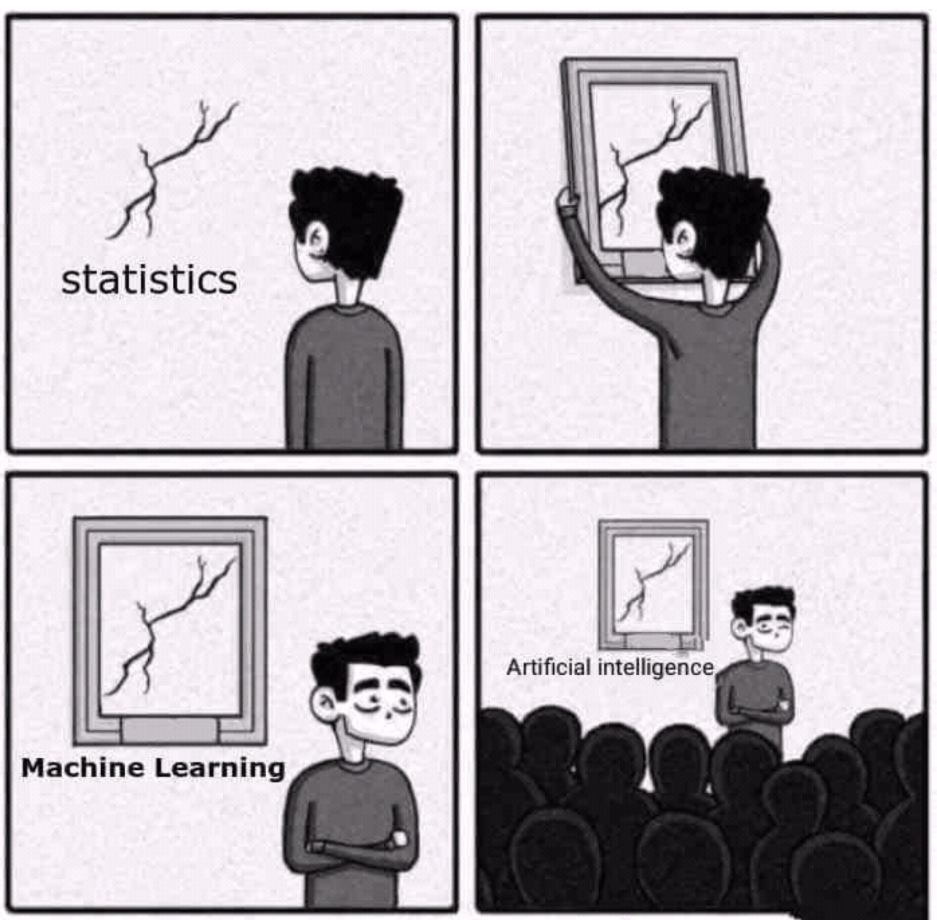
“What influences a client to subscribe to a term deposit?”



CONTENTS

Report……………………………………………………………………………………………...……….1

* Introduction………………………………………………………………………………………...1
* Data preparation................................................................................................................................1
* Description of the dataset.....................................................................................................1
* Data cleansing......................................................................................................................1
* Factor attributes…………………………………………………………………..2
* Integer attributes………………………………………………………………….3
* Numeric attributes………………………………………………………………..3
* Bivariate Analysis.............................................................................................................................4
* Correlations..........................................................................................................................4
* Boxplots...............................................................................................................................5
* Contingency tables...............................................................................................................6
* Model Selection................................................................................................................................7
  + Initial Cleansing...................................................................................................................7
* Problematic values in the model………………………………………………….7
* Hypothesis testing of the model.............................................................................7
* Visualization……………………………………………………………………...8
  + Variable Selection and Multi-collinearity testing................................................................9
  + Visualizations…………………………………………………………………………….11
  + Explanation of the model………………………………………………………………...12
  + Further Analysis.................................................................................................................14
* Conclusions and Discussions..........................................................................................................15
* Comment on whether the model can be used for prediction………………………….…15
* Conclusions……………………………………………………………………………....16
* Bibliography...................................................................................................................................17
* Appendix………………………………………………………………………………………….18

1. **Introduction**

In today’s fast-paced evolving world, marketing plays a crucial role in a lot of businesses and organizations. Improving the marketing communications process will lead to growth and more profits. A lot of banks are using marketing campaigns in order to promote their products. Telemarketing is when the marketing of products is implemented by means of telephone calls. Contacts can be divided in inbound and outbound, based on who contacted who. For example, if an agent calls a client to sell a product, then it is considered outbound, but if the client calls the contact center, it is considered inbound.

1. **Data preparation**

**1.1 Description of the dataset:**

The data refers to telemarketing phone calls and are collected from one of the retail bank, from May 2008 to June 2010. There are 39883 phone contacts and 22 attributes in this data set. The purpose of the “CODE” variable is to select the data that corresponds to the final digit of our registration number. After doing that, we remove this column from our data and thus our new reduced dataset has 4019 rows and 21 columns.

Our aim is to examine which variables contribute to a successful contact (the client subscribes to the product). Beside our output variable y (subscribed), we have 20 attributes which refers to the following 4 groups:

Table 1: Variables in the dataset

|  |  |
| --- | --- |
|  | Variables Names |
| Bank client data | age, job, marital, education, default, housing, loan |
| Last contact of the current campaign | contact, month, day\_of\_week, duration |
| Other attributes | pdays, previous, poutcome |
| Social and economic context attributes | emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed |

The descriptions of these attributes can be found in the Table 1 in Appendix.

**1.2 Data cleansing:**

Our original dataset is a data frame and at first glance it does not look to have any missing values (na, null, nan) or infinite values. Also, we do not have any duplicate rows in our data.

For a better analysis, we decided to divide our data according to their data type into factors, numeric and integers in order to have a clearer view of the results.

Table 2: Variables according to their data type

|  |  |
| --- | --- |
| Variable data type | Variables Names |
| Factor | job, marital, education, default, housing, loan, contact, month, day\_of\_week, poutcome, SUBSCRIBED |
| Numeric | emp.var.rate, cons.price.idx, euribor3m, nr.employed |
| Integer | age, duration, campaign, pdays, previous |

We will use the summary function to see some basic descriptive measures for the numeric and integers variables and some frequency tables for the factors, which will display the distribution of counts per category.

**1.2.1 Factor attributes:**

At first we see that there are some missing records which were kept and labeled as “unknown”.

We observe that the default attribute is binary (yes/no) and all of its non-missing values corresponds to only one of the two possible outcomes, which renders it useless. So, we removed this variable from our dataset.

In the summary of job and education variables we saw a category named “(Other)”. The summary function does not show all the levels of the factors. By further exploration we observed that these variables had also missing values hidden inside them. Specifically, there were 34(0.9%) and 147(3.8%) observations in job and education variables accordingly. Currently, we will leave the unknown values of the variables intact and examine later in our analysis whether they are helpful or not.

**Grouping**

We decided to group the levels of some variables into broader categories, which will also be proven useful in terms of model interpretation.

At first, we recoded the levels of education variable into the following categories, according to the education stage:

Table 3: Grouping of Education variable according to the education stage

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Education** | | | | | |
| illiteracy | primary | secondary | professional course | university | unknown |

These educational stages, just like the days of week, are considered to have an order in their values. So, for clearer and simpler results in our analysis, we made sure that the levels will be ordered from Monday to Friday and from illiterate to university accordingly.

Next, we reclassified the levels of job variable from 12 to 3, because some of those had very small frequencies compared to others. So, we decided to create broader groups of white and blue collar workers, along with other types of jobs.

Table 4: Grouping of Job types

|  |  |  |
| --- | --- | --- |
| **Job** | | |
| White collar | Blue collar | Other |

We also grouped the month column into four groups based on the seasons).

Table 5: Grouping of months, according to seasons

|  |  |  |  |
| --- | --- | --- | --- |
| **Month** | | | |
| Spring | Summer | Fall | Winter |

We also viewed the contrasts associated with a factor. These will show us how the variables have been dummyfied by R and how to interpret them in a model, based on their reference level. For example, we saw that our dependent variable uses the "no" level as reference.

**1.2.2 Integer attributes:**

Next, we continued our analysis by computing some descriptive measures for the integer variables in our dataset.

At first, we observed some strange results in the summary of “pdays” variable. The reason was the existence of a value (“999”) which was used as a descriptor for a client who was not previously contacted. Almost 97% of the values of “pdays” variable are equal to “999”, which renders it unusable.

In order to make better use of this variable, we created a new binary coded one, depending on whether the client was previously contacted or not. We made sure to assign labels in this new factor variable, in order to give meaning to the coded levels. Then we removed “pdays” from our dataset.

Table 6: Frequency table for p\_contact variable

|  |  |
| --- | --- |
| Was the client previously contacted or not?(p\_contact) | |
| yes | no |
| 101 | 3918 |

We also observed that the “previous” variable had only five unique values. A lot of them were equal to zero, but due to the fact that we also had a lot of observations with different values – almost 450 rows (12%) – , we decided that we could make use of this variable and we did not removed it.

**Grouping**

We could make better use of some integer variables, by converting them to categorical. Specifically, we decided to observe the effect of age and duration across some general categories, as shown below. For creating the categories, we set some roughly breakpoints, by taking into account the unique values and the range of each variables.

Table 7: Grouping of age variable in years

|  |  |  |  |
| --- | --- | --- | --- |
| Age groups (in years) | | | |
| 18 - 30 | 30 - 45 | 45 - 60 | Over 60 |

Table 8: Grouping of duration variable in minutes

|  |  |  |  |
| --- | --- | --- | --- |
| Duration groups (in minutes) | | | |
| 0 - 5 | 5 - 10 | 10 - 15 | Over 15 |

**1.2.3 Numeric attributes:**

**Normalization (Min-Max Scalar):**

The numeric variables are all measured in different scales with different ranges. For example, the employment variation rate is used as a quarterly indicator while the euribor 3 month rate is used as a daily indicator. So we decided to bring them into a common scale (between 0 and 1), without distorting differences in the ranges of values by normalizing them.

In the summary of the model we will fit later on, the results will be the same, but, if we did not normalize the numeric variables of our dataset then their estimated coefficients will have smaller values due to the fact that they are on a different scale and we could not understood the exact effect a variable has on the target variable.

After all these modifications, we created a file with the updated data, which we will use further on for our analysis.

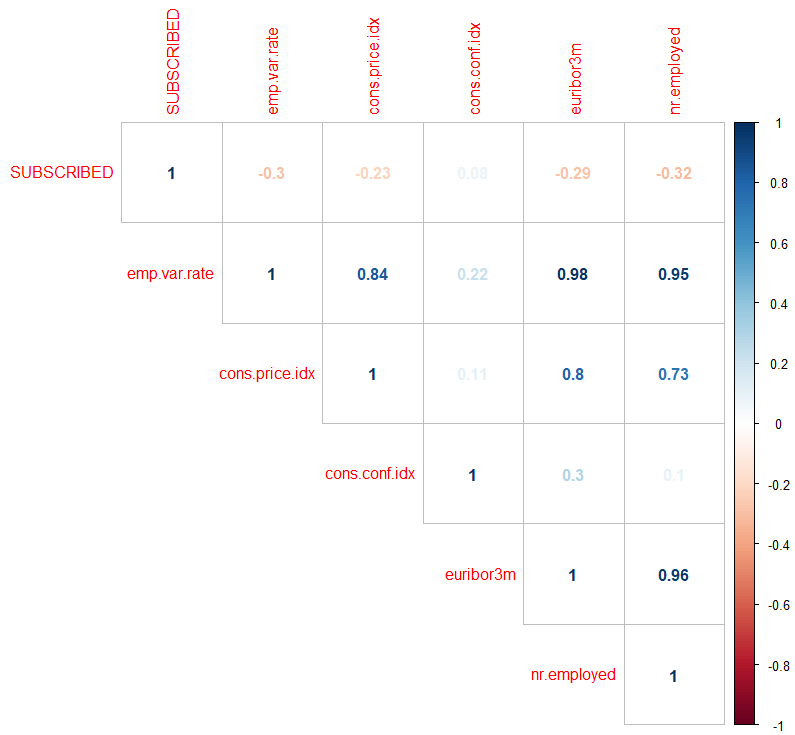
1. **Bivariate Analysis**

**2.1 Correlations:**

We computed correlations and made correlation plots of our dependent variable with the continuous variables to investigate if there are any associations implied by the dataset. We used the empirical values of correlations and we sorted the computed correlations by those who had strong and medium correlation with subscription.

In the plot below we show the correlation coefficients between our target variable and the numeric attributes.

Table 9: Correlation coefficients for the numeric variables with the dependent



Based on the results from the correlations we concluded that the employment variation rate, euribor 3 month rate and the number of employees are almost perfectly correlated. So we should include only one of them in our final model, because if we include collinear variables then the estimated coefficients and standard errors will have really big values and thus their interpretation will not have any meaning.

These three attributes appear to have the higher effect on subscription. They are also negative correlated with our target variable, e.g. the number of clients who subscribe to a term deposit will decrease as the number of employments increases.

We also observed a high correlation between the employment variation rate and the consumer price index.

Finally, as we see in the table 10, some variables seem to be irrelevant to our analysis, because they had almost zero correlation with the subscription variable. So we expect those attributes to not be included in our final model. These attributes are shown below:

Table 10: Variables with very "weak" correlation with the dependent variable

|  |  |
| --- | --- |
| **Variable** | **Correlation coefficient** **with output variable** |
| campaign | -0.08 |
| cons.conf.idx | 0.08 |

**2.2 Boxplots:**

We also created some boxplots of the numeric variables for each level of our dependent variable, as shown below. Thanks to the normalization we did before, we can easily compare those boxplots together according to their structure and median value.

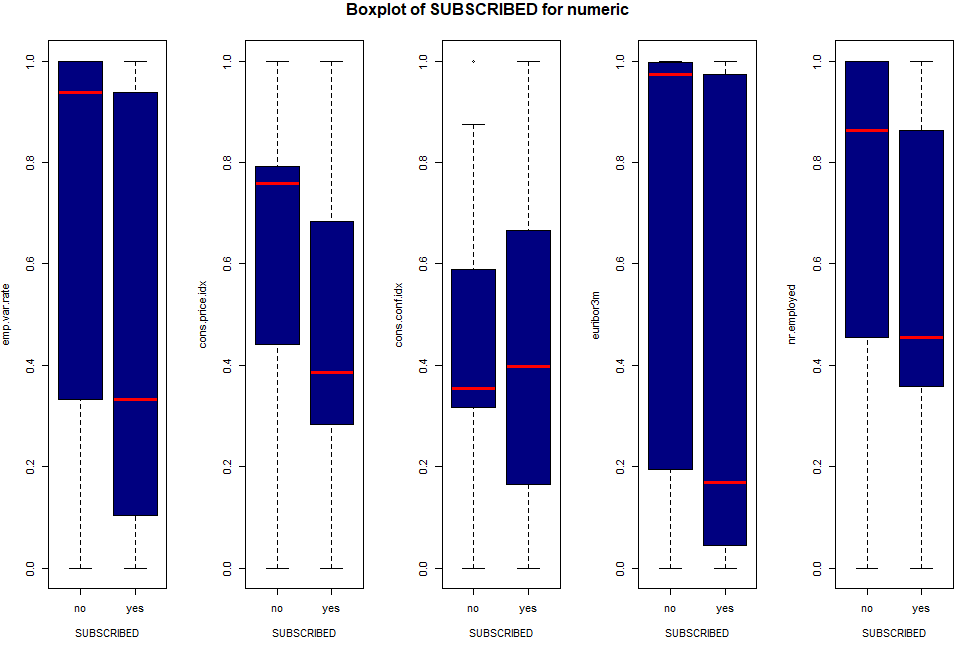


Figure 1: Boxplots between the numeric variables with the dependent

From these boxplots we understood that the variables named “emp.var.rate” , “cons.price.idx”, “euribor3m”, “nr.employed” may be proven useful to our analysis, because their median values seem to differ a lot for each level of our target variable, in contrast with “cons.conf.idx” variable. We can also see that these variables have also the same skewness, negative (mean<median) for the “no” level and positive (mean>median) for the “yes” level of subscription.

Finally, we observed that “emp.var.rate” and “euribor3m” have almost the same dispersion, because of their interquartile ranges (that is, the box lengths), just as we expected due to their high correlation.

We also created some boxplots for the integer variable in our dataset for each level of our dependent variable, as shown below.

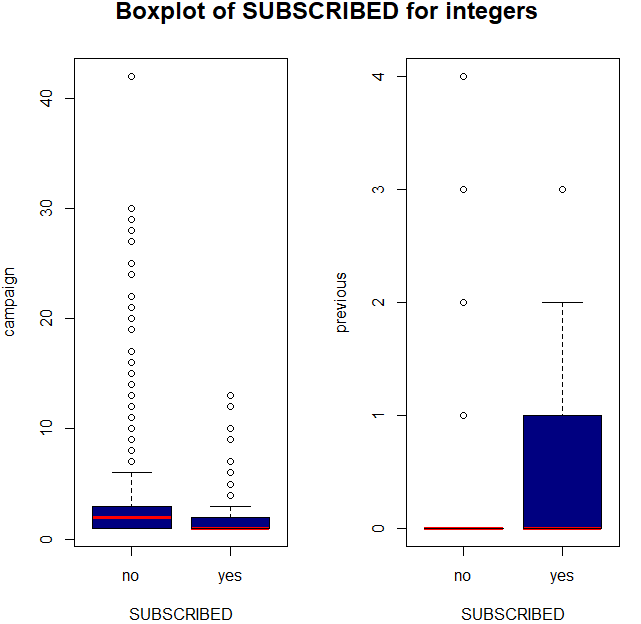


Figure 2: Boxplots of the integer variables with the dependent

So from these boxplots, we understood that the “cons.conf.idx” and “campaign” variables do not add any information to the subscription of a term deposit, because their median values are almost equal for both levels of our target variable.

**2.3 Contingency tables:**

For the factor variables we decided to create some contingency tables in order to examine their relationships with our target variable. Our conclusions from these tables were based on chi-square test.

We observed that the existence of a personal or housing loan and the weekday the last contact was performed, do not affect the outcome of subscription. We also observed that whether a client has a housing or a personal loan is related to each other, just like the previous contact of client is related to the outcome of the previous marketing campaign.

To sum up, from the correlations coefficients and the contingency tables we concluded that the result of the subscription does not seem to be affected by the following variables:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| campaign | cons.conf.idx | housing | loan | day\_of\_week |

1. **Model Selection**

**3.1 Initial Cleansing:**

**3.1.1 Problematic values in the model:**

Our dependent variable is binary and examines whether a client has subscribed to a term deposit or not. That’s why we decided to use a logistic regression model, in order to understand the factors that would influence a client to subscribe.

At first, we fitted a model which contained all the variables in the dataset, along with all the unknown values in some factor’s levels. We computed the summary of the model and we observed that the unknown level of loan variable had NA as values. This is usually occurred due to strong correlation between our independent variables and can be avoided by having one less dummy variable. So, we excluded the unknown level from the loan variable.

In the next model we fitted, we observed that the estimations of the coefficients and the standard errors of the unknown level for marital variable and the illiterate level for the education variable had really high values. That means that we cannot estimate them and that their existence in the model does not affect the rest of the coefficients. By further investigation we observed that the marital variable has 8 unknown values, all of which corresponds to only one of the two levels of our dependent variable. The same occurs with the 3 unknown observations for the illiterate level of the education level. This is the reason why these variables has very high estimates and standard errors. This is known as full separation problem[[1]](#footnote-1).

So, we dropped the rows of the dataset which corresponded to these levels, because they did not add any information to the model.

The fact that we had grouped some variables into broader categories do not change the results taken from the logistic regression. But we should be careful of the reference level used in those categories. For example, we fitted a model where the illiterate level was used as reference and that resulted into high values for all the other levels of the education variable.

After removing the observations that corresponded to the previous levels from our dataset, we fitted another model, which from now on we are going use as our starting point.

**3.1.2 Hypothesis testing of the model:**

**Wald test:**

At first we tested whether a variable has a statistically significant effect in our model. We made use of the Wald test, which corresponds to the same p-value with the one in the summary function, but we can also use it in order totest for an overall effect of a factor. The order in which the coefficients are given in the table of coefficients is the same as the order of the terms in the model.

**Goodness of fit test (GOF)**

At first, we compared our model with the saturated model[[2]](#footnote-2) and concluded that our model fits “well”. Then we compared our model against a model containing only the intercept (null model). The first approach we used to test this was via the likelihood ratio test, where the difference between the null deviance and the model's deviance is distributed as a chi-squared with degrees of freedom equal to the null df[[3]](#footnote-3) minus the model's df. For the second approach we used analysis of variance (anova). Both approaches concluded that our model as a whole fits significantly better than an empty model.

**3.1.3 Visualization**

Before continuing with our analysis, it will be helpful to visualize the relationship between the factor variables and the dependent variable using our updated dataset.

At first we should get a gist of the distribution of our target variable, before examining relationships between this and the independent factor variables.

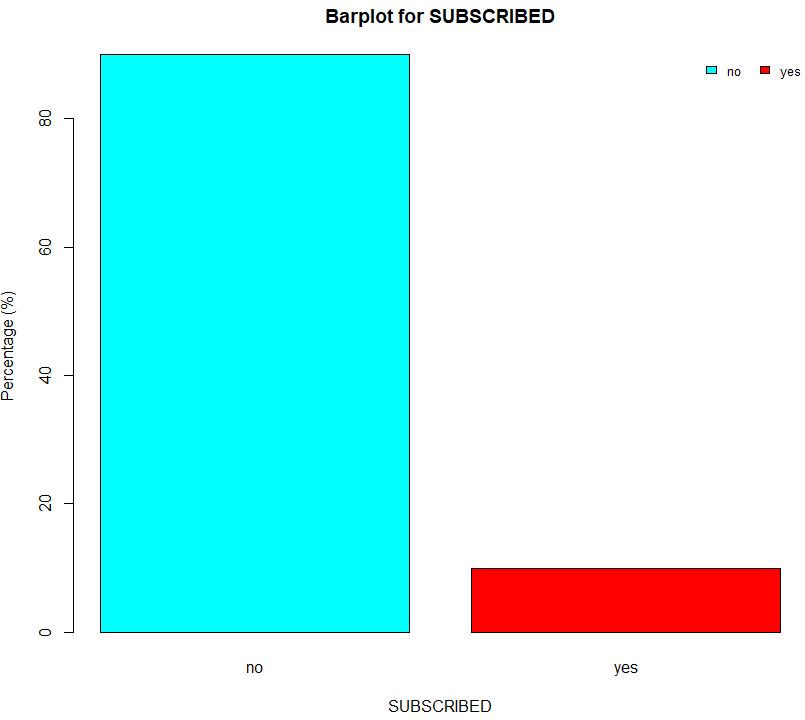


Figure 3: Boxplot for the dependent variable

From the diagram above, we see that almost 90% of the clients had not subscribed to the term deposit.

We also created some bar-plots, in order to see the effect a factor had on whether a client has subscribed a term deposit or not. The most significant changes were observed for the age, job, duration, marital, contact, previous contact, seasons and education attributes.

Based on the bar-plots and the contingency tables, we observed that those who had tertiary education, were single, not blue-or-white collar workers, over 60 years old, had been previously contacted – ideally via cellular phone in fall – and had talked over 15 minutes were more likely to subscribe.

In contrary, when the target group was people who were between 30 and 45 years old, worked as blue-collar, married and were not previously contacted or talked less than 5 minutes the success rate of subscription was way lower.

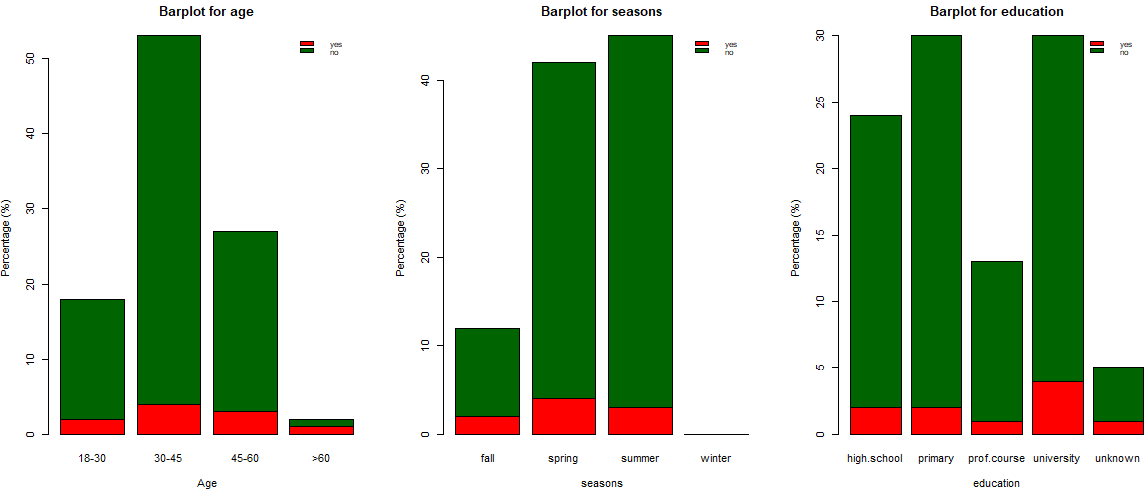


Figure 4: Bar plots for age groups, seasons and education stages with the "SUBSCRIBED" variable

**3.2 Variable Selection and Multi-collinearity testing:**

In order to find the best model for explaining the data, we selected the appropriate features using stepwise procedures. The stepwise methods allows us to remove variables from our model that are not statistically significant. The goal is to find the predictors which will help us minimize either the AIC or the BIC.

We used the stepwise methods according to Akaike’s(AIC) and Bayesian(BIC) information criteria in order to find better models.

Then we checked for multicollinearity between our variables. If two variables are collinear then they carry the same information, so we should not add them both in our model.

We did not face any multi-collinearity problem for the model selected with the BIC criterion, but we faced to the one selected from AIC. In order to avoid such problems we had to exclude two more variables ("euribor3m" and "nr.employed) from the model suggested by AIC.

After fitting the model again we observed that the variable “contact” did not have a statistically significant level anymore.

The p-value for each coefficient of the independent variables tests the null hypothesis that the variable had no correlation with the dependent variable. It is standard practice to use the coefficient p-values to decide whether to include variables in the final model. Keeping variables that are not statistically significant can reduce the model’s precision. That is why we fitted another model, keeping only the statistically significant variables (with the variables named “contact” and “cons.conf.idx” removed).

For all the models shown below, we checked the statistical significance of their variables proposed by via the Wald test and the p-value shown in summary.

The results of our efforts are presented in Table 11.

Table 11: Trials of models for the selection process

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Selection Process** | | | |
| **Models** | **AIC** | **Residual Deviance/Df[[4]](#footnote-4)** | **Pseudo R2 (Mc Fadden)** |
| Full model | 1532 | .3765 | . 4222827 |
| **Stepwise** | | | |
| With AIC | 1517.3 | .3777 | . 4186176 |
| With BIC | 1554.2 | . 3934 | . 3921716 |
| **Multi-collinearity correction** | | | |
| model with BIC | No problem | | |
| model with ΑIC | 1532.3 | 0.3824 | . 4111026 |
| **Keeping only the statistical significant variables** | | | |
| model with ΑIC without "contact” variable | 1532 | . 3832 | . 4104242 |
| model with ΑIC without “contact” and "cons.conf.idx" variables | 1532.3 | . 3832 | . 4094919 |

We observed that the model suggested by BIC, is nested to the one proposed by AIC.

Now was the time to choose between the model suggested by AIC after the modifications and the model suggested by BIC. The first one described the target variable using nine variables and the second one using four.

Generally we want our model to have a small AIC value. As s rule of thumb we also know that McFadden's pseudo R-squared ranging from 0.2 to 0.4 indicates very good model fit. For choosing the best model we were based on the AIC value, on R-squared suggested by Mc Fadden and on the fraction of Residual deviance with the degrees of freedom (the higher the better).

According to the above, both models fit well. But we should prefer the one suggested by BIC, because it is almost equivalent to the first one and has smaller number of parameters.

We also used compared those two models using anova, where we also saw that the model selected using BIC, fits our data better compared to the other.

So the model we are going to use is the one selecting from the stepwise method using the BIC criterion.

According to BIC, the selected model contains the following attributes:

|  |  |  |  |
| --- | --- | --- | --- |
| Job | duration | p\_contact | emp.var.rate |

The reference levels used for the factors in the model are shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| ***Variable*** | job | duration | p\_contact |
| ***Reference level*** | “Blue Collar” | “0 – 5 minutes” | “no” |

**3.3 Visualizations**

We used some bar plots in order to see the distribution of our outcome variable, along the factor variables suggested by our final model.

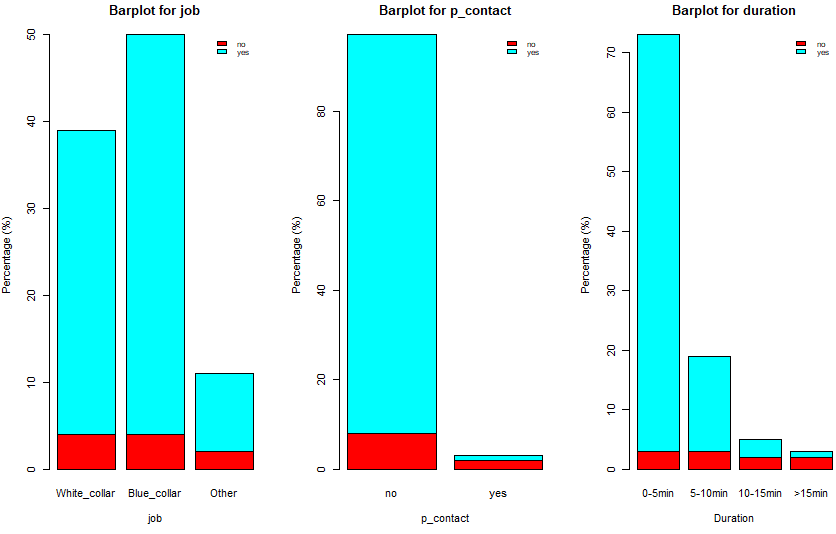


Figure 5: Bar plots for the proportion of client’s job type, successful previous contact and their duration with the "SUBSCRIBED" variable

From the first bar-plot we can see that the bank is mostly associated with blue collar workers. In fact, almost 90% of its clients are blue or white collars. As for the second graph, we observe that most of the clients were not previously contacted. From those who did, 3 out of 4 people (75%) had talked less than 5 minutes.

**3.4 Explanation of the model:**

Mathematical formulation of the model:

{\displaystyle \ell =\log \_{b}{\frac {p}{1-p}}=\beta \_{0}+\beta \_{1}x\_{1}+\beta \_{2}x\_{2}}We concluded to a model with four predictors and one binary response variable named “SUBSCRIBED”, which we denote p=P(“SUBSCRIBED”=”yes”). This is the probability of subscribing to the term deposit and it will be in the unit interval [0,1]. We assume a linear relationship between the predictor variables and the log-odds of the event that “SUBSCRIBED”=”yes”. This linear relationship can be written in the following mathematical form:

odds = P(SUBSCRIBED) / (1 – P(SUBSCRIBED))

log(odds) = logit(SUBSCRIBED) = -1.8 + 0.9\*job2+ 0.4\*job3 + 1.5\* duration2 + 3.6\* duration3 + 4.7\* duration4 + 2.3\*p\_contact- 3.3 \*emp.var.rate

Dummy variables:

* job2 =
* job3 =
* duration2 =
* duration3 =
* duration4 =
* p\_contact =

Obviously job1+ job2, duration1 + duration2 + duration3, p\_contact all ∈ {0,1}.

Baseline category for job is “Blue Collar”, for duration is “0-5min” and for previous contact is “no”.

Table 12: Summary of the chosen logistic regression model

|  |  |  |  |
| --- | --- | --- | --- |
|  | Estimate | Std.Error | p-value |
| Intercept | -1.80 | 0.17 | < 2e-16 \*\*\* |
| job2 | 0.91 | 0.19 | 2.98e-06 \*\*\* |
| job3 | 0.41 | 0.15 | .00635 \*\* |
| duration2 | 1.53 | 0.16 | < 2e-16 \*\*\* |
| duration3 | 3.57 | 0.21 | < 2e-16 \*\*\* |
| duration4 | 4.70 | 0.25 | < 2e-16 \*\*\* |
| p\_contact | 2.34 | 0.26 | < 2e-16 \*\*\* |
| Employment variation rate (emp.var.rate) | -3.33 | 0.23 | < 2e-16 \*\*\* |
|  | | | |
| Residual deviance: 1538.2 on 3910 degrees of freedom | | | |
| AIC: 1554.2 | | | |

We observe that job, duration and previous contact variables have a positive impact – positive coefficients - to our target variable, meaning that a rise in those variables will increase the probabilities of a client to subscribe to a term deposit. In contrary, the employment variation rate has a negative relationship with the outcome variable.

Interpretation of the coefficients of the model

Our dependent variable is computed using the logit transformation; thus, we are now using a different scale.

Intercept β0: If the employment variation rate does not change, then the log odds of subscription for a client who worked as Blue collar and was not previously contacted are equal to -1.8. This multiplies the actual odds by approximately 0.16 units.

Coefficient emp.var.rate: One unit increase in employment variation rate will result the log odds of subscription to increase by 3.3 units or multiply the actual odds of subscribing by e3.3 ≈ 27.1 units, assuming that all the other variables remain constant.

Coefficient job2: β0 + 0.9: If the employment variation rate does not change, then the log odds of subscription for a client who was not a blue- or white-collar worker are equal to -0.9, 0.9 units bigger compared to a blue-collar worker.

Coefficient job3: β0 + 0.4: If the employment variation rate does not change, then the log odds of subscription for a client who was a white-collar worker are equal to -1.4, 0.4 units bigger compared to a blue-collar worker.

Coefficient duration2: β0 + 1.5: If the employment variation rate does not change, then the log odds of subscription for a client whose last contact duration was between 5 and 10 minutes are equal to -0.3, 1.5 units bigger compared to one who did not talk at all or talked less than 5 minutes.

Coefficient duration3: β0 + 3.6: If the employment variation rate does not change, then the log odds of subscription for a client whose last contact duration was between 10 and 15 minutes are equal to 1.8, 1.5 units bigger compared to one who did not talk at all or talked less than 5 minutes.

Coefficient duration4: β0 + 4.7: If the employment variation rate does not change, then the log odds of subscription for a client whose last contact duration was bigger 15 minutes are equal to 2.9, 4.7 units bigger compared to one who did not talk at all or talked less than 5 minutes.

Coefficient p\_contact: β0 + 2.3: If the employment variation rate does not change, then the log odds of subscription for a client who was previously contacted are equal to 0.5, 2.3 units bigger compared to one who was not previously contacted at all.

Extra interpretation of the model

We used the residual deviance for goodness of fit tests.

From the standard error column we see how much the estimation of the coefficients can change. We should look at this column as indication of how stable my estimation is. Here all the standard errors are small and close to the estimation values, so we can assume that all the estimations are stable.

From the final column we see that we reject the null hypothesis, that a coefficient can be assumed to be equal to zero because, the p-value of each coefficient is less than our significance level (a=5%). As a result, all the coefficients in the model are statistically significant in our analysis.

**3.5 Further analysis**

We also created the standard diagnostic plots[[5]](#footnote-5) for our model ('residuals vs fitted', 'normal q-q', 'scale-location', and 'residuals vs leverage' plots). But the interpretations of these plots are not generally valid when used with a logistic regression model. For example, both the 'residuals vs fitted' and the 'scale-location' plots look like there are problems with the model, but in fact there aren't any. So, we will not use them in our analysis, except only for outlier detection.

As for the evaluation of the fit of our model, we could say from the residuals against fitted values plot[[6]](#footnote-6) that we have a pretty decent fit. We also saw that by plotting[[7]](#footnote-7) the Pearson or Deviance residuals against each explanatory variable in our model separately, where we did not observe any unusual pattern. This was also seen by the high value of McFadden's pseudo R-squared.

**Assumptions of logistic regression**

First, binary logistic regression requires the dependent variable to be **binary**, just like our subscribe variable is, which has only two possible outcomes (“yes” vs “no”)

Second, logistic regression requires the observations to be **independent** of each other.  In our case this is false, because the observations come from matched data. This is because our collected data refer to clients from May 2008 to June 2010, during which period more than one contact to the same client was required. This is a major problem, because if we assume that measurements taken in the same client are correlated, the test for a difference in treatments will involve a smaller residual or error variance than that based on a completely randomized design, thereby increasing precision in the analysis.

Third, logistic regression requires there to be little or no **multicollinearity** among the independent variables.  We checked that during our analysis using the GVIF[[8]](#footnote-8) function. As a rule of thumb, a GVIF value that exceeds 5 or the square root of 10 indicates a problematic amount of collinearity and this was the reason we decided to drop some variables from the model suggested by AIC. The final selected model did not have any multicollinearity problems.

Fourth, logistic regression assumes **linearity** of independent variables and log odds. We tested this assumptions by including in the model interactions between the continuous predictors and their logs via the Box-Tidwell test[[9]](#footnote-9). The interaction was significant so the assumption has been violated. However, the size is a factor here too, so we should not be very concerned with just a significant interaction, because our sample size is large.

Finally, logistic regression typically requires a **large sample size**. Here we kept almost 4000 observations from our original sample size, which is considered an adequate number.

We also do not want to have any influential values (extreme values or outliers) in the continuous predictors. From Cook’s distance[[10]](#footnote-10) we saw that there were 3 outliers which may be influential points. By computing the standardized residual error we saw that there were two data points with absolute standardized residuals above 3. So, we could assume that these may be influential points.

So we conclude that not all the logistic regression assumptions are satisfied.

**Link functions**

A link function identifies a function of the mean that is a linear function of the explanatory variables. Our binary dependent variable belongs to the binomial exponential family, so we can use logit or probit as link functions.

Table 13: Comparison between the "probit" and the "logit" link functions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Link Function | Residual Deviance | Residual Deviance/df | AIC | Pseudo R2 |
| Probit | 1516.8 | 0.3879 | 1532.8 | 0.4006193 |
| Logit | 1538.2 | 0.3934 | 1554.2 | 0.3921716 |

We see that the differences between those two link functions are minor and the two models lead to very similar results. So, we could choose either one of them, but due to the fact that our dependent variable is considered to be a truly qualitative and binomial character we prefer the logit modelling.

1. **Conclusions and Discussions**

**6.1 Comment on whether the model can be used for prediction**

From Figure 3, we observed that most of the observations of our target variable, almost 90%, correspond to one of its two levels. This means that we have imbalanced data, because the classes are not represented equally.

Generally, imbalanced classifications pose a challenge for predictive modeling[[11]](#footnote-11), as most of the machine learning algorithms used for classification were based on the assumptions of an equal number of examples for each class. This will result in models that have poor predictive performance, specifically for minority class. Considering that we want to examine which variables contribute to a successful contact, where 90% of the clients in our dataset had not subscribed a term deposit, it will be really hard for our model to predict the chance that a client would subscribe.

We also observed that the “yes” values (coded value=1) of the dependent variable are really spread out in the plot[[12]](#footnote-12) in contrast with the “no” values. That means that our model could not predict well the clients who subscribed a term deposit, just as we expected due to the imbalance of our data.

So, it is advisable to use our model for descriptive purposes instead of predictions.

**6.2 Conclusions**

The analysis of our datasets, showed us that we could identify the attributes that contribute to a successful contact.

The factors which contribute most to the subscription of a term deposit from a client are their job type and the existence and duration of a previous contact. Specifically, if a client was previously contacted then he was more likely to subscribe. We also saw that the more time a client spent talking the better the chances of subscription.

In contrary, we observed that a drop in the employment variation rate will make it a lot more difficult for a successful subscription. This was expected because when the number of people who are hired decreases and the number of those who are fired increases, then we understand that unemployment is an inhibit factor for someone to subscribe to a term deposit.

We also observed that when we removed the observations that corresponded to the “unknown” level of loan variable, the exact same observations corresponded to the “unknown” level of the housing variable. So, we may assume that loan is “unknown” if only housing is “unknown”. By further investigation we saw from the contingency tables that the reason for this was because these two variables were related to each other.

This work has been done considering a part of the dataset. For future works we could use the dataset to its fullest potential, by taking all the observations into consideration and create even more accurate results.

**Bibliography**

1. **WEB SOURCES**

* Anon, Separation, *Wikipedia*. Available at <https://en.wikipedia.org/wiki/Separation_(statistics)>

# [Jason Brownlee](https://machinelearningmastery.com/author/jasonb/), A Gentle Introduction to Imbalanced Classification. Available at <https://machinelearningmastery.com/what-is-imbalanced-classification/>

# Anon, Assumptions of Logistic Regression, *StatisticsSolutions*. Available at <https://www.statisticssolutions.com/assumptions-of-logistic-regression/>

* Anon, Logistic Regressions Assumptions and Diagnostics in R, *STHDA*. Available at <http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/>

# [Lisa M. Sullivan](https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.107.654350?fbclid=IwAR1RZ67L0FhcZ24MXK2nfeATgKtHAv2uczcA8jJn98BF9EDOmguSmYlsyt8&), Repeated Measures, Circulation. Available at <https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.107.654350?fbclid=IwAR1RZ67L0FhcZ24MXK2nfeATgKtHAv2uczcA8jJn98BF9EDOmguSmYlsyt8&>

* [EVAN TARVER](https://www.investopedia.com/contributors/53886/), Marketing Campaign, Investopedia, Available at <https://www.investopedia.com/terms/m/marketing-campaign.asp>
* Jim Marcus, 10 Marketing Trends The Banking Industry Can’t Ignore. Available at <https://thefinancialbrand.com/57325/ten-financial-marketing-priorities-trends/>

1. **OTHER SOURCES**

* Karlis D., allregression (2020) Slides.

**APPENDIX:**

Table1: Data description

|  |
| --- |
| **Retail Bank Data** |
| **Input variables:** |
| **Bank client data**  1 - age (numeric) 2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown') 3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed) 4 - education (categorical: basic.4y','basic.6y','basic.9y', 'high.school','illiterate','professional.course', 'university.degree','unknown') 5 - default: has credit in default? (categorical: 'no','yes','unknown') 6 - housing: has housing loan? (categorical: 'no','yes','unknown') 7 - loan: has personal loan? (categorical: 'no','yes','unknown')  **related with the last contact of the current campaign:** 8 - contact: contact communication type (categorical: 'cellular','telephone')  9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec') 10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri') 11 - duration: last contact duration, in seconds (numeric).  **other attributes:** 12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact) 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted) 14 - previous: number of contacts performed before this campaign and for this client (numeric) 15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')  **social and economic context attributes** 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric) 17 - cons.price.idx: consumer price index - monthly indicator (numeric)  18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)  19 - euribor3m: euribor 3 month rate - daily indicator (numeric) 20 - nr.employed: number of employees - quarterly indicator (numeric) |
| **Output variable (desired target):** |
| 21 - SUBSCRIBED - has the client subscribed a term deposit? (binary: 'yes','no') |

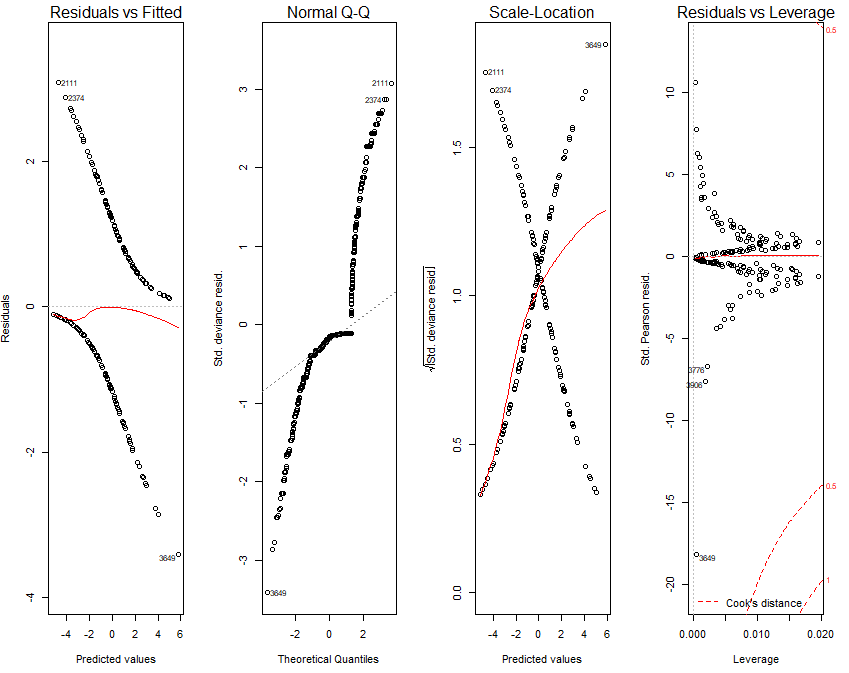


Figure 1: Standard Diagnostic Plots

It seems like the 2111 and 2374 data points are outliers and could very well affect our model.

Note: We should actually be using **Pearson or deviance residuals** to gauge our models fit so the fact that we don’t have anything close to a straight line in the upper left plot is fine. Also, **Q-Q plots** are irrelevant for this kind of model as well.

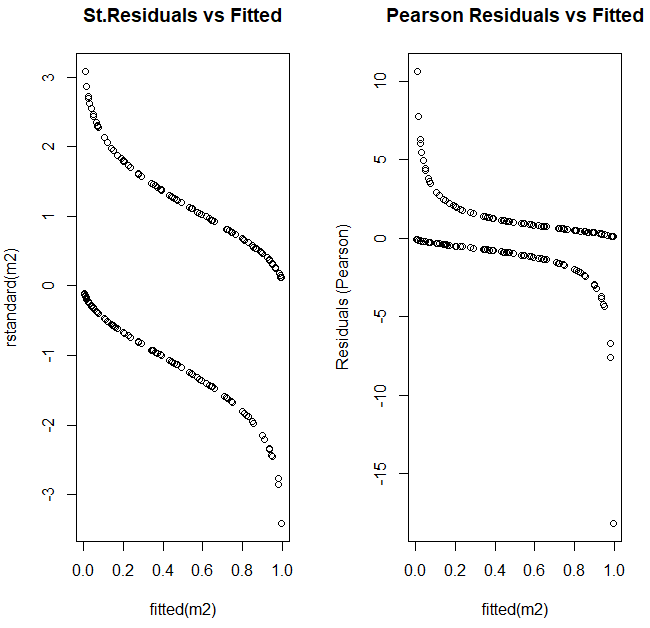


Figure 2: Residuals against Fitted values

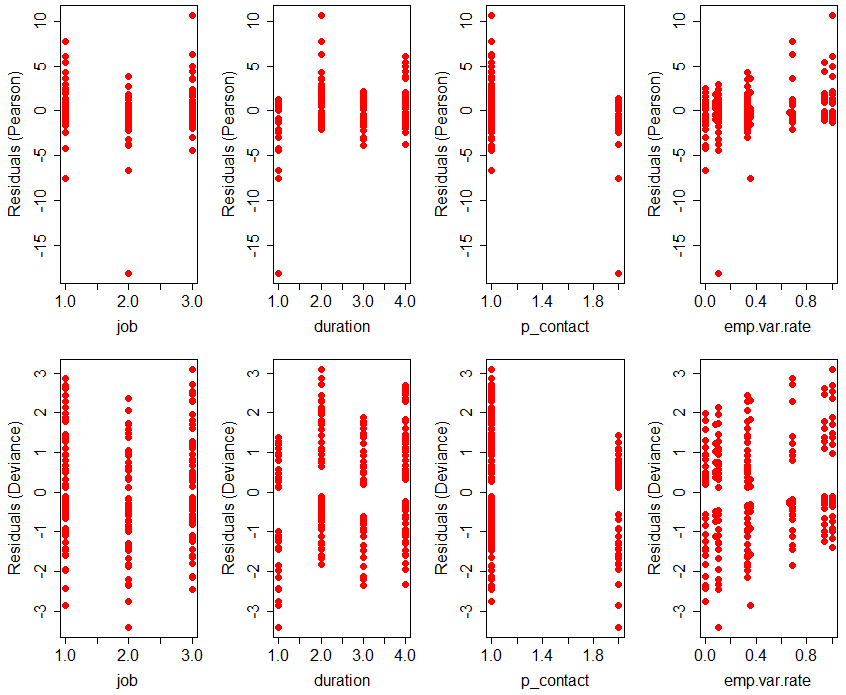
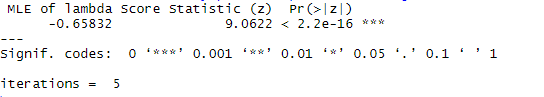


Figure 3: Plot of Pearson and Deviance residuals against the explanatory variables of the model

Table 1: Box-Tidwell test



We see that we reject the null hypothesis (a=.05>p-value). As a result, the interaction is significant, so the linearity assumption cannot be assumed to be true.

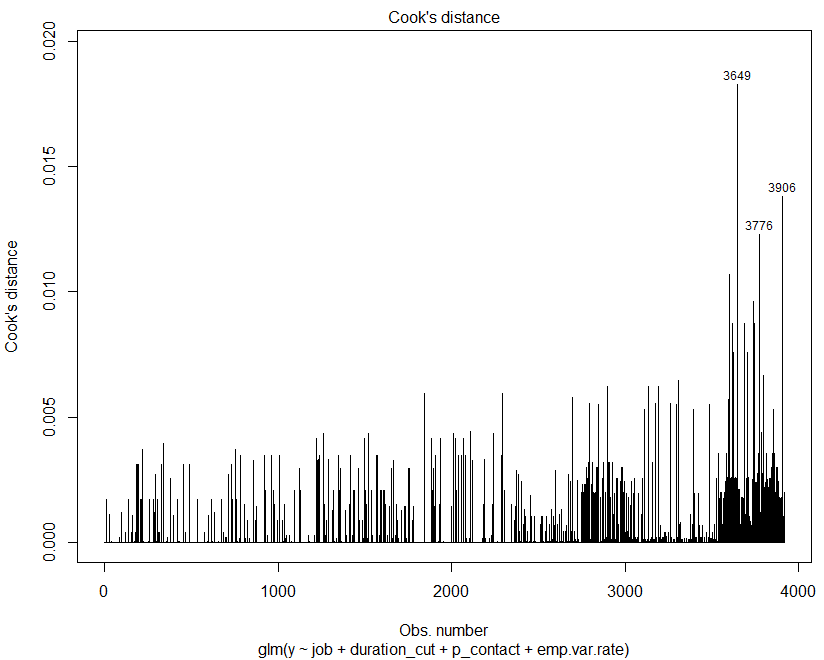


Figure 4: Cook’s distance plot for examining for influential values

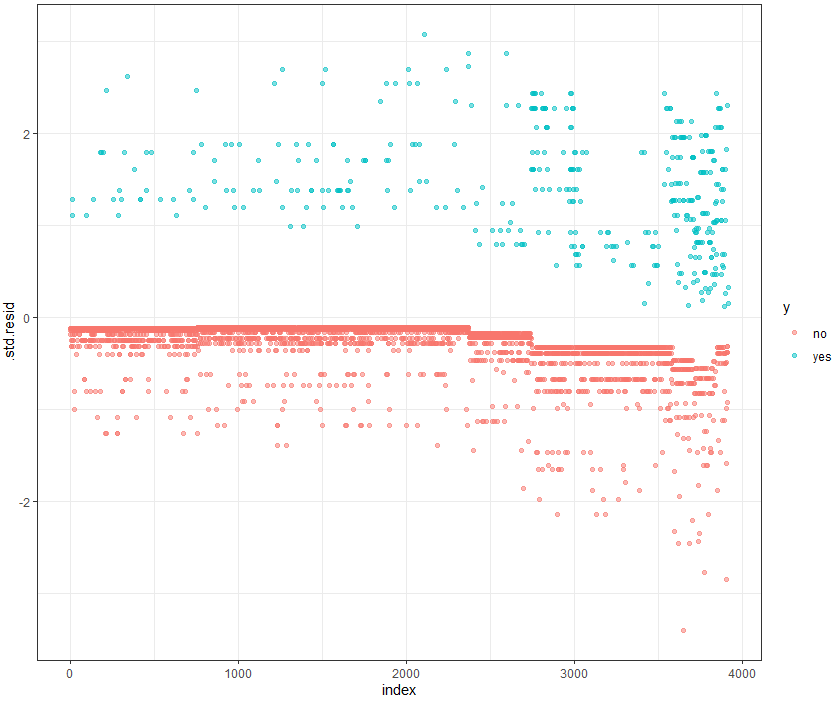


Figure 5: Plot of the Standardized residuals

The “yes” values (coded value=1) of the dependent variable are really spread out in the plot in contrast with the “no” values. That means that our model could only predict well the “no” values.

1. <https://en.wikipedia.org/wiki/Separation_(statistics)> [↑](#footnote-ref-1)
2. A saturated model is one in which there are as many estimated parameters as data points. By definition, this will lead to a perfect fit, but will be of little use statistically, as we have no data left to estimate variance. [↑](#footnote-ref-2)
3. Degrees of freedom [↑](#footnote-ref-3)
4. Degrees of Freedom [↑](#footnote-ref-4)
5. See Figure 1 in APPENDIX [↑](#footnote-ref-5)
6. See Figure 2 in APPENDIX [↑](#footnote-ref-6)
7. See Figure 3 in APPENDIX [↑](#footnote-ref-7)
8. GVIF stands for Generalized Variance Inflation Factor [↑](#footnote-ref-8)
9. See Table 1 in APPENDIX [↑](#footnote-ref-9)
10. See Figure 4 in APPENDIX [↑](#footnote-ref-10)
11. <https://machinelearningmastery.com/what-is-imbalanced-classification/> [↑](#footnote-ref-11)
12. See Figure 5 in APPENDIX [↑](#footnote-ref-12)