Analysis of a Direct Marketing Campaign
Analysis of Discrete Data

Anastasia Nica m20210516 | Bruna Ribeiro m20210226 | Maria Machado m20210371
Nova Information Management School

Introduction

The group decided to analyze a direct marketing campaign of a Portuguese banking institution, that had the goal to communicate a new product, more precisely a bank term deposit. The purpose of this investigation is to understand how the characteristics of a certain costumer affect their reaction to the campaign, more precisely the probability of subscription of a bank term deposit.
Additionally, it was observed that only a few costumers decide to subscribe the new product, so it is also intended to give some tips of how to proceed in next campaigns and to build the profile that is more likely to be interest in this king of product.
Given this goal, we defined our research questions:

- How do the characteristics of the costumers explain the probability of subscription of a bank term deposit?
- Is there a pattern in the behavior of costumers that are more likely to subscribe a bank term deposit?

Materials

First, it was needed to clean the data so, we decided to cut some variables and deleted the observations that were classified as “unknown”. It was decided to do that because such classifications would impact the results.

In our database, there were more explanatory variables. However, we decide to focus only in ones that are important to our study.

After that, it’s important to clarify the description of the variables presents in the model.

Firstly, we have the explanatory variables:

Variables	Type	Description	Values
Age	Numeric	age of the costumer	[17, 98]
Job	Categorical	type of job	'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed'
Marital	Categorical	marital status	'divorced', 'married', 'single'
Education	Categorical	Level of education	'illiterate', 'basic.4y', 'basic.6y', 'basic.9y', 'professional.course', 'high.school', 'university.degree'
Housing	Categorical	has housing loan?	'no', 'yes'
Loan	Categorical	has personal loan?	'no', 'yes'
Campaign	Numeric	number of contacts performed during this campaign and for this client	[1, 43]

Table 1: Explanatory variables

Then, we have the response variable:

Variables	Type	Description	Values
Y	Categorical	has the client subscribed a term deposit?	'no', 'yes'

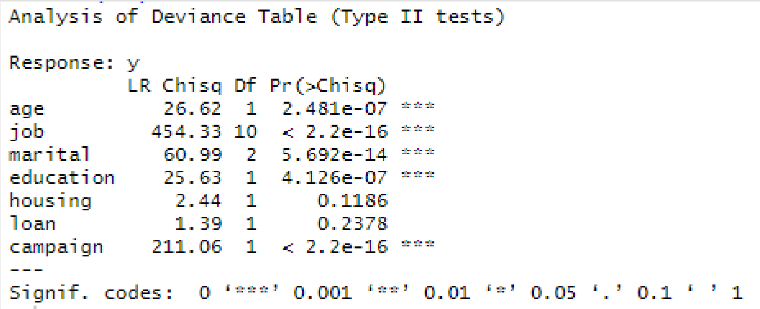
Table 2: Response variable

Methodology

The results of the tests were obtained trough codes in the R Studio.

Logistic regression was used to model the probability of the client to subscribe a term deposit.

To test what explanatory variables are significant through the analysis, it was performed the LR test:



```
Analysis of Deviance Table (Type II tests)

Response: y
          LR Chisq Df Pr(>Chisq)    
age          26.62  1  2.481e-07 ***
job         454.33 10  < 2.2e-16 ***
marital      60.99  2  5.692e-14 ***
education    25.63  1  4.126e-07 ***
housing       2.44  1     0.1186    
loan          1.39  1     0.2378    
campaign    211.06  1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Output 1: LR test

By this output, we can conclude that for most of the variables, the p-value is a number lower than 0.05. Because of this conclusion, we should reject $H0$ at the 5% significance level.
By this, we mean that we reject the hypothesis that the corresponding variable is not significant. Only the p-values of 0.1186 and 0.2378 are higher than 0.05, so we conclude that the variables loan and housing are not significant. Therefore, these variables will not be in the final model.

Analysis/ Results

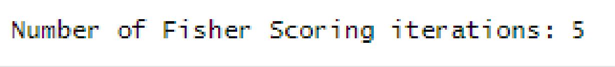
Final Model

In the definition of the final model, the software defined the baseline categories for the variable job as the administrator and for the variable marital as the category divorced. So, the final model is presented below.

$$logit[\pi(x)] = -2{,}57 + 0{,}01 \times age - 0{,}47 \times jobbluecollar - 0{,}39 \times jobentrepreneur - 0{,}13 \times jobhousemaid - 0{,}16 \times jobmanagement + 0{,}79 \times jobretired - 0{,}15 \times jobselfemployed - 0{,}45 \times jobservices + 1{,}03 \times jobstudent - 0{,}16 \times jobtechnician + 0{,}21 \times jobunemployed + 0{,}08 \times maritalmarried + 0{,}38 \times maritalsingle + 0{,}067 \times eductation - 0{,}12 \times campaign = z$$

$$\pi(x) = \frac{e^z}{1 + e^z}$$

In this model the fisher number is 5, which means that were necessary 5 iterations before finding the maximum likelihood estimates.

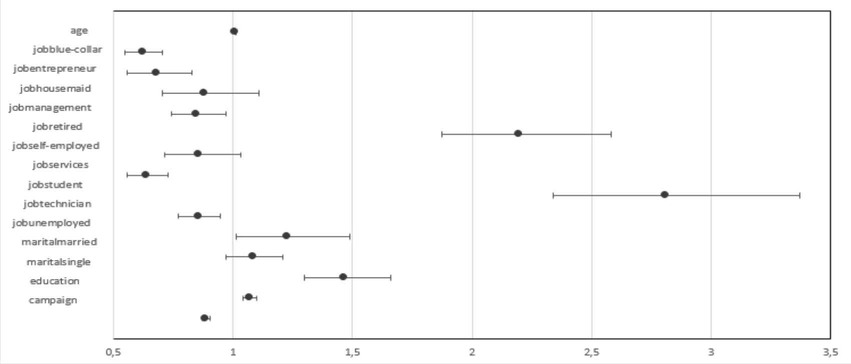


```
Number of Fisher Scoring iterations: 5
```

Odds Ratio

Having in mind the final model it is, now, possible to make the odds ratio analysis for each variable, holding constant the remaining ones. For the variables age, education and campaign it will be analysed the effect of an unit increase, for example, for the variable age, the odds of subscription of the bank term deposit increases in 1,01.

For the variables job and marital status, the analysis in the graphic, have in consideration the comparison between baseline category and each category.



Graphic 1: Odds Ratio and respective 95% confidence interval

Comparing the 2 jobs that have higher odds (retired and student), we conclude that the students are the ones with higher odds of subscription of the bank term deposit. And, applying the same rationale to the jobs with smaller odds (blue-collar and services), we conclude that the costumers with blue-collar jobs are the ones with smaller odds of subscription of the product.

$$odds(student) = e^{1{,}03-0{,}79} \times odds(retired) = 1{,}27 \times odds(retired)$$

$$odds(blue - collar) = e^{-0{,}47-(-0{,}45)} \times odds(services) = 0{,}98 \times odds(services)$$

By the same logic, we can conclude that the odds of subscription of the bank term deposit are higher for the singles costumers, keeping the remaining variables constant. The divorced costumers show smaller odds.

$$odds(married) = e^{0{,}082-0{,}38} \times odds(single) = 0{,}74 \times odds(single)$$

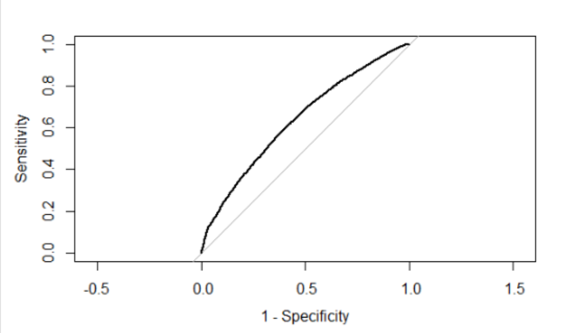
Classification Table

Sample proportion of 1’s for y variable: 𝑝@◌ↄ

Actual	Prediction, $\pi_o = 0,1113348$		Total
	y@◌ↄ=1	y@◌ↄ=0	
y=1	2548	1710	4258
y=0	13658	20329	33987

Table 3: Classification Table

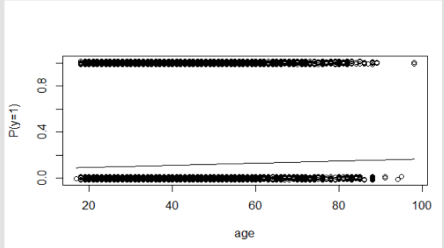
Of the 33987 cases with y = 0, the model predicts 𝑦@◌ↄ= 0 for 20329.
In relation to the 4258 cases with y = 1, the model predicts 𝑦@◌ↄ = 1 for 2548.
Sensitivity = 0,59840
Specificity = 0,59814
We can also obtain the proportion of correct classifications = 0,5982, which means that the model predicts correctly 60% of the time.



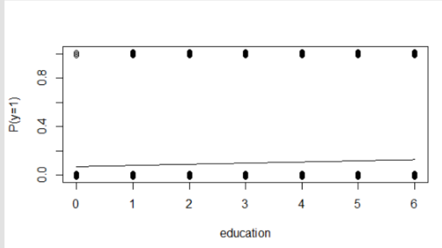
Graphic 2: ROC curve for logistic regression model

This curve shows the sensitivity and the specificity of the predictions for all possible cutoffs π_0. The area under the curve, that summarizes the predictive power, is equal to 0,6362, which gives us the estimated probability that the predictions and the outcomes are concordant.

Logistic Regression Function

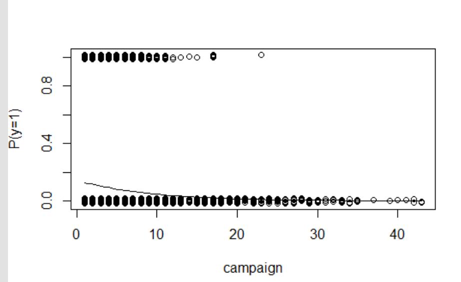


Graphic 3: Logistic regression function with age as factor predictor



Graphic 4: Logistic regression function with education as factor predictor

By these graphics, we can confirm that β_1 and β_{14} are positive, which means that when the ages and level of education increase, the probability of subscription of a bank term deposit also increases.



Graphic 5: Logistic regression function with campaign as factor predictor

By this graph, we can confirm that β_{15}, is negative. This means that when the number of campaigns increase, the probability of subscription of a bank term deposit decreases.
In these graphs we see that the probability of subscription of a bank term deposit never gets a value above 40%. Logically, the symmetric point in the age/campaign corresponds to a value that doesn’t make sense (248.2712/-21,40432, respectively), because to reach a probability of 50% we need to have a huge increase in age and decrease in campaign.

Conclusion

In conclusion, it was performed a logistic regression model with the goal of understanding the impact of the explanatory variables in the response variable.
We concluded that the marketing campaign wasn’t successful because only a low percentage of the clients shows interest in subscribe a bank term deposit.

Recommendations

Since the conversion rate of this marketing campaign was very low (0,11), we suggest for the next campaign to investment in the profiles with the highest estimated probability of subscription of a bank term deposit: older, student, single, high level of education, contacted few times.

References

Database:https://www.kaggle.com/henriqueyamahata/bank-marketing
Livro: Agresti, A., 2019. An Introduction to Categorical Data Analysis. 3rd ed. Wiley.