Athens University of Economics and Business
Department of Management Science and Technology

Statistics for BA II
Professor: D. Karlis

Project I
Telemarketing Dataset

Marianna Konstantopoulou
A.M: P2822122

M.Sc. Business Analytics
Part Time 2021-2023

Athens, 16/02/2022

Table of Contents

Chapter 1

Introduction
- description of the problem, data, aim, background information

This information is about telemarketing phone calls to sell long-term deposits. During a campaign, the agents call a list of customers to promote the product (outbound) or, if the client phones the contact-center for any other purpose, he is urged to subscribe to the product (inbound). As a consequence, the outcome is a binary variable indicating whether the interaction was successful or failed.

This study takes into account real data gathered from a retail bank from May 2008 to June 2010, in a total of about 40K phone interactions. More than one contact with the same consumer was frequently necessary to determine if the product (bank term deposit) will be subscribed to ('yes') or not ('no').

The aim of our project was to determine which elements lead to a successful interaction (the client subscribes to the product). Our dataset contains variables about the bank client data such as his age, education, marital and job status, his housing and personal loans, as well as details about the type and frequency of contact (about current or previous campaigns). Important indexes and rates are also included (employment variation rate, employment variation rate, employment variation rate, employment variation rate and the number or employees).

Chapter 2

Descriptive analysis and exploratory data analysis

To start our analysis, we will load our data in R and check the data types. We will remove a column that is not useful for our research: pdays (we can extract the same information from the "previous" variable so there is no need to keep both). We need to change the rest of the variables' data types to the correct type. Age of the client (age), duration of last contact (duration), the number of contacts performed during this campaign (campaign), the number of contacts performed before this campaign (previous) and the indexes/rates employment variation rate, consumer price index, consumer confidence index, euribor 3 month rate (emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m) as well as the number of employees (nr.employed) are all numeric. We also have a number of categorical variables: job, marital status, education, default credit, housing and personal loan, contact communication type, last contact month and day of the year, day of the week, outcome of previous marketing campaign and of course our response variable whether the client subscribed a term deposit or not (i.e. Table 1). It is also important to perform a check for potential missing values. In our case there are no missing values in our data set.

Checking the describe output (i.e. Table 2) for our numeric variables we can make some quick observations:

- The average age is approximately 40 while the minimum and maximum ages are 17 and 98
- The duration has an average of approximately 257 seconds peaking at 4918 seconds and the average number of contacts performed during this campaign is approximately 3
- The number of contacts performed before this campaign and for this client has a maximum of 5 contacts and a minimum 0 (which means that some clients haven't been contacted before)
- The average employment variation rate (quarterly) is 0.13 with a maximum 1.4 and a minimum of -3.4
- The monthly consumer price index and consumer confidence index have an average of 93.55 and -40.46 respectively while the daily mean Euribor 3 month rate is 3.71

- The average number of employees is approximately 5173 with a maximum of 5228 and a minimum of 4992

Checking the bar plots (i.e. Figure 1) for our categorical variables we can make the following observations:

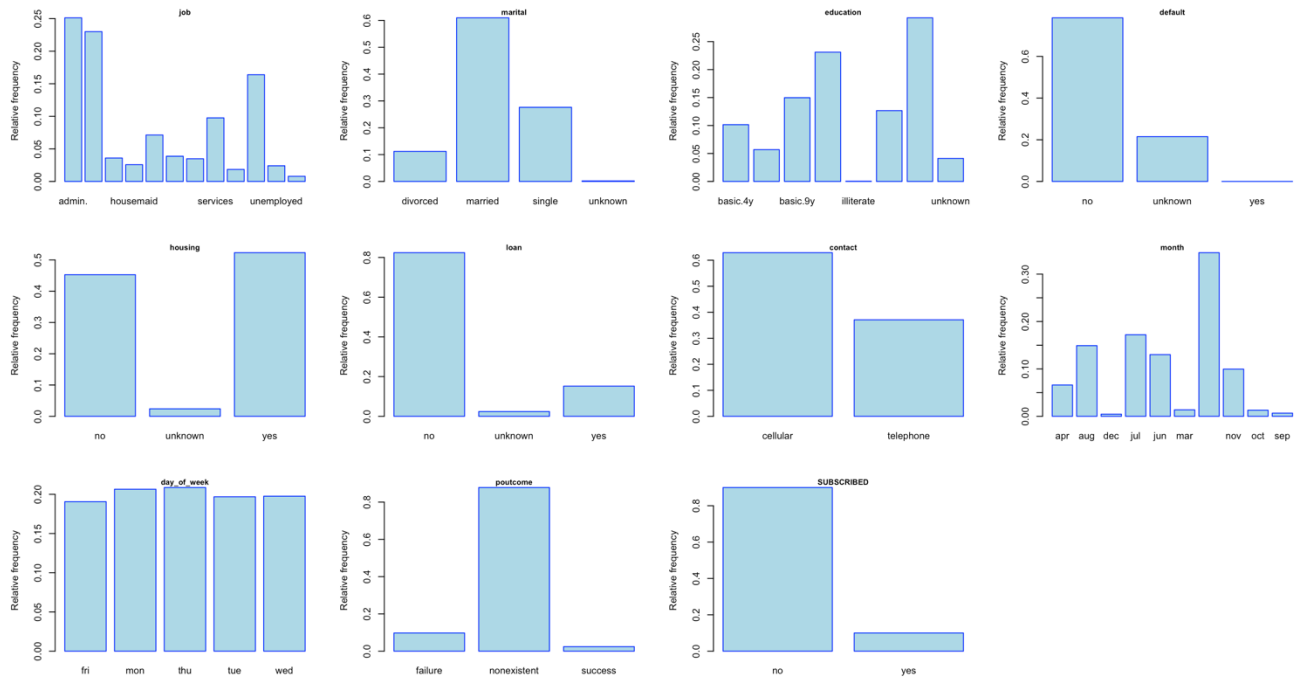
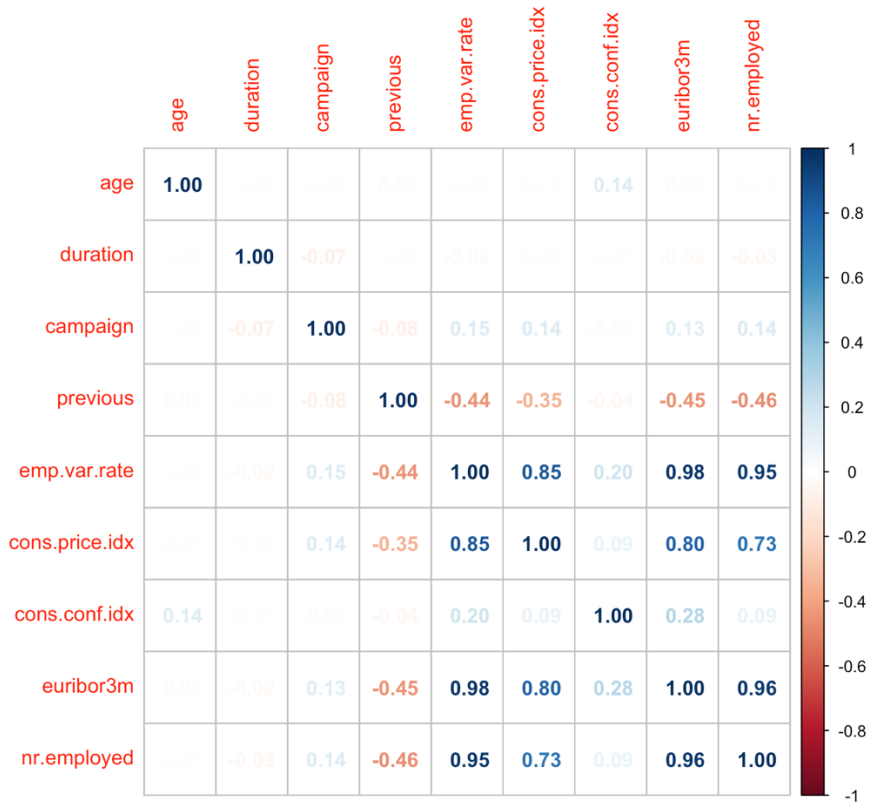


Figure 1: Bar plots for factor variables

- The job with most observations is admin. and most observations' marital status is married
- The education with the most observations is university degree, and most people have no default credit
- The majority of observations has a housing loan but no personal loan
- Most contacts are performed through cellular, while the month and day of the week with the most observations are May and Thursday
- Most outcomes of previous marketing campaigns are nonexistent which means that most clients are contacted for the first time and the most client have not subscribed a term deposit

After studying our variables separately, the next step would be to study their relationships.



	age	duration	campaign	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	1.00						0.14		
duration		1.00	-0.07						
campaign		-0.07	1.00	-0.08	0.15	0.14		0.13	0.14
previous			-0.08	1.00	-0.44	-0.35		-0.45	-0.46
emp.var.rate			0.15	-0.44	1.00	0.85	0.20	0.98	0.95
cons.price.idx			0.14	-0.35	0.85	1.00		0.80	0.73
cons.conf.idx	0.14				0.20		1.00	0.28	
euribor3m			0.13	-0.45	0.98	0.80	0.28	1.00	0.96
nr.employed			0.14	-0.46	0.95	0.73		0.96	1.00

Figure 2: Correlation plot for numeric variables

We use the correlation plot to check the correlation between the pairs of numeric variables. (i.e. Figure 2). Employment variation rate has a strong connection with consumer price index, euribor 3 month rate, number of employees of 0.85, 0.98 and 0.95 respectively. Euribor 3 month rate has also strong connection with number of employees (0.96) and consumer price index (0.80). Lastly, there is a rather strong connection of number of employees with consumer price index (0.73), while the rest of numeric variables have a rather weak correlation. The Pearson's correlation coefficient for these connections which is greater than 0.7 confirms a strong positive linear relationship and indicates a possible multicollinearity issue between these variables.

Chapter 3

Descriptive models

Since we studied the relationships between our variables it's time to try and construct some models so that we can create a model for finding which variables contribute to a successful contact.

We will construct the full model (i.e. Table 3) using logistic regression since our Y response variable can only take 2 possible values (0: unsuccessful and 1: successful contact). This means that:

$$Y_i \sim B(1, p_i) \text{ independent for } i=1,\ldots.39883,$$

$$p_i = P(Y_i = 1): \text{successful contact}$$

In order to implement Logistic Regression, we will need the following transformation as link function, which is the logit function:

$$\log \frac{pi}{1-pi} = \beta_0 + \beta_1 x_i$$

We will conduct LASSO as a variable selection technique. We use cross validation and select the largest value of lambda such that error is within 1 standard error of the minimum (i.e. Figure 3). Using the lambda we selected, we receive the estimated coefficients under the lambda.1se. Our model has subscribed as a response and as predictors all variables and intercept except for age, previous, consumer confidence index and euribor 3 month index. (i.e. Table 4)

To select our final model, we are using stepwise methods as well. Since we are focusing on inference, we used the Stepwise procedure according to BIC. Our final model has subscribed as response and the predictors are default, contact, month, duration, campaign, poutcome, emp. var. rate, cons. price idx and nr employed (i.e. Table 5).

Next step would be to check for multi-collinearity. We used the GVIF test and decided to remove variable emp. var. rate since it caused multicollinearity issues to our model. (i.e. Table 6). In addition to this test, we decided to check the existence of the rest of the variables and we figured out (using Wald test, p-value=.92>.05) (i.e. Table 8) that the variable consumer price index can be removed from our model.

This is the final model (i.e. Table 7) we selected:

logit (*subscribed*) = 79.4 – 0.37 * *defaultuknown* – 7.45 * *defaultyes* – 0.23 * *contacttelephone* + 0.63 * *monthaug* + 0.19 * *monthdec* + 0.49 * *monthjul* + 0.60 * *monthjun* + 1.25 * *monthmar* – 0.73 * *monthmay* + 0.039 * *monthnov* + 0.38 * *monthoct* – 0.18 * *monthsep* + 0.004 * *duration* – 0.04 * *campaign* + 0.44 * *poutcomenonexistent* + 1.77 * *poutcomesuccess* – 0.01 * *nr.employed*

We will start to interpret our final model:

- Since 0 is out of range for some of our numeric variables we will need to center our covariates so that we can interpret the intercept (i.e. Table 10). After centering our intercept means that the odds of successful subscriptions when there is no default credit, the contact is cellular, month is April, outcome of previous marketing campaign is a failure and average duration is 256 seconds, average number of contacts performed during this campaign is 3 and average index of number of employees are about 5173 is equal to 0.029
- 1 unit increase in duration means that the actual odds of successful subscriptions are multiplied by 1, assuming that all other variables are constant
- 1 unit increase in number of contacts performed during this campaign and for this client means that the actual odds of successful subscriptions are multiplied by 0.96, assuming that all other variables are constant
- 1 unit increase in number of employees means that the actual odds of successful subscriptions are multiplied by 0.99, assuming that all other variables are constant
- for an unknown default credit, the actual odds of successful subscriptions are multiplied by 0.69 when all numeric variables are constant and the contact is cellular, month is April, outcome of previous marketing campaign is a failure.
- for default credit the actual odds of successful subscriptions are multiplied by 0.0005 when all numeric variables are constant and the contact is cellular, month is April, outcome of previous marketing campaign is a failure.
- for telephone contact the actual odds of successful subscriptions are multiplied by 0.79 when all numeric variables are constant and there is no default credit, month is April, outcome of previous marketing campaign is a failure.
- from month April to month August the actual odds of successful subscriptions are multiplied by 1.87 when all numeric variables are constant and there is no default credit,

the contact is cellular, outcome of previous marketing campaign is a failure. We would follow the same procedure to interpret the rest of the months.

- for no previous outcome of the previous campaign (so it's the first contact with the client) the actual odds of successful subscriptions are multiplied by 1.55 when all numeric variables are constant and there is no default credit, the contact is cellular and month April.
- for successful previous outcome of the previous campaign the actual odds of successful subscriptions are multiplied by 5.87 when all numeric variables are constant and there is no default credit, the contact is cellular and month April.

As a next step we will perform Goodness of fits tests to check if our model fits as well as the "saturated" model. Our test result proved that we fail to reject the null hypothesis that the model fits "well". One more indicator could be the Deviance residuals.

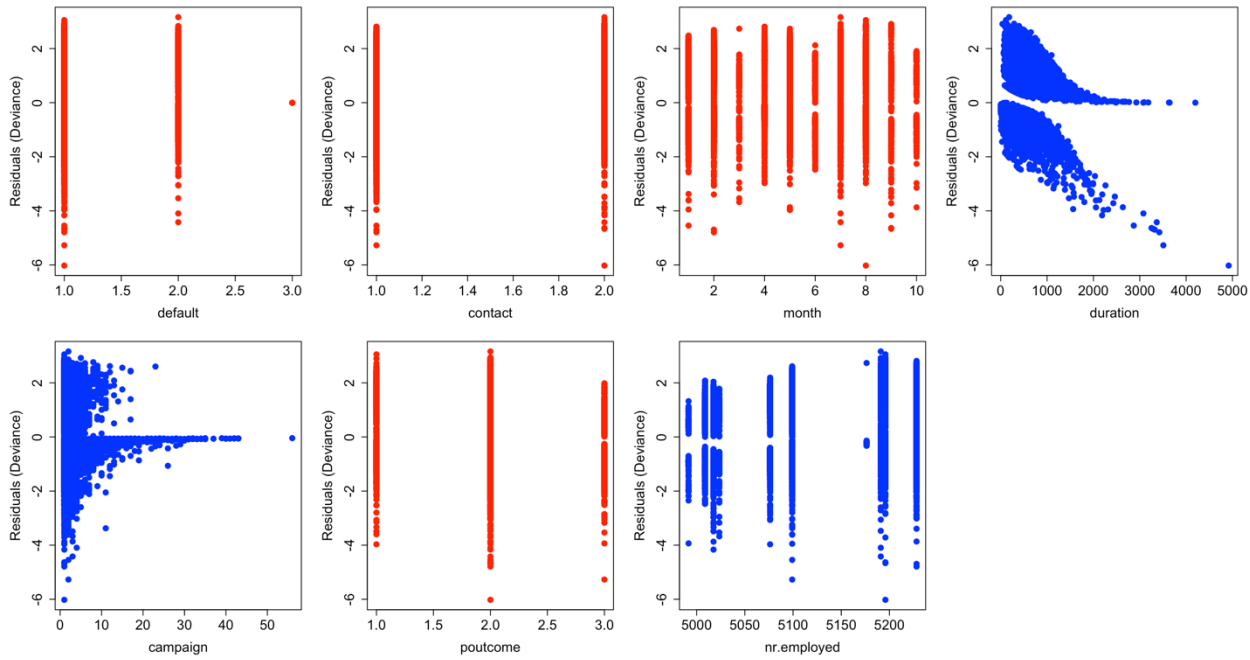


Figure 4: Deviance Residuals

Deviance residuals represent the square root of the distribution that each data point has to the overall Residual Deviance. (i.e. Figure 4) If we were able to zoom in a little bit more into our graphs we would be able to see that our residuals are separated by a line in 0 and we can see that we do not have signs of perfect separation.

Lastly, we can test whether our final model fits significantly better than the null model so we will do a test between the residual deviance for the model with predictors and the

null model. The result showed that there is significant difference between our model and the null model (p-value= 0 < .05). Additionally, we can perform a Likelihood Ratio test to make sure that the variable we removed using Wald test doesn't affect our model negatively. The result showed that there is no statistical difference between our final model and the model including the extra variable (cons.price.idx) (LRT: p-value= 0.92> .05) (i.e. Table 9)

APPENDIX

Tables

TABLE 1. Structure of our final data set

```
$ age           : num [1:39883] 56 57 37 40 56 45 59 41 24 25 ...
$ job           : Factor w/ 12 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 2 10 8 ...
$ marital       : Factor w/ 4 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...
$ education     : Factor w/ 8 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 8 6 4 ...
$ default       : Factor w/ 3 levels "no","unknown",..: 1 2 1 1 1 2 1 2 1 1 ...
$ housing       : Factor w/ 3 levels "no","unknown",..: 1 1 3 1 1 1 1 1 3 3 ...
$ loan          : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 3 1 1 1 1 1 ...
$ contact       : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...
$ month         : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7 7 7 7 ...
$ day_of_week   : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2 2 2 2 ...
$ duration      : num [1:39883] 261 149 226 151 307 198 139 217 380 50 ...
$ campaign      : num [1:39883] 1 1 1 1 1 1 1 1 1 1 ...
$ previous      : num [1:39883] 0 0 0 0 0 0 0 0 0 0 ...
$ poutcome      : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...
$ emp.var.rate  : num [1:39883] 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
$ cons.price.idx: num [1:39883] 94 94 94 94 94 ...
$ cons.conf.idx : num [1:39883] -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...
$ euribor3m     : num [1:39883] 4.86 4.86 4.86 4.86 4.86 ...
$ nr.employed   : num [1:39883] 5191 5191 5191 5191 5191 ...
$ SUBSCRIBED    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

TABLE 2. Describe table for numeric variables

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

	nr.employed
vars	9.00
n	39883.00
mean	5173.22
sd	64.63
median	5191.00
trimmed	5182.16
mad	55.00
min	4991.60
max	5228.10
range	236.50
skew	-0.96
kurtosis	-0.33
se	0.32

Descriptive measures for numeric variables age, duration, campaign, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m and nr.employed.

TABLE 3. Summary of the full model

```
Call:
glm(formula = SUBSCRIBED ~ ., family = binomial(link = "logit"),
    data = my_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.0959  -0.2857  -0.1825  -0.1332   3.5899

Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.370e+02  4.378e+01  -7.698 1.38e-14 ***
age                         -3.974e-04  2.574e-03  -0.154 0.877306
jobblue-collar              -2.342e-01  8.303e-02  -2.820 0.004799 **
jobentrepreneur             -1.547e-01  1.292e-01  -1.197 0.231296
jobhousemaid                 5.011e-03  1.538e-01   0.033 0.973998
jobmanagement               -6.590e-02  8.986e-02  -0.733 0.463383
jobretired                   2.902e-01  1.140e-01   2.546 0.010894 *
jobself-employed            -1.197e-01  1.212e-01  -0.988 0.323099
jobservices                 -1.331e-01  9.010e-02  -1.478 0.139505
jobstudent                   2.633e-01  1.220e-01   2.159 0.030866 *
jobtechnician               -3.466e-02  7.523e-02  -0.461 0.644953
jobunemployed               -1.240e-02  1.375e-01  -0.090 0.928187
jobunknown                  -1.385e-01  2.603e-01  -0.532 0.594687
maritalmarried              -1.234e-02  7.252e-02  -0.170 0.864853
maritalsingle                9.847e-02  8.215e-02   1.199 0.230689
maritalunknown               1.642e-01  4.169e-01   0.394 0.693587
educationbasic.6y            9.212e-02  1.244e-01   0.741 0.458933
educationbasic.9y           -2.814e-02  9.875e-02  -0.285 0.775645
educationhigh.school        -6.231e-03  9.674e-02  -0.064 0.948642
educationilliterate          9.825e-01  7.568e-01   1.298 0.194200
educationprofessional.course 6.939e-02  1.075e-01   0.645 0.518605
educationuniversity.degree   1.436e-01  9.681e-02   1.483 0.137943
educationunknown             1.353e-01  1.264e-01   1.070 0.284691
defaultunknown              -2.674e-01  6.852e-02  -3.902 9.56e-05 ***
defaultyes                  -7.270e+00  1.135e+02  -0.064 0.948909
housingunknown              -1.884e-01  1.520e-01  -1.240 0.215099
housingyes                  -3.464e-03  4.348e-02  -0.080 0.936508
loanunknown                         NA         NA      NA       NA
loanyes                     -6.800e-02  6.069e-02  -1.120 0.262526
contacttelephone            -4.200e-01  8.324e-02  -5.046 4.52e-07 ***
monthaug                     2.890e+00  2.042e-01  14.156  < 2e-16 ***
monthdec                     8.852e-01  2.165e-01   4.089 4.32e-05 ***
monthjul                     8.818e-01  1.214e-01   7.264 3.76e-13 ***
monthjun                    -8.682e-01  1.455e-01  -5.969 2.39e-09 ***
monthmar                     2.195e+00  1.508e-01  14.556  < 2e-16 ***
monthmay                    -9.592e-02  8.861e-02  -1.082 0.279043
monthnov                     2.769e-01  1.445e-01   1.917 0.055262 .
monthoct                     1.224e+00  1.888e-01   6.484 8.94e-11 ***
monthsep                     8.773e-01  2.312e-01   3.795 0.000148 ***
day_of_weekmon              -9.672e-02  6.965e-02  -1.389 0.164918
day_of_weekthu               5.312e-02  6.760e-02   0.786 0.431949
day_of_weektue               8.879e-02  6.920e-02   1.283 0.199473
day_of_weekwed               1.494e-01  6.936e-02   2.154 0.031213 *
duration                     4.744e-03  7.642e-05  62.075  < 2e-16 ***
campaign                    -3.427e-02  1.181e-02  -2.903 0.003697 **
previous                    -9.159e-04  8.331e-02  -0.011 0.991228
poutcomenonexistent          4.031e-01  1.190e-01   3.386 0.000708 ***
poutcomesuccess              1.693e+00  9.836e-02  17.212  < 2e-16 ***
emp.var.rate                -2.679e+00  1.707e-01 -15.692  < 2e-16 ***
cons.price.idx               3.622e+00  2.984e-01  12.137  < 2e-16 ***
cons.conf.idx               -6.851e-02  1.191e-02  -5.751 8.86e-09 ***
euribor3m                    1.027e+00  1.478e-01   6.948 3.69e-12 ***
nr.employed                 -2.536e-03  3.738e-03  -0.678 0.497489
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25925  on 39882  degrees of freedom
Residual deviance: 15610  on 39831  degrees of freedom
AIC: 15714
```

TABLE 4. Coefficients selected with lambda 1se

	s1
(Intercept)	-80.112937140
age	.
jobblue-collar	-0.218472635
jobentrepreneur	-0.065848385
jobhousemaid	.
jobmanagement	.
jobretired	0.313310479
jobself-employed	-0.019647869
jobservices	-0.105653675
jobstudent	0.338879239
jobtechnician	.
jobunemployed	.
jobunknown	.
maritalmarried	.
maritalsingle	0.091512842
maritalunknown	.
educationbasic.6y	.
educationbasic.9y	-0.041378346
educationhigh.school	-0.010053253
educationilliterate	0.602785360
educationprofessional.course	.
educationuniversity.degree	0.107057868
educationunknown	0.046434873
defaultunknown	-0.266590811
defaultyes	.
housingunknown	-0.069867613
housingyes	.
loanunknown	-0.002124561
loanyes	-0.030289230
contacttelephone	-0.227632272
monthaug	0.859017919
monthdec	0.278700536
monthjul	0.362479670
monthjun	.
monthmar	1.511246070
monthmay	-0.604089421
monthnov	.
monthoct	0.539809122
monthsep	0.053377973
day_of_weekmon	-0.095935378
day_of_weekthu	.
day_of_weektue	0.046003002
day_of_weekwed	0.096032851
duration	0.004602963
campaign	-0.029420354
previous	.
poutcomenonexistent	0.363692106
poutcomesuccess	1.707615030
emp.var.rate	-0.749076330
cons.price.idx	1.056024827
cons.conf.idx	.
euribor3m	.
nr.employed	-0.004467736

TABLE 5. Summary of the model selected with Stepwise procedure

```
Call:
glm(formula = SUBSCRIBED ~ default + contact + month + duration +
    campaign + poutcome + emp.var.rate + cons.price.idx + nr.employed,
    family = binomial(link = "logit"), data = my_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.0114  -0.2940  -0.1846  -0.1341   3.4419

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.076e+02  3.698e+01 -13.726  < 2e-16 ***
defaultunknown      -3.474e-01  6.618e-02  -5.249 1.53e-07 ***
defaultyes          -7.189e+00  1.134e+02  -0.063  0.94946
contacttelephone    -5.127e-01  7.284e-02  -7.038 1.95e-12 ***
monthaug             2.696e+00  1.605e-01  16.800  < 2e-16 ***
monthdec             1.221e+00  1.974e-01   6.185 6.23e-10 ***
monthjul             8.240e-01  9.781e-02   8.425  < 2e-16 ***
monthjun            -1.041e+00  1.342e-01  -7.757 8.71e-15 ***
monthmar             2.652e+00  1.427e-01  18.581  < 2e-16 ***
monthmay            -5.083e-02  8.556e-02  -0.594  0.55247
monthnov             5.872e-01  1.027e-01   5.715 1.10e-08 ***
monthoct             1.562e+00  1.365e-01  11.442  < 2e-16 ***
monthsep             1.446e+00  1.891e-01   7.643 2.12e-14 ***
duration             4.708e-03  7.577e-05  62.129  < 2e-16 ***
campaign            -3.783e-02  1.179e-02  -3.209  0.00133 **
poutcomenonexistent  4.568e-01  6.679e-02   6.840 7.93e-12 ***
poutcomesuccess      1.723e+00  9.729e-02  17.715  < 2e-16 ***
emp.var.rate        -2.645e+00  1.655e-01 -15.984  < 2e-16 ***
cons.price.idx       4.372e+00  2.820e-01  15.500  < 2e-16 ***
nr.employed          1.816e-02  2.200e-03   8.252  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25925  on 39882  degrees of freedom
Residual deviance: 15748  on 39863  degrees of freedom
AIC: 15788

Number of Fisher Scoring iterations: 10
```

TABLE 6. GVIF test

GVIF before removing emp. var. rate

	GVIF	Df	GVIF^(1/(2*Df))
default	1.1	2	1.0
contact	1.9	1	1.4
month	29.1	9	1.2
duration	1.2	1	1.1
campaign	1.0	1	1.0
poutcome	1.3	2	1.1
emp.var.rate	191.9	1	13.9
cons.price.idx	68.6	1	8.3
nr.employed	64.3	1	8.0

GVIF after removing emo. var. rate

	GVIF	Df	GVIF^(1/(2*Df))
default	1.1	2	1.0
contact	1.7	1	1.3
month	3.9	9	1.1
duration	1.2	1	1.1
campaign	1.0	1	1.0
poutcome	1.3	2	1.1
cons.price.idx	3.9	1	2.0
nr.employed	4.2	1	2.0

TABLE 7. Summary of our final model

```
Call:
glm(formula = SUBSCRIBED ~ default + contact + month + duration +
    campaign + poutcome + nr.employed, family = binomial(link = "logit"),

    data = my_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.0217  -0.3019  -0.1926  -0.1388   3.1623

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          7.945e+01  2.184e+00  36.385  < 2e-16 ***
defaultunknown      -3.717e-01  6.580e-02  -5.649 1.62e-08 ***
defaultyes          -7.451e+00  1.132e+02  -0.066 0.947513
contacttelephone    -2.305e-01  6.317e-02  -3.648 0.000264 ***
monthaug             6.366e-01  8.595e-02   7.406 1.30e-13 ***
monthdec             1.904e-01  1.859e-01   1.024 0.305767
monthjul             4.942e-01  9.486e-02   5.209 1.90e-07 ***
monthjun             6.064e-01  8.838e-02   6.861 6.83e-12 ***
monthmar             1.257e+00  1.176e-01  10.693  < 2e-16 ***
monthmay            -7.251e-01  7.285e-02  -9.954  < 2e-16 ***
monthnov             3.913e-02  9.280e-02   0.422 0.673251
monthoct             3.838e-01  1.228e-01   3.125 0.001776 **
monthsep            -1.828e-01  1.573e-01  -1.162 0.245207
duration             4.650e-03  7.471e-05  62.244  < 2e-16 ***
campaign            -4.356e-02  1.186e-02  -3.674 0.000239 ***
poutcomenonexistent  4.458e-01  6.666e-02   6.687 2.27e-11 ***
poutcomesuccess      1.771e+00  9.700e-02  18.262  < 2e-16 ***
nr.employed         -1.624e-02  4.327e-04 -37.544  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25925  on 39882  degrees of freedom
Residual deviance: 15978  on 39865  degrees of freedom
AIC: 16014

Number of Fisher Scoring iterations: 10
```

TABLE 8. Wald test for the variable cons.price.idx

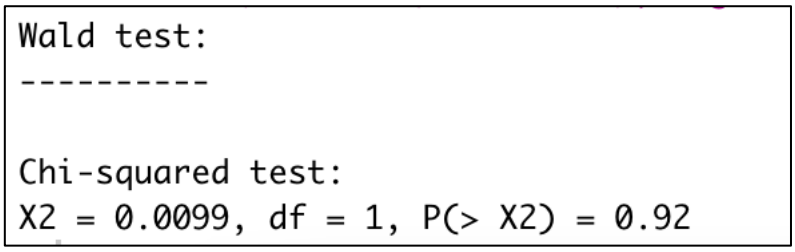


TABLE 9. Likelihood ratio test between our final model and the model with the extra variable we removed with the Wald test

```
Analysis of Deviance Table

Model 1: SUBSCRIBED ~ default + contact + month + duration + campaign +
    poutcome + cons.price.idx + nr.employed
Model 2: SUBSCRIBED ~ default + contact + month + duration + campaign +
    poutcome + nr.employed
  Resid. Df Resid. Dev Df   Deviance Pr(>Chi)
1     39864      15978
2     39865      15978 -1 -0.0098766   0.9208
```

TABLE 10. Model with centered covariates

```
Call:
glm(formula = SUBSCRIBED ~ default + contact + month + duration +
    campaign + poutcome + nr.employed, family = binomial(link = "logit"),
    data = my_datacentered)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.0217  -0.3019  -0.1926  -0.1388   3.1623

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.501e+00  9.228e-02 -37.943  < 2e-16 ***
defaultunknown      -3.717e-01  6.580e-02  -5.649 1.62e-08 ***
defaultyes          -7.451e+00  1.132e+02  -0.066 0.947513
contacttelephone    -2.305e-01  6.317e-02  -3.648 0.000264 ***
monthaug             6.366e-01  8.595e-02   7.406 1.30e-13 ***
monthdec             1.904e-01  1.859e-01   1.024 0.305767
monthjul             4.942e-01  9.486e-02   5.209 1.90e-07 ***
monthjun             6.064e-01  8.838e-02   6.861 6.83e-12 ***
monthmar             1.257e+00  1.176e-01  10.693  < 2e-16 ***
monthmay            -7.251e-01  7.285e-02  -9.954  < 2e-16 ***
monthnov             3.913e-02  9.280e-02   0.422 0.673251
monthoct             3.838e-01  1.228e-01   3.125 0.001776 **
monthsep            -1.828e-01  1.573e-01  -1.162 0.245207
duration             4.650e-03  7.471e-05  62.244  < 2e-16 ***
campaign            -4.356e-02  1.186e-02  -3.674 0.000239 ***
poutcomenonexistent  4.458e-01  6.666e-02   6.687 2.27e-11 ***
poutcomesuccess      1.771e+00  9.700e-02  18.262  < 2e-16 ***
nr.employed         -1.624e-02  4.327e-04 -37.544  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25925  on 39882  degrees of freedom
Residual deviance: 15978  on 39865  degrees of freedom
AIC: 16014

Number of Fisher Scoring iterations: 10
```

Figures

Figure 1: Bar plots for factor variables

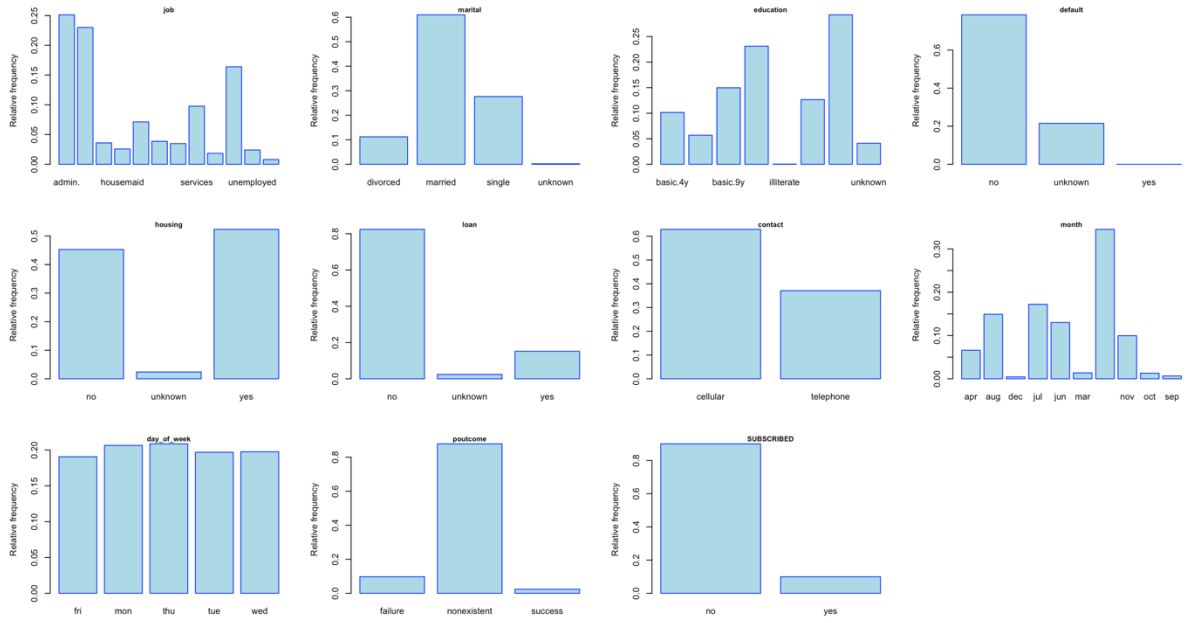


Figure 2: Correlation plot for numeric variables

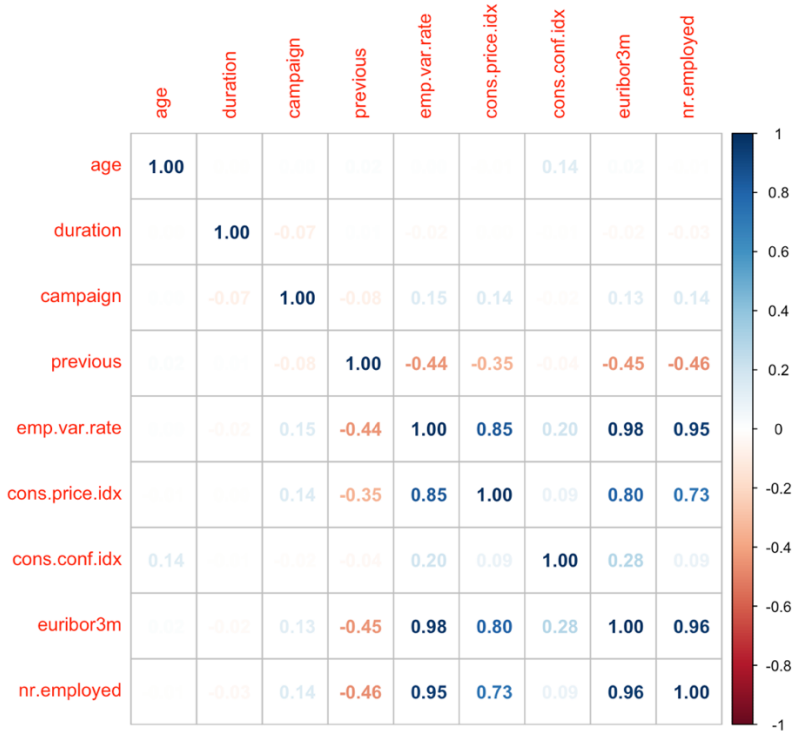
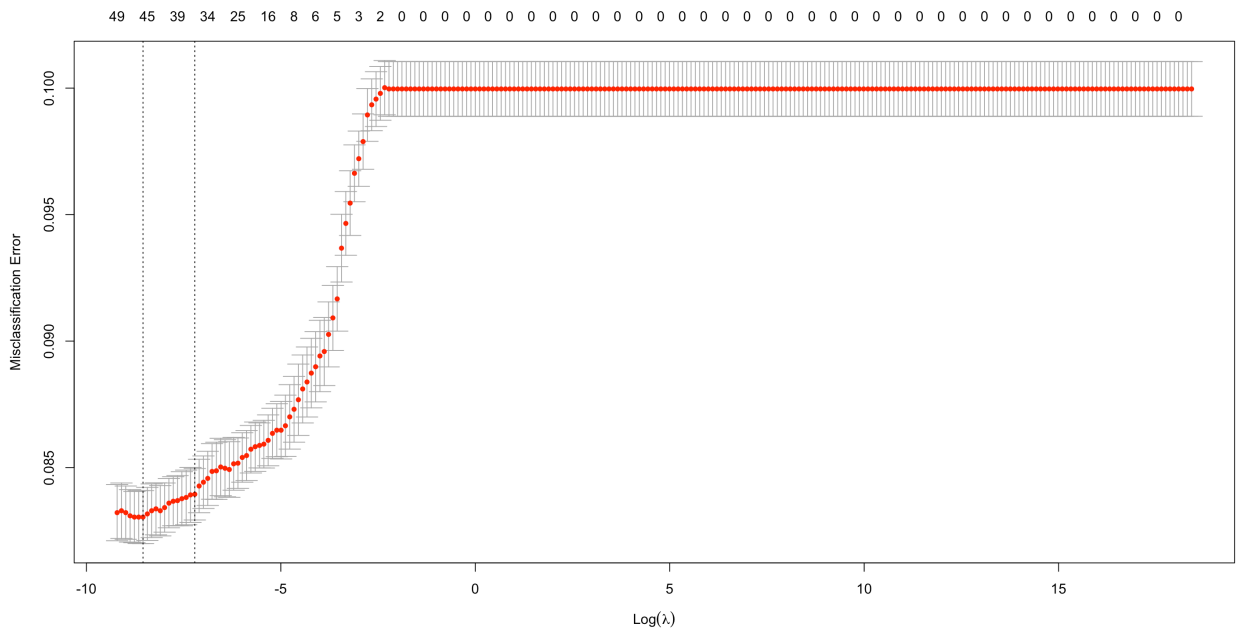


Figure 3: The cross-validation curve (red dotted line) along with upper and lower standard deviation curves along the λ sequence (error bars)



We use the cross validation to select our λ. First λ will be the lambda.min which is the one with the minimum CV-MSE and the second one is largest value of lambda such that error is within 1 standard error of the minimum. On the top part of the graph, we can see the number of coefficients dropping as log(λ) increases as well as the Misclassification Error.

The vertical lines depict the lambdas we mentioned above, the first one is the lambda.min and the second one is the lambda.1se. It is better to choose the lambda.1se because we have less coefficients at this stage.

Figure 4: Deviance Residuals