STATISTICS FOR BUSINESS ANALYTICS II

Project I 2021-2022

Name: Ioannis Dekoulakos

A.M.: P2822110

Professor: Karlis Dimitrios

Academic Year 2021 - 2022

Part time program
2nd Quarter

Table of Contents

1. INTRODUCTION – DESCRIPTION OF THE PROBLEM

Marketing campaigns basically constitute a technique of outsourcing by organizations with the goal of improving the financial posture of their businesses. Telemarketing is way of straight marketing where a dealer come up to the buyer directly or through telephonic calls and influences them to purchase the products. Nowadays, telephone has been broadly used. It is cost effective and keeps the customers up to date.

In Banking sector, marketing is the backbone to sell its product or service. Banking advertising and marketing is mostly based on an intensive knowledge of objective information about the market and the actual client needs for the bank profitable manner. Banking advertising and marketing is mostly based on an intensive knowledge of objective information about the market and the actual client needs for the bank profitable manner.

Screening of targeted customers for telemarketing that are more likely to subscribe products will reduce the cost of marketing. Using available information and customer metrics, it is possible to build and establish automated protocols for selecting customers in advance. Such a protocol allows one to reduce the time and costs of campaigns and performing fewer and more effective phone calls will diminish client stress and intrusiveness.

The scope of this project is to find a model (using logistic regression) that can explain success of a contact. Such a model can increase campaign efficiently by identifying the main characteristics that affect success, statistical modeling could be a useful tool for a better management of the available resources and support decision making. So, the primary focus will be understanding as better as possible the parameters and characteristics of customers that affect the campaign and not prediction of any results.

For this purpose, a real dataset collected from one of the retail banks, from May 2008 to June 2010, in a total of near 40K phone contacts. These campaigns consist of information on phone calls where people were asked if they want to subscribe to a bank term deposit. The phone call is considered a success if a term deposit is made.

The dataset encompasses 4 main groups of features:

- Demographic Information of the clients — age, job, marital, education, default, housing, balance, loan
- Time Characteristics of the Call — day, month, duration
- Characteristics of the Campaign — contact, campaign, pdays, previous, poutcome
- Social and Economic context attributes — emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed

SUBSCRIBED (target) — Indicates whether client has subscribed to a term deposit

2. Descriptive analysis and exploratory data analysis

2.1 APPROPRIATE DATA TRANSFORMATIONS AND CLEANING

Importing the records of telemarketing phones call created an R data frame which has 21 attributes and 39883 observations (columns & rows as well).

Upon importing the dataset, actions regarding changes in data types of variables, exclusion of specific columns without valuable information, as well as merging some levels of categorical variables were considered. All performed transformations will be analyzed below.

Some variables that were given as numeric may have been more useful as categorical variables. For instance, it makes more sense to see how *age* might influence the decision to subscribe across a few broad categories (more details with a histogram can be founded in Appendix 6, Figure 6.1). Additionally, the variable education will be merged into 4 levels and (illiterate, basic, intermediate, advance) as well.

According to the variable *pdays* more than 75% of the data set wasn't contacted previously by another campaign (more details with a histogram can be founded in Appendix 6, Figure 6.2), 999 means no previous contact, so it will be removed because it offers the same information as the variable *poutcome*, it does not offer any extra valuable information (more details with a barplots can be founded in Appendix 6, Figure 6.3).

Table 1. List of the attributes from the dataset after cleaning

variable	*Description*	*Type*
age	Age of the client	Factor
job	Type of job	Factor
marital	marital status	Factor
education	education of the client	Factor
default	has credit or default?	Factor
housing	Indicates whether client has a housing loan	Factor
loan	Indicates whether client has a personal loan	Factor
contact	Type of call contact communication	Factor
season	Season of call	Factor
day_of_week	day of call	Factor
duration	duration of call (sec)	num
campaign	Number of contacts made during current campaign for this client	num
poutcome	outcome of the previous marketing campaign	Factor

emp.var.rate	Employment variation rate	num
cons.price.idx	Consumer price index	num
cons.conf.idx	Consumer confidence index	num
euribor3m	Euribor 3 month interest rate	num
nr.employed	Number of employed citizens in Country	num
previous	number of contacts performed before this campaign	num
SUBSCRIBED	has the client suscribed the term deposit?	Factor

2.2 PAIRWISE COMPARISONS

2.2.1 Numeric Variables

Relationship between the numeric variables

Having studied each variable separately the next step in exploratory analysis is to detect possible linear relationships between pairs of independent variables, as well as between independent variables and the dependent variable (SUBSCRIBED). For this purpose, the following correlation matrix has been created.

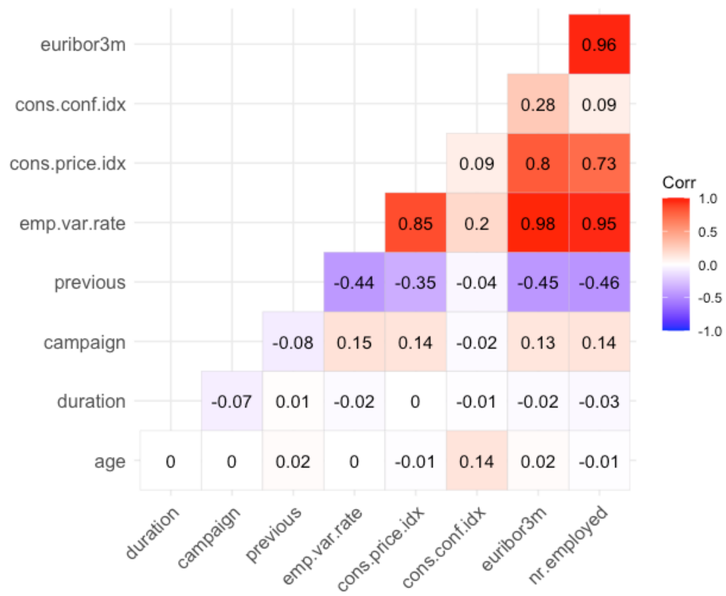

Figure 1 Correlation matrix of numeric variables

The Euribor 3-month interest rate correlated perfectly with Employment variation rate and Number of employed citizens in Country. Additionally, the consumer price index correlated strong with Euribor 3-month interest rate and Number of employed citizens in Country as well. So, only one of these correlated variables will enter in the model for the interpretation of successfully call as they offer similar information.

Relationship between the numeric variables ~ Subrcribed

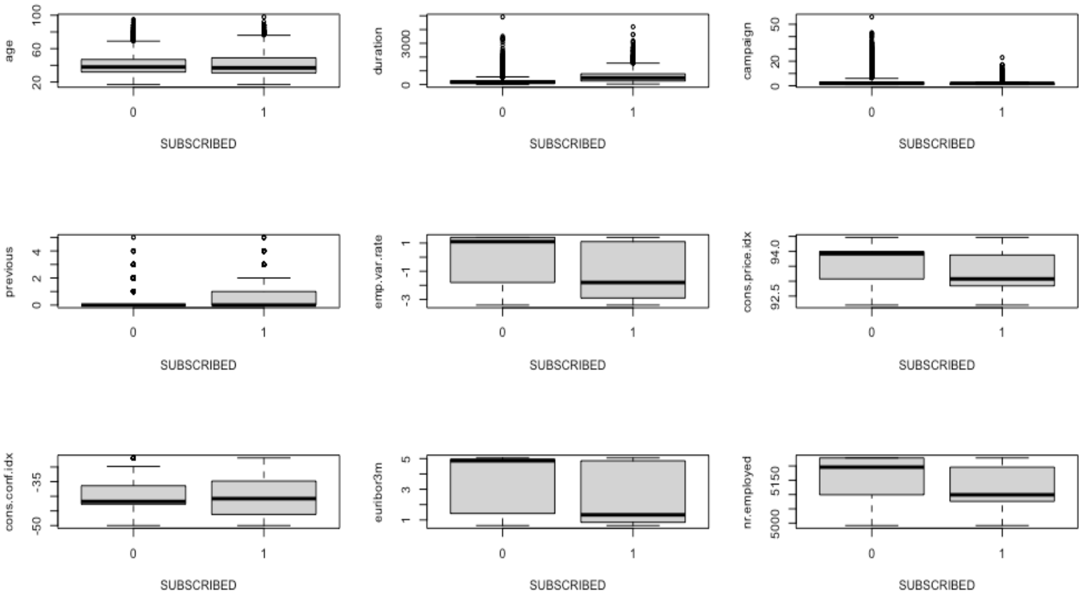

Figure 2. Boxplots Numeric ~ Subscribed

It seems that the relationship between the numeric variables and Subscribed are statistically significant, besides the Campaign and Consumer confidence index. When the rates are low, the clients make a subscription easier.

There is no difference in the median Consumer confidence index & Campaign of clients between those who subscribed and those who doesn’t.

Some describe information about the numeric variables can be found in Appendix 6 (Table 1)

2.2.4 Categorical Variables

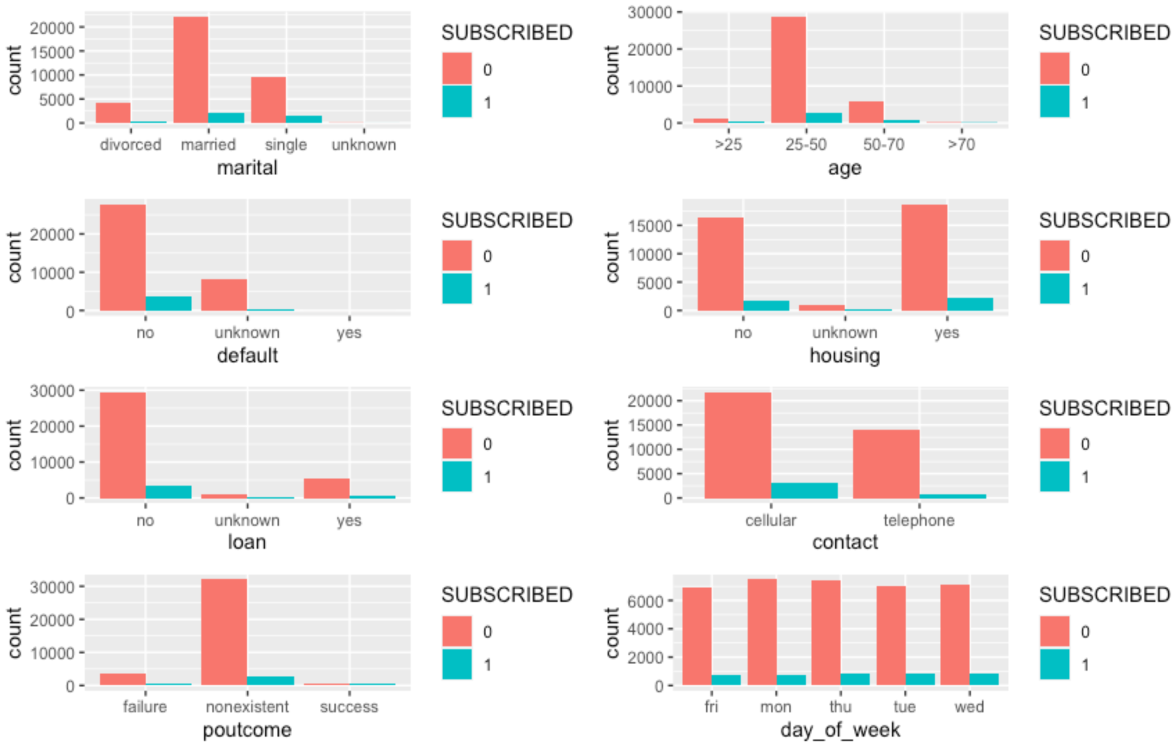

Figure 3.1 Bar charts for Categorical~ Subscribed

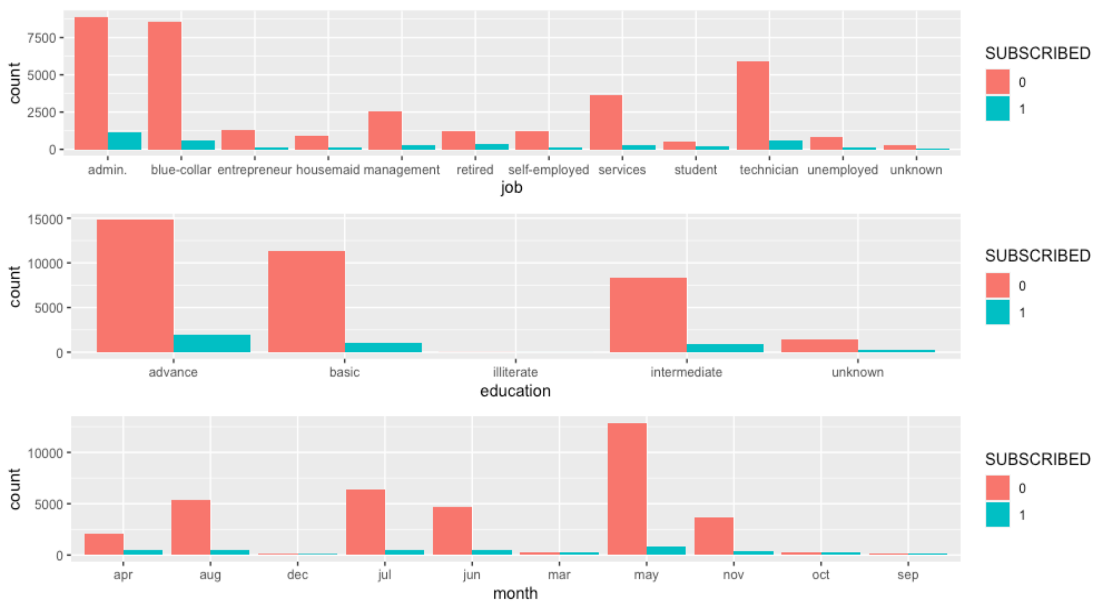

Figure 3.2 Bar charts for Categorical~ Subscribed

The audiences of these campaigns target mostly:

- Age: at the range of 25-50 years old. So, it is logical to mentioned more successful call in this range

- Marital status: Most of the target group are married. Specifically, the married clients are twice as single people. But, in proportion to the size of the sample in each category, single people seem to have more subscriptions (proportional)

- Job: Most clients are admins, blue collars & technicians. So, the most subscriptions appear in these levels

- Education: Most people have an advance education level while illiterate people are very less. It is logical the most subscription appear in this level.

- Default: Most people have no default stay on their credit history. Same the subscriptions

- Loan: Most people have no personal loan. Same the subscriptions

- Housing: Most of the people has a housing loan. Same the subscriptions

- Contact: The common way of communication is cellular. Same the subscriptions

- Month: May is the most activated month of the year and December is the least activated month, because of the holidays season. Same the subscriptions

- Day of week: Monday & Thursday are the busiest day of week but there isn't any significant difference between the days of week, maybe this variable will not included in the model

- Poutcome: Most clients do not have any previous contact. It seems that when a client has a previous contact make a subscription. ('Yes' is bigger in this level)

- SUBSCRIBED: only 3987 calls were successfully

General, the categorical variables, besides the day of week which it is expected not to be used in the final model, can affect the result of a successful call. For instance, the job of a user can influence the result of the subscripted because they obviously earn different salaries. Intuitively, a user who has administration job has a high tendency of subscribing to the next campaign than an unemployed person. Also, if there is a previous contact can convent the client easier to descripted.

3. MODEL SELECTION

3.1 USE OF STEPWISE PROCEDURES ONLY

Variable selection techniques should be used to select the most important features for the model. Two approaches have been implemented (a) use of the stepwise procedures only (AIC & BIC), (b) Variable screening with LASSO prior to the stepwise procedures. Finally, some Wald Tests will be used to confirm if some explanatory variables in a model are significant or not.

At first, the stepwise procedure with AIC method generated a model ('model_aic') using 15 of 20 input variables (dropping the *variables job, housing, loan, previous, nr.employed*) , with residual deviance: 15619 on 39848 degrees of freedom and AIC: 15689. However, there is strong multicollinearity in the independent variables, it is expected due to the strong correlation between *euribor3m,* and other numeric variables as mentioned in the section 2.2.4 (Figure 1).

So, after removing the variables *emp.var.rate* and *Euribor* generated a new model ('**model_aic3'**) with residual deviance: 16340 on 39850 degrees of freedom & AIC: 16406 (The summary of the model can be found at Appendix Figure 6.4). It seems that there are some not statistically significant variables.

So, the next step was to repeat the process using stepwise selection with penalty = log(n) or BIC, generated a model ('model_bic') using 10 of 20 input variables. However, there is multicollinearity in the independent variables so after removing the correlated variables generated a new model ('**model_bic3'**) included the following independent variables*: default, contact, month, duration, poutcome, cons.conf.idx euribor3m* with Residual deviance: 16105 on 39865 degrees of freedom & AIC: 16141. (The summary of the model can be found at Appendix Figure 6.5).

3.2 MODEL WITH LASSO SCREENING

In this implementation, LASSO approach was used to perform an initial screening of the input variables (Figure 4). Using lambda.1se as the λ value, the method provided the following independent variables (*age, job, marital, education, default, housing, contact, month, day_of_week, duration, campaign poutcome, emp.var.rate, cons.price.idx, nr.employed*). These variables were used as the starting point (the "full" model) of the stepwise procedures.

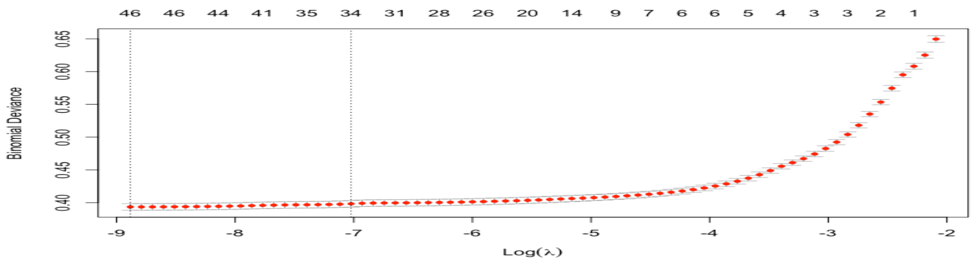

Figure 4 Output of Lasso. Variable screening with lambda.min & lambda.1se)

In this model ('model_lasso'), there is multicollinearity, and a lot of variables are not statistically significant.

The next model was generated directly from the implementation of the stepwise procedure (AIC) on the 15 variables selected by LASSO. After, removing the variable *emp.var.rate* due to the multicollinearity reasons the model (**'model_lasso_aic'**) has the following independent variables: *age, marital, education, default, contact, month, day_of_week , duration, campaign, poutcome, cons.price.idx, nr.employed* with Residual deviance: 15893 on 39850 degrees of freedom & AIC: 15959. (The summary of the model can be found at Appendix Figure 6.6).

Furthmore, using stepwise selection with penalty = log(n) on the variables selected by LASSO as well generated a new model but it seems that there is multicollinearity again. So, after removing the variable *emp.var.rate* the new model ('model_lasso_bic') contains the independent variables: *default, contact, month, duration, campaign, poutcome, cons.price.idx, nr.employed* with Residual deviance: 15978 on 39864 degrees of freedom & AIC: 16016.

Finally, checking with Wald tests the variables which were left out by BIC while lasso and AIC included them, we found that the variables *age, education, cons.conf.idx, job* should be added in the final model because they are statistically significant *(Wald test, $P < 0.05$)* for the interpretation of a success call. Also, it seems that the variables *cons.price.idx & contact* are not statistically significant and does not offer any valuable information, so it will be removed.

As expected, same decisions about the variables' selections confirmed with Likelihood ratio test. (The code & output can be found in R file)

The final model (**'model_lasso_bic7'**) has the independent variables *default, month, duration, campaign, poutcome, nr.employed, age, education, cons.conf.idx, job* with Residual deviance: 15878 on 39847degrees of freedom & AIC: 15950, Likelihood: -7938.799. The best AIC, Residual Deviance, likelihood from all the previous models.

The form of the Model:

$\text{Logit } E(Y_i) = \beta_0 + \beta_1 x_i = \beta' x_i$

$\beta' = (\beta_0, \beta_1), x'_i = (1, x_i).$

So, the most important features for our response (Y_i) variable (subscribed) and the independent variables (x_i) of model as well are : *default, month, duration, campaign, poutcome, nr.employed, age, education, cons.conf.idx, job*

Final Model

	Dependent variable:
	SUBSCRIBED
jobblue-collar	*-0.285*** (0.081)*
jobentrepreneur	*-0.204 (0.127)*
jobhousemaid	*-0.084 (0.152)*
jobmanagement	*-0.097 (0.088)*
jobretired	*0.084 (0.118)*
jobself-employed	*-0.153 (0.120)*
jobservices	*-0.192** (0.089)*
jobstudent	*0.303** (0.126)*
jobtechnician	*-0.094 (0.067)*
jobunemployed	*-0.015 (0.135)*
jobunknown	*-0.069 (0.255)*
cons.conf.idx	*-0.032*** (0.007)*
age25-50	*-0.251** (0.099)*
age50-70	*-0.111 (0.114)*
age> 70	*0.267 (0.192)*
educationbasic	*-0.168** (0.068)*
educationilliterate	*0.915 (0.735)*
educationintermediate	*-0.147** (0.060)*
educationunknown	*-0.025 (0.107)*
defaultunknown	*-0.313*** (0.068)*
defaultyes	*-7.412 (113.207)*
monthaug	*0.972*** (0.138)*
monthdec	*0.376* (0.199)*
monthjul	*0.705*** (0.114)*
monthjun	*0.673*** (0.093)*
monthmar	*1.240*** (0.116)*
monthmay	*-0.625*** (0.075)*
monthnov	*0.276* (0.116)*
monthoct	*0.795*** (0.168)*
monthsep	*0.088 (0.184)*
duration	*0.005*** (0.0001)*
campaign	*-0.048*** (0.012)*
poutcomenonexistent	*0.423*** (0.067)*
poutcomesuccess	*1.772*** (0.097)*
nr.employed	*-0.017*** (0.001)*

Constant	*80.839*** (2.503)*

Residual deviance: 15875 on 39846 degrees of freedom
AIC: 15950
Likelihood: -7938.799

Table 2. Summary of Final Model

For:*jobblue-collar, age25-50, educationbasic, defaultunknown, monthaug, poutcomesuccess*

jobblue-collar

The probability of a client to make a subscription when his job is blue-collar is less than another with an admin Job

cons.conf.idx

As the Consumer confidence index increase the probability of making a subscription decrease.

Age25-50

The probability of a client 25-50 years old to make a subscription is less than another who is 17-24 years old

Educationbasic

The probability of a client to make a subscription when his education lever is basic is less than another with university education level.

Defaultunknown

The probability of a client to make a subscription when it is not known if he has credit is less than another who hasn't any credit for sure.

Monthaug

The probability of a client to make a subscription on Aug is higher than the probability to make in April

Duration

As the seconds of call increase the probability of making a subscription increase as well.

Campaign

As Number of contacts made during current campaign for a client increase the probability of making a subscription decrease.

Poutcomesuccess

The probability of a client to make a subscription when his outcome of the previous campaign is success is higher than another whose is failure.

nr.employed

As Number of employed citizens in Country increase the probability for a client to make a subscription decrease.

For Instance,

$P_{SUBSCRIBED} =$

*exp (80.839- -0.285--0.032*cons.conf.idx - 0.251*

–0.168 - 0.313- 0.972+

*0.005*duration - 0.046* campaign + 0.429 - 0,016* nr.employed)*/

1+*((79.05 - 0.283 -0.026*cons.conf.idx - 0.251*

–0.168 -0.314 - 0.112 + 0.870*

*+ 0.005*duration - 0.048* campaign + 1.772 - 0,017* nr.employed)*

4. Goodness of fit

The residual deviance of the final model (15878) and the df (39847) indicates that the model fits well the data.

Specifically, there is not any evidence to reject Null Hypothesis that the model is not (1-pchisq (res.dev, df), *p-value = 1, >5%*).

It is expected since that a Bernoulli Model has only two possibilities 0 or 1 with probability p and (1-p) as well. So, a model with 10 independent variables fits well. McFaden R2: 0.387 is very close to 0.4 so model seems to has a very good fit of the data.

The model differs significantly from the null (pchisq (null. Deviance - deviance, df.null - df.residual), =0) , tell us that the final model as a whole fit significantly better than the null model.

In comparison, the final model has a better AIC, Residual Deviance than the rest model so it has a better fit as well. Finally, likelihood tests provide p-value less than 0.05 we can reject the null hypothesis and assume that the this model has a statistically significant improvement in terms of goodness of fit.)

5. ASSUMPTIONS

There are the following Assumptions:

-Linearity Assumption

For this assumption we should check the linear relationship between continuous predictor variables and the logit of the outcome (scatterplots)

-Influential Values

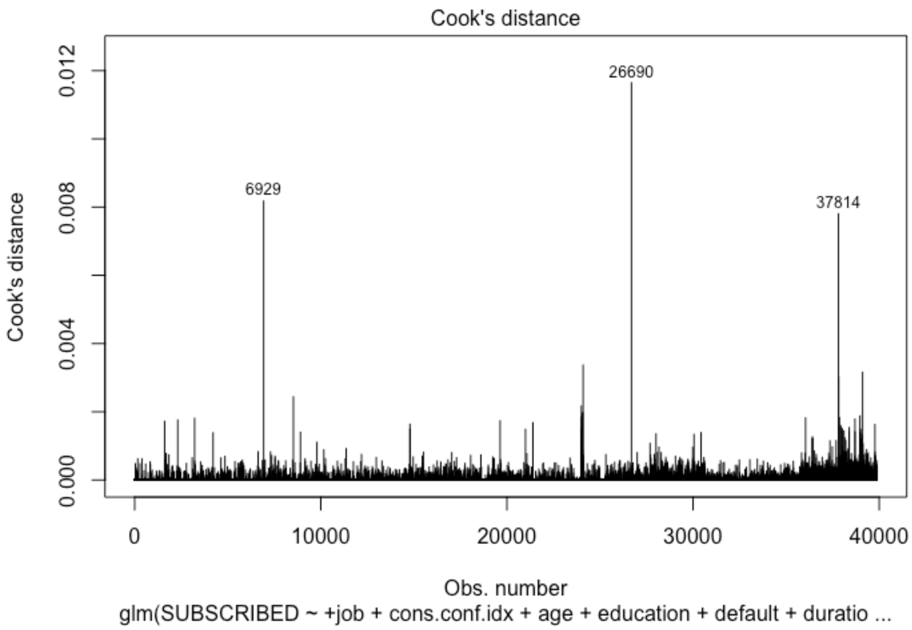

Figure 5. The most extreme values in the data

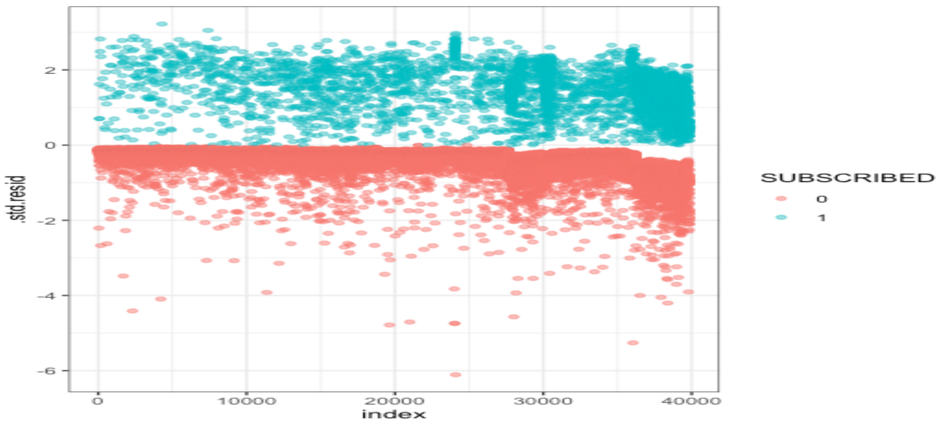

Figure 5. There is Independence of residuals (The points are randomly scattered without any pattern, around zero)

So, Applying the filter abs(std.res) > 3 seems that there are very few data points with an absolute standardized residuals above 3 so maybe indicate possible outliers but they are very little compared to the dataset, in any case deserve closer attention. Maybe, they need to be removed or transformed into log.

.

-Multicollinearity

This assumption has already check it at section 3. There is not multicollinearity in the final model.

- Target Variable is binary 0-1

5. CONCLUSIONS

The main goal of this project was to identify the variables which affect sell long-term deposits in a telemarketing campaign ending up in a model that can explain success of a contact. Demographic Information of the clients, time characteristics of the call as well as social and economic context attributes affect the subscription. More specifically it was found that default, month, duration, campaign, poutcome, number of employees, age, education, consumer confidence index and *job* significantly affect the behaviour of people when it comes to make a long-term deposit.

The final model fits very well in the data, and it is the best model which presented in this report. General, it is a good model with simple interpretation.

For future work it would be important to look for other parameters that affect sell long-term deposits or make some other interactions between independent variables, improving the model fit.

REFERENCES

- *ALAN AGRESI (2015). Foundations of Linear and Generalized Linear Models. (E-book)*

- *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2017 edition), An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics). (E-book)*

- *Course slides, STATISTICS FOR BUSINESS ANALYTICS II 2022, (aueb)*

6. APPENDIX

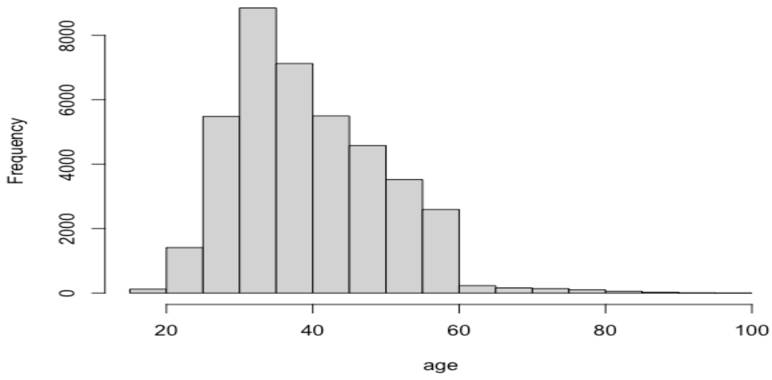

figure 6.1 Histogram for Variable -Age

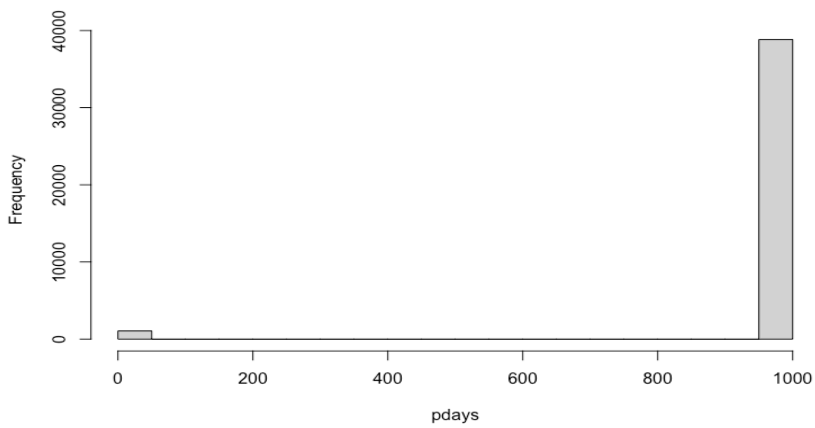

figure 6.2 Histogram for Variable -Pdays

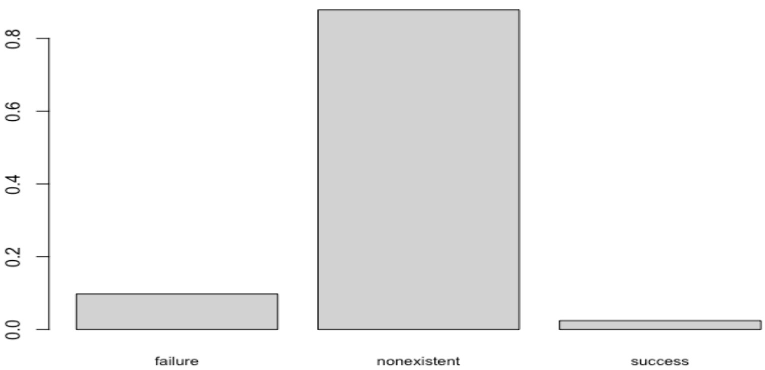

figure 6.3 Bar plot for Variable -Poutcome

variables	*n*	*mean*	*sd*	*median*	*trimmed*	*mad*	*min*	*max*	*range*	*skew*	*kurtosis*	*se*
duration	39883	256.70	258.84	177.00	208.78	136.40	0.00	4918.00	4918.00	3.26	20.04	1.30
campaign	39883	2.59	2.80	2.00	2.01	1.48	1.00	56.00	55.00	4.73	36.31	0.01
emp.var.rate	39883	0.13	1.57	1.10	0.33	0.44	-3.40	1.40	4.80	-0.81	-0.94	0.01
cons.price.idx	39883	93.55	0.57	93.44	93.56	0.82	92.20	94.47	2.26	-0.21	-0.84	0.00
cons.conf.idx	39883	-40.46	4.61	-41.80	-40.60	6.52	-50.00	-26.90	23.10	0.36	-0.40	0.02
euribor3m	39883	3.71	1.69	4.86	3.91	0.16	0.63	5.04	4.41	-0.81	-1.24	0.01
nr.employed	39883	5173.22	64.63	5191.00	5182.16	55.00	4991.60	5228.10	236.50	-0.96	-0.33	0.32

Table 2 Description Analysis for Numeric Variables

model_aic3

	Dependent variable:
	SUBSCRIBED
age25-50	-0.405^{***} (0.091)
age50-70	-0.147 (0.105)
age> 70	0.648^{***} (0.163)
maritalmarried	0.031 (0.071)
maritalsingle	0.196^{**} (0.078)
maritalunknown	0.132 (0.410)
educationbasic	-0.279^{***} (0.055)
educationilliterate	0.841 (0.755)
educationintermediate	-0.151^{***} (0.054)
educationunknown	0.023 (0.101)
defaultunknown	-0.426^{***} (0.067)
defaultyes	-7.648 (113.418)
contacttelephone	-0.733^{***} (0.079)
monthaug	-2.056^{***} (0.110)
monthdec	-0.725^{***} (0.203)
monthjul	-1.011^{***} (0.094)
monthjun	0.275^{***} (0.090)
monthmar	1.331^{***} (0.116)
monthmay	-1.181^{***} (0.075)
monthnov	-1.702^{***} (0.100)

monthoct	-1.323*** (0.170)
monthsep	-1.736*** (0.189)
day_of_weekmon	-0.067 (0.068)
day_of_weekthu	0.023 (0.067)
day_of_weektue	0.098 (0.068)
day_of_weekwed	0.151** (0.068)
duration	0.005*** (0.0001)
campaign	-0.041*** (0.012)
poutcomenonexistent	0.255*** (0.066)
poutcomesuccess	2.063*** (0.098)
cons.price.idx	-1.240*** (0.055)
cons.conf.idx	0.098*** (0.006)
Constant	117.213*** (5.120)
Null deviance: 25925 on 39882 degrees of freedom Residual deviance: 16340 on 39850 degrees of freedom AIC: 16406	
Log Likelihood -8,169.853	

Figure 6.4 Summary of Model_aic3

Model_BIC3

	Dependent variable:
	SUBSCRIBED
defaultunknown	-0.391*** (0.066)
defaultyes	-7.363 (113.357)
contacttelephone	-0.283*** (0.076)
monthaug	-0.085 (0.123)
monthdec	0.183 (0.199)
monthjul	0.171 (0.107)
monthjun	0.464*** (0.091)
monthmar	1.525*** (0.117)
monthmay	-0.803*** (0.073)
monthnov	-0.163 (0.111)
monthoct	0.214 (0.166)
monthsep	-0.198 (0.183)
duration	0.005*** (0.0001)
poutcomenonexistent	0.427*** (0.066)
poutcomesuccess	1.876*** (0.096)
cons.conf.idx	0.053*** (0.006)
euribor3m	-0.592*** (0.019)
Constant	-0.066 (0.300)
Null deviance: 25925 on 39882 degrees of freedom Residual deviance: 16105 on 39865 degrees of freedom AIC: 16141	

Figure 6.5 Summary of Model_Bic3

model_lasso_aic

	Dependent variable:
	SUBSCRIBED
age25-50	-0.309*** (0.092)
age50-70	-0.098 (0.106)
age> 70	0.402** (0.163)
maritalmarried	0.007 (0.072)
maritalsingle	0.153* (0.079)
maritalunknown	0.183 (0.413)
educationbasic	-0.249*** (0.055)
educationilliterate	0.858 (0.734)
educationintermediate	-0.145*** (0.055)
educationunknown	-0.040 (0.104)
defaultunknown	-0.327*** (0.067)
defaultyes	-7.431 (112.991)
contacttelephone	-0.238*** (0.068)
monthaug	0.500*** (0.098)
monthdec	0.117 (0.187)
monthjul	0.409*** (0.096)
monthjun	0.572*** (0.090)
monthmar	1.202*** (0.118)
monthmay	-0.722*** (0.074)
monthnov	-0.039 (0.098)
monthoct	0.305** (0.127)
monthsep	-0.262 (0.162)
day_of_weekmon	-0.105 (0.069)
day_of_weekthu	0.025 (0.067)
day_of_weektue	0.074 (0.068)
day_of_weekwed	0.140** (0.068)
duration	0.005*** (0.0001)
campaign	-0.043*** (0.012)
poutcomenonexistent	0.434*** (0.067)
poutcomesuccess	1.765*** (0.098)
cons.price.idx	-0.037 (0.069)
nr.employed	-0.015*** (0.001)
Constant	78.938*** (5.117)

Null deviance: 25925 on 39882 degrees of freedom
Residual deviance: 15893 on 39850 degrees of freedom
AIC: 15959
Log Likelihood: -7,946.544

Figure 6.6 Summary of model_lasso_aic