	45211	45211	45211	45211	45211	45211	45211	45211
unique	**12**	**3**	**4**	**2**	**2**	**2**	**3**	**12**	**4**	**2**
top	blue-collar	married	secondary	no	yes	no	cellular	may	unknown	no
freq	9732	27214	23202	44396	25130	37967	29285	13766	36959	39922

According to the table:

- **Job**: there are 12 categories of job recording and the most common one is the blue-collar because 9732 clients are from this category.
- **Marital**: the majority of clients are married with almost 28000 records
- **Education**: more than 23000 people have a secondary level

- **Default**: there are 45211 clients, and less than 815 has credit default.
- **Housing**: more than half of the customers have a housing loan.
- **Loan**: almost 38000 clients do not have any personal loans.
- **Contact**: more than half of customers were contacted via a cellular.
- **Month**: almost one third of customers were contacting in May
- **Poutcome**: there is no information about the outcome of any previous marketing campaign.
- **target**: from the database, most of clients have not subscribed to a term deposit (39922 customers)

The table above was a brief description of the categorical variables, but all the details can be seen in the visualization part of those variables. That means we are going to do further and deeper description in order to better understand the data and see which category is more suitable and to target.

3.2.1 Categorical variable: job

The categorical variable 'job' has 12 unique values as seen on the figure below, and the blue-collar is obviously the most common professional job but is also the category which contains the greatest number of bad clients, while for the management, there is certainly a larger number of bad clients, but it is also the category who has the highest number of good clients. In addition to that, there is a class labeled as 'unknown' that should be considered a missing value and will be removed in the next section, Data Wrangling: Cleaning and Feature Engineering.

Figure 15: professional occupation

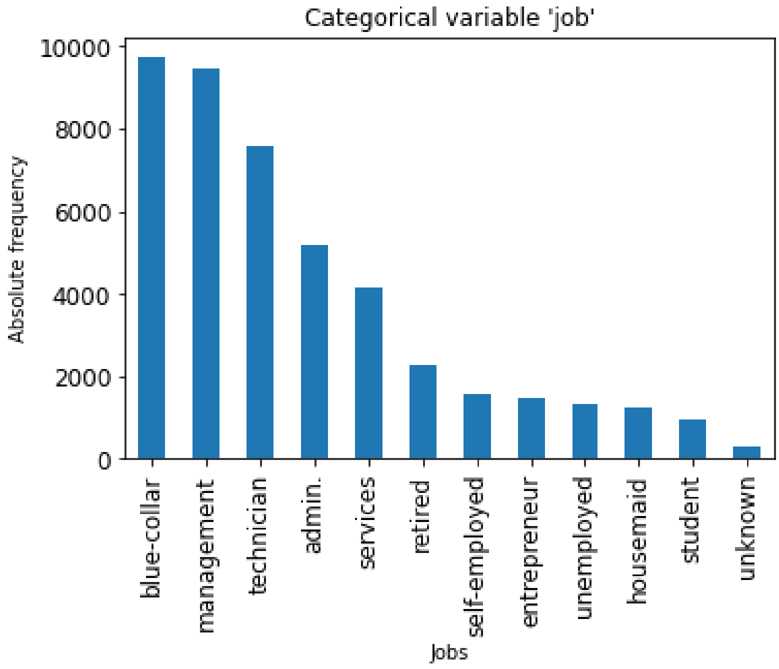
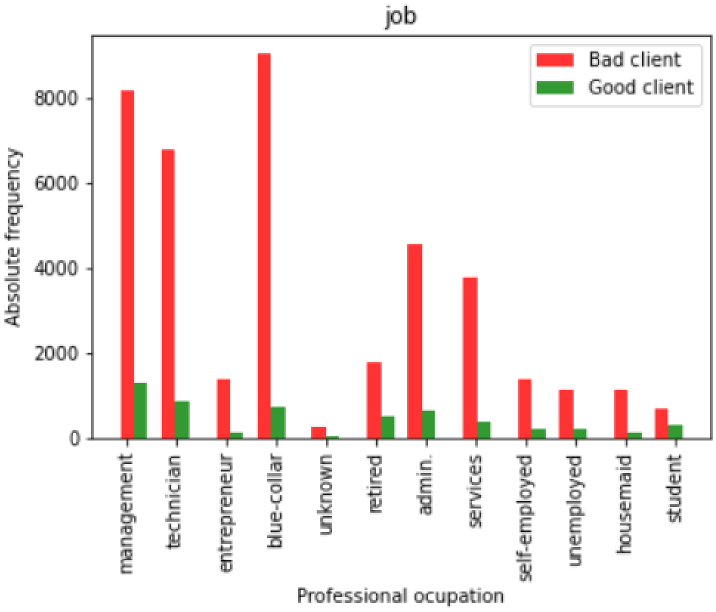
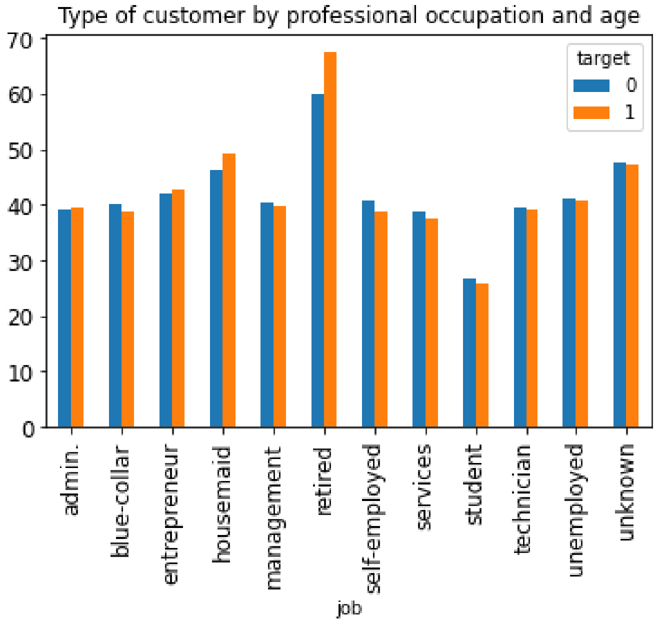


Figure 16: professional occupation toward good and bad clients



By classifying each professional occupation by age (yes = 1, no = 0) as seen on the figure 17, the retired and housemaid categories over 40 years old are more likely to subscribe to new product because the number of acceptances is higher compared to the other categories.

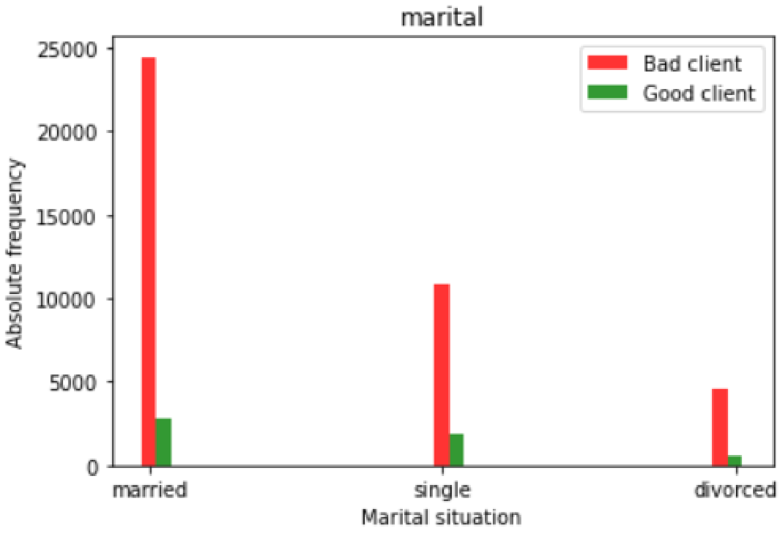
Figure 17: clients by professional occupation and age



3.2.2 Categorical variable: marital

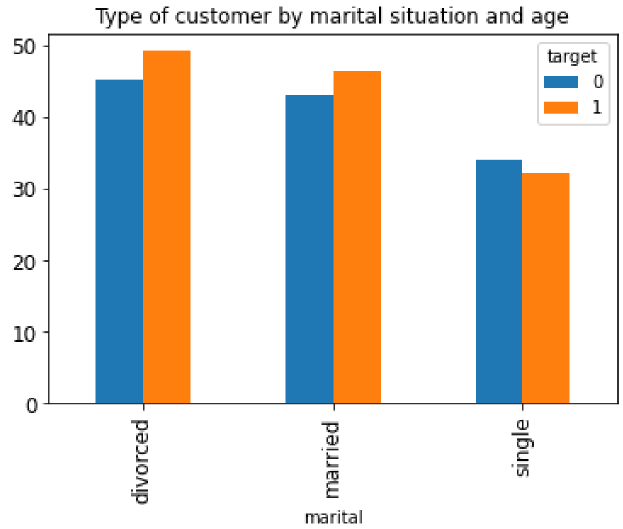
The variable 'marital' has 3 unique values and regarding the overall subscription seen on the chart below, the class which subscribe the most is the married class but is also the class who contains the highest number of bad clients i.e., the clients who did not subscribe at all.

Figure 18: Marital situation chart



However, regarding the age as shown on the figure 19, not only the married clients but also the divorced clients agree the most to the new product.

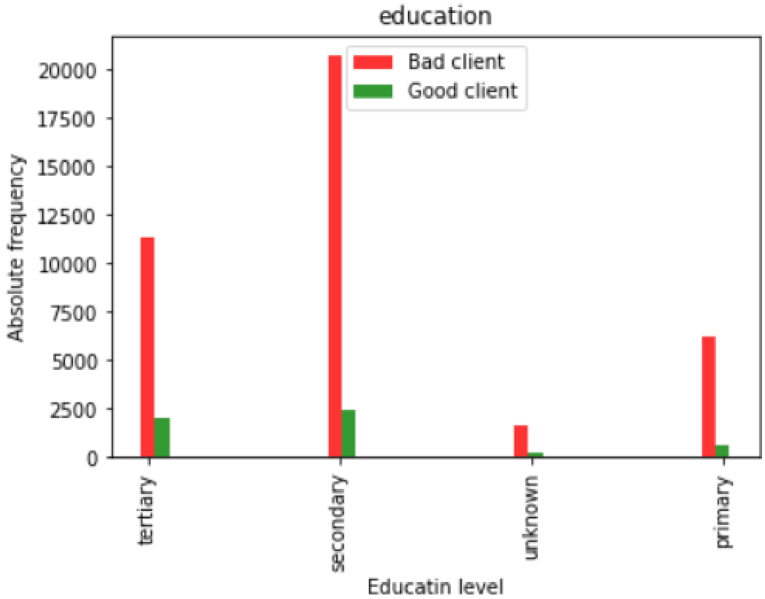
Figure 19: clients by marital situation and age



3.2.3 Categorical variable: education

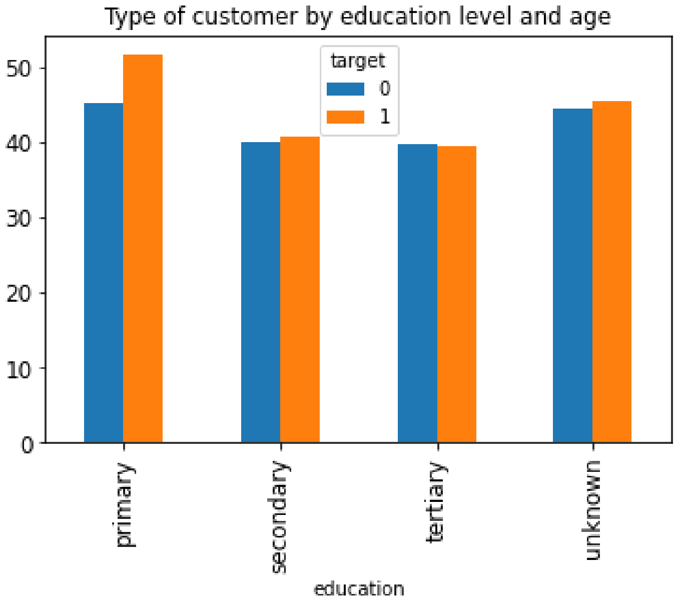
The variable 'education' has 4 unique values as well clients with at least a primary, the secondary the tertiary levels, and the unknown class. On the following figure we can see that most of clients with the secondary level are more likely to subscribe but also reuse the most compared to the clients from another level of education.

Figure 20: Education level



According to the age regarding the figure 21, most of clients who prone to subscribe and have at least the primary level as education. But the next class that has most of subscriptions is unknow.

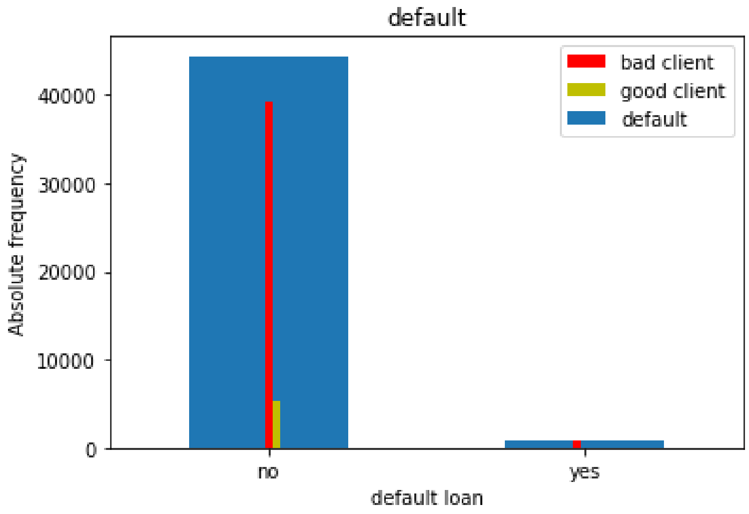
Figure 21: clients by education level and age



3.2.4 Categorical variable: default

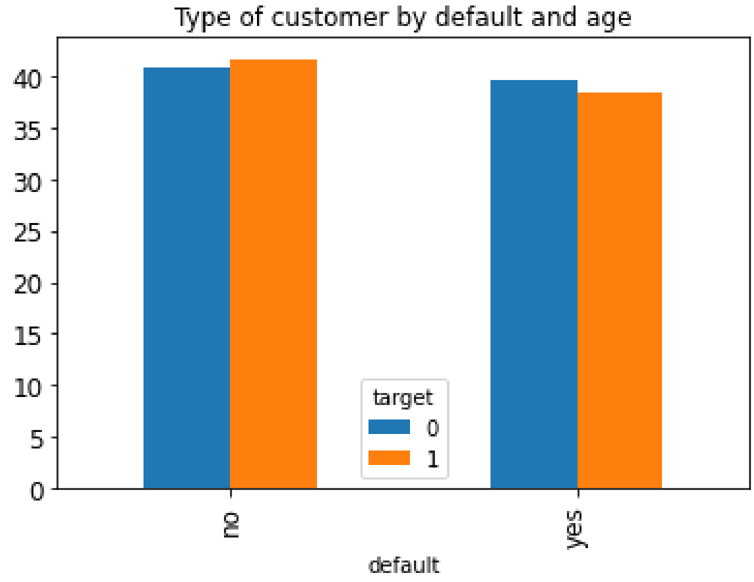
According to the chart below, there are more clients with no credit default than with credit default. Furthermore, clients with no credit default certainly subscribe more but also reject the most the new product proposal than those who has credit default.

Figure 22: good and bad clients according to their default loan



Therefore, regarding the age on the figure 23, most of clients with no credit default agree to the new term deposit, while the number of subscriptions for those with credit default is a bite smaller than the number of refusals.

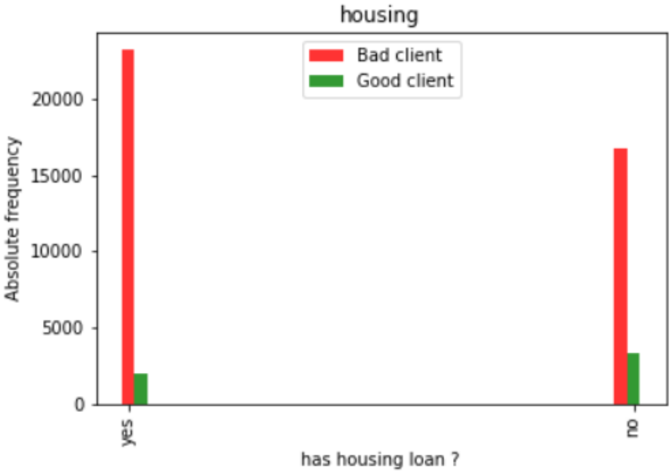
Figure 23: clients by default loan and age



3.2.5 Categorical variable: housing

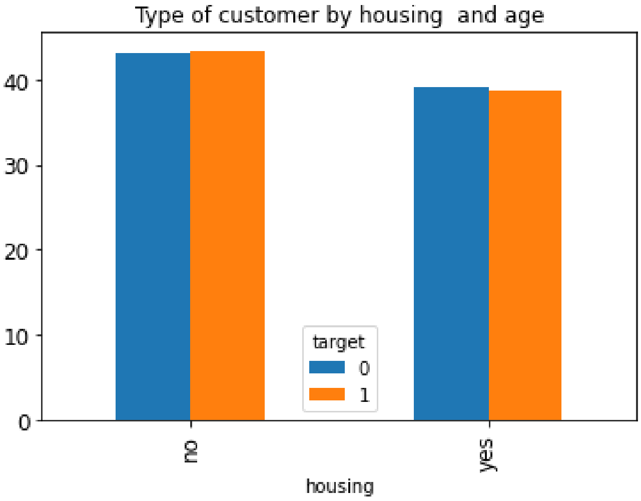
For the housing variable has two unique values as well yes for clients who have housing loan and no for client without housing loan. The following chart shows that the no category has the highest number of subscriptions but also the smallest number of refusals, while the yes category has the smallest number of subscriptions and the highest number of refusals.

Figure 24: housing loan situation



In addition to the above description, the next chart displaying to the clients by housing and the age confirms that clients without housing loan are more likely to subscribe than those with housing loan.

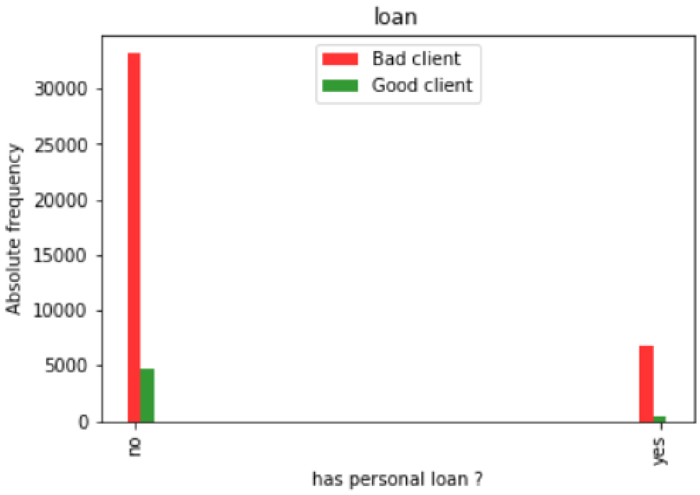
Figure 25: clients by housing and age



3.2.6 Categorical variable: loan

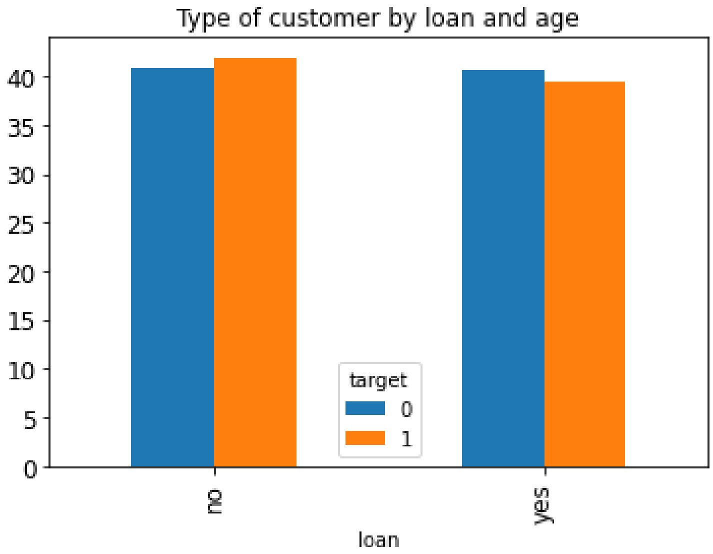
The loan variable has two unique values (no for clients without personal loan and yes for client with personal loan). According to the following figure, the class with the highest number of bad but also of good clients is the no class.

Figure 26: personal loan toward good and bad clients



Therefore, regarding the age as seen on the figure 27, the number of subscriptions is higher for the no category than for the yes category, while the number of refusals in both categories is almost equal.

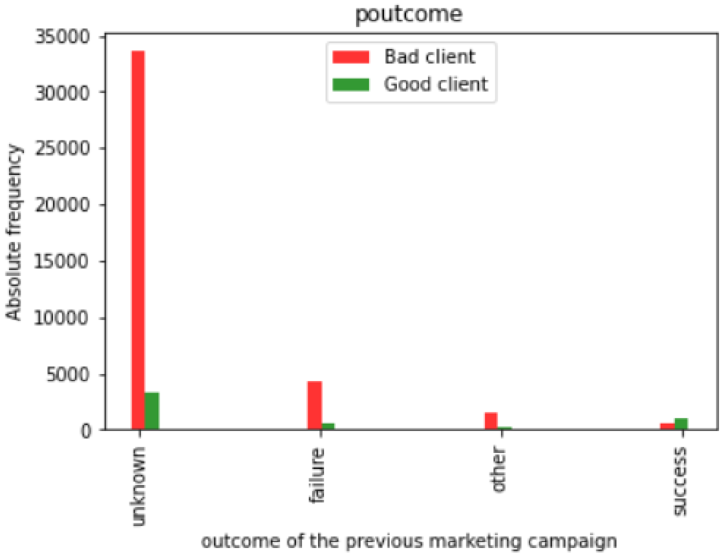
Figure 27: clients by loan and age



3.2.7 Categorical variable: poutcome

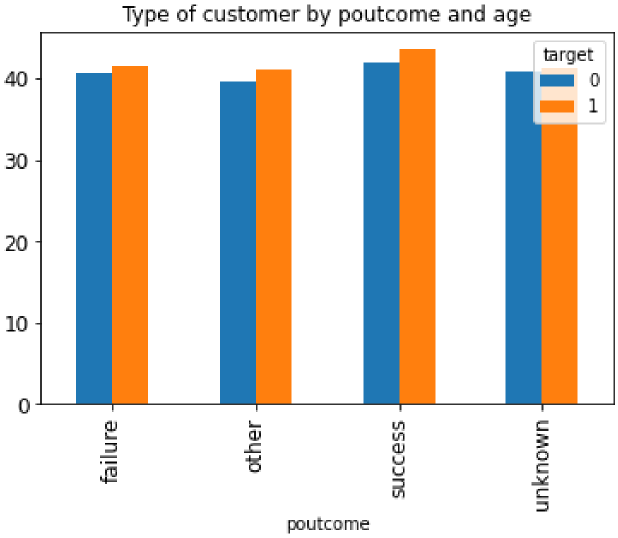
The poutcome variable has 4 unique variables with the class labeled as ‘unknow’ Regarding the chart below, the class with the highest number of good and bad clients is unknow.

Figure 28: Outcome of the previous marketing campaign



unlike the figure 28, which shows that the category that subscribes the most is unknown, on the figure below we can see that the category success has the greatest number of subscriptions according to the age. On the other hand, the number of refusals per client is almost equal for all four categories.

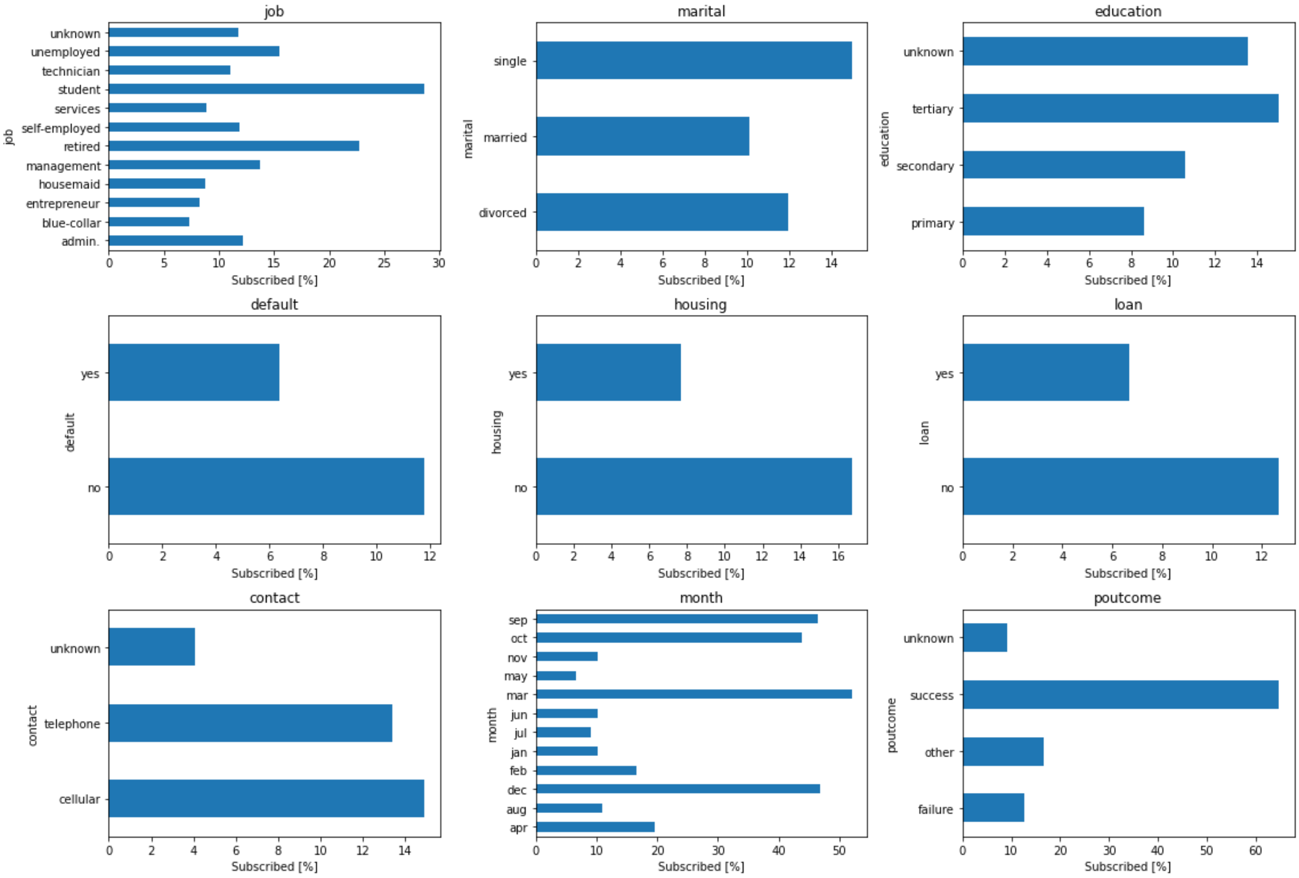
Figure 29: clients by previous outcome and age



3.2.8 Summary of categorical variables

The graphic below illustrates the percentage of subscriptions for each categorical variable.

Figure 30: percentage of subscription for categorical variables



4 Data wrangling

Data modelling is the most important step. For this work, we first divided the dataset into the training and test dataset. And then, we used the data training to train the models using the classification methods in order to find which model with the highest accuracy or precision. Before doing this, an important phase consists in cleaning up the data, because they may contain outliers or missing values that may lead to a poor estimation of the model. In this context, we inputted the unknow values to the classes with the greatest frequency.

4.1 Numerical variables

According to the above description there is no unknown and missing values in the numeric variables. Yet, to find which variable can be selected and which can be dropped for the data modelling, we calculated Pearson correlation and the p-value. Concerning the Pearson correlation, we already know from the data description part that all numeric variables are linearly independent. Concerning the P-value, if it is ≤ 0.05, the correlation is very low and therefore significant and the variable can be kept, otherwise, the variable must be dropped.

Table 6: P-value and Pearson correlation

Variable	Pearson Correlation	p-value
age	0.03	0
balance	0.05	0
day	-0.03	0
duration	0.39	0
campaign	-0.07	0
pdays	0.10	0
previous	0.09	0

Since the p values of numeric variables toward the target are all null and there is no linear dependency between them according to the table above, we kept them all for modelling.

4.2 categorical variables

As for the numerical variables, we determined the p-value of each categoric variable and toward the target variable in order to find which variable can be keep or dropped. We have then seen there is a low correlation between all categoric variables as well as with the target variable (Pearson correlation and p-value) so that we can keep them all except the variable “default” because it is strongly unbalanced. By referring to the part of the description of categorical variables, we noted that there were unknown classes and we considered them as missing values and to handle this issue, we imputed those unknown classes to the category with the greatest frequency. However, for the variable poutcome, we will not input the unknown category because, it is the class with the highest frequency. Finally, we convert each categoric variable into numeric by using the dummies method (One Hot Encoding and Binary Encoding) and numeric transformation.

After the step of data cleaning, we finally obtain a database with 45 variables. Before applying the different classification methods, we first determined if these variables are linearly independent to each other and with the target variable. For it we calculated their p values and displayed the results in the following table by only considering variables with p-values greater than 0.05 (variables we dropped) because the p-value of other variables are less than 0.05, thus strongly significant.

Table 7: Pearson correlation and p-value

variable	Pearson Corr	p-value
Self-employed	0	0.86
Education	0	0.71
Divorced	0	0.56
Admin.	0.01	0.23
Aug	-0.01	0.07
Jan	-0.01	0.06
technician	-0.01	0.06

5 Experiments and results

5.1 Machine learning algorithm

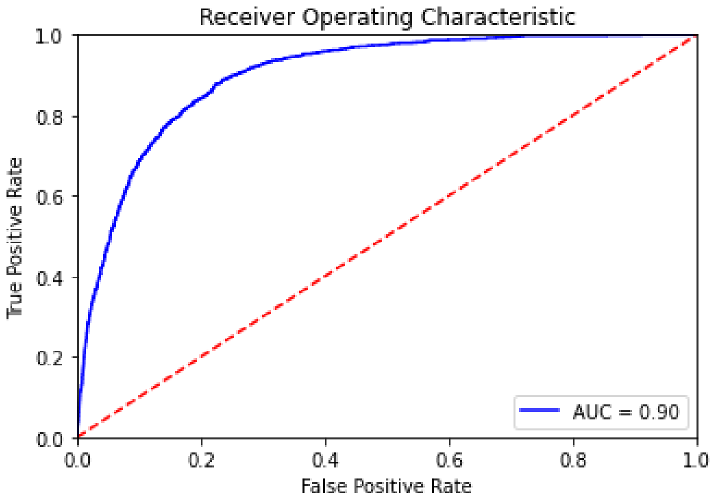
We build models using different classification methods such as logistic regression, the K Nearest Neighbour (KNN), the Support Vector Machine (SVM), the Decision Trees, the Random Forest, and the Neural Network in order to find, which model can predict the best whether the client will subscribe to the new term deposit or not. For it, we split the dataset obtained after cleaning and transformation into two subsets namely 80% for the training set and 20% for the test set. Then we normalized and scale the training and test set in order to remove outlier and have better accuracy using the command “MinMaxScaler()”. Thus, we build and tested each classification models on the training set and tested them on the test set using the cross validation for the evaluation. Therefore, the model with the greatest accuracy and AUC (area under the Curve) will be considered as the best to predict if the client will or not subscribe to a term deposit.

5.2 Model training and results

5.2.1 Logistic regression (LR)

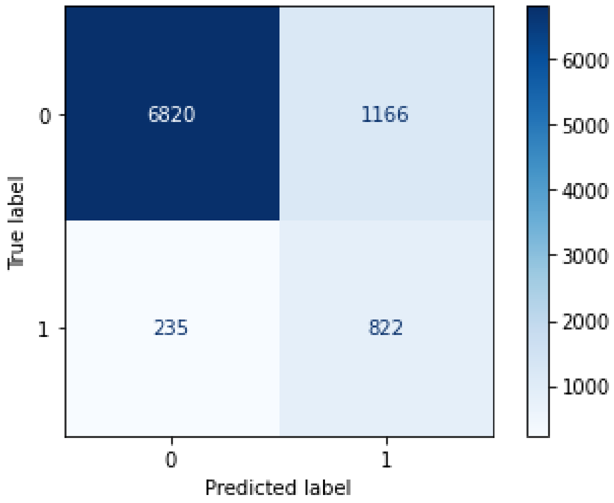
The training of the LR model depends on the parameters. So, we established a grid of five parameters namely 0.001; 0.01; 0.1; 10; 100. Thus, to find the best parameter with the best accuracy, we defined 500 iterations as well as 5-folds cross-validations (number of groups that a given data sample is split into). After training the model, it came out that the best parameter is equal to 0.1 because the accuracy achieved 84,30% and AUC 90%. The figure below shows the evolution ROC Curve of the LR model.

Figure 31: **ROC curve**



After that, we determined from the predicted values the number of clients identified as good, the number of clients identified as bad, the number of bad clients identified as good, and the number of good clients identified as bad (false positives and false negatives) as it is displayed in the following matrix.

Figure 32: Prevision according to the LR model

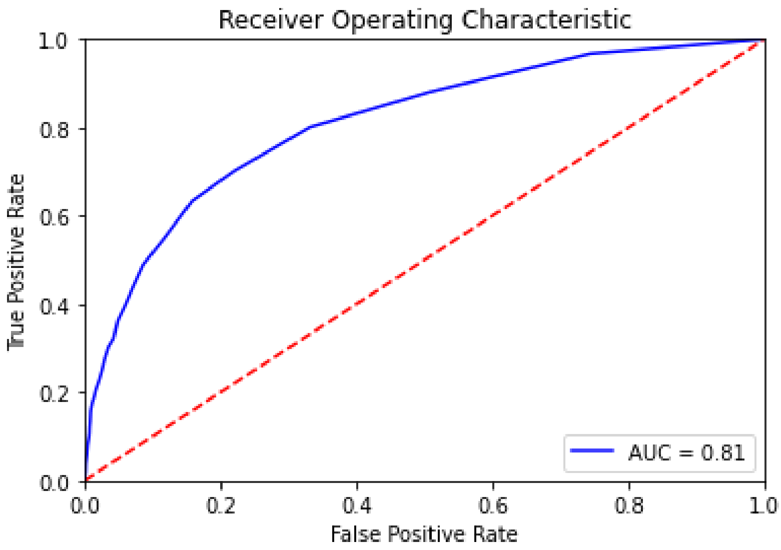


This matrix shows that from the predicted values 6820 clients are identified as bad and will not subscribe to the new term deposit, 822 as good clients who are likely to subscribe, 1166 good clients as bad as well as 235 bad clients as good toward the new product.

5.2.2 K Nearest Neighbor (KNN)

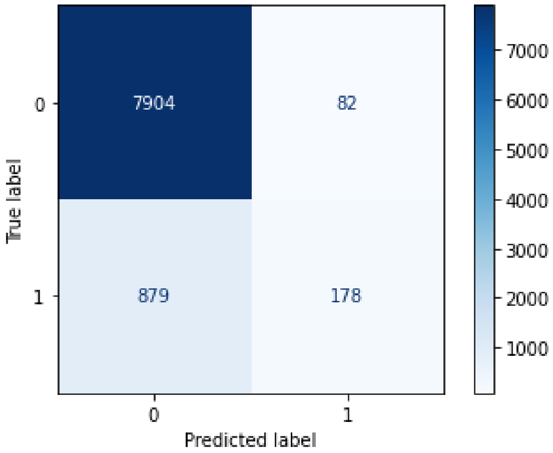
In KNN process, each object is classified by many votes of its neighbour, and it is thus classified to the class with the greatest frequency among its k nearest neighbours *(Fix et all, 1951).* In the same logic as in the case of logistic regression, we must find which parameter k hast the highest accuracy among the k nearest neighbours. For it, we first specified the range of k parameters from 1 to 200 with an interval of 5 between 2 parameters and then trained the KNN model on this range. The best parameter was k equal to 37 because it achieved 89.37% accuracy and the associated AUC was about 81.07%. For the KNN classifier, we obtained the following ROC curve.

Figure 33: ROC curve KNN



From the predicted values, we identified the number of clients who are more likely to subscribe or not based on the hypothesis that, if the predicted value is greater than 0.5, we can then consider that the client is good and can subscribe. The result is presented in the following matrix.

Figure 34: Matrix of predicted values KNN



According to this matrix, 7904 clients are identified as bad clients and will not subscribe to a term deposit while only 178 can subscribe to the term deposit. Last but not least, there are 82 good clients identified as bad candidates (false negative) as well as 879 bad clients identified as good clients for the subscription to the term deposit (false positive).

5.2.3 Support Vector Machine (SVM)

The SVM is a supervised machine learning algorithm with the purpose of classification and regression *(Vapnik et al, 1997).* From the data training with two classes, the SVM helps to build a model that classified the data test to one or other *(Vapnik et al, 1997).* By applying the SVM algorithm on the data training, we set a grid of three parameters as well [0.1, 1, 10] and 5 folds Cross Validation in order to find the best parameter for which the accuracy is the highest. The best parameter for this model is equal to 10 with 84.85% of precision and 89.49% AUC.

Figure 35: ROC curve SVM

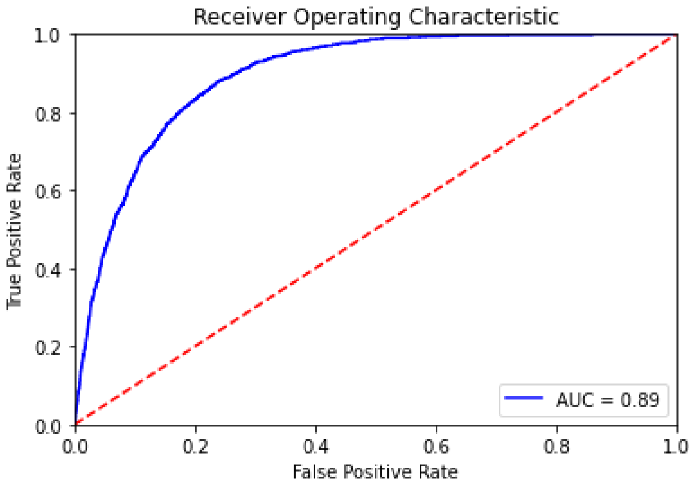
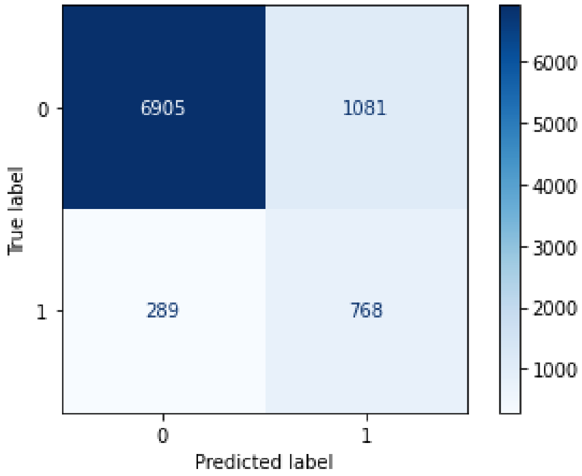


Figure 36: Matrix of predicted values SVM



Based on the same hypothesis as for the LR and KNN, the above matrix shows that 6905 clients are identified as bad while 768 clients are identified as good. 1081 good clients are identified as bad candidates (false negative) and 289 bad clients are identified as good (false positive).

5.2.4 Decision Tree (DT)

The Decision tree is a tree in which each internal node is classified with an input factor and each leaf node is labelled with the class or probability over the classes (*Rokach et al., 2008).* After the estimation on the training set, the prediction on test set, and by classifying the variables by importance, it came out that the duration is the most important one toward the target and the model achieved 90.16% accuracy and 86% AUC. On the order hand, the Decision tree algorithm also helps to determine the importance of each variables for the prediction on the test set.

Figure 37: ROC curve DT

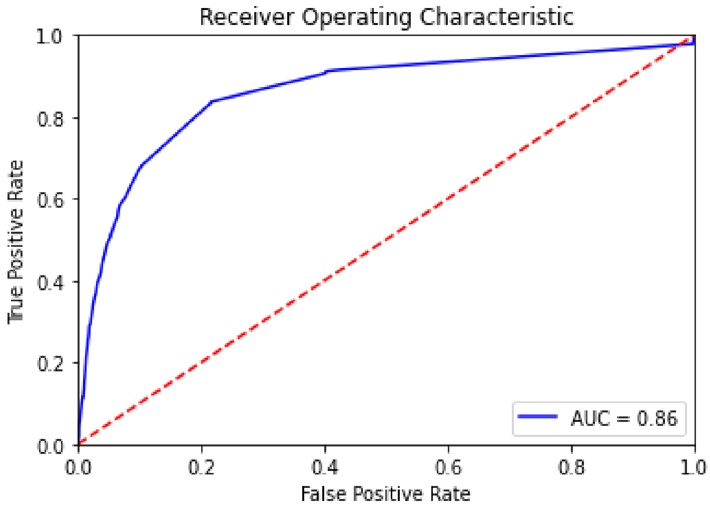
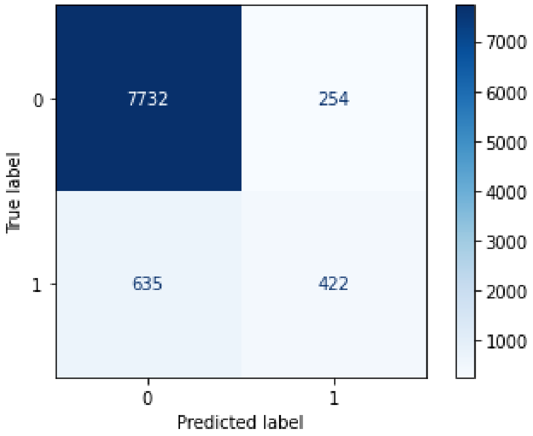


Figure 38: Matrix of predicted values DT



Based on the same hypothesis as for the above models, 7732 clients are bad clients, and 422 clients are good. However, 254 clients are considered as good but subscribe not to the term deposit (false positive) and 635 clients as bad but can subscribe to the term deposit (false negative).

5.2.5 Random Forest (RF)

In the Random Forest, the n-tree sample are generated randomly from the training set and for each sample based on the number of variables randomly selected at each split, the best split is selected *(Zeng Tang et all, 2019).* To build the RF model, we defined a set of estimators representing the number of variables randomly selected form the training set and determined the best estimator with the highest precision using 5-folds cross-validation and the set [10, 100, 300, 1000] as grid of parameters. The best estimator was 100 with 90.37% accuracy with an AUC of about 92.11%.

Figure 39: ROC curve RF

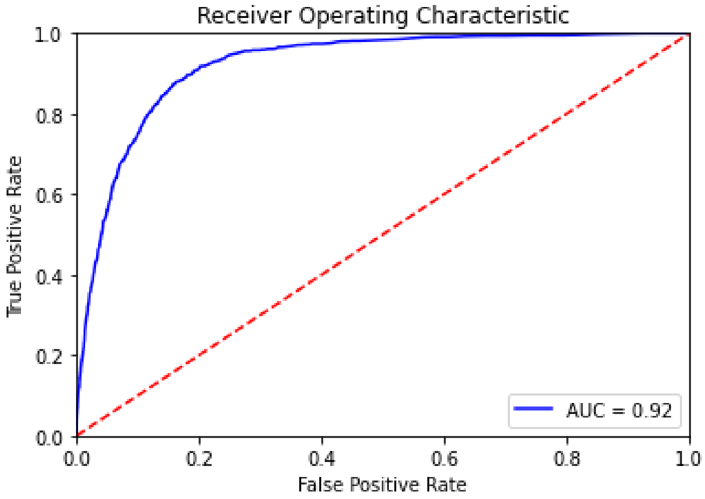
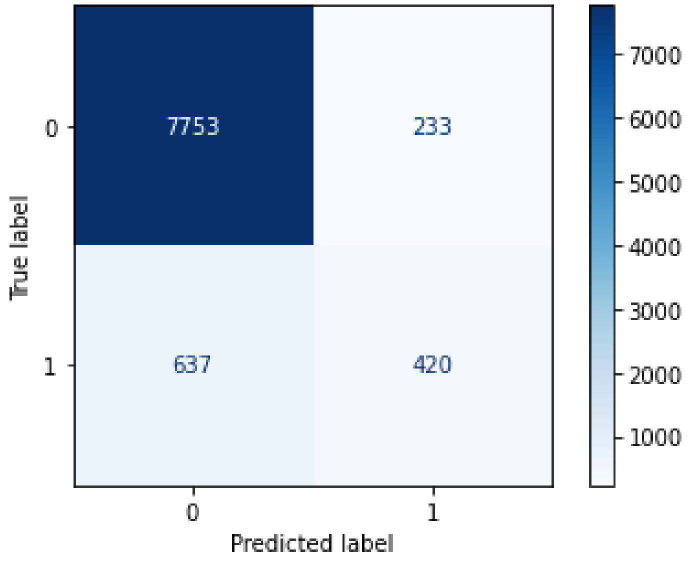


Figure 40: Matrix of predicted values RF



According to the predicted values on the matrix and based on the same hypothesis as for the DT, 7753 clients are identified as bad and will not subscribe to a term deposit, while 420 clients are considered as good and subscribe to a term deposit. 233 clients are considered as good but subscribe not to the new product (false positive), and 637 clients are considered as bad but are more likely to agree to the new term deposit (false negative).

5.2.6 Neural Network (NN)

The prediction of the target using the Neural Network depends on the number of input variables, the number of hidden layers and different learning rates *(Mahmoud K. Okasha, 2014).* By applying the NN algorithm on the training set, we specified a set of number of hidden layer as well as the number of neurons to each hidden layers and different learning rate. The purpose was to find the matching number of hidden layers which performs the best the training set and for which the accuracy is the greatest. Finally, the model achieved 89.89% accuracy and 89.61% AUC. The shape of the ROC curve and the predicted subscription matrix are as follows:

Figure 41: ROC curve NN

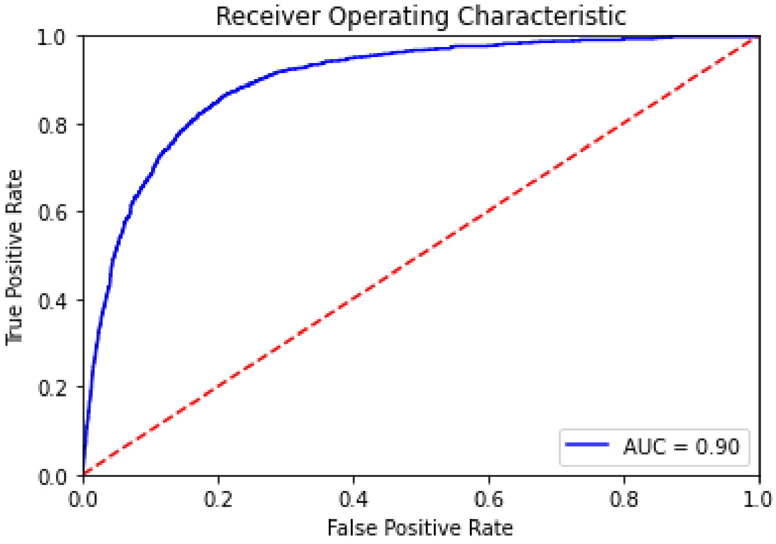
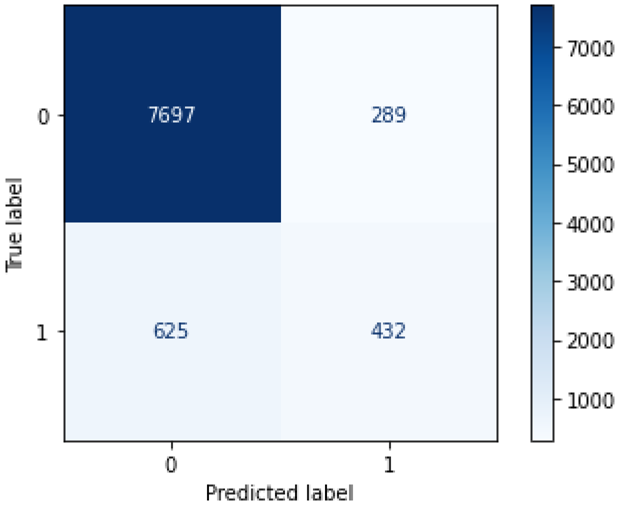


Figure 42: Matrix of predicted values NN



Based on the same hypothesis as for the RF, 7697 clients are considered as bad, 432 clients as good, while 289 clients are though considered as good but will probably not agree to the new product (false positive) and 625 are considered as bad clients but can nevertheless subscribe to the new term deposit (false negative).

5.2.7 Comparison of all models

The results of all accuracy of the classifiers can be display in the following table.

Table 8: summary classification models

Classifier	**Accuracy (%)**	**AUC (%)**	**TN (%)**	**FP (%)**	**FN (%)**	**TP (%)**
LR	84.30	90	75.42	12.89	2.60	9.09
KNN	89.37	81.07	87.40	0.91	9.72	2.97
SVM	84.85	89.49	86.09	2.22	8.44	3.25
DT	90.17	86.20	85.50	2.81	7.02	4.67
RF	90.38	92.11	85.73	2.58	7.04	4.64
NN	89.89	89.61	84.12	3.20	6.91	4.78

- **FN** represents the false negative (good clients identified as bad)
- **FP** represents the false positive (bad clients identified as good)
- **TN** represents the percentage of bad customers
- **FN** represents the percentage of good customers

From this table, the RF model is obviously the most appropriate for the prediction of the subscription to the new term deposit because it has the greatest accuracy (90.38%). Moreover, its AUC ROC is also the highest (92.11%).

6 Discussion and limitations

Direct marketing, as a discipline, has evolved into an integrated and systematic field used by companies and more particularly by banks. In this project, we were able to highlight from the database of the Portuguese bank the application of data mining in the field of banking using different machine learning algorithms. From the best prediction model, the bank will first be able to identify and target the category of customers who will be most likely to subscribe to the new term deposit and to adapt the direct marketing policy in relation to each category of target customer, as well. Thus, we expect that the costs linked to the direct marketing campaign can be reduced.

the study based on direct marketing (via telephone call) has certainly advantages but can also be costly for the bank in terms of money and time because there is no certainty that the contacted client will agree to the new term deposit. On the other hand, the classification model chosen as best may not be the better one given that some variables in the database have missing values for example for variables poutcome and contact where there were unknown classes. Moreover, the model could generate enormous costs for the bank as part of the marketing campaign.

Conclusion

The most important goal for banks is to satisfy the customer and attract new clients. But this takes much time and money. To overcome this issue more and more banks use machine learning algorithms in order to target by category. In this paper we explained how the data mining can perform the database of the Portuguese bank. For it, first applying the EDA process in order to understand the data better and see the different percentage of subscription of each class per categoric variables. Then we used the classification models as well LR, KNN, SVM, DT, RF, NN to build the model prediction. After applying the machine learning algorithm and training the models, we had to test them using 5-folds cross-validation for each model and the evaluation metrics as well the accuracy or precision, the AUC, and the perform. We led finally to the conclusion that the best model was the RF because it achieved 90.38% accuracy, 92.11% AUC.

References

S. Moro, R. Laureano and P. Cortez, 2011. ***Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology***

Moro et al., Jun 2014, Moro P., Corterz and Rita P. A., ***Data Driven Approach to predict the success of Bank.*** Telemarketing, Decision Support System, Elsevier, 62:22-31.

Fix, Evelyn; Hodges, Joseph L., 1951. Discriminatory Analysis. ***Nonparametric Discrimination: Consistency Properties***

Vapnik, Vladimir N., ,1997***. Support-vector networks. Machine Learning***. 20 (3): 273–297.

Lior Rokach, Oded Maimon, 2008, ***Data Mining with Decision Trees: Theroy and Applications***

Zheng Tan, Zigin Yan, Guangwei Zhu, (2019), ***Stock Selection with Random Forest: An exploitation of excess return in the Chinese stock market***, Journal homepage: www.heliyon.com, Volume 5

Mahmoud K. Okasha, (2014), ***Using Support Vector Machines in Financial Time Series***. International Journal of Statistics and Applications, 4(1),28-39, DOI: 10.5923/j.statistics.20140401.03.

Goncalo Guimaraes Gomes, 2020. ***Bank Marketing Campaign: Cost Prediction*** https://github.com/goncaloggomes/cost-prediction/blob/master/ML_fullproject_bankmktcampaign.ipynb (source code).

Vapnik, Vladimir N., ,1997. ***Support-vector networks. Machine Learning***. 20 (3): 273–297.

Mahmud T., Mehmet E., Özdal K., Ertugrul T., Januar 2019. ***Data Mining in Digital Marketing***. DOI: 10 1007/978-3-319-92267-6_4.

Dejana P., Marija R., Sonja J., April 2014. Industrija 42(1): 189-201 ***Application of Data Mining in direct marketing in banking sector***. DOI: 10.5937/industrija-42-5087.

Lilian S., and Jiayang W., IJCSI International Journal of Computer Science Issue. Vol. 10, issue 2, No 2, Marche 2013. ***Data Mining Framework for Direct Marketing: A Case Study of Bank Marketing***. ISSN(Print): 1694-0814 | ISSN (online): 1694-0784. www.IJCSI.org.

https://github.com/nickr007/Bank-Marketing (source code)

Statutory declaration

We hereby affirm that we, Fotso Tenku Joel Cedric and Djankou Nana Sonia have authored this thesis independently, that we have not used other than the declared sources, and that we have explicitly marked all material which has been quoted either literally or by content from the used sources.

This thesis has not been submitted or published either in whole or part, for a degree at this or any other university or institution.

Paderborn, 26.04.2021

Fotso Tenku Joel Cedric

Djankou Nana Sonia