# Client Settings as a Determinant of Bank Telemarketing Success Rate on Long-Term Deposits

Anastasios Moraitis

p2822124

Statistics for Business Analytics II, MSc in Business Analytics

School of Business, AUEB, Athens, Greece

February 2022

Abstract

This study aims to determine the factors affecting the decision of a given bank customer towards selecting a long-term deposit product. It concentrates the key factors that best describe the possibility the customer subscribes to the product and, in general, which variables contribute to a successful contract. To this end, we draw 49 months of real data collected from one of the retail banks, in a total of near forty thousand phone contacts. These data are split between bank client data, data related with the last contact of the current campaign and other attributes, and social and economic context attributes. Overall, we find that a mix of characteristics impacts customer acquisition.

*Keywords*: Telemarketing sales, Marketing strategy, Long-Term Deposits, Feature Selection

## 1. Introduction

To effectively approach clients ready to invest in new products businesses include marketing selling campaign, and, more precisely, direct marketing. Telemarketing is highly likely to be included in this set of tactics. Telemarketing has always been a popular way to expand a business, but the use of personalized scripts and the wide availability of phone numbers for consumers has been an easy way for individuals to end calls before they begin. When done correctly, however, telemarketing can be a useful advertising medium for businesses.

This project aims to help banks to increase the accuracy of their customer profiling through understanding as well as identifying a group of customers who have a high probability to subscribe to a long-term deposit. The remainder of this paper proceeds as follows. Section II and III presents the materials and methods used to achieve the objective of this research, Section IV presents the experimental results, and finally Section V concludes with some indication for future work.

## 2. Data

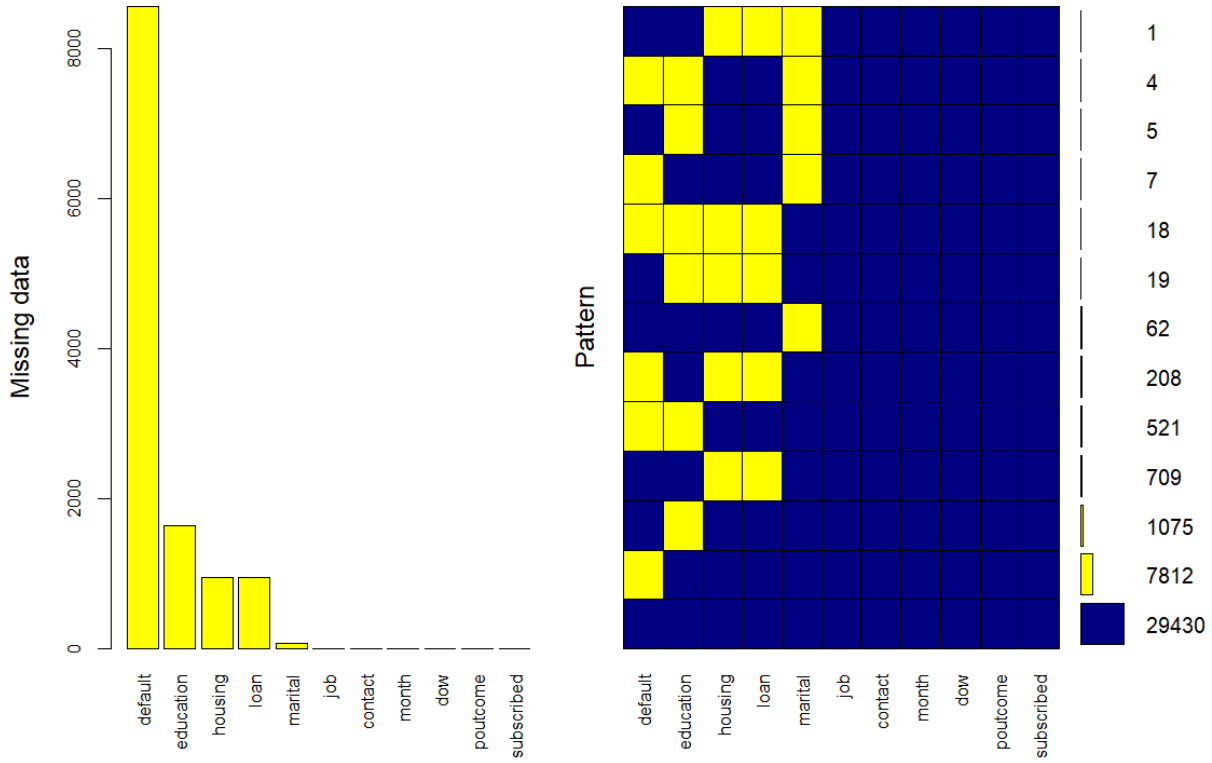
Bank telemarketing data has been made available. The dataset contains subscription statuses obtained after each call to the targeted customer, thus making it easy to attach bank client attributes, features related with the last contact of the current campaign and other attributes, with social and economic ones among them.

*Table 1*- Descriptive statistics of numeric variables

| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|
| campaign | 39,871 | 2.589 | 2.801 | 1 | 1 | 3 | 56 |
| emp_var_rate | 39,871 | 0.131 | 1.572 | -3.400 | -1.800 | 1.400 | 1.400 |
| cons_price_idx | 39,871 | 93.553 | 0.572 | 92.201 | 93.075 | 93.994 | 94.465 |
| cons_conf_idx | 39,871 | -40.463 | 4.611 | -50.000 | -42.700 | -36.400 | -26.900 |
| euribor3m | 39,871 | 3.710 | 1.690 | 0.634 | 1.405 | 4.961 | 5.045 |
| nr_employed | 39,871 | 5,173.223 | 64.626 | 4,992 | 5,099.1 | 5,228.1 | 5,228 |

During the Data Cleaning and Manipulation process, we transformed variables in a way that they will provide a meaningful interpretation if included in a regression model and, also, cleaned unknown values using imputation. First, we deleted any duplicates and cleaned variable names. We then deleted the *duration* attribute because this information is provided after the call. To perform EDA, we split the dataset into 2 new, one with each factorial and one with each numeric feature. We treated “unknown” values as NAs in the final dataset and we used imputation on the columns containing such values. We can now look at the features (or combination of features marked as NA). This will give us enough insight as to which variables are present in the dataset. The figure displays only the factorial variables, the numeric ones are omitted, because those didn’t contain any unknown cells. We notice that *default* had the most

*Figure 1*- Distribution of missing values in the dataset




missing values (almost 20% of the data set), and education is the column with the second most missing values. As described earlier, we used the Multivariate Imputation via Chained Equations package to impute missing values on the dataset. The package takes care of uncertainty in missing values by creating multiple imputations as compared to a single (such as a mean). We used logistic regression for the binary variables (loan, default, housing) and Bayesian polytomous regression on factor variables with more than two levels (job, marital, education).

Once we had a clean data set (in terms of missing values), we tried to transform some of our variables (some numeric and some factorials) to meaningful attributes with less levels. We further analyze the analysis results in the *Descriptive Analytics* section.

## 3. Descriptive Analysis

### Univariate

We are now looking to each variable, respectively. It seems that *pdays* is ranging between 1 and 21, and that a lot of values are 999, which means that the user has never been called again. We transform the variable into a factorial one, with 3 levels, so that it will give us a better insight in terms of explaining our response variable.

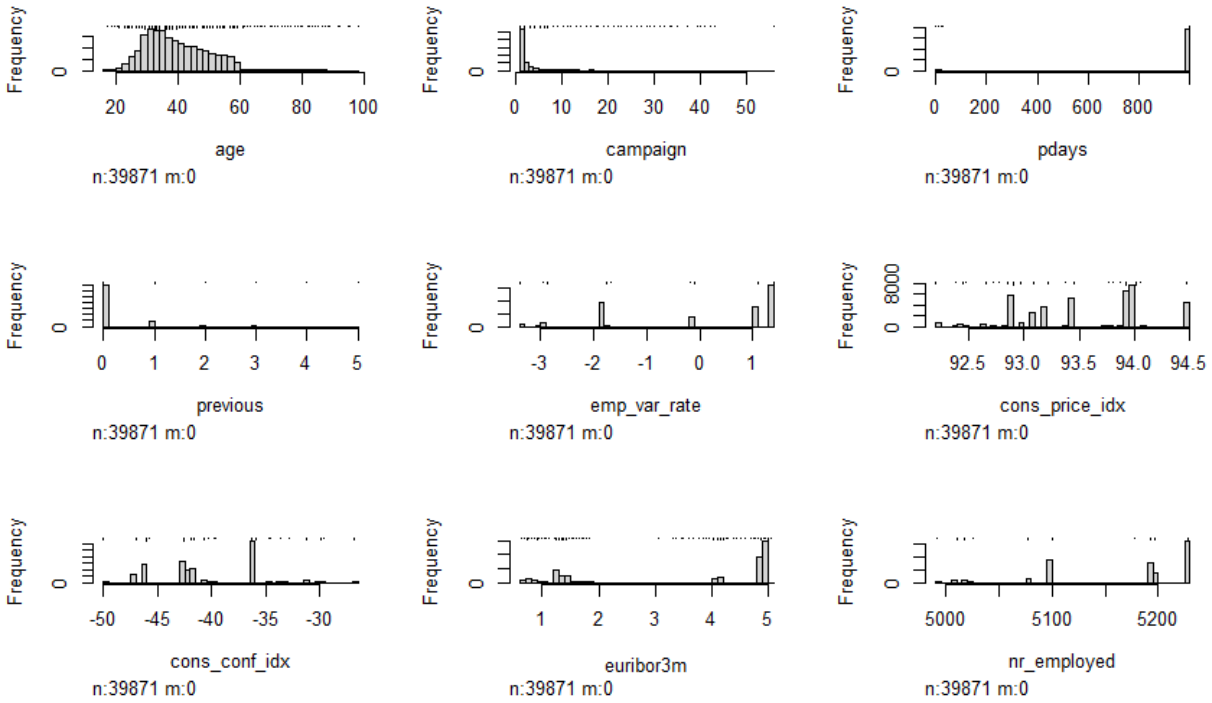
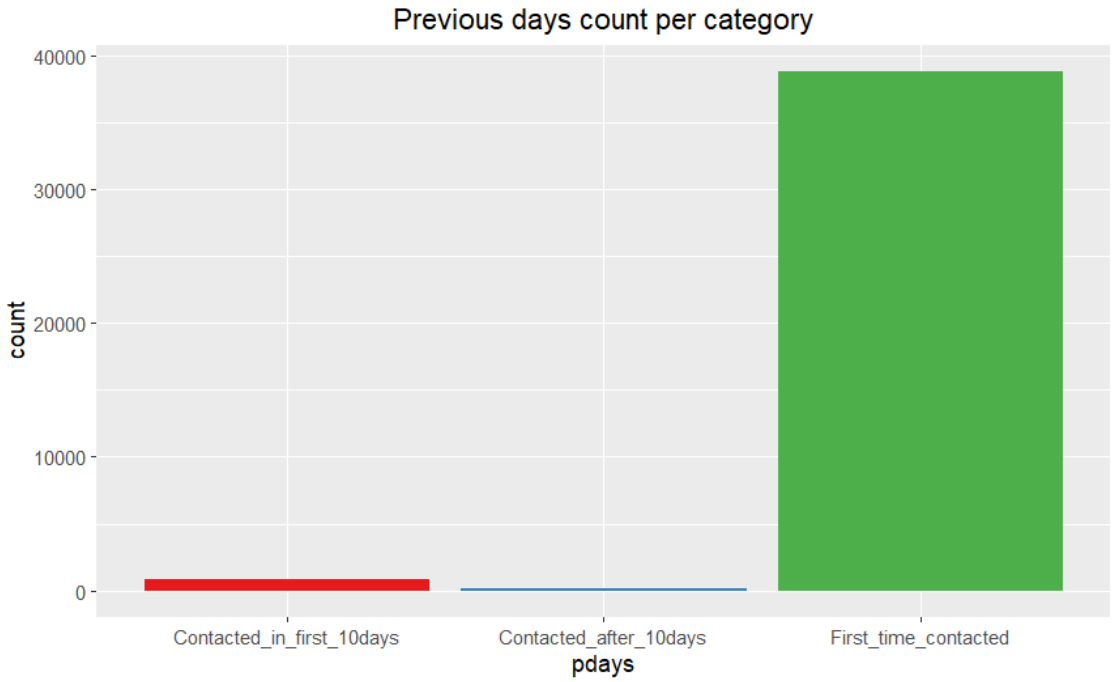



*Figure 2* – Frequency plots for each numeric feature

We end up having 3 groups assigned to the *pdays* feature as shown in *Figure 3*, in which we notice that the response rate is significantly higher for prospects that have been contacted for first time. We now see that for our model to explain whether a user is subscribed or not, we will need a more meaningful representation of the age of the client. Thus, we are transforming that

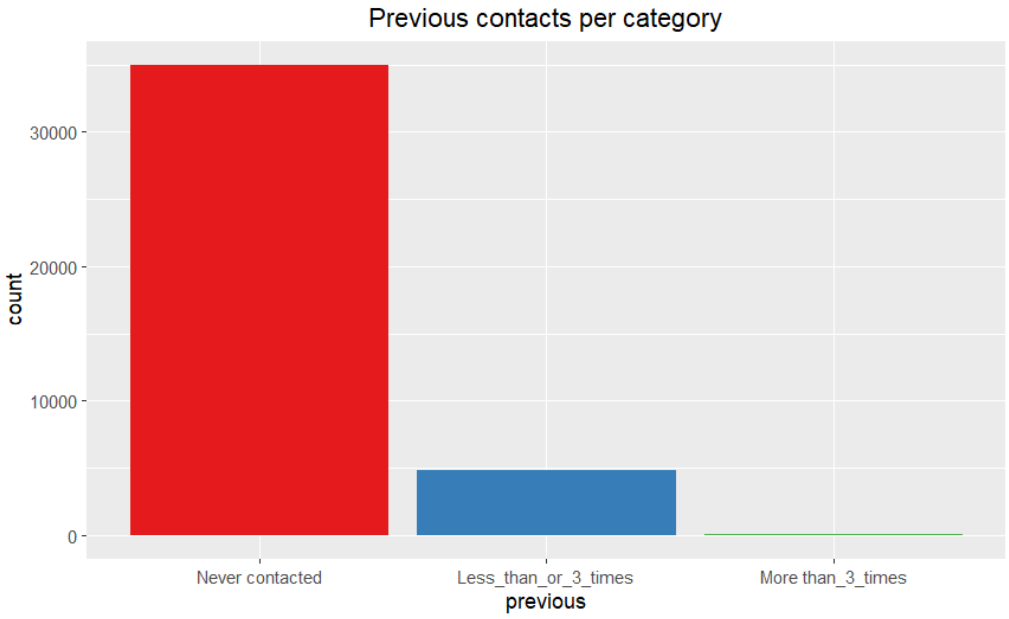
feature to a factorial, split into different decades, except from 70+ bin which contains the outliers, as shown in Figure 6 (Appendix).

*Figure 3*




In addition, we have *previous* (number of contacts performed before in this campaign) variable, which seems to have only 6 values. We are converting that to a factorial variable with the categories as shown in *Figure 4.*

*Figure 4*




We will also group the months by quarters. This will result to fewer categories, and it will also reduce the degrees of freedom in our model. The result is shown in Figure 9 (Appendix) and we notice that we don't have a lot of contact records in Q4. We touch the *education* parameter now and we also reduce that into 3 main categories (Primary, Secondary and Tertiary). That will give us enough insight as to whether a user of a specific level will subscribe to the service. In Figure 7 (Appendix), we notice a significant number of contacts for people in tertiary education.

**Bivariate**

We are now using the Hoeffding's-D measure as the test statistic to test dependency of the numeric variables, defined by Hoeffding, W. (1948). We don't use Pearson's correlation coefficient because our data are not normally distributed. We test normality of the variables using the Kolmogorov-Smirnov test as defined by Thode Jr. (2002). The tests on whether the D statistic is significantly different from zero gives evidence for non-dependence among the numeric

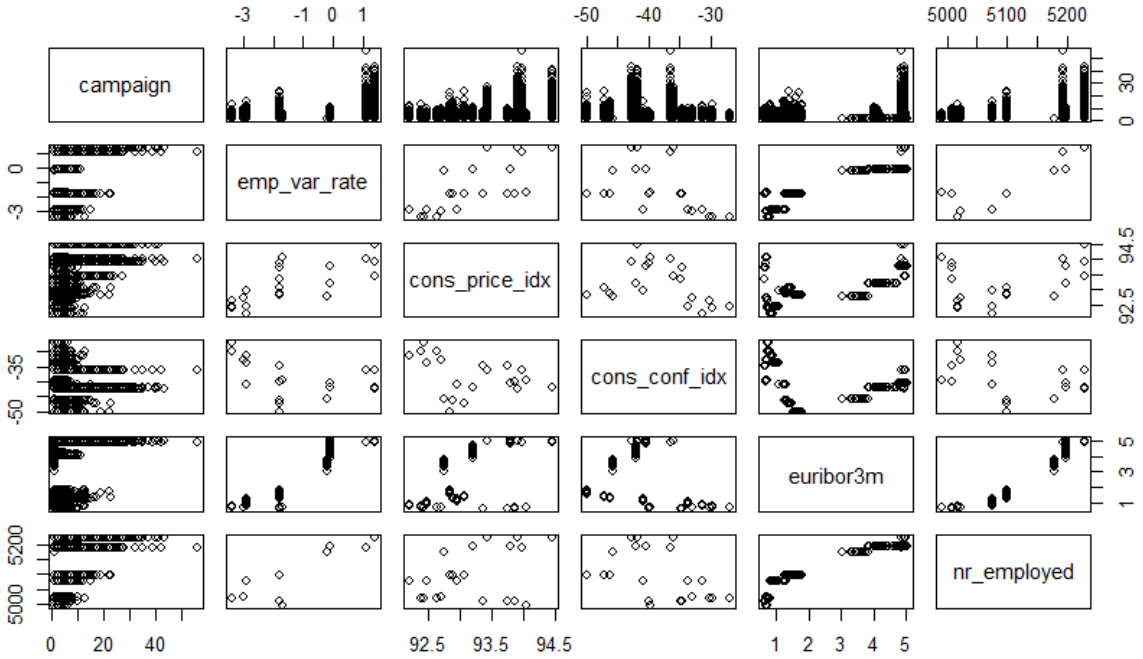
variables. Thus, we cannot remove any feature from our further analysis at this point. For evidence, we provide the *Table 1*, which provide the test results.

*Table 2*- Hoeffding's D test

| Feature | campaign | emp_var_rate | cons_price_idx | cons_conf_idx | euribor3m | nr_employed |
|---|---|---|---|---|---|---|
| campaign | NA | 1.00E-08 | 1.00E-08 | 1.76E-07 | 1.00E-08 | 1.00E-08 |
| emp_var_rate | 1.00E-08 | NA | 1.00E-08 | 1.00E-08 | 1.00E-08 | 1.00E-08 |
| cons_price_idx | 1.00E-08 | 1.00E-08 | NA | 1.00E-08 | 1.00E-08 | 1.00E-08 |
| cons_conf_idx | 1.76E-07 | 1.00E-08 | 1.00E-08 | NA | 1.00E-08 | 1.00E-08 |
| euribor3m | 1.00E-08 | 1.00E-08 | 1.00E-08 | 1.00E-08 | NA | 1.00E-08 |
| nr_employed | 1.00E-08 | 1.00E-08 | 1.00E-08 | 1.00E-08 | 1.00E-08 | NA |

Also, here, in *Figure 5* we should also provide the relationships between numerical features graph-wise.

*Figure 5*



## 4. Model Engineering

### Feature Selection

From our previous analysis in sections 2 and 3, our understanding is that we can reduce the number of variables in model building. In this section, we will focus on modeling of our problem, which is to find a function that describes whether the user subscribes in campaign, social and economic context, and client's personality. As a last resort to filtering out unwanted variables will be to run Near Zero Variance diagnostics on our dataset. That will exclude predictors having one unique value (i.e., zero variance), or predictors having very few unique values relative to the number of samples and the ratio of the frequency of the most common value to the frequency of the second most common value is large. We decide to remove the features with (near) zero variance, because we have already treated other possible features in an earlier step of our analysis (Section 3). Using Table 2 as a reference, we will not exclude any of the variables. We would proceed with omission of a variable if that had **exact** zero variance.

*Table 3* – Zero variance diagnostics

| | Frequency Ratio | Percent of Unique Values | Zero Variance | Near Zero Variance |
|---|---|---|---|---|
| age | 1.5853 | 0.0176 | No | No |
| campaign | 1.6558 | 0.1053 | No | No |
| pdays | 43.4217 | 0.0075 | No | Yes |
| previous | 7.2600 | 0.0075 | No | No |
| emp_var_rate | 1.7674 | 0.0226 | No | No |
| cons_price_idx | 1.1618 | 0.0527 | No | No |
| cons_conf_idx | 1.1618 | 0.0527 | No | No |
| euribor3m | 1.0984 | 0.6320 | No | No |
| nr_employed | 1.9020 | 0.0251 | No | No |
| job | 1.0881 | 0.0276 | No | No |
| marital | 2.2080 | 0.0075 | No | No |
| education | 1.3375 | 0.0075 | No | No |

| | Frequency Ratio | Percent of Unique Values | Zero Variance | Near Zero Variance |
|---|---|---|---|---|
| default | 13289.3333 | 0.0050 | No | Yes |
| housing | 1.1524 | 0.0050 | No | No |
| loan | 5.4287 | 0.0050 | No | No |
| contact | 1.6956 | 0.0050 | No | No |
| dow | 1.0101 | 0.0125 | No | No |
| poutcome | 9.0036 | 0.0075 | No | No |
| subscribed | 9.0028 | 0.0050 | No | No |
| quarter | 1.4490 | 0.0100 | No | No |

Before running any regression model, we need to check the assumptions of the General Linear Regression framework.

- The dependent variable should be binary:
    - This is true because *subscribed* is 0 or 1.
- Observations should be independent of each other:
    - Customers are unrelated individuals in our case, this is also true.
- No multicollinearity among independent variables:
    - Our first attempt in the previous section didn't show anything important in terms of correlation.
    - Let's now try computing the Generalized Variance inflation factor to verify if that stands true. We are running the process as defined by Fox, J. et al (2018). This is based on the full model with all the variables considered. It turns out that we cannot compute GVIF for this model, because a perfect relationship exists in our data. The variable that is causing the issue is the *poutcome,* of which the coefficient is not compute neither reported for this model. Removing this from the model will give us the following table with GVIF and

Degrees of Freedom (normalized GVIF is reported, with values greater than 5 reporting collinearity, we will not use it for our analysis) values:

*Table 4*- GVIF diagnostic for the full model, minus the *poutcome* predictor

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| age | 3.7215 | 6.0000 | 1.115732 |
| campaign | 1.0424 | 1.0000 | 1.021001 |
| pdays | 1.5770 | 2.0000 | 1.120617 |
| previous | 1.7362 | 2.0000 | 1.147897 |
| emp_var_rate | 47.6325 | 1.0000 | 6.901629 |
| cons_price_idx | 16.0564 | 1.0000 | 4.007044 |
| cons_conf_idx | 4.8732 | 1.0000 | 2.207531 |
| euribor3m | 102.0445 | 1.0000 | 10.101707 |
| nr_employed | 72.1614 | 1.0000 | 8.494785 |
| job | 5.9465 | 10.0000 | 1.093234 |
| marital | 1.4317 | 2.0000 | 1.093859 |
| education | 2.1465 | 2.0000 | 1.210415 |
| default | 1.0000 | 1.0000 | 1 |
| housing | 1.0105 | 1.0000 | 1.005222 |
| loan | 1.0035 | 1.0000 | 1.001763 |
| contact | 1.8671 | 1.0000 | 1.366429 |
| dow | 1.0483 | 4.0000 | 1.005918 |
| quarter | 2.8877 | 3.0000 | 1.193326 |

As a second step, we will try removing the value with the biggest GVIF from Table 3. And re-run the diagnostic:

*Table 5* - GVIF diagnostic for model of Table 3, minus the *euribor3m* predictor

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| age | 3.7102 | 6.0000 | 1.115449 |
| campaign | 1.0383 | 1.0000 | 1.018946 |
| pdays | 1.5750 | 2.0000 | 1.12026 |
| previous | 1.7323 | 2.0000 | 1.147243 |
| emp_var_rate | 44.5797 | 1.0000 | 6.676801 |
| cons_price_idx | 14.1923 | 1.0000 | 3.76727 |

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| cons_conf_idx | 1.6606 | 1.0000 | 1.288646 |
| nr_employed | 22.2175 | 1.0000 | 4.713545 |
| job | 5.9083 | 10.0000 | 1.092881 |
| marital | 1.4321 | 2.0000 | 1.093944 |
| education | 2.1471 | 2.0000 | 1.210496 |
| default | 1.0000 | 1.0000 | 1 |
| housing | 1.0103 | 1.0000 | 1.005123 |
| loan | 1.0035 | 1.0000 | 1.001758 |
| contact | 1.9313 | 1.0000 | 1.389715 |
| dow | 1.0428 | 4.0000 | 1.005256 |
| quarter | 2.3195 | 3.0000 | 1.150533 |

It seems that we are not done and that we must remove the *emp_var_rate* predictor, which is the variable with the greatest GVIF>10 reported in Table 4.

*Table 6* - GVIF values for the model reported in Table 4, minus the *emp_var_rate* predictor

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| age | 3.7093 | 6.0000 | 1.115425 |
| campaign | 1.0330 | 1.0000 | 1.016387 |
| pdays | 1.5703 | 2.0000 | 1.119424 |
| previous | 1.7333 | 2.0000 | 1.147404 |
| cons_price_idx | 3.3818 | 1.0000 | 1.838963 |
| cons_conf_idx | 1.3113 | 1.0000 | 1.145107 |
| nr_employed | 2.5981 | 1.0000 | 1.611862 |
| job | 5.8838 | 10.0000 | 1.092655 |
| marital | 1.4324 | 2.0000 | 1.094 |
| education | 2.1494 | 2.0000 | 1.210824 |
| default | 1.0000 | 1.0000 | 1 |
| housing | 1.0095 | 1.0000 | 1.004757 |
| loan | 1.0035 | 1.0000 | 1.001758 |
| contact | 1.6383 | 1.0000 | 1.279948 |
| dow | 1.0408 | 4.0000 | 1.005011 |
| quarter | 1.7397 | 3.0000 | 1.096682 |

This is what we will consider as our full model for any further Feature Selection tactics.

We notice that all the GVIF values are less than 10 and that all the GVIF^(1/(2*Df)) values are less than 5.

- Large dataset
    - The data has near 40 thousand rows, which is a very large data set for this problem.

Now we are ready to start the process of finding a good subset of predictors considering the model reported in the previous paragraph, for which there is no multicollinearity. To reduce the number of features in the model, so that we will not have overfitting in place, we used three different processes. The function that describes our full model, is:

$$\text{subscribed} \sim Bernoulli\left(\text{prob}_{\text{subscribed =yes}} = \hat{P}\right)$$

$$\log\left[\frac{\hat{P}}{1-\hat{P}}\right] = 64.33 - 0.25\left(\text{age}_{(20,30]}\right) - 0.38\left(\text{age}_{(30,40]}\right) - 0.45\left(\text{age}_{(40,50]}\right) -$$
$$0.3\left(\text{age}_{(50,60]}\right) - 0.08\left(\text{age}_{(60,70]}\right) + 0.08\left(\text{age}_{(70,100]}\right) - 0.05(\text{campaign}) -$$
$$0.17\left(\text{pdays}_{\text{Contacted \_after \_10days}}\right) - 1.72\left(\text{pdays}_{\text{First \_time \_contact ed}}\right) - 0.5\left(\text{previous}_{\text{Less \_than \_or \_3\_times}}\right) - 0.22\left(\text{previous}_{\text{More than \_3\_times}}\right) -$$
$$0.01(\text{cons\_price\_idx}) + 0.02(\text{cons\_conf\_idx}) - 0.01(\text{nr\_employed}) - 0.22\left(\text{job}_{\text{blue –collar}}\right) -$$
$$0.05\left(\text{job}_{\text{entrepreneur}}\right) - 0.14\left(\text{job}_{\text{housemaid}}\right) - 0.03\left(\text{job}_{\text{management}}\right) + 0.08\left(\text{job}_{\text{retired}}\right) -$$
$$0.03\left(\text{job}_{\text{self –employed}}\right) - 0.18\left(\text{job}_{\text{services}}\right) + 0.19\left(\text{job}_{\text{student}}\right) - 0.07\left(\text{job}_{\text{technician}}\right) -$$
$$0.02\left(\text{job}_{\text{unemployed}}\right) + 0.03\left(\text{marital}_{\text{married}}\right) + 0.08\left(\text{marital}_{\text{single}}\right) + 0.03\left(\text{education}_{\text{Secondary \_Education}}\right) +$$
$$0.1\left(\text{education}_{\text{Tertiary \_Education}}\right) - 8.53\left(\text{default}_{\text{yes}}\right) - 0.01\left(\text{housing}_{\text{yes}}\right) - 0.04\left(\text{loan}_{\text{yes}}\right) -$$
$$0.38\left(\text{contact}_{\text{telephone}}\right) - 0.2\left(\text{dow}_{\text{mon}}\right) + 0.06\left(\text{dow}_{\text{thu}}\right) + 0.03\left(\text{dow}_{\text{tue}}\right) +$$
$$0.11\left(\text{dow}_{\text{wed}}\right) + 0.39\left(\text{quarter}_{\text{Q2}}\right) - 0.51\left(\text{quarter}_{\text{Q3}}\right) - 0.56\left(\text{quarter}_{\text{Q4}}\right)$$

At first, we run stepwise methods that would provide us with a basic set of variables to use. AIC and BIC stepwise processes, based on AIC and BIC metric respectively, gave us two models. We also applied LASSO Regression on the sparse matrix of our data and used the $\lambda=\exp(x)$, x is the 1 standard error value of the plot Binomial Deviance-Log($\lambda$), which is reported from cross validation. We took the features of which the beta coefficients were not equal to 0. We see that the models are nested. The model from BIC is nested in the model found using LASSO and that, in turn, is nested in the one from AIC. We will not be evaluating models created from

stepwise methods on the “LASSO model”, because those will drive to the initial ones ran on the full. Here are the model descriptions for each process:

$$\begin{aligned}\text{subscribed} &\sim Bernoulli\left(\text{prob}_{\text{subscribed =yes}} = \hat{P}\right)\\ \log\left[\frac{\hat{P}}{1-\hat{P}}\right] &= 63.18 - 0.26\left(\text{age}_{(20,30]}\right) - 0.41\left(\text{age}_{(30,40]}\right) - 0.5\left(\text{age}_{(40,50]}\right) -\\ &\quad 0.35\left(\text{age}_{(50,60]}\right) - 0.12\left(\text{age}_{(60,70]}\right) + 0.03\left(\text{age}_{(70,100]}\right) - 0.05(\text{campaign}) -\\ &\quad 0.17\left(\text{pdays}_{\text{Contacted \_after \_10days}}\right) - 1.72\left(\text{pdays}_{\text{First \_time \_contacted}}\right) - 0.5\left(\text{previous}_{\text{Less \_than \_or \_3\_times}}\right) - 0.23\left(\text{previous}_{\text{More than \_3\_times}}\right) +\\ &\quad 0.02(\text{cons\_conf\_idx}) - 0.01(\text{nr\_employed}) - 0.23\left(\text{job}_{\text{blue -collar}}\right) - 0.06\left(\text{job}_{\text{entrepreneur}}\right) -\\ &\quad 0.15\left(\text{job}_{\text{housemaid}}\right) - 0.04\left(\text{job}_{\text{management}}\right) + 0.08\left(\text{job}_{\text{retired}}\right) - 0.04\left(\text{job}_{\text{self -employed}}\right) -\\ &\quad 0.19\left(\text{job}_{\text{services}}\right) + 0.2\left(\text{job}_{\text{student}}\right) - 0.08\left(\text{job}_{\text{technician}}\right) - 0.02\left(\text{job}_{\text{unemployed}}\right) +\\ &\quad 0.03\left(\text{education}_{\text{Secondary \_Education}}\right) + 0.11\left(\text{education}_{\text{Tertiary \_Education}}\right) - 0.39\left(\text{contact}_{\text{telephone}}\right) + 0.38\left(\text{quarter}_{\text{Q2}}\right) -\\ &\quad 0.5\left(\text{quarter}_{\text{Q3}}\right) - 0.55\left(\text{quarter}_{\text{Q4}}\right)\end{aligned}$$

LASSO:

AIC:

$$\begin{aligned}\text{subscribed} &\sim Bernoulli\left(\text{prob}_{\text{subscribed =yes}} = \hat{P}\right)\\ \log\left[\frac{\hat{P}}{1-\hat{P}}\right] &= 66.59 - 0.05(\text{campaign}) - 0.19\left(\text{pdays}_{\text{Contacted \_after \_10days}}\right) - 1.74\left(\text{pdays}_{\text{First \_tim e\_contacted}}\right) -\\ &\quad 0.51\left(\text{previous}_{\text{Less \_than \_or \_3\_times}}\right) - 0.19\left(\text{previous}_{\text{More than \_3\_times}}\right) + 0.02(\text{cons\_conf\_idx}) - 0.01(\text{nr\_employed}) -\\ &\quad 0.42\left(\text{contact}_{\text{telephone}}\right) + 0.39\left(\text{quarter}_{\text{Q2}}\right) - 0.55\left(\text{quarter}_{\text{Q3}}\right) - 0.58\left(\text{quarter}_{\text{Q4}}\right)\end{aligned}$$

BIC:

$$\begin{aligned}\text{subscribed} &\sim Bernoulli\left(\text{prob}_{\text{subscribed =yes}} = \hat{P}\right)\\ \log\left[\frac{\hat{P}}{1-\hat{P}}\right] &= 66.59 - 0.05(\text{campaign}) - 0.19\left(\text{pdays}_{\text{Contacted \_after \_10days}}\right) - 1.74\left(\text{pdays}_{\text{First \_time \_contacted}}\right) -\\ &\quad 0.51\left(\text{previous}_{\text{Less \_than \_or \_3\_times}}\right) - 0.19\left(\text{previous}_{\text{More than \_3\_times}}\right) + 0.02(\text{cons\_conf\_idx}) - 0.01(\text{nr\_employed}) -\\ &\quad 0.42\left(\text{contact}_{\text{telephone}}\right) + 0.39\left(\text{quarter}_{\text{Q2}}\right) - 0.55\left(\text{quarter}_{\text{Q3}}\right) - 0.58\left(\text{quarter}_{\text{Q4}}\right)\end{aligned}$$

## Models Comparison

Using the models, we ended up in the Feature Selection process, we will evaluate them, and the features selected. Firstly, we run the likelihood ratio test with ANOVA to compare between the 4 models (the full included). We see which model will explain better our response variable.

*Table 7* - ANOVA LR test

| | Resid. Df | Resid. Dev | Df | Deviance | Pr(>Chi) | |
|---|---|---|---|---|---|---|
| Full Model | 39831.0000 | 21354.0000 | | | | |
| AIC | 39837.0000 | 21357.0000 | -6 | -2.72 | 0.8431 | |
| LASSO | 39841.0000 | 21389.0000 | -4 | -32.279 | 1.68E-06 | *** |
| BIC | 39859.0000 | 21488.0000 | -18 | -98.945 | 3.46E-13 | *** |

As we see in Table 6, The model created using the BIC stepwise procedure is significantly better than the full model and the others. We reject the null hypothesis, that the model produced by BIC has the same explanatory ability with the others.

In Table 8 we see the metrics for each model.

*Table 8* - Statistics of different models

| Process | AIC | BIC | Tjur.s.R2 | Overdispersion Ratio |
|---|---|---|---|---|
| BIC | 21512.16 | 21615.28 | 0.16571 | 0.53910435 |
| AIC | 21424.93 | 21717.11 | 0.17095 | 0.536108055 |
| Lasso | 21449.21 | 21707.01 | 0.16952 | 0.536864417 |
| Full Model | 21434.21 | 21777.95 | 0.17098 | 0.536120533 |

## 5. Results and Discussion

We will choose the model produced by the BIC stepwise procedure to explain our response variable, which has the lowest BIC value and a good $R^2$(Pseudo).

The model has the best overdispersion ratio among the others.

References

Lahmiri, S. (2017). A two-step system for direct bank telemarketing outcome classification. *Intelligent Systems in Accounting, Finance & Management*, *24*(1), 49–55. https://doi.org/10.1002/isaf.1403

Fox, John, and Georges Monette (1992). "Generalized Collinearity Diagnostics." *Journal of the*

Americ Tjur, T. (2009). *Coefficients of determination in logistic regression models - A new proposal: The coefficient of discrimination. The American Statistician, 63(4), 366-372.an Statistical Association* 87 (417): 178–83. http://www.jstor.org/stable/2290467

Moro, S., Laureano. R., and Cortez. P (2012). Enhancing bank directmarketing through data mining, Proceedings of the Forty-First International Conference of the European Marketing Academy, European Marketing Academy. pp. 1–8.

Hoeffding, W. (1948) *"A Class of Statistics with Asymptotically Normal Distribution." The Annals of Mathematical Statistics*, 19(3) 293-325. https://doi.org/10.1214/aoms/1177730196

Thode Jr., H.C. (2002): *Testing for Normality*. Marcel Dekker, New York.

Yamashita, T., Yamashita, K., & Kamimura, R. (2007). A stepwise AIC method for variable selection in linear regression. *Communications in Statistics - Theory and Methods*, *36*(13), 2395–2403. https://doi.org/10.1080/03610920701215639

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling,* New York, NY: Springer

Fox, J. and Weisberg, S. (2018) *An R Companion to Applied Regression*, Third Edition, Sage.

Figure 6- *Age distribution/Outliers*

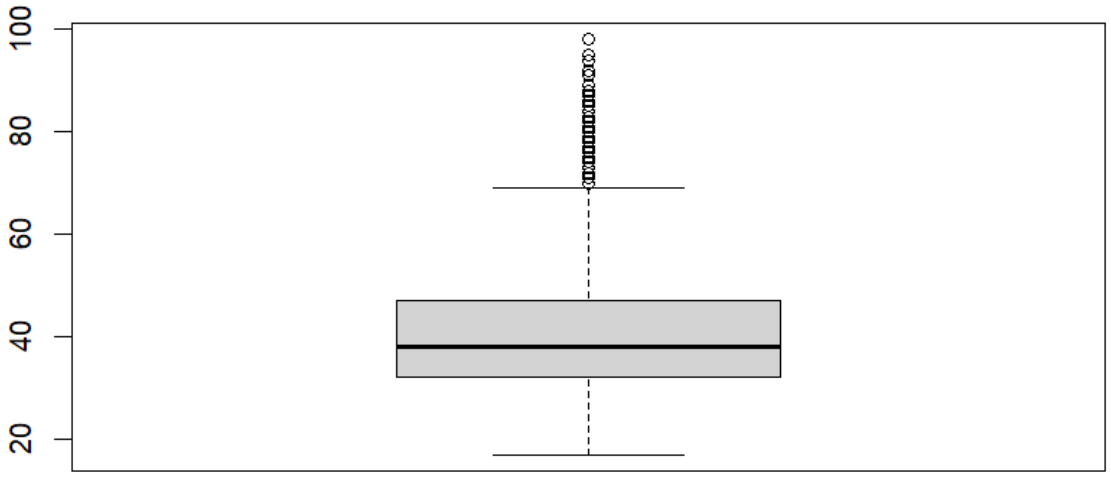


*Figure 7* - Count contacts for each education level

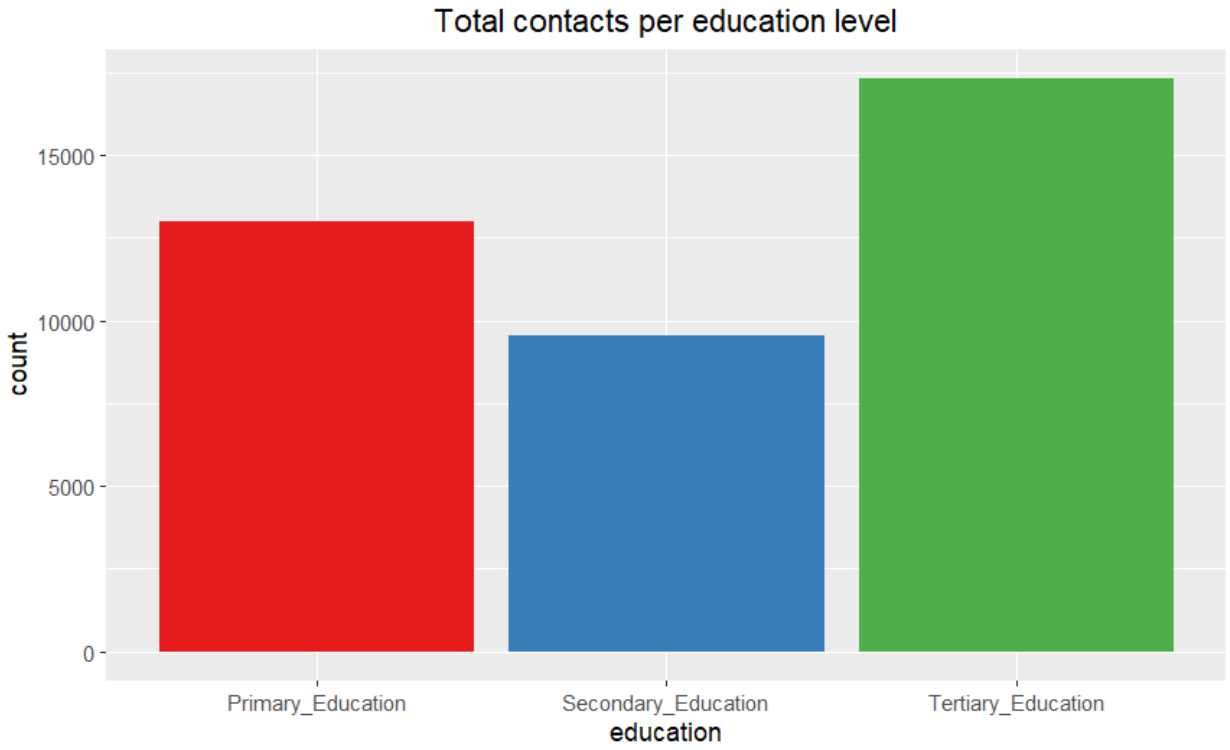


*Figure 8* - Contacts per factorial variable