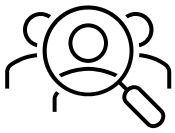
MARKETING CAMPAIGN



1. Analysis objective

The 'Marketing Campaign' dataset contains data about the outcome of a marketing campaign, with information about some characteristics regarding the potential customers contacted. The purpose of this report is to identify the effectiveness of this campaign in terms of actual users acquired, ie those who have actually signed up to the contractual proposal. There are also personal explanations and considerations aimed at improving the final results. In light of the intended goal, it is believed that the implementation of a logistic regression model is appropriate to obtain the desired information.

2. Description of the dataset

The dataset is made up of 41,188 records and 19 features that are very varied and informative, categorical and continuous, not very correlated with each other. The preliminary exploratory analysis provided an overview of how the data are distributed among the variables considered:

- On average, the subjects contacted are forty years old, they are mostly workers and technicians, most of whom are married.
- On the economic side, most of them do not have overdue credits and have no personal debts, while more than 50% of these still have an active mortgage on the property.
- contact with these subjects took place via mobile phone while for some of them also via landline.
- 86% of users have never been contacted before while 11% have been contacted once in some previous campaign.

3. Analysis process

The path that led to the definition of the problem starts from a correlation matrix of the variables in question to identify which of them are dependent. There were no highly correlated variables, only a slight positive dependence between the rate of change in employment, the three-month 'Euribor' index and the consumer price index, as shown in Figure 1.

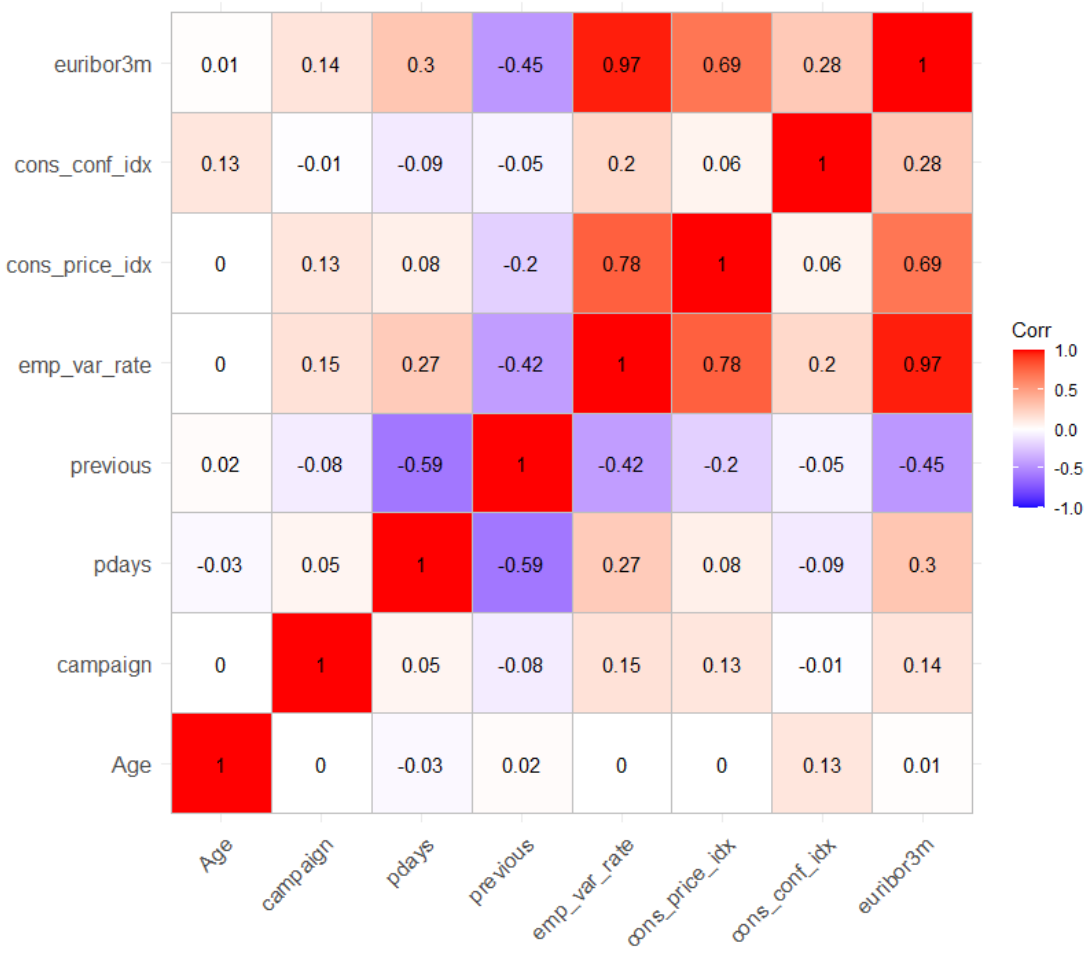


Figure 1- correlation analysis

A Likelihood-Ratio test was performed to verify that the model with multiple regressors is actually better than the model with only intercept. Looking at the results in Figure 2, it is possible to confirm that the regressors are able to explain the dependent variable better than the intercept-only model.

```
Likelihood ratio test

Model 1: Target ~ Age + Job + Maritial_Status + Education + Previous_Default +
    House_Ownership + Existing_Loans + Contact_Channel + Month +
    Day_Of_Week + campaign + pdays + previous + poutcome + emp_var_rate +
    cons_price_idx + cons_conf_idx + euribor3m
Model 2: Target ~ 1
  #Df LogLik  Df  Chisq Pr(>Chisq)
1  19 -11556
2   1 -14499 -18 5886.3  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Figure 2- Likelihood ratio test

Now, however, the question arises as to which are the best regressors, that is, those who best fit the forecast. To that question, the answer as follows:

to identify the most significant variables to be used in the logistic regression model, we proceed with a stepwise selection based on the lowest AIC value. The results of this selection allow to eliminate insignificant variables, so as to create the model with only the highly qualifying variables.

The features selected concern the variable 'Job', 'Contact_Channel', the 'Euribor' index, the consumer confidence index, the consumer price index, the rate of change in employment, the last month and contact day, 'pdays', 'campaign', 'previous default' and finally, the outcome of the previous campaign, for a total of twelve variables.

Subsequently, the analysis involves the creation of the training set consisting of approximately 28,000 records and the use of the latter to train a logistic regression model, using both the 'logit' and the 'probit' links as hyper-parameters, with almost similar results, with some variations as regards the significance of the individual coefficients.

What results were obtained from the logistic regression model?

```
Call:
glm(formula = Target ~ ., family = binomial(link = "logit"),
    data = Marketing.training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9932  -0.3813  -0.3190  -0.2694   3.0712

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.737537   0.168399 -10.318  < 2e-16 ***
Job               0.002526   0.005830   0.433 0.664807
Previous_Default -0.331895   0.066420  -4.997 5.83e-07 ***
Contact_Channel  -0.884389   0.065150 -13.575  < 2e-16 ***
Month            -0.072738   0.008377  -8.683  < 2e-16 ***
Day_Of_Week       0.054848   0.015313   3.582 0.000341 ***
campaign         -0.150176   0.032154  -4.671 3.00e-06 ***
pdays            -0.162254   0.018520  -8.761  < 2e-16 ***
poutcome          0.544754   0.057180   9.527  < 2e-16 ***
emp_var_rate     -1.359462   0.111667 -12.174  < 2e-16 ***
cons_price_idx    0.812320   0.040125  20.244  < 2e-16 ***
cons_conf_idx     0.231284   0.019496  11.863  < 2e-16 ***
euribor3m         0.067140   0.096562   0.695 0.486861
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20397  on 28962  degrees of freedom
Residual deviance: 16065  on 28950  degrees of freedom
AIC: 16091

Number of Fisher Scoring iterations: 6
```

Figure 3- Logistic regression model

First of all, observing Figure 3 we can see how the residuals have a negative median but close to zero, equal to -0.31.

Among the numerous variables taken into consideration, it would seem that the most significant are represented by the Euribor index, consumer price index, rate of change in employment, the success of the previous campaign, the number of times the operator has contacted the customer in this campaign, the method of contact and, to a lesser extent, the type of work that the customer carries out.

The confusion-matrix on the test-set shows how the model is able to predict failure cases well, while a certain difficulty arises when trying to predict success cases. The model predicts well 11,000 cases out of about 12,000, the accuracy is 89%, but at the same time the specificity is 19%, which means that the model does not predict successful cases well, i.e. the cases in which the customer subscribes. the contract offered. The sensitivity calculation instead shows how the model achieves a 98% forecast of failure cases.

These results are obtained with a theshold of 0.56. what happens if the threshold is reduced or increased?

By reducing the theshold to 0.2, the accuracy is slightly reduced to 85% but at the same time, the specificity significantly increases up to 56%. If the objective were to correctly predict those who actually signed the contract, it is advisable to set a low threshold, more tending to zero.

By increasing the theshold to 0.8, on the other hand, poor results are obtained both on the accuracy side which is reduced, and on the specificity side which drops to 5%.

Having analyzed the results obtained, it is assumed that a theshold lower than 0.5, can significantly improve the ability of our model to predict the users who have actually signed the contract.

The Mc-Fadden R-Square is 21%, and indicates how well the model is able to explain the outcome of the campaign with the chosen regressors. It is not an optimal result, but it is still the best value obtained with the following data. The Pseudo-R-Square instead is further reduced to 14%.

Next, the ROC curve and the Under Curve Area are shown, in figure 4.

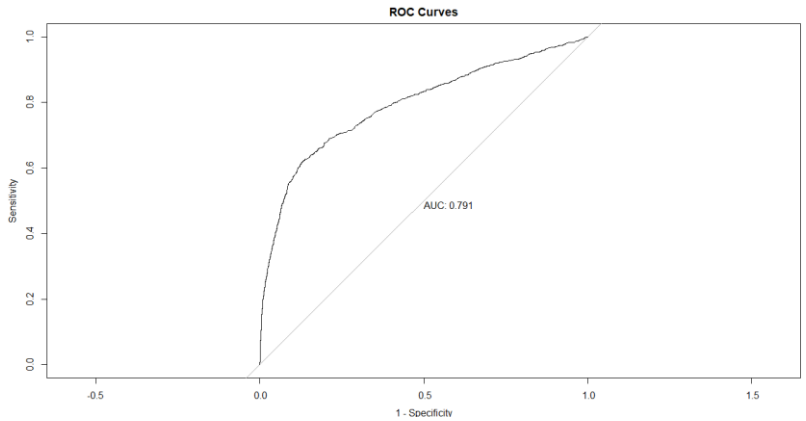


Figure 4- ROC Curve and AUC

Given the results obtained and the shape that the ROC curve takes, it is possible to say that the model is all in all good, the area under the curve (AUC) is 79%.

4. Interpretation of the results

The outcome of the campaign is successful for 11% of cases, while for the remainder it is unsuccessful. Is it a good result? Depends.

Surely, through the research carried out, it is possible to obtain more indications to significantly improve the percentage of success cases.

It can be seen that as the three-month Euribor index and the consumer price index increase, the possibility of customers signing the contract decreases. This is because, as noted by the exploratory analysis of the data, the majority of people still have a mortgage on the property, and obviously the more the interest on the mute increases, the more the user will be less inclined to subscribe to another policy. It may be good practice to contact these specific clients at times when the index in question is minimal, so that the chance of success increases.

In the same way, however, as the rate of change in employment and the number of times that the operator has contacted the customer in this specific campaign increases, the possibility of the latter signing the proposed contract significantly increases. Evidently, the higher the employment rate, the greater the number of people with jobs and the greater the positive sentiment towards these proposals. As for the number of times the subject has been contacted, it is likely that those who are contacted over and over again is because they initially showed interest in the offer but still undecided and not really convinced about what to do; therefore, further contact could facilitate the agreement between the parties and conclude the agreement.

Finally, it would seem that those who perform the duties of blue collar and service clerk are more likely to sign the contract, rather than those who perform managerial or entrepreneurial roles. This result may prompt the bidder to focus on these subjects, focusing their efforts on improving their success rate.

Work done by Ridolfi Lorenzo