ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
SCHOOL OF BUSINESS
ΜΕΤΑΠΤΥΧΙΑΚΟ στην ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΑΝΑΛΥΤΙΚΗ
MSc in BUSINESS ANALYTICS

Course:

# Statistics for Business Analytics II

Professor:

**Dimitris Karlis**

First Assignment:

# Telemarketing phone calls Analysis

Student:

Name: Maria Skoli

AM: p2822131

Feb 2022

# Table of Contents

# Introduction

Nowadays, a common problem that concerns retail banks is the low total balances in long-term deposits. An individual with a long-term deposit agrees not to withdraw the cash from the bank until a certain date agreed between the individual and the financial institution. Knowing that, banks can have a better understanding on their liquidity and can analyze the amount of cash they can use for investments. More term deposits offer greater stability in a bank.

Our division has been contributed to the increase of banking products through numerous campaigns that reach customers by phone. Often, more than one contact to the same client was required, to access if the product would be or not subscribed.

On the occasion of the next campaign that is to be conducted with the aim of attracting customers to subscribe to long term deposits, it was deemed necessary to explore the factors that influence our customers' choice.  Such capabilities will allow our bank to optimize the parameters and to focus the next campaign to groups of people that meets certain characteristics.

To facilitate these needs, historical data from our bank system is used documenting near 40K phone contacts in order to sell long-term deposits for the years 2008 and 2010.

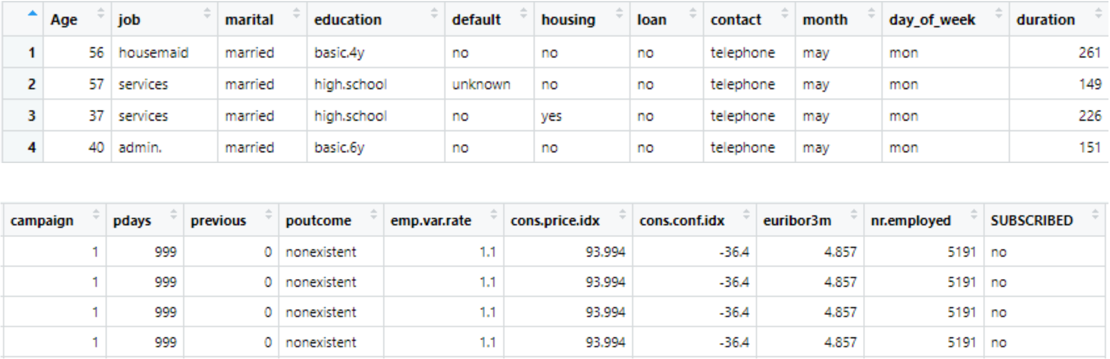
The aim of this assignment is to determine which among the available features influence the possibility that the phone call will be successful, namely that the client will subscribe to a long-term deposit. In the pages below, certain particularities of the features will first be explored in an exploratory analysis based on visual elements. Then, models are constructed in an attempt to identify the key variables that contribute to or affect the outcome of the phone contact. Once the optimal model is determined, we will check how well this model can interpret our data.

# Descriptive Analysis and Exploratory data Analysis

The analysis begins by loading the data and giving an overview of the structure of the dataset.

The dataset contains 39883 observations and for each of them 21 attributes are recorded including some personal characteristics of the client (e.g., age, job, education level, marital status, acquiring a personal or a housing loan, credit in default), details of the last contact (e.g., communication type, month, day of week and duration of the last contact), and other attributes concerning the campaign (e.g., number of performed contacts during this campaign and during previous campaigns, the outcome of the previous campaign for this client, whether the client was previously contacted). We also have information about some economic and social indices for this period (e.g., employment variation rate, consumer price index, consumer confidence index, Euribor 3-month rate and number of employees). While this dataset comes from real data some cleaning was still required before moving into the exploratory analysis. Although in this dataset we do not have missing values, for several observations some personal characteristics of the clients remain unknown.

Before moving into the analysis some variables are of no interest for the analysis presented here and therefore, we dropped them. Specifically, the number of employees and the employee's variation rate keep the same information and therefore, we do not need both. We dropped the number of employees variable.



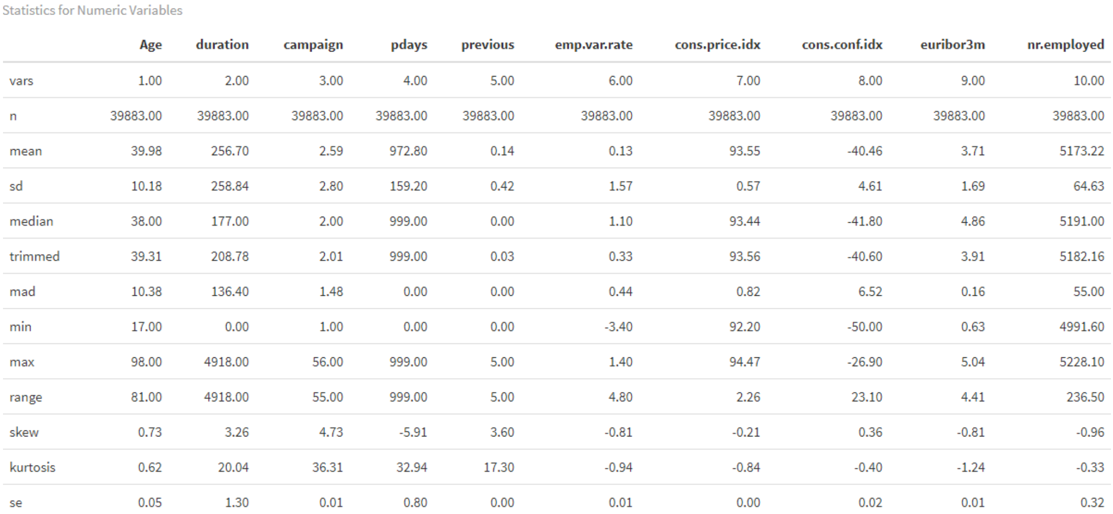
|  | Age | job | marital | education | default | housing | loan | contact | month | day_of_week | duration | campaign | pdays | previous | poutcome | emp.var.rate | cons.price.idx | cons.conf.idx | euribor3m | nr.employed | SUBSCRIBED |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 56 | housemaid | married | basic.4y | no | no | no | telephone | may | mon | 261 | 1 | 999 | 0 | nonexistent | 1.1 | 93.994 | -36.4 | 4.857 | 5191 | no |
| 2 | 57 | services | married | high.school | unknown | no | no | telephone | may | mon | 149 | 1 | 999 | 0 | nonexistent | 1.1 | 93.994 | -36.4 | 4.857 | 5191 | no |
| 3 | 37 | services | married | high.school | no | yes | no | telephone | may | mon | 226 | 1 | 999 | 0 | nonexistent | 1.1 | 93.994 | -36.4 | 4.857 | 5191 | no |
| 4 | 40 | admin. | married | basic.6y | no | no | no | telephone | may | mon | 151 | 1 | 999 | 0 | nonexistent | 1.1 | 93.994 | -36.4 | 4.857 | 5191 | no |

*Table 1: A brief view of the dataset*

This dataset consists of both numeric and categorical variables. The duration of the phone call, and the economic indices are continuous variables and hence numeric while attributes like Job, marital status and education are recorded in the form of categorical variables with multiple levels. It must be noted here that the attribute we are going to analyze (variable *"Subscribed"*) is also categorical as we have two different levels: "No" for clients that did not subscribe to a long-term deposit and "Yes" for those they did. In the exploratory analysis presented below, the numeric variables will be examined through histograms and the categorical variables through frequency tables and bar plots.

## Numeric Variables – Descriptive Statistics

As seen in Table 2 the range of the age that this dataset examines is between 17 and 98 years old. The mean age of the clients that have been asked for this product is 40-year-old. Moreover, we have a high deviation in duration as a phone call can be very short or very long. The average duration of a phone call is almost 4 minutes. Moreover, a client has been called at most 5 times before the running campaign. Finally, looking to the skewness and kurtosis metrics we can conclude that none of the below variables is distributed normally.



Statistics for Numeric Variables

| | Age | duration | campaign | pdays | previous | emp.var.rate | cons.price.idx | cons.conf.idx | euribor3m | nr.employed |
|---|---|---|---|---|---|---|---|---|---|---|
| vars | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 | 6.00 | 7.00 | 8.00 | 9.00 | 10.00 |
| n | 39883.00 | 39883.00 | 39883.00 | 39883.00 | 39883.00 | 39883.00 | 39883.00 | 39883.00 | 39883.00 | 39883.00 |
| mean | 39.98 | 256.70 | 2.59 | 972.80 | 0.14 | 0.13 | 93.55 | -40.46 | 3.71 | 5173.22 |
| sd | 10.18 | 258.84 | 2.80 | 159.20 | 0.42 | 1.57 | 0.57 | 4.61 | 1.69 | 64.63 |
| median | 38.00 | 177.00 | 2.00 | 999.00 | 0.00 | 1.10 | 93.44 | -41.80 | 4.86 | 5191.00 |
| trimmed | 39.31 | 208.78 | 2.01 | 999.00 | 0.03 | 0.33 | 93.56 | -40.60 | 3.91 | 5182.16 |
| mad | 10.38 | 136.40 | 1.48 | 0.00 | 0.00 | 0.44 | 0.82 | 6.52 | 0.16 | 55.00 |
| min | 17.00 | 0.00 | 1.00 | 0.00 | 0.00 | -3.40 | 92.20 | -50.00 | 0.63 | 4991.60 |
| max | 98.00 | 4918.00 | 56.00 | 999.00 | 5.00 | 1.40 | 94.47 | -26.90 | 5.04 | 5228.10 |
| range | 81.00 | 4918.00 | 55.00 | 999.00 | 5.00 | 4.80 | 2.26 | 23.10 | 4.41 | 236.50 |
| skew | 0.73 | 3.26 | 4.73 | -5.91 | 3.60 | -0.81 | -0.21 | 0.36 | -0.81 | -0.96 |
| kurtosis | 0.62 | 20.04 | 36.31 | 32.94 | 17.30 | -0.94 | -0.84 | -0.40 | -1.24 | -0.33 |
| se | 0.05 | 1.30 | 0.01 | 0.80 | 0.00 | 0.01 | 0.00 | 0.02 | 0.01 | 0.32 |

*Table 2: Descriptive Statistics of numeric variables*

From the histograms below we can observe that the the Age parameter has a long right tail but the frequency of contacts with people older than 60 years old is lower. For the purposes of our analysis it would be more useful to divide the Age attribute into 3 Age groups 17-35, 35-45 and 45+. About the duration of the contact we can conclude that long-term calls are more rarely observed. Additionally, most people have been contacted once during this campaign.

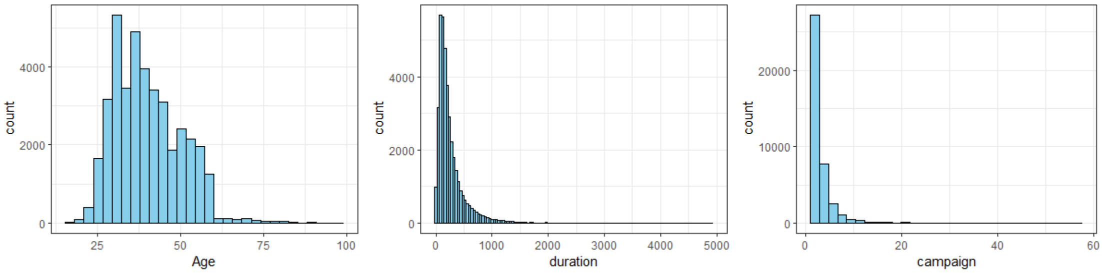


*Table 3:Histograms of Age, duration and campaign*

In the histograms below we can observe the economic and social indices. We can observe that we have high variation between the prices. This is reasonable as most of the indices refer to quarterly data. Moreover, years 2008-2010 were detrimental for the economy if we consider the global financial crisis in 2008. Therefore, price fluctuations seem logical.

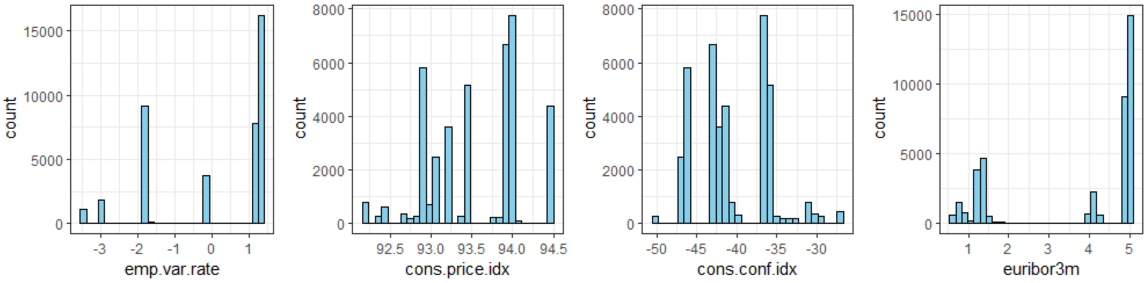


*Table 4: Histograms of the economic indices*

Furthermore, many outliers can be spotted in the boxplots corresponding to variables of Age and Duration.

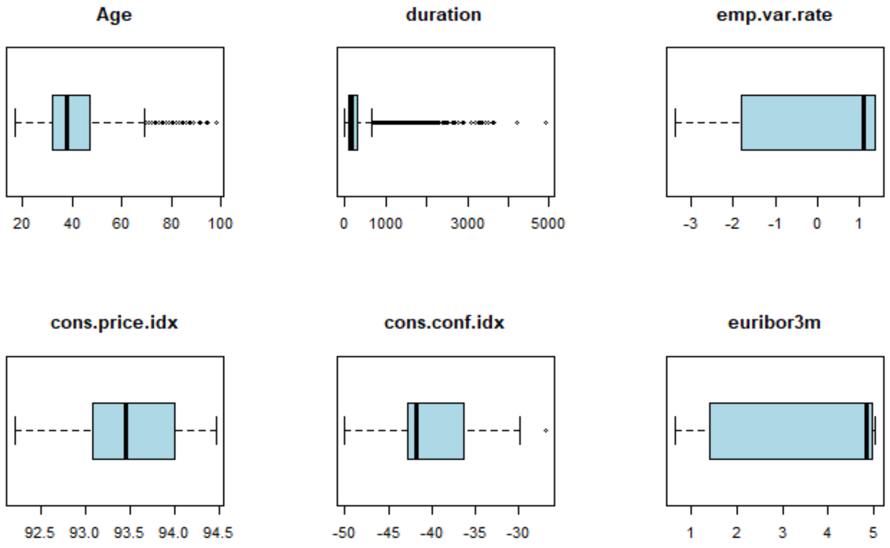


*Table 5: Present of outliers in the variables*

## Categorical Variables – Descriptive Statistics

First, is important to have an insight into the variable of interest of this analysis, the decision of the client to subscribe or not to the product. Only 10% of the clients that have been contacted in this dataset have decided to subscribe to a long-term deposit.

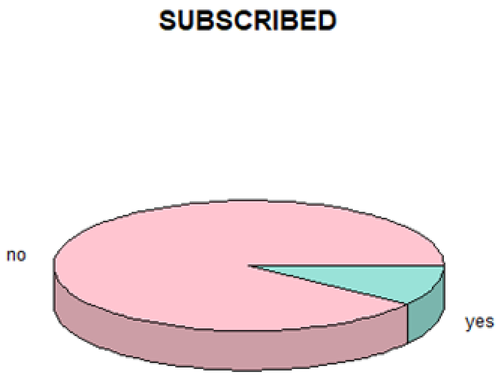


*Table 6: The proportion of subscribers in the dataset*

Let’s have a look into personal characteristics like Education and Marital status of the clients that have been asked.

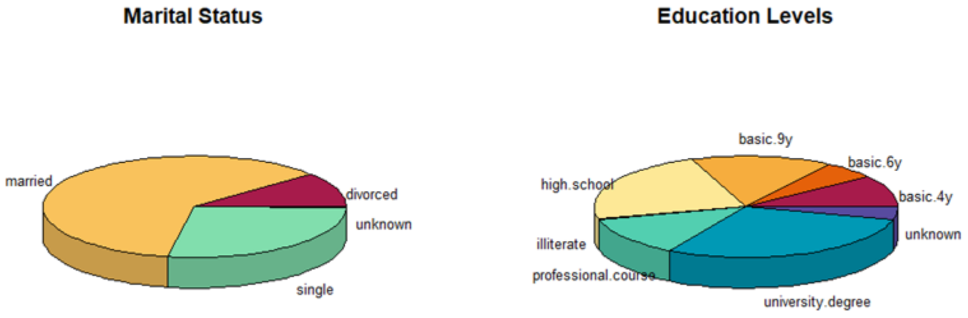


*Table 7: The proportions of Marital and Education status in the dataset*

Additionally, from the plot below we see that 15% of the clients have a personal loan and more than 50% of the clients have a housing loan. Also, only 3 clients were in default and therefore it is not observable in the plot below. For this reason, the variable default is not useful for our analysis as we don’t have enough information to examine whether the default status of a client can affect his decision and therefore, we will not conclude the default parameter into the descriptive model.

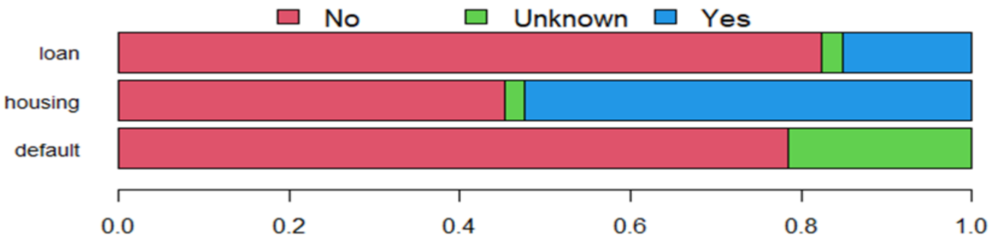


*Table 8: Visualization of Loan Housing and Default status*

## Associations of the attributes with the Subscribers

Below we will examine how some attributes associate with the final outcome of the campaign.

### *Job and Subscribed*

From the plot below we can conclude that it is more possible for students and for the retired ones to subscribe to a long-term deposit. This is reasonable as many students select to keep money for their studies and the retired ones have a few obligations and therefore it is more possible to access a long-term deposit. The percentages of the subscription per Job Title are shown in the Table 9.

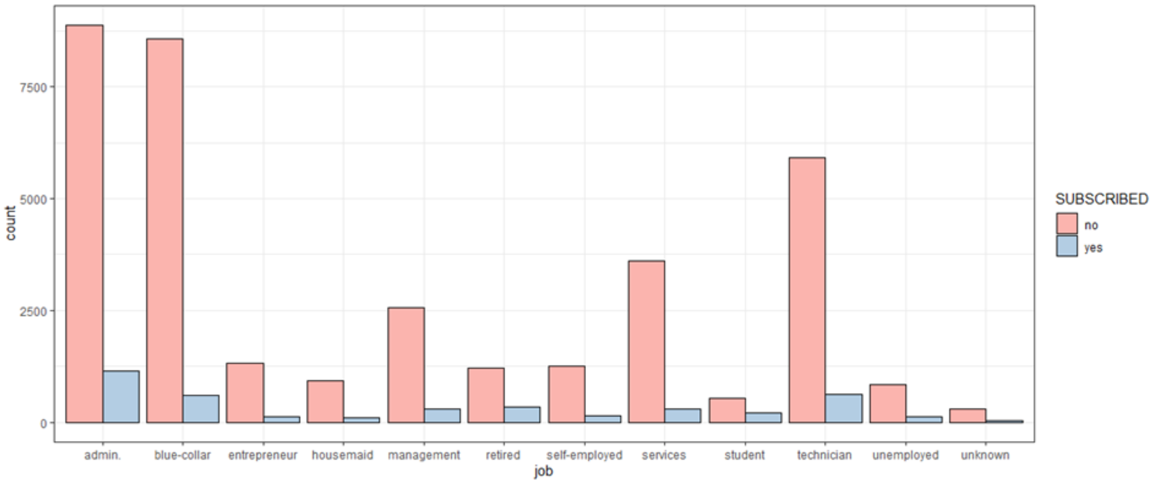
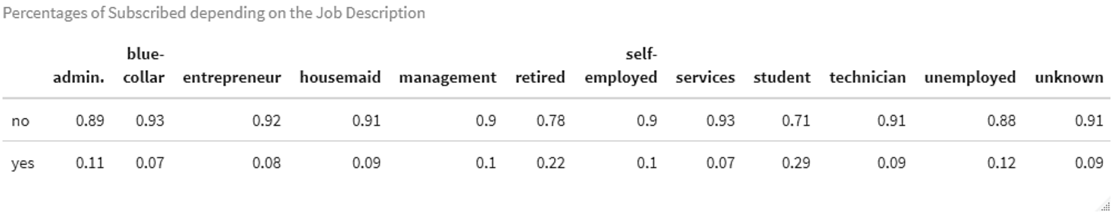


*Table 9: Associations of Job and Subscribers*



Percentages of Subscribed depending on the Job Description

| | admin. | blue-collar | entrepreneur | housemaid | management | retired | self-employed | services | student | technician | unemployed | unknown |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| no | 0.89 | 0.93 | 0.92 | 0.91 | 0.9 | 0.78 | 0.9 | 0.93 | 0.71 | 0.91 | 0.88 | 0.91 |
| yes | 0.11 | 0.07 | 0.08 | 0.09 | 0.1 | 0.22 | 0.1 | 0.07 | 0.29 | 0.09 | 0.12 | 0.09 |

*Table 10: Table of proportions for Job vs Subscribers*

### *Age Group and Subscribed*

As we mentioned before we divide the Age into 3 Age groups. From the plot below we can observe that for each Age group the possibility of subscription to the product is almost the same.

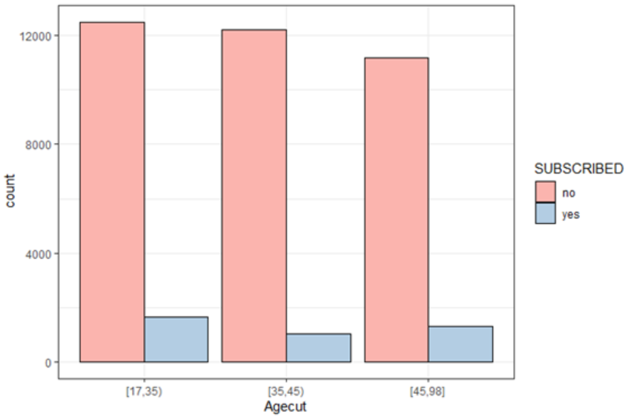


*Table 11: Table of proportions per Age Group for Subscribers*

### *Duration and Subscribed*

From the boxplot below we can conclude that the median duration of a phone call for someone that selected to subscribe is larger that someone who did not. This is reasonable because if someone is interested in the product will ask more details and the phone call will last longer. The difference in the medians between subscribers and non-subscribers can be an indicator that the variable Duration will be significant for the interpretation of the model.

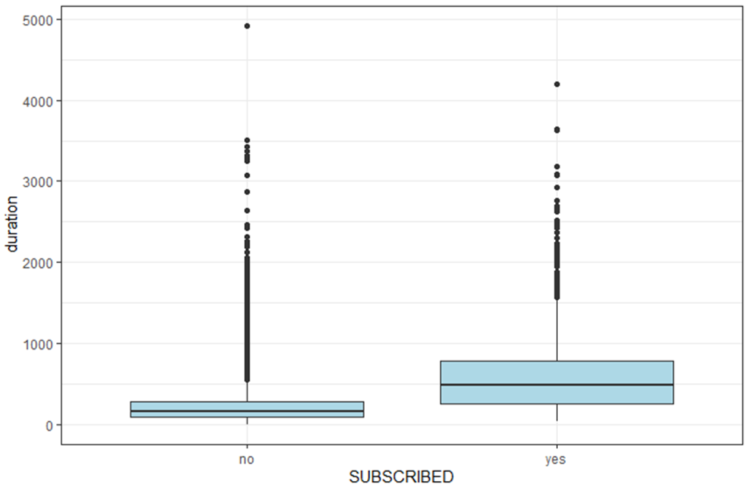


*Table 12: Boxplot presenting the mean duration per outcome of campaign*

### *Time attributes and Subscribed*

From the graphs below we observe that all days of the week hold the same possibility to achieve a successful campaign. This can be an indicator that the day of week will not be significant for our model as it does not affect the outcome. On the other hand, we can December and March are the months with the highest successful rate.

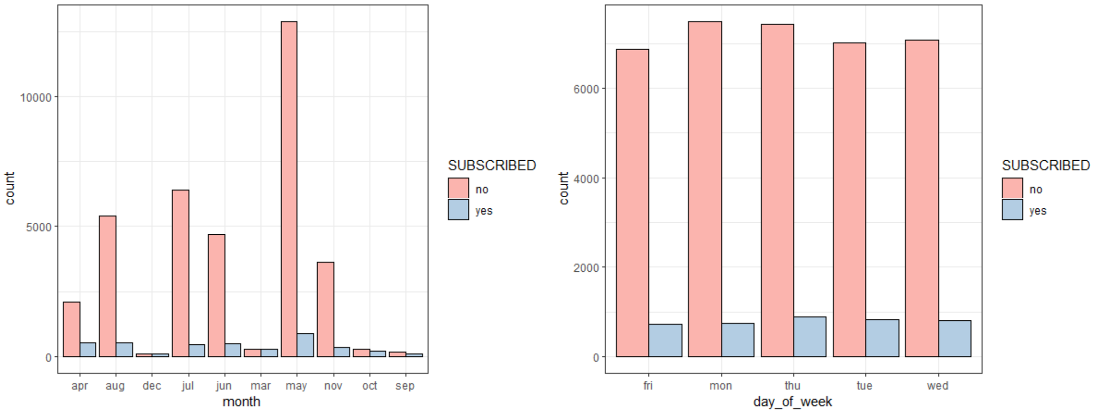
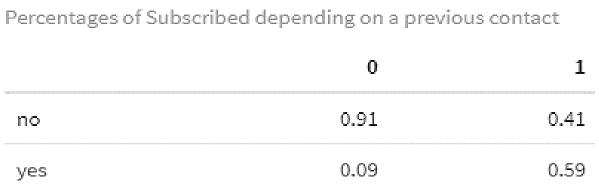


*Table 13: Months and Weekdays VS Subscribers*

### *Previous contact and Subscribed*

As we described earlier, we changed the variable pdays (number of days that passed by after the client was last contacted from a previous campaign) into categorical with two levels. Level “0” if a client was never contacted before and Level “1” if a client was contacted at least once before. From the table below we examined that the percentage of clients that was never contacted before hold a small possibility into subscribing to the product (9%). On the other hand, if a client has been contacted before it holds the same possibility into subscribing or not (59%).



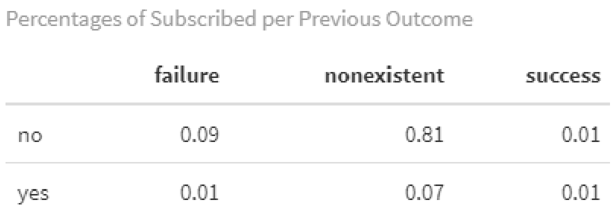
Percentages of Subscribed depending on a previous contact

| | 0 | 1 |
|---|---|---|
| no | 0.91 | 0.41 |
| yes | 0.09 | 0.59 |

*Table 14:Percentage of Subsribed depending in a previous contact*

### *Previous Outcome and Subscribed*

We can observe that the percentage of people that the failure or the success of a previous campaign in a client has no effect in the running campaign as the percentage of success is the same with the percentage of failure.



Percentages of Subscribed per Previous Outcome

| | failure | nonexistent | success |
|---|---|---|---|
| no | 0.09 | 0.81 | 0.01 |
| yes | 0.01 | 0.07 | 0.01 |

*Table 15: Percentages of subscribed based on the previous outcome*

## Pairwise comparisons

In this part the existing relationships among the explanatory variables will be examined. In the Table 16 (see below) the Pearson correlation between the variables is presented. The Pearson correlation is a linear correlation and as such it will not always capture with accuracy the relationship between variables related in non-linear fashion.

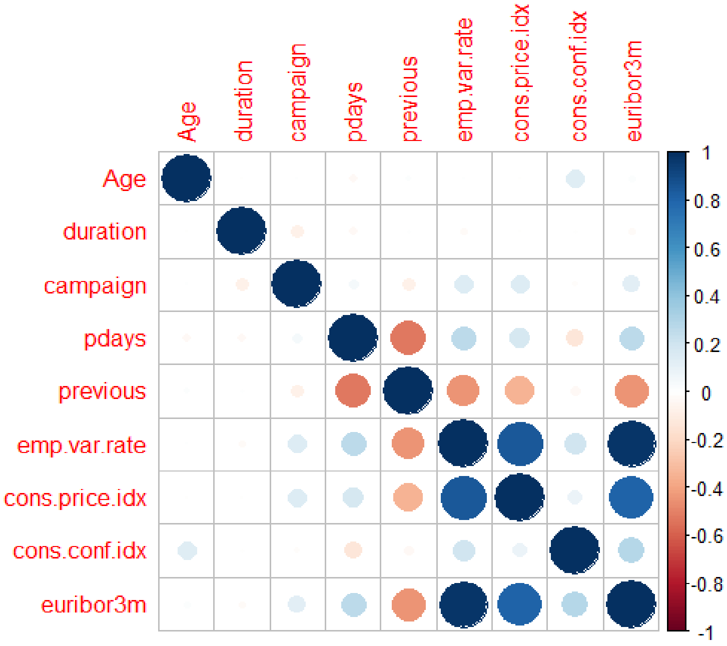


*Table 16: Correlation Table*

Let’s now highlight some key observations based on the above table:

- Previous and pdays have a high negative correlation
- Euribor has a high positive correlation with employee variation rate, consumer price index.
- Consumer price index has a high positive correlation with employees variation rate and Euribor 3m

In order to avoid the problem of multicollinearity in the model we will not include all of these economic indices. However, we selected to include into the model the parameter of Euribor 3m as it refers in daily values and it obtains more information than the other with is referred to quarterly data.

## Descriptive Models

Taking into consideration all the above results, it is now possible to make an educated selection of the variables that should be part of the initial model. As we examined in the sector above, the variable corresponding to the default situation of a client cannot be included in the model because this attribute does not provide us with significant information as only 3 of the 40K observations concern clients who were in default.

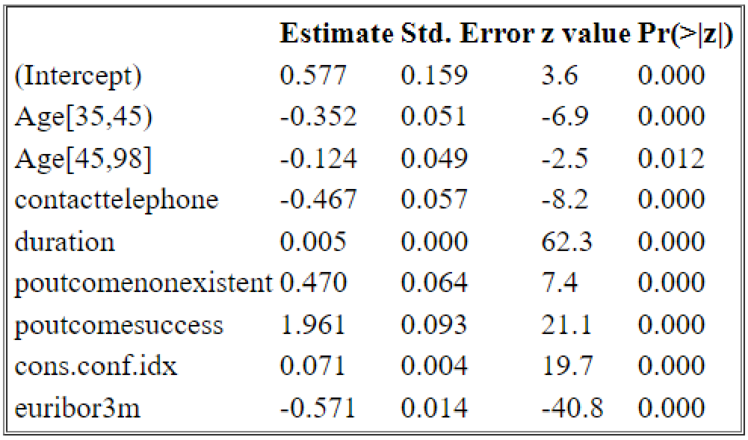
The variable in interest in this analysis is a binary variable and presents the possibility of someone to subscribe to a long-term deposit. For this reason, we cannot assume a linear regression. However, we are going to use the logistic regression as we want to interpret a possibility.

The aim of this analysis is to interpret the parameters that affect the selection of the client to subscribe or not to the long-term deposit, and therefore, the BIC methodology is recommended. Other useful methodologies are the AIC methodology and the Lasso. The key difference between lasso and stepwise regression is that the latter shrinks all coefficients towards zero, while the former has the potential to remove predictors from the model by shrinking the coefficients completely to zero. We examined all of them, but the best model is achieved by BIC methodology. We conducted a model from AIC methodology also but it has the same proportion of residuals deviance/degrees of freedom and many more variables and that's why we selected the BIC method which provide us with less parameters. Moreover, we should take into consideration the multicollinearity into the models concerning the high correlations of the economic indices we observed in the previous sector.

After implementing the BIC methodology, we extracted the following model.

***Subscribed ~ Age + contact + month + duration + poutcome + emp. Var rate +cons.price.idx +cons.conf.idx +euribor3m***

However, testing for multicollinearity (with Gvif test) we conclude that the variables month, employment variation rate and consumer price index should be excluded as they are highly associated and therefore the final model is the following.



| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 0.577 | 0.159 | 3.6 | 0.000 |
| Age[35,45) | -0.352 | 0.051 | -6.9 | 0.000 |
| Age[45,98] | -0.124 | 0.049 | -2.5 | 0.012 |
| contacttelephone | -0.467 | 0.057 | -8.2 | 0.000 |
| duration | 0.005 | 0.000 | 62.3 | 0.000 |
| poutcomenonexistent | 0.470 | 0.064 | 7.4 | 0.000 |
| poutcomesuccess | 1.961 | 0.093 | 21.1 | 0.000 |
| cons.conf.idx | 0.071 | 0.004 | 19.7 | 0.000 |
| euribor3m | -0.571 | 0.014 | -40.8 | 0.000 |

*Table 17: The estimated model*

The final model can be written in the following way:

$$Log\ \left(\frac{Possibility\ of\ Subsribe}{1-Possibility\ of\ Subscribe}\right) = 0{,}577\ -0{,}352*Age[35,45)\ -\ 0{,}124*Age[45,98]- 0{,}467*contacttelephone + 0{,}005*duration + 0{,}470*poutcomenonexistent +1{,}961*poutcomesuccess\ + 0{,}071*\ cons.conf.idx\ -0{,}571*euribor3m$$

A way to measure the goodness of fit of this model is the proportion of the Residuals Deviance with the Degrees of freedom of this model. From the table above we can calculate that:

$$\frac{Residual\ Deviance}{Degrees\ of\ Freedom} = \frac{16685}{39863} = 0{,}4185$$

This percentage is lower from that of the null model (0,65) and therefore the model has been ameliorated.

## Examine the assumptions of the final model

The response variable must follow a binomial distribution.

Logistic Regression assumes a linear relationship between the independent variables and the link function (logit).

The dependent variable should have mutually exclusive and exhaustive categories.

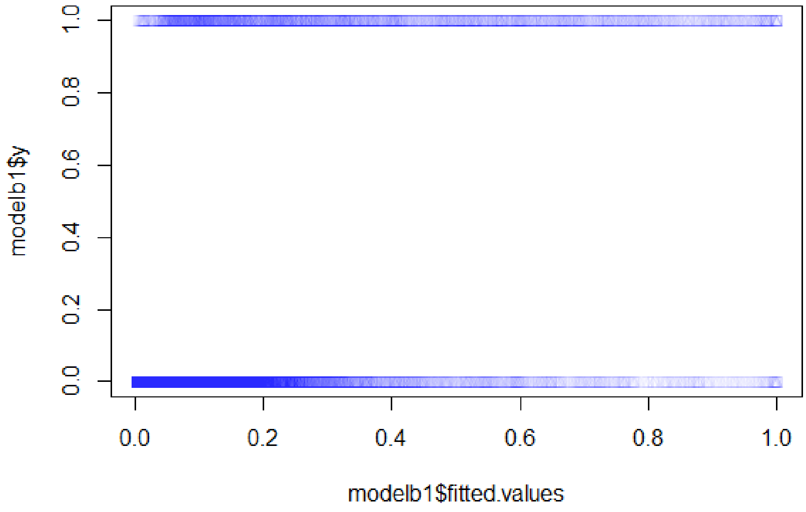


*Table 18: Fitted values and the actual values*

In the plot above we can assume that there are observations where our model assigns a low probability, but the actual outcome was client subscription. In the same way, there are observations where our model is almost sure of client subscription, but the actual outcome was that the client rejects the subscription. So, we conclude that the model is not perfect but it is useful to gain important insight.

## Interpretation of the final model

The model that is deemed optimal was created with the use of the BIC methodology. It is now time to try to interpret the parameters of that model.

- The odd of successful campaign is 1,78 ($e^{0,577}$) times higher when all the attributes are zero, meaning that the Age group is 17-35 years old, the contact conducted via cellular, previous outcome was failure and all continuous variables are zero. Practically is not useful this interpretation but the intercept contribute to the other parameters.

- The odd of successful campaign is 0,70($e^{-0,352}$) times higher for someone who is in the age group of ages 35-45 than someone who is not with all the other attributes constant.

- The odd of successful campaign is 0,88 ($e^{-0.124}$) times higher for someone in the age group of 45-98 than someone who is not with all the other attributes constant.

- The odd of successful campaign is 0,63 ($e^{-0,467}$) times higher for someone we contacted via telephone than someone who is contacted via cellular with all the other attributes constant.

- The odd of successful campaign is 1,34 ($e^{60*0.005}$) time higher for someone whose phone call last 1 minute longer than someone else with all the other attributes constant.

- The odd of successful campaign is 1,60 ($e^{0,470}$) times higher for someone who’s the outcome of the previous campaign is nonexistent than someone from who we know the result with all the other attributes constant.

- The odd of successful campaign is 7,11 ($e^{1,961}$) times higher for someone who’s the outcome of the previous campaign is successful than the others with all the other attributes constant.

- The odd of successful campaign is 1,07 ($e^{0,071}$) times higher when the consumer confidence index is increased by 1 unit with all the other attributes constant.

- The odd of successful campaign is 0,57 ($e^{-0,571}$) times higher when the Euribor 3m index is increased by 1 unit with all the other attributes constant.

## Conclusion

The model proposed has a relatively good fit and is easy to understand and interpret.

From the above analysis we can conclude that some of the major factors that greatly affect the probability to access the product is the duration of the call, the successful or the non-existent outcome of the previous campaign and the consumer confidence index.

Although, our model selects specific variables as important, we can have also important insight from the graphs in the first section of this analysis. This analysis allow us to gain insight on the methods that we can follow in order to maintain higher percentages of successful campaigns.

First, we can approach clients that are students or retired. Also, we can reach customers that derived from the age group of 45+ in order to increase the possibility of a successful campaign. Economic indices are not factors that we can intervene. Moreover, the duration of the phone call is a measurement that is known after the end of a campaign. However, we can use questionnaires or more person centralized questions to make the call last longer. We can decrease the campaign calls as the more we call someone the more possible is to reject the deposit. Finally, it would be useful to target a group of clients that their outcome of a previous campaign was successful.

By combining all these strategies and simplifying the market audience the next campaign should address, it is likely that the next marketing campaign of the bank will be more effective than the current one.

In future analysis, it would be useful to include additional variables such as the year of each contact because the economic situation between the years is not stable and therefore, the analysis may be not precis. Another variable can be the balance that the client has already in our bank.