INTRODUCTION

The dataset used for the present project is part of the investigation conduit by Moro et al. (2011), who collected data from a Portuguese bank from the process of 17 marketing campaigns from May 2008 to November 2010, which had the increment of long term deposits, as it main objective of study. The original dataset englobes 59 variables relate with the client information, some information about the first and last contact in the campaign, historical information of previous campaigns, and the results of each campaign in terms of the invested resources. The total number of contacts made by the bank during this time was of 79,354, with a success rate of 8%, when measuring the number of contacts that terminated in deposits effectuated.

The dataset public available for research, contains only 17 variables, and 45211 number of instances, which can be understood as clients contacted by the bank in the one campaign. This dataset can be found in the UCI repository (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing). The goal of the original study was to increase the efficiency of direct campaigns for long-term deposits by using Naïve Bayes, Decision Trees and Support Vector Machine. Moreover, the aim of the present project is to test Statistical Classification tools on the reduce dataset, in order to obtain a model which could predict, based on the historical information of contact and the information of the client, if he is suitable to performed a long term deposit, and in general, determined which client profile should the bank pursue in their contacts in order to improve the ratio of success in these campaigns.

The main problem is to determine whether the client is going to make a deposit in the long term or not, this is important because these conformed banks liabilities, which are transform to loans and other financial products that are part of the assets of the banks, and these ones generate the profit. If a bank can determine with some certainty which are the clients who are more likely to make term deposits, it can focus its marketing campaigns on those clients and with less resources gain those capitals by creating attractive strategies.

Statistical learning is related to a large set of tools that allow the understanding of data. These tools can be classified as supervised or unsupervised. Unsupervised statistical learning can provide an understanding of the relationships and the structure of a dataset, but it does not give a supervised outcome or response (James, G. et al., 2015, p.15). On the contrary, supervised tools, predict or estimate a response from various inputs that can be understood as predictors. Seeing that there is an interested in a prediction method, because there is an associated response variable, which is the historical result of one of the campaigns for each of the contacted clients, called “y”, this is a problem that can be solve through supervised tools. Moreover, cause the response is a qualitative variable, the problem can be treated as a classification one, since each observation has to be assign to a category or class. In the following sections, two type of statistical classification tools have been performed: logistic regression and Bayes classifiers.

A model accuracy will be given by the proportion of mistakes that result when comparing the real response of the training data, vs the response obtained through the model. This could be explained by the next equation: $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$. Where: the $I(y_i \neq \hat{y}_i)$ can be understood as the indicator variable, that takes values of 1 if the response variable predicted $\hat{y}_i$ differs from the veridic y_i response, or 0 if they are the same. This calculation gives a result that can be show in matrix form, and where the diagonal observations are the ones misclassified. The best method will minimize this classification error and give a more precise response at an overall level considering also other metrics. In the case of the estimation of general liabilities from customer deposits, there can be identify two errors define as: one, the clients which the bank expects them to make a long-term deposit and spend resources by contact them, and in the end, did not effectuated a deposit, and in the other hand, one less shocking to the cost of the bank the ones that where estimated not to perform a deposit and in the end, make one. These errors will be the ones that the bank will take in account, in order to be able to predict the amount of clients that have to contact and estimate the amount of money that they could have available from the clients at a certain period. Deriving these errors in terms of this marketing campaign point of view, can be resume as: the clients that the bank classified as proper to contact for this campaign and in the end, did not make a deposit, and on the contrary, the clients that did not call and make a deposit. It is no easy to determine the economic cost derived from these errors, because they will be derived from the actual cost of contact and the opportunity cost of contact someone else, which profile is most accurate to be an elected candidate for the campaign and could make a deposit. Besides this, a matrix cost has been determined just for an example purpose for this analysis and which will be used after defining the best method to predict the probability of a contact client making a long-term deposit. All the calculations, and some further plots can be found in the R. file attach.

DATA PREPROCESSING

#	Name	Description	Type	Categories
		Identification of the client		
1	age	Age measure in years	Numerical	
2	job	Type of job	Categorical	"admin.", "unknown","unemployed","management" "housemaid","entrepreneur","student","blue-collar" "self-employed","retired","technician","services"
3	marital	Marital status	Categorical	"married","single" "divorced" as divorced or widowed
4	education	Level of education	Categorical	"unknown","secondary","primary" "tertiary"
		Bank information of the client		
5	default	Client with credit in default	Categorical	"yes","no"
6	balance	Average yearly balance, in euros.		
7	housing	If the person has a housing loan.	Categorical	"yes","no"
8	loan	If the person has a personal loan.	Categorical	"yes","no"
		Last contact information		
9	contact	Channel of communication type.	Categorical	"unknown","telephone","cellular"
10	day	Last contact day of the month.	Categorical	
11	month	Last contact month of year.	Categorical	"jan","feb","mar","apr","may","jun" "jul","aug","sep","oct","nov","dec"
12	duration	Last contact duration, in seconds.	Numerical	
		Campaign information		
13	campaign	Number of contacts performed during this campaign and for this client.	Numerical	
14	pdays	Number of days that passed by after the client was last contacted from a previous campaign.	Numerical	-1 = client was not previously contacted
		Previous campaign information		
15	previous	Number of contacts performed before this campaign and for this client.	Numerical	
16	poutcome	Outcome of the previous marketing campaign.	Categorical	"unknown","other","failure","success"
		Result of the campaign		
17	y	If the client suscribed a long term deposit	Categorical	"yes","no"

The information given in the data set, has been classified in 6 categories, which can be observed in Table 1. This dataset was partitionated in two parts, an 80% as training data and the rest as testing data, with the final objective of proving the accuracy of each type of model, since we are not interested on how the model works for the same data from which the model was obtain, but on how well it works with unseen data. This partition was made with the package *caret*, from which the **createDataPartition** function can be used to create balanced splits of the data, since variable "y" is a factor, the random sampling occurs within each class and preserves the overall class distribution of the data (Kuhn, 2019).

TREATMENT OF MISSING VALUES

After, reviewing the training dataset, it was determined that there are no missing variables, but there are in fact categories label as unknow which can be traduced as missing values of the dataset. The variables affected are: "poutcome" with an 81.70%, "contact" 28.72%, "education" 4.06%, and "job" 0.65% of missing values. Particularly, the variable "poutcome" cannot be reassign values through any imputation method, because of the small percentage of data known for this sample (less than 20%), so this variable will not be considered for the present analysis. Additionally, the variable "contact" does not seem important since the categories available are telephone or cellular, but both give no extra information, as for example will give if the process of contact involves different channels as visiting the clients or emailing, or courier, in consequence, this variable was also taken off from the analysis.

Furthermore, "education" and "job" are considered as important variables, and because of the small percentage of missing value, there was effectuated just a cutting of the observations with missing values, which turn out to be 4.47% of the training data. On the other hand, there is an assignation already performed in the variable "pdays" which contains the data of the days that passed from a previous contact to the contact to each client at the campaign observed, and which has a -1, whenever this contact was the first approach to the client. This consideration has been taken through the following analysis.

TRANSFORMATION OF VARIABLES

The variable age was categorized in generations because there is evidence that there is a common behavior in each group, and different strategies should be used to approach them. The Center for Generational Kinetics defines these group as 5. Considering that the data is from 2008 to 2010, and the exact year of the sample is not available, the aggrupation has been done over the age assuming that the data is from 2008. The data was grouped in Millennials or Gen Y: Born 1977 – 1995, generation X: born 1965 – 1976, Baby Boomers: born 1946 – 1964 and, traditionalists or silent Generation: Born 1945 and before. It will be important to the bank to estimate, which category is more suitable for each campaign and change strategies to approach different groups, and working with this categories will give them a clear path.

In order to minimize job categories, some jobs groups have been created. "admin.", "services", "technician", and "management" will be categorized as "white collar", which is a common term to categorized employment performed in

offices and that require formal education. To the category defined by the bank as “blue collar” the group housemaid was added. And, the entrepreneurs and self-employees have been jointed as only one group.

The variable default was drop from the dataset, since it is no logical that there is an expectation of a long-term investment from a client that is already in debt with the bank.

Another change effectuated, was the categorization of the variable “day”, which contains the day of the month when the client was contacted for the campaign, these were divided in two parts, as the first 15 days of the month and the second part the rest of the days.

The variable “month” was also aggrouped in four groups dividing the month of the last contact in quarters of the year, in order to determine seasons where is more likely to have a positive outcome from the client.

The variable “duration” has some marked positive skewness, in order to fix this problem and to make the variable more suitable to be significant for the models, has been transformed to a log form in search of symmetricity (Figure 1).

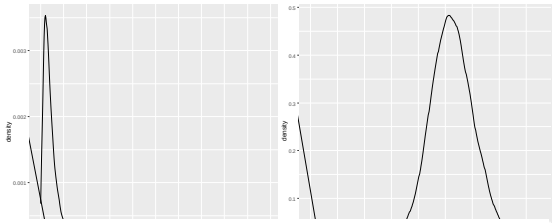


Figure 1. “duration” transformation

Additionally, “pdays” which englobes the number of days that passed after the client was contacted for previous campaigns has also been categorized in 6 groups based on the quintiles of the variable, where the first group will be the clients that were not contacted before.

Lastly, the “previous” variable was transform into a categorical type, by dividing the number of days that the client was contacted for previous campaigns onto five groups, less than 5 times , between 6 and 10 , and more than 10.

The rest of the variables, “marital, “education”, “balance”, “housing”, “loan”, and “campaign” have been not transformed. These variables seem to have not so much categories, and the data is balance between them. The variable “balance” presents some symmetry, but in order to log transform it, it would be necessary to add a constant, some transformation were tested but there was a considerable loss of observations without the constant, and when introduce it, again the plot do not show any improvement. This variable captures the true essence of the data, since it measures the income and expenses plus debt, giving a positive or negative outcome that can vary a lot from client to client, and will be maintained with its original measures.

CORRELATIONS BETWEEN VARIABLES

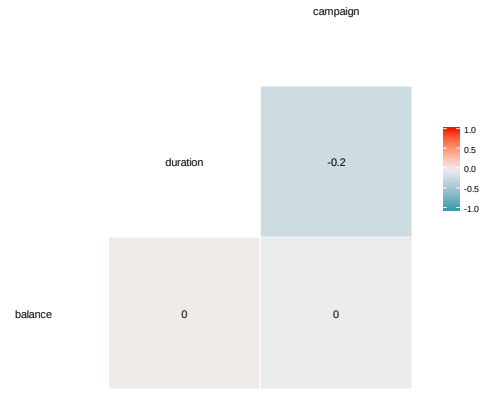


Figure 2. Correlation matrix

Between continuous variable, by performing a correlation analysis, it can be determined that there are no high correlations that could affect the analysis (Figure 2). Moreover, between the rest of the categorical variable and considering the response variable “y” it can be said by reviewing the following plots that there are no major tendencies that could affect the estimation (Figure 3). Interactions between variables were proved and just the variable “duration” seems to have a different behavior between groups of the variable “previous”. This will be treated in the modelization section.

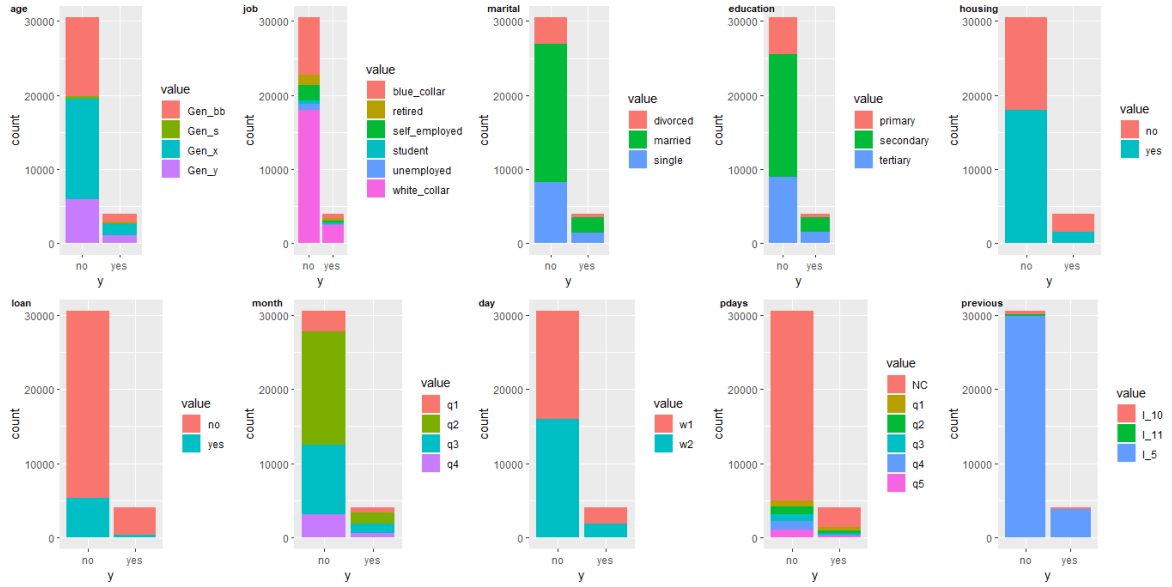


Figure 3. Qualitative variable plot, axis y = response variable

TREATMENT OF OUTLIERS

For all the variables that have not been categorized, there has been performed some boxplots in order to determined if there are some observations that show an abnormal behavior.

As seen before there are in fact some observations that could be to far from the mean and that need to be treated. The determination of outliers will be considered from two points of view, one considering each variable independent, and the other one considering the variable as if there were one multivariate group. The first approach considers an outlier to every observation points which lie beyond the extremes of the whiskers of the boxplot, these calculation were made through the **boxplot.stats()$out** from R. The other approach considers the Mahalanobis distance to calculate the observations that in an overall level differ from a normal behavior given by the rest of the observations. Considering that this calculation gave a less percentage of excluded data, and that considers the behavior of all the variables, this approach was followed to remove the outliers, which were 706 observations. After this process the plot shows better the variables.

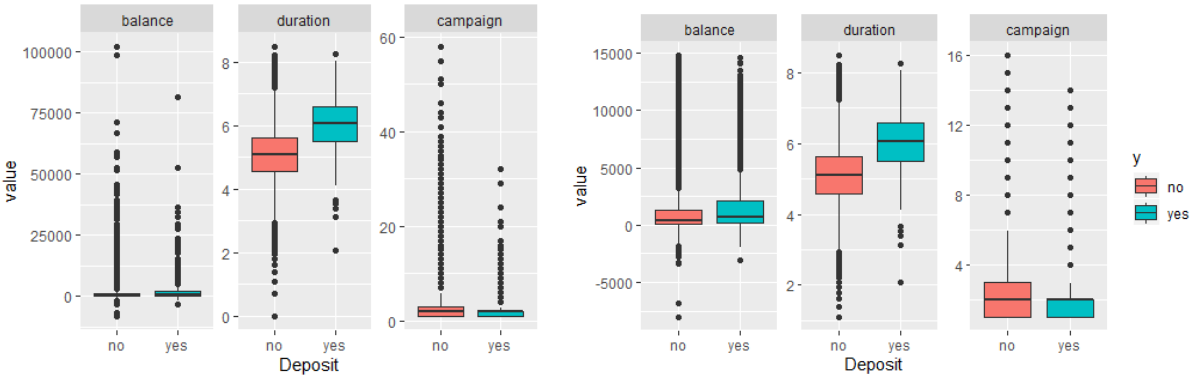


Figure 4. Quantitative variables, removal of outliers (Right without removal) (Left removal performed)

CLASSIFICATION METHODS

LOGISTIC REGRESSION

Logistic regression will model the probability that the response variable, in this case “y”, belongs to a category “yes” or “1” if there was effectuated a long-term deposit and “no”, which is the control group denoted by “0”, otherwise. Since, the response variable is a binary one, then:

$$E(y|x_1,\ldots,x_p)=p=P(y=1|X=x)=F(\beta_0+\beta^T x)$$

P, is a function that gives outputs between 0 and 1 for all values of x that are the predictors, then we can use the logit over this function called odds.

$$\text{logit}(p)=\log\frac{p}{1\text{-}p}=\beta_0+\beta^T x$$

The final goal is that the estimates for the β's putted onto the model for the logistic function p, expressed by the equation below, gives a result close to one for the clients that made a long-term deposit and zero for those who didn't.

$$p=F(\beta_0+\beta^T x)=\frac{\exp(\beta_0+\beta^T x)}{1+\exp(\beta_0+\beta^T x)}=\frac{1}{1+\exp(-\beta_0\text{-}\beta^T x)}$$

To find these estimates for the β's, the Maximum Likelihood is used.

$$\max\prod_{i=1}^{n} p_i(x_i|\beta)^{y_i}\left(1\text{-}p_i(x_i|\beta)\right)^{1\text{-}y_i}$$

Then, the estimate probability will be given by:

$$\hat{\rho}(x)=P(y=1)=\frac{\exp\left(\widehat{\beta_0}+\widehat{\beta}\ x\right)}{1+\exp\left(\widehat{\beta_0}+\widehat{\beta}\ x\right)}=\frac{1}{1+\exp\left(-\widehat{\beta_0}\text{-}\widehat{\beta}\ x\right)}$$

There is a boundary that will classify the observations, based on the resulted probability, in the classes, this could be 0.5, if it is assumed the symmetricity on the importance of both classes "yes" and "no" deposit. Consequently, y estimated will be classify as it will make a long-term deposit, when $\widehat{\beta_0}+\widehat{\beta}\ x\geq 0$, since in this case $\hat{\rho}(x)\geq 0.5$.

The final logistic model for the problem it is given by the variables: "age", "job", "marital, "education", "balance", "housing", "loans","month","duration","campaign","day","pdays", "previous", and an interaction between "previous" and "duration", which was the only interaction between categorical and continuous variables that result significant in an overall level. The results from the model are presented in the next table:

```
Call:
glm(formula = y ~ . + duration:previous, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7441  -0.4169  -0.2305  -0.1105   4.0164

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -8.928e+00  9.229e-01  -9.674  < 2e-16 ***
ageGen_s              1.153e+00  1.287e-01   8.959  < 2e-16 ***
ageGen_x             -2.097e-02  5.134e-02  -0.408 0.682988
ageGen_y              2.736e-01  6.441e-02   4.248 2.16e-05 ***
jobblue_collar       -1.953e-01  6.025e-02  -3.241 0.001191 **
jobself_employed     -3.490e-01  8.653e-02  -4.033 5.50e-05 ***
jobretired            1.612e-02  1.058e-01   0.152 0.878882
jobstudent            5.830e-01  1.215e-01   4.798 1.60e-06 ***
jobunemployed         2.021e-02  1.103e-01   0.183 0.854651
maritalmarried       -1.645e-01  6.451e-02  -2.549 0.010793 *
maritalsingle         4.574e-02  7.369e-02   0.621 0.534758
educationsecondary    2.480e-01  7.004e-02   3.541 0.000398 ***
educationtertiary     5.831e-01  7.755e-02   7.518 5.55e-14 ***
balance               4.180e-05  9.370e-06   4.461 8.18e-06 ***
housingyes           -8.950e-01  4.612e-02 -19.408  < 2e-16 ***
loanyes              -5.147e-01  6.437e-02  -7.996 1.29e-15 ***
monthq3               2.626e-02  5.268e-02   0.498 0.618151
monthq4               2.583e-01  6.789e-02   3.805 0.000142 ***
monthq1               3.454e-01  6.559e-02   5.266 1.40e-07 ***
duration              1.308e+00  1.629e-01   8.031 9.68e-16 ***
```

```
campaign               -9.493e-02  1.288e-02  -7.369 1.72e-13 ***
dayw2                  -1.995e-01  4.183e-02  -4.770 1.85e-06 ***
pdaysq1                 1.780e+00  7.702e-02  23.110  < 2e-16 ***
pdaysq2                 1.055e+00  8.573e-02  12.301  < 2e-16 ***
pdaysq3                 1.085e+00  9.290e-02  11.678  < 2e-16 ***
pdaysq4                 5.228e-01  1.035e-01   5.052 4.36e-07 ***
pdaysq5                 1.073e+00  9.778e-02  10.969  < 2e-16 ***
previousl_11           -6.807e+00  2.395e+00  -2.842 0.004488 **
previousl_5            -3.294e+00  9.316e-01  -3.536 0.000407 ***
duration:previousl_11   1.136e+00  4.196e-01   2.708 0.006770 **
duration:previousl_5    5.245e-01  1.655e-01   3.170 0.001527 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

To see if a coefficient is significant for the prediction of the response, a Wald Test is performed. By the command **glm()** in R we can obtained all the test for this model. By the statistic Z, the null hypothesis which tests that the coefficient is cero, can be rejected or not:

$$H_0: \beta_j = 0 \quad Z_0 = \frac{\widehat{\beta_j}}{\sqrt{\mathrm{Var}(\widehat{\beta_j})}}$$

Based on the p-value, we can conclude some effects by seeing which coefficient are significant and in which categories these effects differ.

By analyzing the variables that correspond to the client profile, we can conclude that, in fact, there are some difference on the effects between generations, without taking in account the rest of the predictors and the generation Baby Boomers as the base variable, it is seen that the group S, is more likely to invest, also the Y generation but in less proportion. In addition, there is a negative effect by comparing the white collar category of job from the blue collar, and the self employed and entrepreneurs, which is expected at first sight, on the contrary, a client which is a student is more likely to make a deposit than the others. Additionally, there is no difference if the client is retired or single in the decision of making a deposit, but there is indeed, a negative effect if the person is married. Another expected effect, isolating the rest of the variables, is the one from the level of education, since the fact that the client has a high level of education is expected to have a positive effect on the log-odds of the response.

From the information derived from the bank, we could say that with all the variables controlled, if the client has a housing loan or any kind of loan, the client is less likely to make a long-term deposit. On the contrary, an increase of the balance will all the variables fixed, increases the log-odds of the response variable, which makes the client more likely to make a long-term deposit.

The same effect in different proportions but in the same direction, has been identified for the variable “duration”, also, a combine effect of the variables “duration” and “previous” could be prove, and which can be interpreted as the effect of the duration of the call in the response variable, which will be different between the clients that have been contacted for previous campaigns, in consequence, if the person of the call center makes the call to last longer, and the person has been contacted several times before, then it is more likely to make a long-term deposit. Moreover, the number of contacts previous for different campaigns, which turn out to be favorable in order to increase the odds, lays between 5 and 9 times. On the contrary the variable “campaign”, which records the times that the client has been contacted for the same campaign turn out to have a negative effect, these could be translated in the fact that the client seeks to be inform of different content but the odds of a positive result go down if the client is stunned or bombarded with much of the same content.

In terms of the timing, it can be said that the effect from contact the client in the first 15 days of the month is positive, in terms of a better result. Also, the first and the fourth quarters of the year result more effective, at least compared to the second quarter. Lastly, there is a positive effect from contacting the client after it has already been contacted for previous campaigns, and this effect decreases as the days go by.

When testing the model with the 20% of the data, the results are the following:

LOGIT

Prediction	Reference: no	Reference: yes	Total
no	7428	735	8163
yes	194	264	458
Total	7622	999	8621

	no	yes
no	97.45%	73.57%
yes	2.55%	26.43%
Total	100.00%	100.00%

Accuracy	89.22%
Kappa	31.27%
Sensitivity	97.45%
Specificity	26.43%
Precision	91.00%
F1	94.11%

As it can be seen the accuracy, with a boundary of 0.5, is of 89.22%. Considering, that the classes for the response variable is "y" are unbalanced, it is needed to see more measurements in order to determine if it is a correct model, not just accuracy. As it has been already said, there are two mistakes that lead to a higher error and a less efficiency of the model, in this case there is a low level of specificity (26.43%) , which gives the percentage of the clients identify as the ones who will do a long-term deposit, from the total clients which indeed make a deposit. This means that the model will not be accurate when determining most of the deposits, but on the other hand, the level of sensitivity is high (97.45%), this shows that this model could predict in a very good way the clients that will not make a deposit. Additionally, there is a rate of 2.55% of the error understood as the one with higher cost. Moreover, based on what was found in the original study with all the data, the bank only had a rate of 8% of success, if we could avoid contacting all the clients that do not seem to have a suitable profile to make a deposit, then the rate of success could go up to 26%, this demonstrates the validity of the model for the purpose of determining the clients that need to be contact.

SHRINKAGE METHOD

Through the package caret and the function **glmnet()** a penalized regression was performed. This method optimizes the parameters α, the elasticnet mixing parameter, and λ, which determines the amount of shrinkage, in order to improve performance and found the best parameter that will minimize the cross-validation error. Since the classes are unbalanced, the default metric accuracy will not be a good metric, consequently, the performance metric was set to kappa, which adjusts accuracy by the possibility of a correct prediction obtained by chance. The **trainControl** object from the package *caret* allows to specify a re-sampling method, in this case, it was used a non-repeated cv with 5-fold cross validation. The α was set to 0.4, and λ=0, so the penalty term has no effect, in this case. The results are closer to the ones from the logistic regression, this could be attributed to the fact that the shrinkage measure was zero, these results are showed below.

PENALIZED LOGIT

Prediction	Reference: no	Reference: yes	Total
no	7427	740	8167
yes	195	259	454
Total	7622	999	8621

	no	yes
no	97.44%	74.07%
yes	2.56%	25.93%
Total	100.00%	100.00%

Accuracy	89.15%
Kappa	30.63%
Sensitivity	97.44%
Specificity	25.93%
Precision	90.94%
F1	94.08%

BAYES CLASSIFIERS

In contrast of the regression, where the conditional probability is given directly, the Bayes classifiers, model separately each predictor x_i for each class of the response variable, then $p(y|x_1, \dots, x_p)$, is calculated by the Bayes formula. The Bayes classifier assigns each observation to the most likely class given its predictor values, and minimizing the classification error (James, G. et al., 2015, p.37). The results from applying some of these classifiers will be show next.

LDA

Prediction	Reference: no	Reference: yes	Total
no	7407	740	8147
yes	215	259	474
Total	7622	999	8621

	no	yes
no	97.18%	74.07%
yes	2.82%	25.93%
Total	100.00%	100.00%

Accuracy	88.92%
Kappa	29.94%
Sensitivity	97.18%
Specificity	25.93%
Precision	90.92%
F1	93.94%

QDA

Prediction	Reference: no	Reference: yes	Total
no	6786	592	7378
yes	836	407	1243
Total	7622	999	8621

	no	yes
no	89.03%	59.26%
yes	10.97%	40.74%
Total	100.00%	100.00%

Accuracy	83.44%
Kappa	26.92%
Sensitivity	89.03%
Specificity	40.74%
Precision	91.98%
F1	90.48%

NAÏVE BAYES

Prediction	Reference: no	Reference: yes	Total
no	7148	661	7809
yes	474	338	812
Total	7622	999	8621

	no	yes
no	93.78%	66.17%
yes	6.22%	33.83%
Total	100.00%	100.00%

Accuracy	86.83%
Kappa	30.06%
Sensitivity	93.78%
Specificity	33.83%
Precision	91.54%
F1	92.64%

The linear discriminant analysis o LDA, is an alternative approach that models the distribution of the predictors x separately in each of the response classes, and then uses Bayes' theorem to convert them into estimates for P($y \in g$ |X = x) (James, G. et al., 2015, p.138). The LDA involves the assumption that all the predictors form a multivariate Gaussian distribution with a mean vector for each class and a common same covariance matrix. The posterior probabilities are given by:

$$p_g(x)=P(y\in g|X=x)=\frac{f_g(x)\pi_g}{\sum_k f_k(x)\pi_k}$$

Where, $f_g(x)$ is the multivariate distribution, and π_g *the* prior probability known, that an observation belongs to group g. π_g can be estimated using the proportion of training observations that belong to class g: $\hat{\pi}_g$ = ng /n. This method also requires the definition of a threshold for the calculation of the posterior probability in order to assign an observation to a class. This threshold could be optimized if it is seen that it is preferred a balance of the errors or if one is more important than the other. With the purpose of comparing, at first glance the threshold assign is 0.5.

Quadratic discriminant analysis or QDA, is another alternative method which unlike LDA, it does not assume that there is a common covariance matrix for all the variables. QDA, involves a quadratic function of the predictors, and the estimates obtain are unbiased but with a higher variance than the ones of LDA. Moreover, the QDA can model a larger range of problems and fits more flexible classifiers and is recommended if the training set is very large, so when the variance of the classifier is not a major concern (James, G. et al., 2015, p.150 & p.151).

On the other hand, Naïve Bayes assumes that variables are independent, so interactions are not considered. This is the reason why the model tested is the used in the previous methods, but without the relation between the variables.

By seeing the results from these three methods, it can be said that the Naïve Bayes method, gives a larger kappa, additionally, a bigger F1 which involves the metric recall, the same as sensitivity, and precision. The metrics Precision and F1 have been calculated for the positives which in this case are the cases of the clients which did not make a deposit, since we are interested in telling which customers the bank should or should not contact. If we contemplate the numeric value of the kappa in all the methods, we see that it is low compare to other metrics, these could be explained by the fact that the classes are un-balanced and by making them equally probable, the good indicators go down. In order to see how the methods, react to a proportional class, instead of using the proportions found in the training set, the prior probabilities were set to be 0.5. The results show that the error which involves a higher cost, that is if a client is classified as one who will make a deposit and at the end he did not, is if fact larger. Besides that, if we consider that a client is equally likely to make a deposit and not, the final rate of clients with a long-term deposit increase, the LDA shows a rate over 27% and the QDA of 18%, which is highly superior than the one the bank had, consequently, it will not be accurate to take in account such a conservative prior rate, since we are overestimating the clients that will indeed make a deposit. This approach will work for another face, once the clients have been discriminated, then there could be a possibility that they have almost the same probability of making or not the deposit that will depend on other variables, but in this case we known no just because of the sample, but even the real rate of deposits for this bank is lower.

CONSERVATIVE LDA

Prediction	Reference: no	Reference: yes	Total
no	6073	180	6253
yes	1549	819	2368
Total	7622	999	8621

	no	yes
no	79.68%	18.02%
yes	20.32%	81.98%
Total	100.00%	100.00%

Accuracy	79.94%
Kappa	38.65%
Sensitivity	79.68%
Specificity	81.98%
Precision	97.12%
F1	87.54%

CONSERVATIVE QDA

Prediction	Reference: no	Reference: yes	Total
no	6496	521	7017
yes	1126	478	1604
Total	7622	999	8621

	no	yes
no	85.23%	52.15%
yes	14.77%	47.85%
Total	100.00%	100.00%

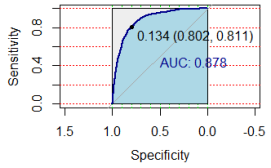
Accuracy	80.90%
Kappa	26.19%
Sensitivity	85.23%
Specificity	47.85%
Precision	92.58%
F1	88.75%

Moreover, considering the results from the Bayes methods, shrinkage and the logistic regression, the one with more accuracy is the logistic one. If we take in account the rest of the measures previously shown, we can see that the sensitivity, the F1, precision and kappa are larger for the last two, and the type one error, which seems the smaller from all the tests. This goes with the theory, which says that the logistic models are the best for binary response variables, and this approach performed well with linear boundaries. Considering that the parameter alpha in the penalized regression was positive, then it is better to use this method.

BAYES RULE AND COST-SENSITIVE LEARNING

Furthermore, the thresholds or the boundaries that are used to classify the observations, can be optimized. Changing the thresholds, is a tool that allows us to control better the errors, by increasing the one with less repercussions. There are two approaches in order to calculate the optimal value, the cost-sensitive learning which requires a domain knowledge of the application and the ROC curve, which is commonly use when there is no information available.

The Receiver Operating Characteristic or ROC curve shows true positives vs false positives in relation with different thresholds, there are several packages in R to calculate it, the one used is **pROC**, the plot shows the exact result which gives an area under the curve of 0.8711, which is a good indicator since measures the overall performance of the classifier, and a threshold of 0.134. Also by using the metric ROC in the package caret, we can determine the best parameters for a penalized regression, the results of this model are much more accurate for this case, as the table shows, the F1 is larger because of the sensitivity and the type one error is smaller also, of course there is a cost which involved the decrease of the specificity.



PENALIZED LOGIT ROC

Prediction	Reference: no	Reference: yes	Total
no	7487	795	8282
yes	135	204	339
Total	7622	999	8621

	no	yes
no	98.23%	79.58%
yes	1.77%	20.42%
Total	100.00%	100.00%

Accuracy	89.21%
Kappa	26.16%
Sensitivity	98.23%
Specificity	20.42%
Precision	90.40%
F1	94.15%

If we classify the probabilities from the penalized linear regression with the threshold of 0.13, given by the ROC curve, the errors are balanced, as it can be seeing, the type one increases to 22% while the type two decreases to 17%. Even though, this approach will be correct if we would like to give up some sensitivity in order to get more specificity, but in this case, the error are not balance because there is one error that it is more expensive in terms of the use of resources, therefore, another approach is needed.

THRESHOLD 0.3 PLR

Prediction	Reference: no	Reference: yes	Total
no	5939	173	6112
yes	1683	826	2509
Total	7622	999	8621

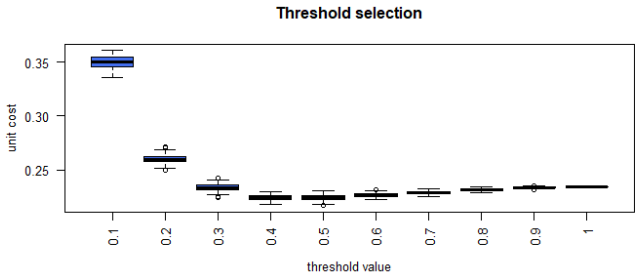
	no	yes
no	77.92%	17.32%
yes	22.08%	82.68%
Total	100.00%	100.00%

Accuracy	78.47%
Kappa	36.58%
Sensitivity	77.92%
Specificity	82.68%
Precision	97.17%
F1	86.49%

The sensitivity cost analysis, on the other hand, will prove exactly how much cost implies the misclassification of the observations, for this, the bank will have to be aware of the costs that implies each error of its call center. For example, we could determine the cost of the marketing campaign if we say that the cost of contact a client is 1€, if this client does not make a deposit at the end, in order to replace this client we will have to contact another one, so the cost of this

error will be 2€. On the other hand, if we know that we did not have to call a client because he is no likely to make a long-term deposit, the bank will not lose anything. Additionally, if we do not contact a client but in fact he did a deposit, the bank do not lose, but maybe if the bank had contacted him, the deposit will be effectuated sooner and one less call could have been done, so the cost will be 1€.

By effectuating 100 test and train divisions and modeling for each train part a penalized logistic regression, which was tested after, and calculating the cost derivate to the misclassification of the prediction error with every possible threshold from 0.1 to 1, we got an estimated mean and median cost for each threshold. The results tell us that the threshold which minimizes the cost taking as a measure the median is 0.4, moreover, the optimal threshold for the mean is 0.5.



Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Median	0.350089	0.259669	0.233392	0.224314	0.224387	0.226971	0.228964	0.231178	0.233688	0.234426
Mean	0.350214	0.260258	0.233248	0.224309	0.22429	0.226952	0.22881	0.231273	0.233695	0.234426

The final prediction uses a combination of both thresholds 0.45, which gives a final cost of 0.2240 € per client after testing the linear penalized regression with the testing dataset the results show good metrics, the sensitivity which measures the rate of success on determining the clients that will not make a deposit is high. Additionally, the specificity is better than other methods, and finally the type one error rate, which in this case corresponds to the clients that were catalog as suitable to make a long-term deposit and in the end didn't, is low, and the most important thing, it minimizes the cost involve, which is the final goal. Through the profile determined as the most likely to make a deposit, in the logistic regression, it is possible for the bank to tell who are the best clients and to improve the selection of costumers to contact, in that way they will get a higher success rate and with less resources involved.

THRESHOLD 0.45 PLR

Prediction	Reference: no	Reference: yes	Total
no	7371	681	8052
yes	251	318	569
Total	7622	999	8621

	no	yes
no	96.71%	68.17%
yes	3.29%	31.83%
Total	100.00%	100.00%

Accuracy	89.19%
Kappa	35.10%
Sensitivity	96.71%
Specificity	31.83%
Precision	91.54%
F1	94.05%

Bibliography

Moro, S., Laureano, R., and Cortez, P. (2011). *Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.* In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October 2011. EUROSIS. Retrieved from: https://core.ac.uk/download/pdf/55616194.pdf

Kuhn, M. (2019). *The caret Package*. Retrieved from: https://topepo.github.io/caret/index.html

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2015). An Introduction to Statistical Learning with Applications in R. Springer. Heidelberg Dordrecht, London.

Center of Generational Kinetics. *Generational Breakdown: Info About All of the Generations.* Retrieved from: https://genhq.com/faq-info-about-generations/