
**Course:** Statistics for Business Analytics II

**Project I**

**Professor:** D. Karlis


**Student**

**Name:** Loukopoulos Orestis

**AM:** f2822104

**Program:** Full Time 2021-2022


ATHENS, 2022

Index

# Chapter 1

# Introduction – Description of the problem

In this assignment we have been given a dataset related to telemarketing phone calls. This data set includes data collected from a retail bank, from May 2008 to June 2010, in a total of approximately 40 thousand phone calls. With these phone calls bank aims to sell long-term deposits. The result of these calls is a binary variable (SUBSCRIBE) describing whether the contact was successful or unsuccessful, which means if the customer has subscribed to the product or not.

The main aim of this assignment is to find what influences customers to subscribe to the product.

# Chapter 2

# Further Analysis

To begin with, we inserted the data set in R. We did not find any NAs, but we had to change the data types to the correct ones. We changed variables job, marital, education, default, housing, loan, contact, month, day_of_week, poutcome, and SUSCRIBED (response variable) from character to factors. Furthermore, we noticed that the strict majority, almost 97% of the observations, of the variable which describes the number of days that passed by after the client was last contacted from a previous campaign (pdays) has 999 as value, which means that almost all of our clients have never been contacted by a previous campaign. We consider this column to be useless, as it does not offer any important information to the problem we are trying to solve. We also figured out that we take the same information by the variable previous which indicates the number of contacts performed before this campaign and for every client. So, we removed the variable pdays.

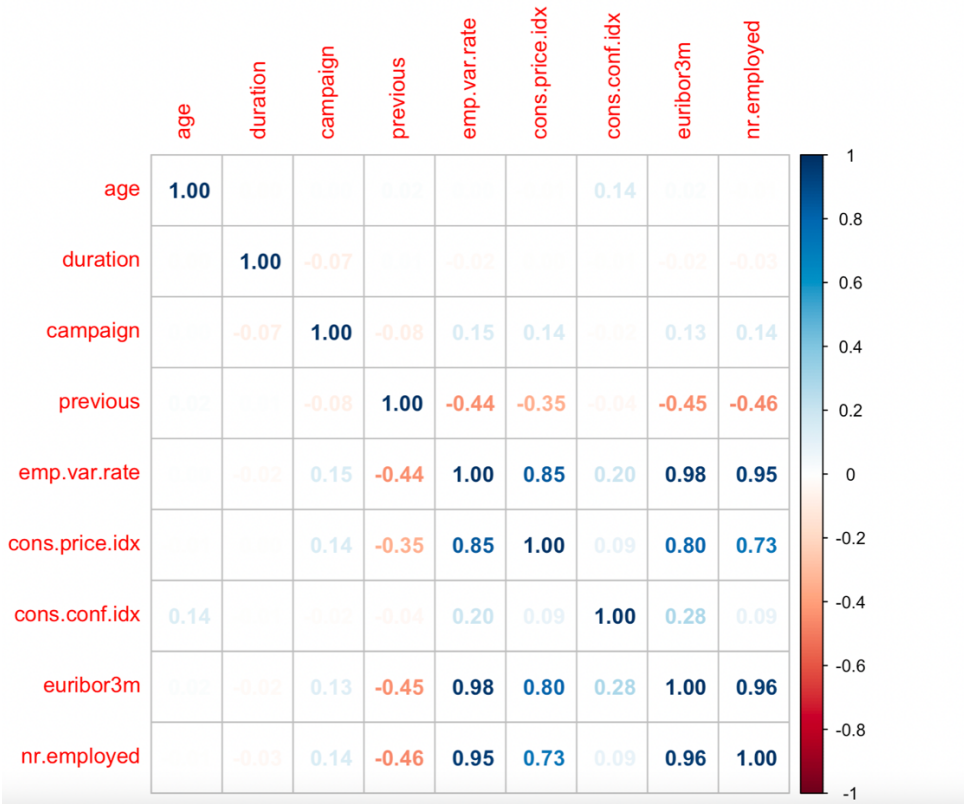


***Figure 1**: Correlation between numeric variables*

Continuing, we checked the correlation between the numeric variables. From the above figure we can see that there is a strong positive linear relationship between:

- consumer price index (cons.price.idx) and employment variation rate (emp.var.rate) with ρ = 0.85
- consumer price index (cons.price.idx) and euribor 3-month rate (euribor3m) with ρ = 0.80
- consumer price index (cons.price.idx) and indicator of number of employees (nr.employed) with ρ = 0.73
- euribor 3 month rate (euribor3m) and employment variation rate (emp.var.rate) with ρ = 0.98
- euribor 3 month rate (euribor3m) and indicator of number of employees (nr.employed) with ρ = 0.96
- indicator of number of employees (nr.employed) and employment variation rate (emp.var.rate) with ρ = 0.95

We took a first insight about the problem of multicollinearity. We will see how we will tackle it further on our analysis.

We will continue our analysis by defining our full model. As we mentioned before, we want to find which variables contribute to a subscription to the product by a client. Hence, we will take the variable SUBSCRIPTION as response variable on our GLM model. As, this variable is binary (takes only two values: YES, NO) we will implement logistic regression with the logit function as a link function.

$$\boldsymbol{log}\,\frac{\boldsymbol{pi}}{\boldsymbol{1-pi}}\ \textbf{, with } p_i \in (0,\,1)$$

With $p_i = P(Y_i = 1)$: the probability of a successful subscription (YES) for the *i*–th client.
With $Y_i \sim B(1, p_i)$ independent for $i = 1,\ldots, 39883$ and $Y_i = SUBSCRIPTION_i$

After the transformation we performed, we will define our full model. In order to remove any unnecessary variables from our model we implemented Lasso selection technique, which will help us select the covariates of our model. After executing Lasso, we will try to select lambda(λ) using cross validation process. The first candidate is the lambda (λ) at which the minimum CV-MSE is achieved but it is likely that this model has many variables. The second

candidate is the largest lambda(λ) at which the CV-MSE is within one standard error of the minimum CV-MSE. It is typical to choose the second, CV-MSE minimized 1se lambda(λ). (Where MSE: Mean Square Error, CV: Cross Validation)

As said before, we selected 1se lambda for our model selection (Lasso). So, after executing Lasso a number of times, we select those variables that are selected more often. The new model after the above process has SUBSCRIBED as response variable and contains the following variables as covariates:

- job
- marital
- education
- default
- housing
- loan
- contact month
- day_of_week
- duration
- campaign
- poutcome
- emp.var.rate
- cons.price.idx
- nr.employed

Subsequently, we used stepwise procedure according to BIC, as we want to do an inference. We applied this method to the model we concluded from Lasso procedure. After the stepwise method, our new model now has SUBSCRIBED as response variable and default, contact, month, duration, campaign, poutcome, emp.var.rate, cons.price.idx, nr.employed variables as covariates.

After concluding to the above-mentioned model, we should start checking for multicollinearity. We removed emp.var.rate variable as it had a very high GVIF and caused a multicollinearity problem. We then execute a summary for our model after removing the variable mentioned above. We found out that cons.price.idx was a statistically insignificant

covariate. We also executed a Wald test in that term and got a very large p-value (e.g., equal to 0.92) which means that in level of significance of 95% we failed to reject the null hypothesis that the coefficient of cons.price.idx is equal to zero ($H_o$: $\beta_{18} = 0$). To wrap things up, the hypothesis that consumer price index does not affect the probability of a subscription by client is not rejected.

Continuing, we performed a Likelihood Ratio Test between the model with cons.price.idx and the new model without that variable. We observed that there is not a statistical difference between those two models, so we decided to remove the above-mentioned variable and continue with the reduced model as our final model.

Our final model is the one below:

$$\log \frac{\hat{P}(SUBSCRIBED)}{\hat{P}(NOT\ SUBSCRIBED)} = 79.452 - 0.372 \cdot defaultunknown - 7.451 \cdot defaultyes - 0.23 \cdot contacttelephone + 0.637 \cdot monthaug + 0.19 \cdot monthdec + 0.494 \cdot monthjul + 0.606 \cdot monthjun + 1.257 \cdot monthmar - 0.725 \cdot monthmay + 0.039 \cdot monthnov + 0.384 \cdot monthoct - 0.183 \cdot monthsep + 0.005 \cdot duration - 0.044 \cdot campaign + 0.446 \cdot poutcomenonexistent + 1.771 \cdot poutcomesuccess - 0.016 \cdot nr.employed$$

The interpretation of this model will follow in the next chapter.

We then performed a Goodness of Fit Test for our final model. The p-value on this test was very high (e.g., p-value = 1). So, in level of significance of 95% we fail to reject the null hypothesis that the model fits well, which means that our final model has a quite good fit. Continuing, we will compare our model against the null model in order to test if our model fits significantly better than the null model. After executing the test, we get a p-value equal to 0, which means that in level of significance of 95% we failed to reject the null hypothesis that there is a significant difference between the two models. So, there is a significant difference between our model and the null model.

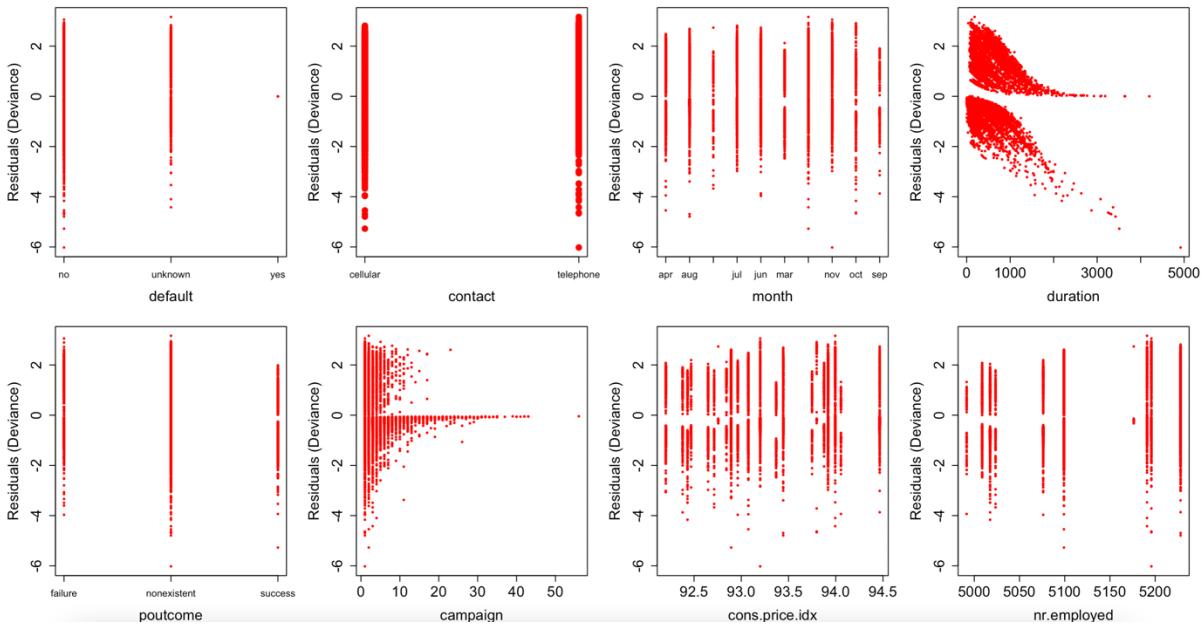


***Figure 2**: Deviance Residuals against all the explanatory variables of the final model*

Now we will focus on the deviance residuals of our final model. In the above figure we see the Deviance residuals against all the variables of our model. First of all, we can see that all values are separated in zero (in some plots the separation is not visible, that is due to the number of observations, if we zoom in a small interval close to zero, we will be able to see the separation). We can observe that there is no perfect separation between residuals as we have areas that the positive and negative residuals are entangled. So, the illustration of deviance residuals is quite good.

# Chapter 3
# Interpretation

In this chapter we will try to interpret our final model. As said before our final model is the following:

$$\log \frac{\widehat{P}(SUBSCRIBED)}{\widehat{P}(NOT\ SUBSCRIBED)} = 79.452 - 0.372 \cdot defaultunknown - 7.451 \cdot defaultyes - 0.23 \cdot contacttelephone + 0.637 \cdot monthaug + 0.19 \cdot monthdec + 0.494 \cdot monthjul + 0.606 \cdot monthjun + 1.257 \cdot monthmar - 0.725 \cdot monthmay + 0.039 \cdot monthnov + 0.384 \cdot monthoct - 0.183 \cdot monthsep + 0.005 \cdot duration - 0.044 \cdot campaign + 0.446 \cdot poutcomenonexistent + 1.771 \cdot poutcomesuccess - 0.016 \cdot nr.employed$$

Subsequently, we are going to interpret our final model. To begin with, we will interpret the intercept. As our numeric variables which describe number of contacts performed during this campaign for this client (campaign) and the index of number of employees (nr.employed) are not taking 0 value, as zero data point is not within the observation space of our dataset, we should center our covariates in order to be able to interpret the intercept. After centering, we can see that the probability of a successful subscription - when the client has not a credit in default, the type of communication is cellular, the month of last contact is April, the outcome of the previous marketing campaigns is a failure, we have average duration in the last contact, average number of contacts performed during this campaign for this client and average value in the index of number of employees - is equal to $\frac{e^{-3.501}}{1+e^{-3.501}} \approx 0.029$.

As far as variable default is concerned, a change in the knowledge of the occupation of a credit card from a client from no occupation, to unknown if the client occupies a credit card, is multiplying the actual odds of successful subscription by $e^{-0.372} \approx 0.69$ with the rest of covariates remained unchanged. Also, a change in the occupation of a credit card from a client from no occupation to a client who occupies a credit card, is multiplying the actual odds of successful subscription by $e^{-7.451} \approx 0.00058$ with the rest of covariates remained unchanged

(The rest covariates remained unchanged means that the contact type was cellular, the month of last contact was April and the outcome of the previous campaign on this client was a failure. Also, the call duration, the number of contacts and the indicator of number of employees do not change).

As far as variable contact is concerned, a change on contact type from cellular to telephone multiplies the actual odds of successful subscription by $e^{-0.23} \approx 0.79$ with the rest covariates remained unchanged.

As far as variable month is concerned, a change of last contact month from April to August multiplies the actual odds of successful subscription by $e^{0.637} \approx 1.89$ with the rest covariates remained unchanged. A change of last contact month from April to December multiplies the actual odds of successful subscription by $e^{0.19} \approx 1.20$ with the rest covariates remained unchanged. With the same pattern we can interpret the rest dummy variables of month.

As far as variable duration is concerned, an extra second on call duration telephone multiplies the actual odds of successful subscription by $e^{0.005} \approx 1.005$ with the rest covariates remained unchanged.

As far as variable campaign is concerned, an increase to the number of contacts performed during this campaign for a client by one unit, multiplies the actual odds of successful subscription by $e^{-0.044} \approx 0.96$ with the rest covariates remained unchanged.

As far as variable poutcome is concerned, a change to the outcome of the previous campaign from failure to nonexistent, is multiplying the actual odds of successful subscription by $e^{0.466} \approx 1.6$ with the rest of covariates remained unchanged. A change to the outcome of the previous campaign from failure to success, is multiplying the actual odds of successful subscription by $e^{1.771} \approx 5.87$ with the rest of covariates remained unchanged.

As far as variable nr.empployed is concerned, an increase to the indicator of number of employees by one unit, multiplies the actual odds of successful subscription by $e^{-0.016} \approx 0.98$ with the rest of covariates remained unchanged.