MSc in Business Analytics at AUEB
Statistics for Business Analytics II

Assignment I: Creating a Statistical Model

Retail Banking Telemarketing Phone Calls

Professor: Karlis Dimitrios
Student: Despotis Spyridon
Student ID: P2822111

2022

Contents

Introduction

In recent years, the retail banking industry has been trying to increase its sales territory with telemarketing phone calls. The main benefit of using telemarketing to promote your business is that it allows you to immediately gauge your customer's level of interest in your product or service. For banking managers it is crucial to understand what the most important parameters are that play significant roles in a successful telemarketing contact. Specifically, to define a telemarketing strategy with more successful contacts and fewer calls, which will boost banks sales and profits. The data that is generated by such telemarketing campaigns is therefore very attractive for research, in order to export useful insights that can be used to influence telemarketing strategies. For this assignment we have analyzed real data that has been collected from one retail bank, from May 2008 to June 2010, totaling nearly 40.000 phone contacts. The dataset that was analyzed contains important customer characteristics related to the direct telemarketing campaigns. In this assignment, our main objective is to find which variables of the dataset contribute to a successful contact (i.e. the client subscribes to the product). Our guiding questions to reach our main objective are:

- Which variables are important?
- Do we need to transform them?
- How good is the model?
- Are there any assumptions that need to be check carefully?

$ age	num	56 57 37 40 56 45 59 41 24 25 ...
$ job	chr	"housemaid" "services" "services" "admin." ...
$ marital	chr	"married" "married" "married" "married" ...
$ education	chr	"basic.4y" "high.school" "high.school" "basic.6y" ...
$ default	chr	"no" NA "no" "no" ...
$ housing	chr	"no" "no" "yes" "no" ...
$ loan	chr	"no" "no" "no" "no" ...
$ contact	chr	"telephone" "telephone" "telephone" "telephone" ...
$ month	chr	"may" "may" "may" "may" ...
$ day_of_week	chr	"mon" "mon" "mon" "mon" ...
$ duration	num	261 149 226 151 307 198 139 217 380 50 ...
$ campaign	num	1 1 1 1 1 1 1 1 1 1 ...
$ pdays	num	999 999 999 999 999 999 999 999 999 999 ...
$ previous	num	0 0 0 0 0 0 0 0 0 0 ...
$ poutcome	chr	"nonexistent" "nonexistent" "nonexistent" "
$ emp.var.rate	num	1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
$cons.price.idx	num	94 94 94 94 94 ...
$ cons.conf.idx	num	-36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4
$ euribor3m	num	4.86 4.86 4.86 4.86 4.86 ...
$ nr.employed	num	5191 5191 5191 5191 5191 ...
$ SUBSCRIBED	chr	"no" "no" "no" "no" ...

Table 0.1 Retail Bank Dataset Structure

1. Data Manipulation & Explanatory

Our preliminary dataset named “project I 2021-2022” contained 21 variables (columns) and 39.883 observations (rows). Before we started our statistical analysis, we had to examine our dataset. We began by looking for missing values and then we proceeded by performing appropriate data type conversions and exclusions. First, by checking our data structure balance we can see that we have a high percentage of missing values especially in variables: default, education, housing, loan and job (figure 1.1 & table 1.1). Those results led us to exclude the “default” variable since the majority of the data points were missing and it did not provide any additional vital information. Then, we checked the data ranges for the columns and we saw that column “pdays” (number of days that passed after the client was last contacted from a previous campaign) ranges from “0” to “27” with an extra level of “999”. The value “999” means that the client was not previously contacted, which the majority of the records contained. We decided to set all of the “999” values to “0”, and since the vast majority of data points were now “0”, we decided to exclude the column. The high imbalance in these variables was also detected using the “nearZeroVar” function of the caret R package, which can be used to identify variables of low variability, and thus of low informative value.

Order	variables	types	missing_count	missing_percent	unique_count	unique_rate
1	age	numeric	0	0	78	0.00196
2	job	character	317	0.795	12	0.000301
3	marital	character	79	0.198	4	0.000100
4	education	character	1643	4.12	8	0.000201
5	default	character	8571	21.5	3	0.0000752
6	housing	character	955	2.39	3	0.0000752
7	loan	character	955	2.39	3	0.0000752
8	contact	character	0	0	2	0.0000501
9	month	character	0	0	10	0.000251
10	day_of_week	character	0	0	5	0.000125

Table 1.1 Amount of Missing Values per Variable

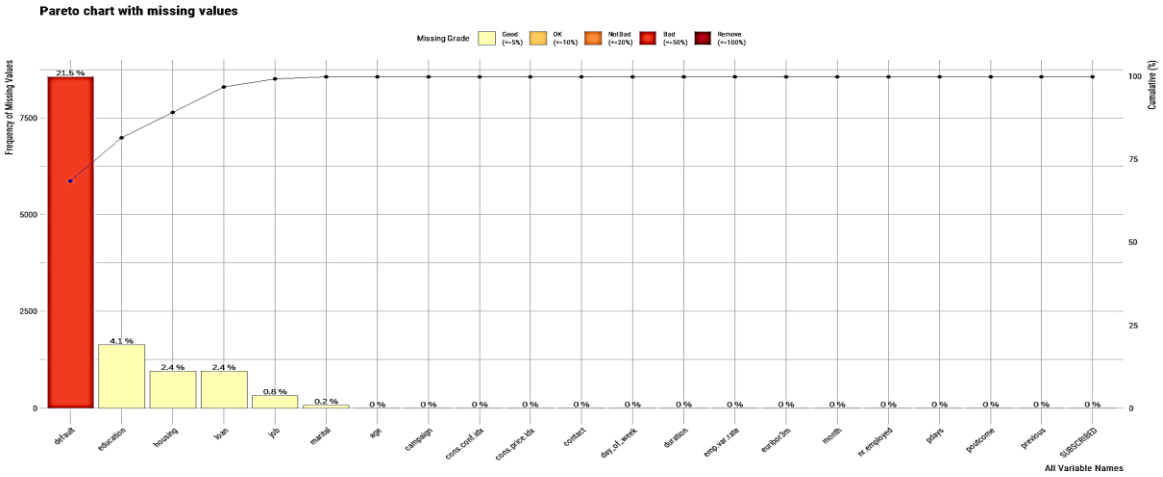

Figure 1-1 Pareto Chart with Missing Values

1.1. Handling Continuous Variables

Visualizing the numeric variables can give us useful insights about the distribution of the data. As we see in figure 1.2, there is evidence of some extreme values in the histograms which could possibly indicate the existence of outliers. Also we have left and right skewness at the most of the variables and the most symmetric seem to be the age variable. The main reason for examining the histograms is to identify imbalances in the variables. We can see that there are some variables where the majority of the observations concerns one or fewer values. This pattern was revealed for the following variables: “emp.var.rate”, “nr.employed”, “previous”, “campaign”, “duration”, and “age”. These variables were binned and converted to factors, with balanced levels to the extent possible.

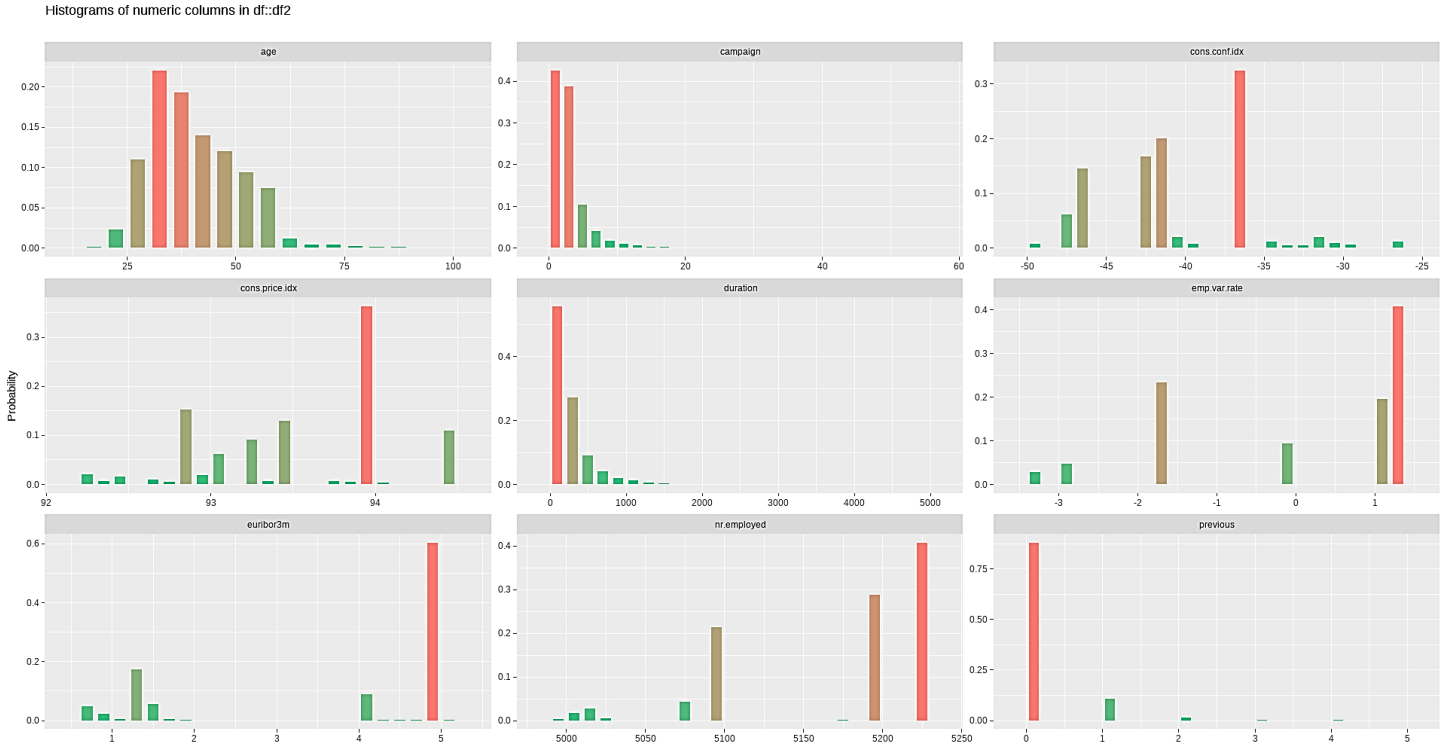

Figure 1-2 Histogram Plots for Numeric Variables

Also, we used pairwise comparisons to have a straight forward view of the correlations between our variables. From the figure 1.3 we can see there is high positive correlation only between the “euribor3m” and “cons.price.idx” . This means that for higher consumer price index, it most possibly to have higher euribor 3 month rate. So, due to high correlation (almost colinear) we decided to exclude “euribor3m” variable.

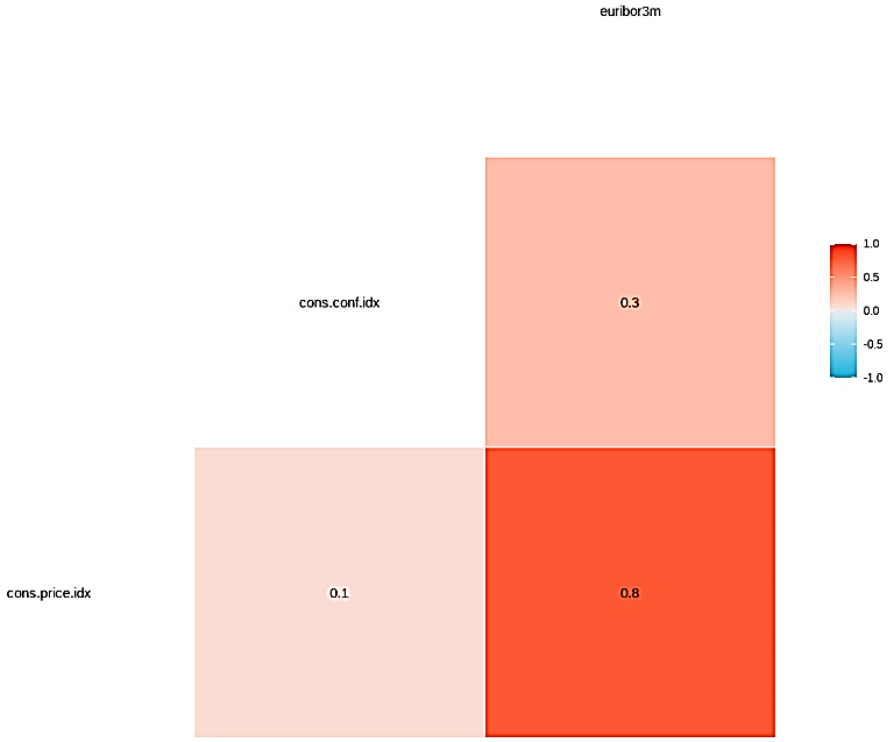

Figure 1-3 Pairwise Comparisons for Continuous Variables

Finally, we visualize the numeric variables that we will keep in our analysis. For the variable "cons.conf.idx" we can see that the majority of data points are located on the left side of the histogram, between -36.1 and -47.1. The highest probability has the value -36.4 and follows the value -42.7 with probability 0.2. For the variable "cons.price.idx:" we see the majority of records at the center of the histogram, while the highest probability has the 93.994 value.

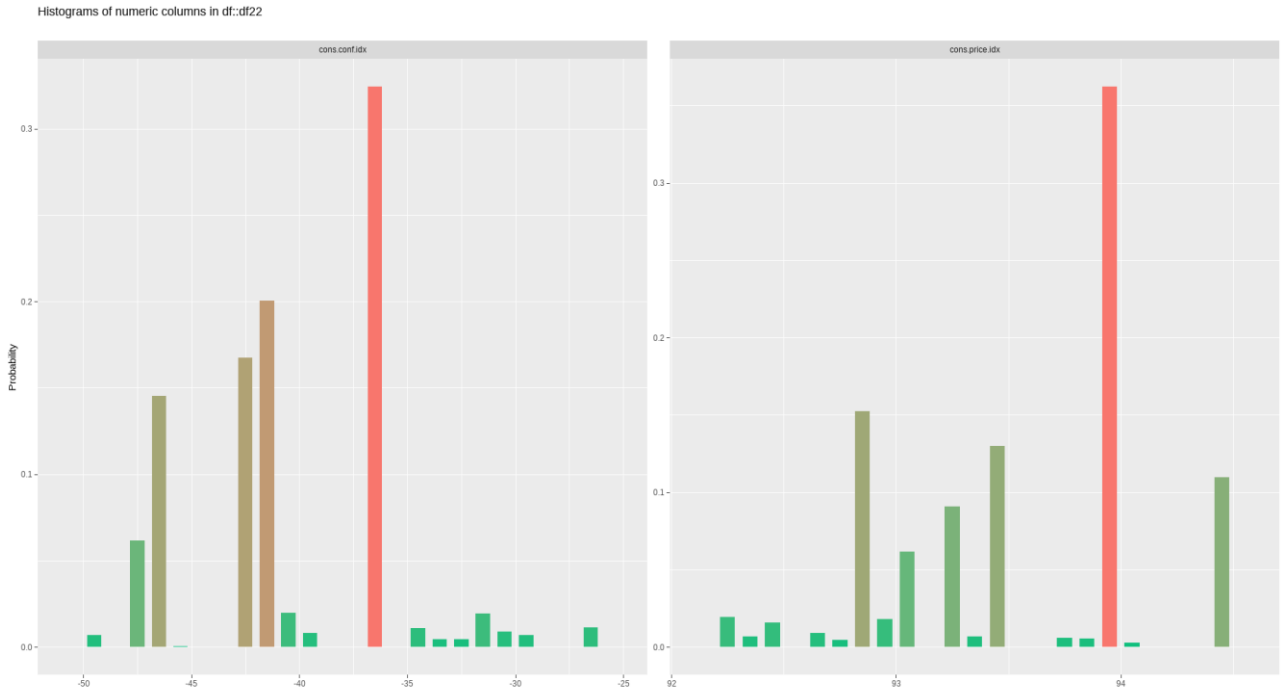

Figure 1-4 Histograms with variables "cons.conf.idx " and "cons.price.idx"

1.2. Handling Categorical Variables

First we converted all character data types to factors in order to store the same values as levels. Then, in order to proceed to the appropriate conversions it is important to visualize the categorical data. The figure 1-5, below summarizes all factor variables with their frequency. The grey colors indicates missing values. For example, we can see that the most frequent value in the response variable “SUBSCRIBED” is no. We also have many values like “"basic.9y", "basic.4y", "basic.6y" in the “education” variable which can be merged since they refer to the same level of education.

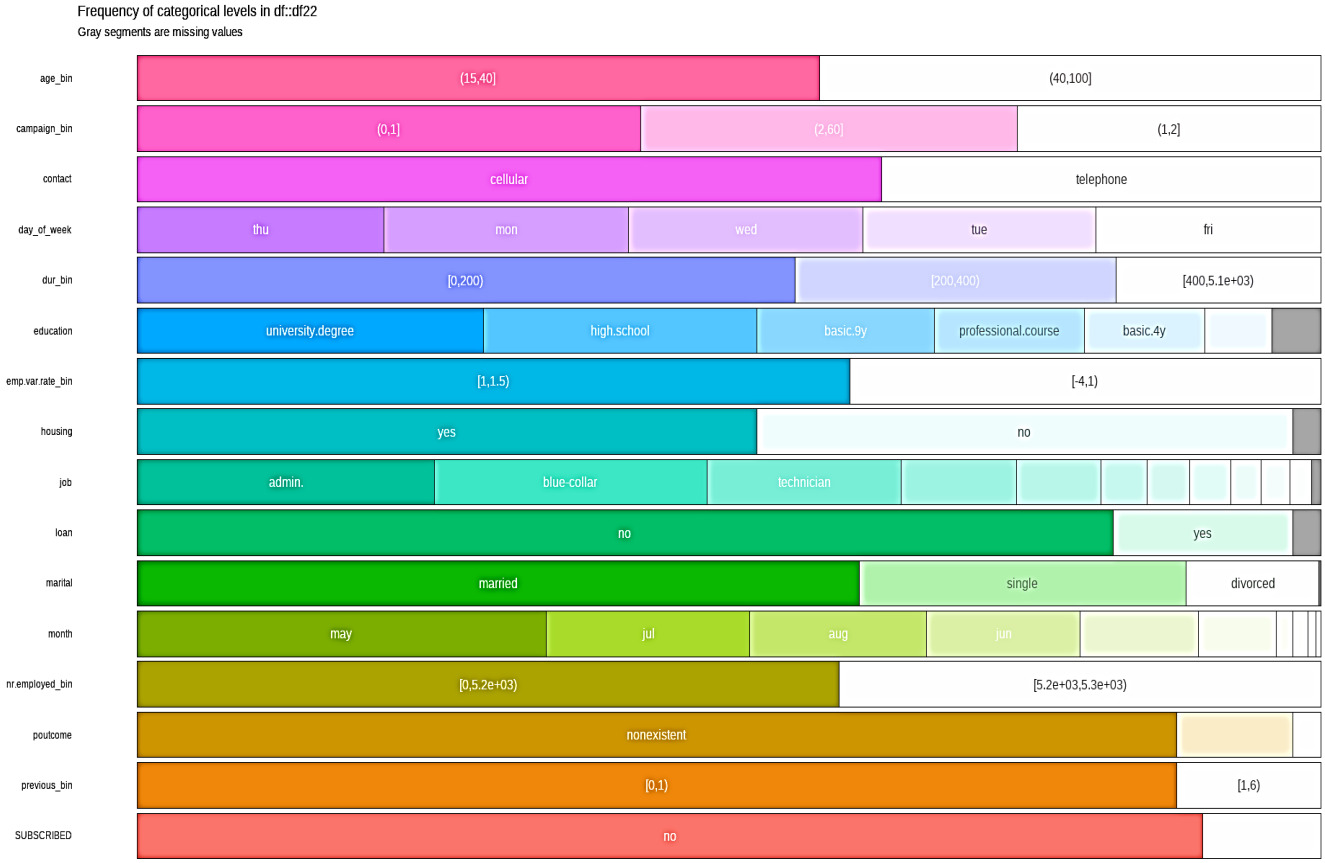

Figure 1-5 Frequency of Categorical Levels

Therefore, we proceed by grouping all basic levels of education into one level, named “basic”. Also from “month” variable, we extracted a new variable named “seasons” (with values: “spring”, “summer”, “winter”,”autum”) since we do not expect a strong linear relationship for each month with the “subscribed” variable. Then, by checking the range of the values, we decided to drop the “winter” level, since it had only 182 records, comparing to the 39883 records of the rest levels. Also, for the “job” variable we adjusted the levels. Specifically, we kept the same levels that contained at least 15% of records while the levels that contained below

the 15% we grouped them as “others”. The bellow Figure, contains all associations between all variables of our dataset.

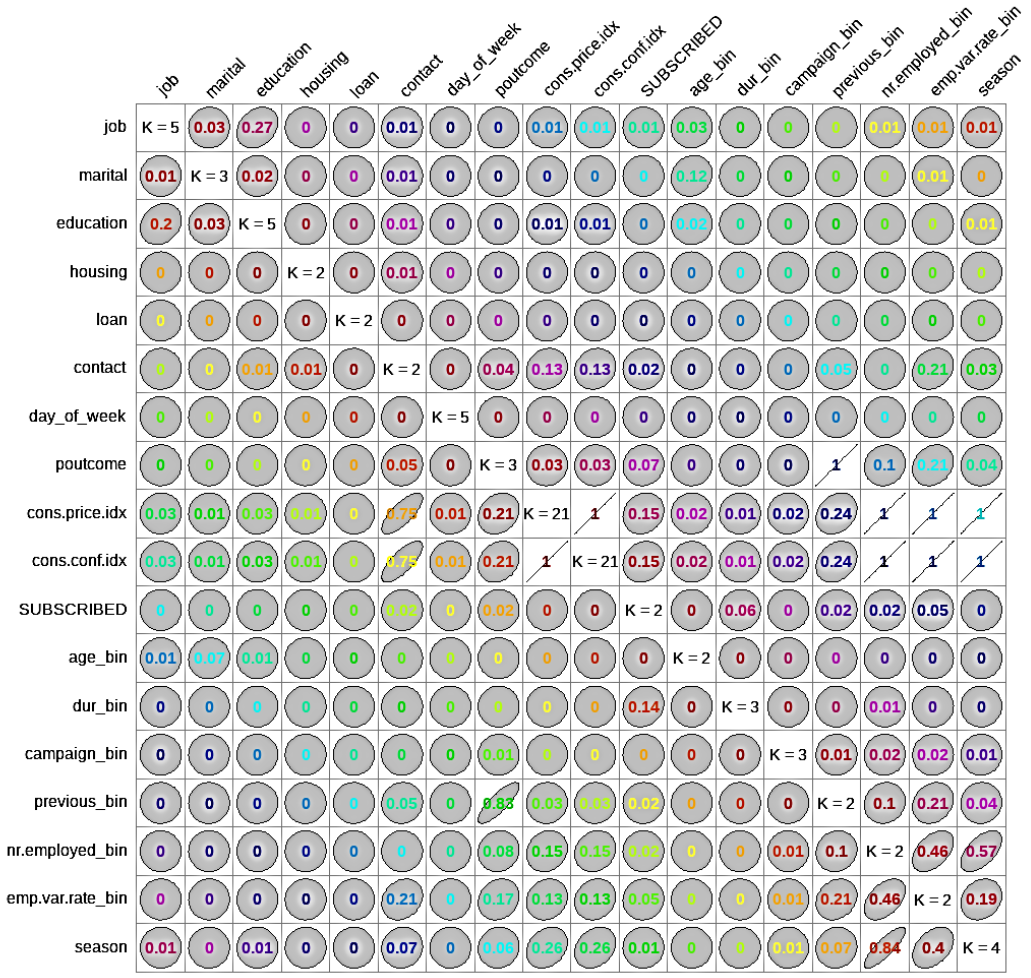

Figure 1-6 Associations Between the Variables of the Dataset

1.3. Final Data Structure

After the process of data cleaning and transformations that we saw in the two previous parts, the final dataset structure contains 18 variables (columns) and 37.068 observations (variables). Specifically, we have 2 continuous columns and 16 discrete columns.

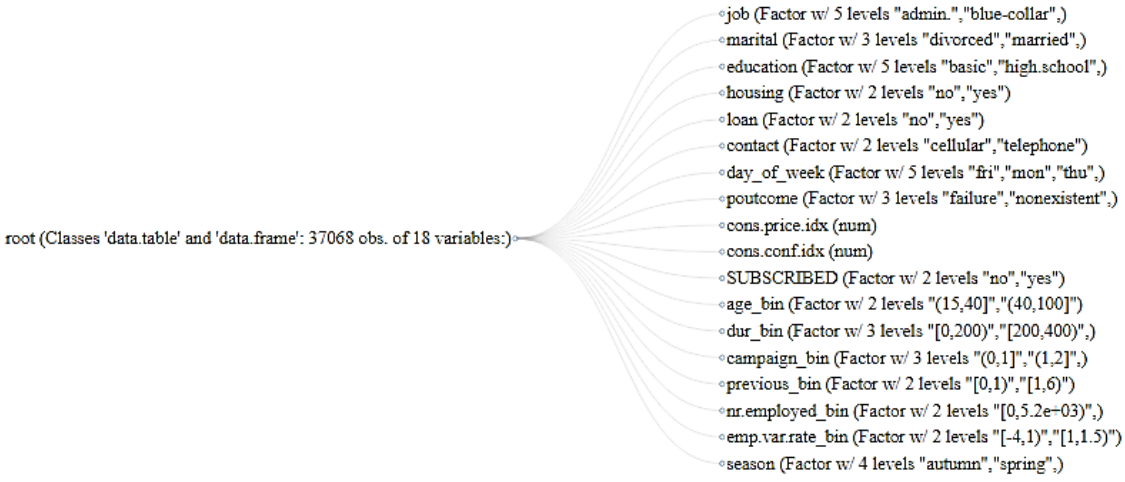

Figure 1-7 Final Structure of the Cleaned Dataset

The bellow Figure 1-8, contains the comparisons of most common levels of the factors for the observations with SUBSCRIBED=yes and SUBSCRIBED=no. For example we can see that there are different prevalent levels between the two subsets for “emp.var.rate”, “duration season”, “education,job”, “day”. For variables like “poutcome” and “previous”, the same level is the most frequent, but with different frequencies, while for “age” and “loan” small differences were found.

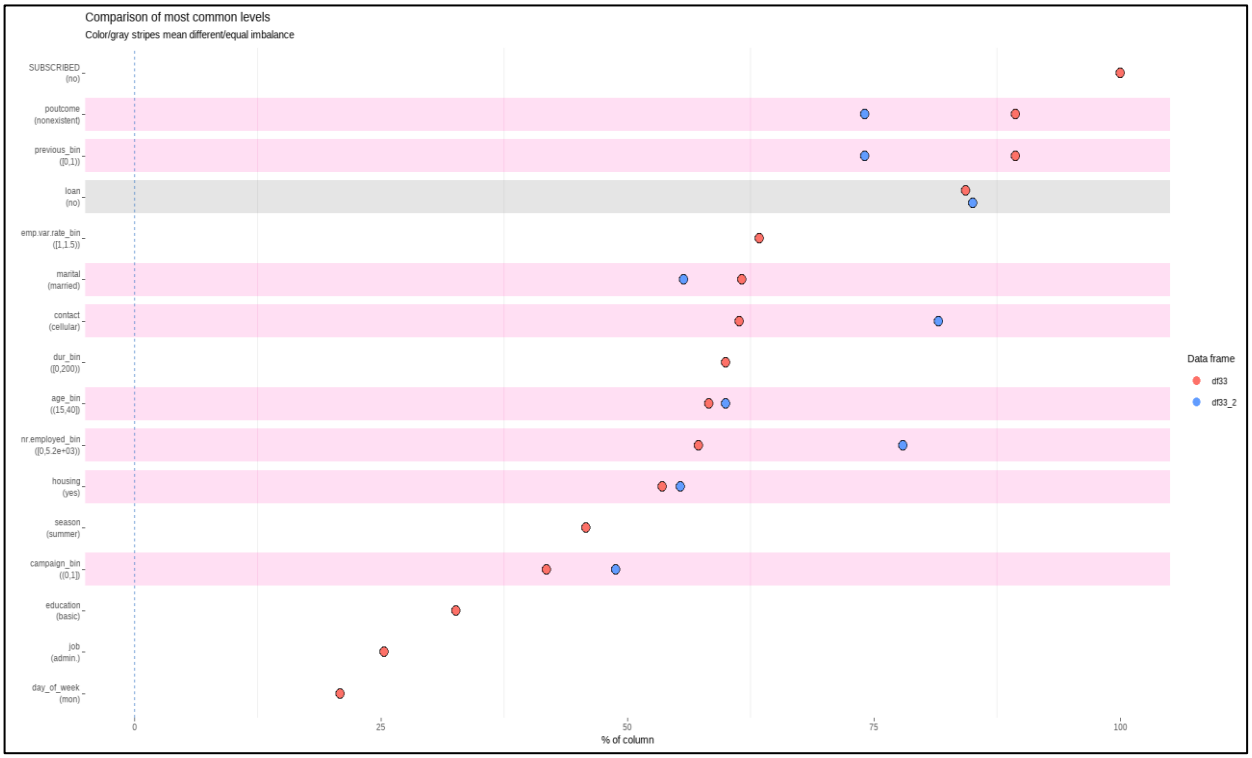

Figure 1-8 Comparison of the Most Common Levels of Factorial Variables

2. Variable Association Testing

We proceed by making the appropriate tests for each data type column. For factors we apply Pearson's Chi-squared test and for numeric variables Linear Model ANOVA test. The results are summarized in the bellow Figure 2-1. We conclude that, as the p-values are below the significant level of alpha (0.05), we reject the null hypothesis that there is no association between the variables. We fail to reject the null hypothesis for "loan" variable, since the 0.233 is way higher than the usual confidence level of 0.05.

	no (N=33396)	yes (N=3672)	Total (N=37068)	p value
job				< 0.001
admin.	8461 (25.3%)	1090 (29.7%)	9551 (25.8%)	
blue-collar	7926 (23.7%)	557 (15.2%)	8483 (22.9%)	
services	3377 (10.1%)	269 (7.3%)	3646 (9.8%)	
technician	5614 (16.8%)	585 (15.9%)	6199 (16.7%)	
Other	8018 (24.0%)	1171 (31.9%)	9189 (24.8%)	
marital				< 0.001
divorced	3787 (11.3%)	382 (10.4%)	4169 (11.2%)	
married	20590 (61.7%)	2045 (55.7%)	22635 (61.1%)	
single	9019 (27.0%)	1245 (33.9%)	10264 (27.7%)	
education				< 0.001
basic	10903 (32.6%)	941 (25.6%)	11844 (32.0%)	
high.school	8086 (24.2%)	868 (23.6%)	8954 (24.2%)	
illiterate	14 (0.0%)	4 (0.1%)	18 (0.0%)	
professional.course	4436 (13.3%)	477 (13.0%)	4913 (13.3%)	
university.degree	9957 (29.8%)	1382 (37.6%)	11339 (30.6%)	
housing				0.034
no	15510 (46.4%)	1638 (44.6%)	17148 (46.3%)	
yes	17886 (53.6%)	2034 (55.4%)	19920 (53.7%)	
loan				0.233
no	28161 (84.3%)	3124 (85.1%)	31285 (84.4%)	
yes	5235 (15.7%)	548 (14.9%)	5783 (15.6%)	
contact				< 0.001
cellular	20501 (61.4%)	2996 (81.6%)	23497 (63.4%)	
telephone	12895 (38.6%)	676 (18.4%)	13571 (36.6%)	
day_of_week				0.005
fri	6349 (19.0%)	669 (18.2%)	7018 (18.9%)	
mon	6981 (20.9%)	687 (18.7%)	7668 (20.7%)	
thu	6917 (20.7%)	818 (22.3%)	7735 (20.9%)	
tue	6515 (19.5%)	756 (20.6%)	7271 (19.6%)	
wed	6634 (19.9%)	742 (20.2%)	7376 (19.9%)	
poutcome				< 0.001
failure	3194 (9.6%)	432 (11.8%)	3626 (9.8%)	
nonexistent	29846 (89.4%)	2721 (74.1%)	32567 (87.9%)	
success	356 (1.1%)	519 (14.1%)	875 (2.4%)	
cons.price.idx				< 0.001
Mean (SD)	93.585 (0.553)	93.207 (0.603)	93.548 (0.570)	
Range	92.201 - 94.465	92.201 - 94.465	92.201 - 94.465	
cons.conf.idx				< 0.001
Mean (SD)	-40.602 (4.371)	-39.593 (6.290)	-40.502 (4.607)	
Range	-50.000 - -26.900	-50.000 - -26.900	-50.000 - -26.900	
age_bin				0.048
(15,40]	19469 (58.3%)	2203 (60.0%)	21672 (58.5%)	
(40,100]	13927 (41.7%)	1469 (40.0%)	15396 (41.5%)	
dur_bin				< 0.001
[0,200)	20035 (60.0%)	566 (15.4%)	20601 (55.6%)	
[200,400)	9086 (27.2%)	950 (25.9%)	10036 (27.1%)	
[400,5.1e+03)	4275 (12.8%)	2156 (58.7%)	6431 (17.3%)	
campaign_bin				< 0.001
(0,1]	13967 (41.8%)	1793 (48.8%)	15760 (42.5%)	
(1,2]	8556 (25.6%)	962 (26.2%)	9518 (25.7%)	
(2,60]	10873 (32.6%)	917 (25.0%)	11790 (31.8%)	
previous_bin				< 0.001
[0,1)	29846 (89.4%)	2721 (74.1%)	32567 (87.9%)	
[1,6)	3550 (10.6%)	951 (25.9%)	4501 (12.1%)	
nr.employed_bin				< 0.001
[0,5.2e+03)	19105 (57.2%)	2863 (78.0%)	21968 (59.3%)	
[5.2e+03,5.3e+03)	14291 (42.8%)	809 (22.0%)	15100 (40.7%)	
emp.var.rate_bin				< 0.001
[-4,1)	12228 (36.6%)	2635 (71.8%)	14863 (40.1%)	
[1,1.5)	21168 (63.4%)	1037 (28.2%)	22205 (59.9%)	
season				< 0.001
autumn	3854 (11.5%)	633 (17.2%)	4487 (12.1%)	
spring	14159 (42.4%)	1568 (42.7%)	15727 (42.4%)	
summer	15299 (45.8%)	1394 (38.0%)	16693 (45.0%)	
winter	84 (0.3%)	77 (2.1%)	161 (0.4%)	

Figure 2-1 Pearson's Chi-squared test

3. Building the Statistical Model

3.1 Logistic Regression

Binary Logistic Regression is used to explain the relationship between the categorical dependent variable and one or more independent variables. When the dependent variable is dichotomous, we use binary logistic regression. However, by default, a binary logistic regression is almost always called logistics regression. In our case, due to the fact that the variable "SUBSCRIPTED2" is binary (Yes or No) we proceed with binomial regression. In the above Table, we can see the summary of the preliminary full model:

	Dependent variable: SUBSCRIBED2
jobblue-collar	-0.288***
	(0.081)
jobservices	-0.239***
	(0.089)
jobtechnician	-0.103
	(0.074)
jobOther	0.093
	(0.059)
maritalmarried	-0.025
	(0.071)
maritalsingle	0.147*
	(0.079)
educationhigh.school	0.005
	(0.070)
educationilliterate	0.934
	(0.720)
educationprofessional.course	0.071
	(0.085)
educationuniversity.degree	0.123*
	(0.069)
housingyes	-0.017
	(0.043)
loanyes	-0.047
	(0.059)
contacttelephone	-0.292***
	(0.074)
day_of_weekmon	-0.112
	(0.069)
day_of_weekthu	-0.010
	(0.067)
day_of_weektue	0.077
	(0.068)
day_of_weekwed	0.038
	(0.068)
poutcomenonexistent	0.472***
	(0.070)
poutcomesuccess	2.018***
	(0.106)
cons.price.idx	0.899***
	(0.080)
cons.conf.idx	0.164***

	(0.006)
age_bin(40,100]	-0.084*
	(0.048)
dur_bin[200,400)	1.299***
	(0.062)
dur_bin[400,5.1e+03)	3.425***
	(0.060)
campaign_bin(1,2]	-0.149***
	(0.052)
campaign_bin(2,60]	-0.065
	(0.052)
nr.employed_bin[5.2e+03,5.3e+03)	1.004***
	(0.142)
emp.var.rate_bin[1,1.5)	-4.068***
	(0.141)
seasonspring	1.332***
	(0.081)
seasonsummer	1.481***
	(0.087)
Constant	-81.055***
	(7.293)
Observations	36,907
Log Likelihood	-7,713.592
Akaike Inf. Crit.	15,489.180
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 3-1 Summary of the Logistic Regression Model

3.2 Stepwise Selection AIC & BIC

The next step in conducting the logistic regression algorithm, is to fit various models by using stepwise on the full model. We used both AIC and BIC to compare their extra terms later. Specifically, below are the models of each stepwise selection:

- **AIC selects (m1 model):**
  SUBSCRIBED2 ~ job + marital + contact + day_of_week + poutcome + cons.price.idx + cons.conf.idx + age_bin + dur_bin + campaign_bin + nr.employed_bin + emp.var.rate_bin + season

- **BIC selects (m2 model):**
  SUBSCRIBED2 ~ job + contact + poutcome + cons.price.idx + cons.conf.idx + age_bin + dur_bin + nr.employed_bin + emp.var.rate_bin + season

Figure 3-2 The Models that Selected the Stepwise Methods

The next step was to conduct Wald Test (W) for the extra terms from the first model, in order to test if the extra terms do not effect of subscribing.

Wald test - Chi-squared test

X2	df	P(> X2)
32.3	8	8.3e-05

Table 3.1 Results of Wald Test

The chi-squared test statistic of 32.3 with 3 degrees of freedom is associated with a p-value of less than 8.3e-05 indicating that the overall effect of the extra terms is statistically significant. So, comparing the two selections, the best model is the "m1".

3.3 Goodness of Fit & Models Comparison

Next, we can test the proportions in a probability model using a variant of $\chi 2$ called Goodness of Fit (GOF). The $\chi 2$ GOF hypothesis compares observed data to theoretical probabilities. The null hypothesis of the test is that the model is adequate and fits the data well. The alternate hypothesis is that the model is not adequate and does not fit the data well. Since the output of model "m1" is "1" it seems that it is fitting well. Specifically, "fitting well" in our case, means the model has a specific set of parameters which can define the problem at hand. In other words the parameters that contains our model, come closest to predicting the data and therefore are close as possible to the ones observed in population.

The next step is to test whether our model "m1" fits significantly better than a model with just an intercept (null model). The test statistic is the difference between the residual deviance for the model with predictors and the null model. 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 minus the model's df. Due to the fact that the p-value is close to "0", we reject the null hypothesis that there is no significant difference between the "m1" model and the null model. So, there is significant difference between our model and the null model.

m1 *Compare to Full Model*	***m1*** *Compare to Null Model*
1	0

Table 3.2 Goodness Of Fit Outputs for Model m1

3.4 Final Model & Interpretation

After choosing the model with AIC selection criteria and conducting the appropriate tests the output from the optimal model as described is the following:

	Dependent variable:
	SUBSCRIBED2
jobblue-collar	-0.352***
	(0.066)
jobservices	-0.300***
	(0.084)
jobtechnician	-0.109
	(0.066)
jobOther	0.082
	(0.057)
maritalmarried	-0.026
	(0.071)
maritalsingle	0.152*
	(0.078)
contacttelephone	-0.294***
	(0.074)
day_of_weekmon	-0.111
	(0.069)
day_of_weekthu	-0.007
	(0.067)
day_of_weektue	0.076
	(0.068)
day_of_weekwed	0.037
	(0.068)
poutcomenonexistent	0.473***
	(0.070)
poutcomesuccess	2.018***
	(0.106)
cons.price.idx	0.898***
	(0.080)
cons.conf.idx	0.165***
	(0.006)
age_bin(40,100]	-0.090*
	(0.048)
dur_bin[200,400)	1.298***
	(0.062)
dur_bin[400,5.1e+03)	3.422***
	(0.060)
campaign_bin(1,2]	-0.150***
	(0.052)
campaign_bin(2,60]	-0.064
	(0.052)
nr.employed_bin[5.2e+03,5.3e+03)	0.999***
	(0.142)
emp.var.rate_bin[1,1.5)	-4.067***
	(0.140)
seasonspring	1.327***
	(0.081)
seasonsummer	1.484***
	(0.087)
Constant	-80.886***
	(7.291)
Observations	36,907
Log Likelihood	-7,717.249
Akaike Inf. Crit.	15,484.500
	*p<0.1; **p<0.05; ***p<0.01

Table 3.3 Final Model coefficients estimates and std. error in parentheses

An important concept for interpreting the logistic beta coefficients, is the odds ratio. An odds ratio measures the association between a predictor variable (x) and the outcome variable (y). It represents the ratio of the odds that an event will occur (event = 1) given the presence of the

predictor x (x = 1), compared to the odds of the event occurring in the absence of that predictor (x = 0). For a given predictor (say x1), the associated beta coefficient (b1) in the logistic regression function corresponds to the log of the odds ratio for that predictor.

The interpretation of our model is the following:

$$\log\left[\frac{P(\widehat{\text{SUBSCRIBED2}} = 1)}{1 - P(\widehat{\text{SUBSCRIBED2}} = 1)}\right]$$
$$= -80.886 - 0.352(\text{job}_{\text{blue-collar}}) - 0.3(\text{job}_{\text{services}}) - 0.109(\text{job}_{\text{technician}})$$
$$+ 0.082(\text{job}_{\text{Other}}) - 0.026(\text{marital}_{\text{married}}) + 0.152(\text{marital}_{\text{single}})$$
$$- 0.294(\text{contact}_{\text{telephone}}) - 0.111(\text{day\_of\_week}_{\text{mon}})$$
$$- 0.007(\text{day\_of\_week}_{\text{thu}}) + 0.076(\text{day\_of\_week}_{\text{tue}})$$
$$+ 0.037(\text{day\_of\_week}_{\text{wed}}) + 0.473(\text{poutcome}_{\text{nonexistent}})$$
$$+ 2.018(\text{poutcome}_{\text{success}}) + 0.898(\text{cons.price.idx}) + 0.165(\text{cons.conf.idx})$$
$$- 0.09\left(\text{age\_bin}_{(40,100]}\right) + 1.298\left(\text{dur\_bin}_{[200,400)}\right)$$
$$+ 3.422\left(\text{dur\_bin}_{[400,5.1e+03)}\right) - 0.15\left(\text{campaign\_bin}_{(1,2]}\right)$$
$$- 0.064\left(\text{campaign\_bin}_{(2,60]}\right) + 0.999\left(\text{nr.employed\_bin}_{[5.2e+03,5.3e+03)}\right)$$
$$- 4.067\left(\text{emp.var.rate\_bin}_{[1,1.5)}\right) + 1.327(\text{season}_{\text{spring}})$$
$$+ 1.484(\text{season}_{\text{summer}}).$$

Table 3.4 Interpretation Of the Final Model

- We observe that the odds of having successful subscription when job=admin, marital=divorced, education=basic, contact=cellular, day_of_week=friday, poutcome=failure, cons.price.id=0, cons.conf.idx=0, age_bin(15,40], dur_bin=[0.200), campaign_bin= (0,1], nr.employed_bin = [0,5.2e+03), emp.var.rate_bin= [-4,1) and season = autum, is **0.00**.
- The effect of job [blue-collar] and job [services] is statistically significant and negative. The odds of having a successful subscription compared to the admin (included in the Intercept), decrease **by 0.70** for every unit increase in job type of blue-collar with the rest variables fixed, while for every unit increase in job type of services decrease by **0.74**, with the rest variables fixed.
- The effect of contact [telephone] is statistically significant and negative. The odds of having a successful subscription compared to the cellular (included in the Intercept), decrease by **0.74** for every unit of telephone contact with the rest variables fixed. That makes sense due to the fact that cellular phones offer more portability, and people are carrying most of the time with them.

- The effect of poutcome [nonexistent] and [success], is statistically significant and positive. The odds of having a successful subscription when the previous marketing campaign does not exist, compared to a previous campaigned with failure (included in the Intercept), increase **by 1.6** for every unit increase in campaign, with the rest variables fixed. Also, when the outcome was success, the odds of having a successful subscription compared to the failure, increases **by 7.5** for every unit increase with the rest variables fixed. That make senses since when a marketing campaign is not successful the clients possibly will have less interest to participate. So the successful previous campaigns, followed by those who'd never experienced a campaign, are more likely to generate successful subscriptions .
- The effect of consumer price index - monthly indicator (CPI) proved to be statistically significant and positive. The odds of having a successful subscription by increasing one unit of the price indicator is **2.45** with the rest variables fixed**.** When clients keep money in the bank, they may earn interest, which balances out some of the effects of inflation. When inflation is high, banks typically pay higher interest rates, meaning people receive a better return on their savings, which may encourage them to save rather than spend. In any case their behavior would depends on many other parameters such as their income.
- The effect of consumer confidence index (CCI) proved to be statistically significant and positive. The odds of having a successful subscription by increasing one unit of the consumer confidence index is **1.17** with the rest variables fixed. That seems unexpected, since, if consumers are optimistic about their financial situations the tend to spend more and save less.
- The effect of the call durations groups between 200 and 400 seconds and over 400 seconds proved to be statistically significant and positive. The odds of having a successful subscription by increasing one second in both duration groups, compared to the duration group 15 up to 40 seconds (included in the Intercept), are increased by **3.66** (for the group 200-400 sec) or **30.63** (for the group over 400 sec) with the rest variables fixed. That seems sensible since may people at seconds be hesitating or they need time to have whole image of the offer.
- The effect of the number of contacts performed during this campaign and for this client for the group (1,2] proved to be statistically significant and positive. The odds of having a successful subscription by increasing one unit in contacts in group (1,2] compared to the group campaign_bin= (0,1] (included in the Intercept), are decreased by **0.86** for every unit increase in number of contacts with the rest variables fixed.

- The effect of the number of employees in group **[5.2e+03, 5.3e+03)** proved to be statistically significant and positive. The odds of having a successful subscription by increasing one unit in group of employees [5.2e+03,5.3e+03), are increased **by 2.72** compared to the group [0,5.2e+03) (included in the Intercept) with the rest variables fixed. That seems logical seems more employees will increase the amount of phone calls and therefore possibly the subscriptions.
- The effect of employment variation rate (the variations od hiring or firing) group [1,1.5) proved to be statistically significant and negative. The odds of having a successful subscription by increasing one unit in contacts in group [1,1.5) competed to the group [-4,1) (included in the Intercept), are decreased by **0.01** for every unit increase with the rest variables fixrd.
- The effect of the seasons (last contact in seasons of the year) Spring and Summer proved to be statistically significant and positive. The odds of having a successful subscription by increasing one unit in both seasons, compared to the Autumn (included in the Intercept) are increased either **3.76** for Spring or **4.41** for Summer, with the rest variables fixed.

Next, visualizing the coefficients of the model can give us useful insights for our model. Specifically, the coefficient plot visualizes the confidence intervals and their corresponding regression estimates. The bellow Figure 3-3, is centered around zero, meaning that any estimate that crosses zero is statistically not significant. For example the variables “day_of_weekend”, “jobtechnician”, “jobOther” and “maritalmarried” are not statistically significant.

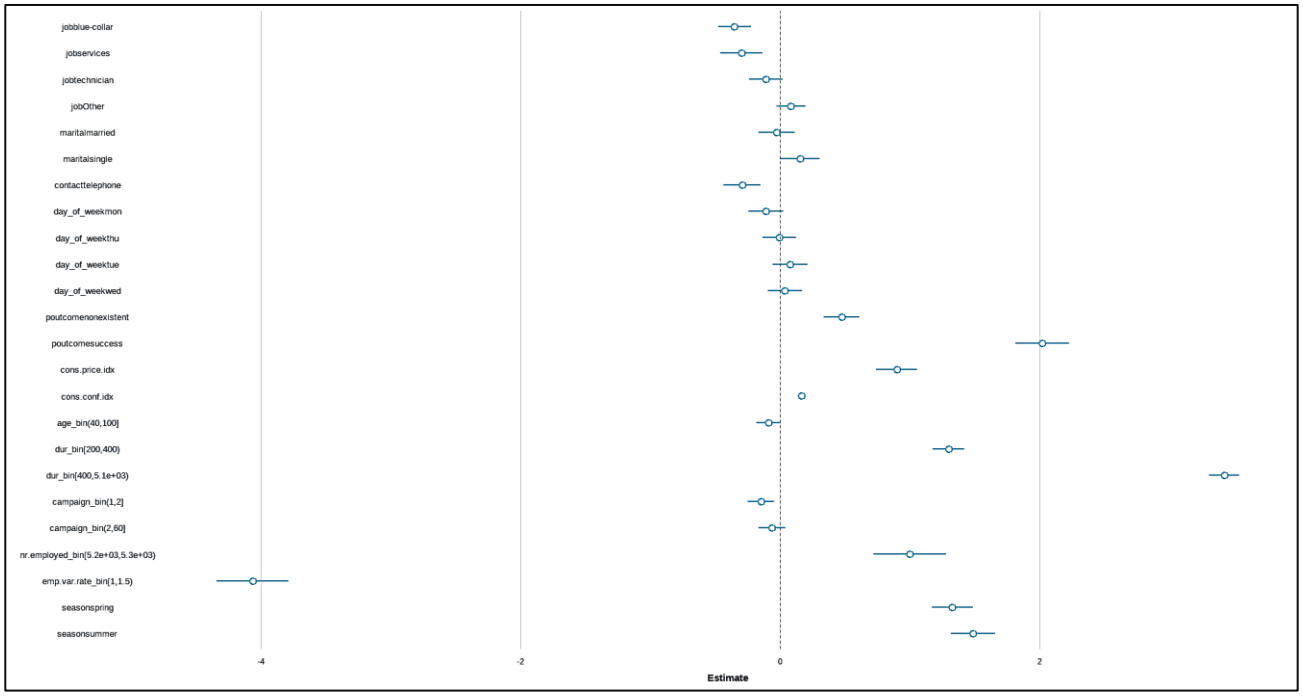

Figure 3-3 Plot with the Coefficients of the Model

Then, we proceed with visualizations of marginal effects (i.e., “average partial effects”) to understand the shape of relationship between the predictors and outcome and the differences in probabilities. For example in the next Figure 3-4, we visualize the relationship between the model predictor “cons.conf.idx” (independent variable) with the model response “SUBSCRIBED2” (dependent variable). Specifically, this plot shows how the probability of “SUBSCRIBED2” varies with the cons.conf.idx variable. We can conclude that the effects is positive and stronger at higher values of the predictor.

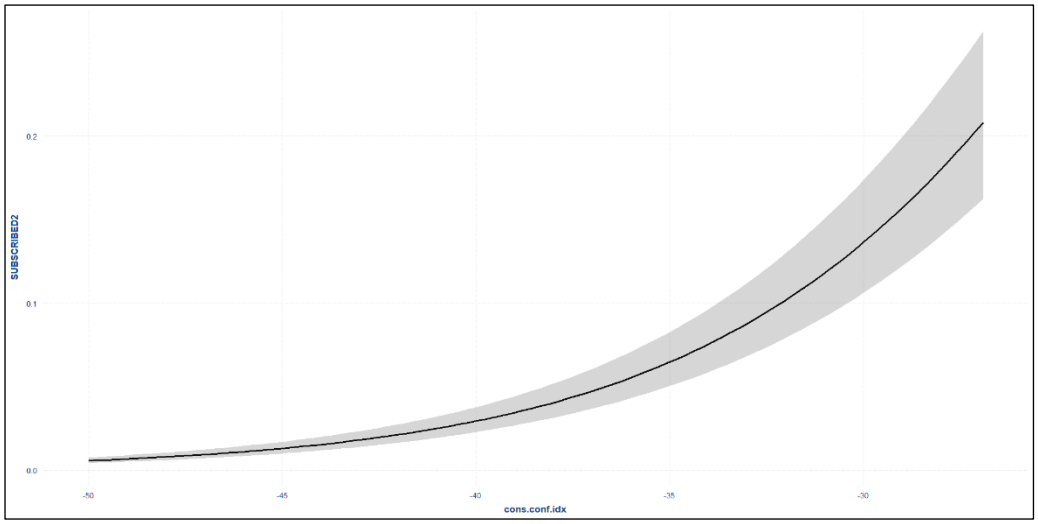

Figure 3-4 How the probability of “SUBSCRIBED2” varies with the “cons.conf.idx” variable

The next Figure 3-5, makes it clear how the consumer confidence index affect the probability of subscription in the two employment groups (-4,1), (1,1.5). The effect of the consumer confidence index is stronger in the employment group (-4,1) compared to the second group (1,1.5).

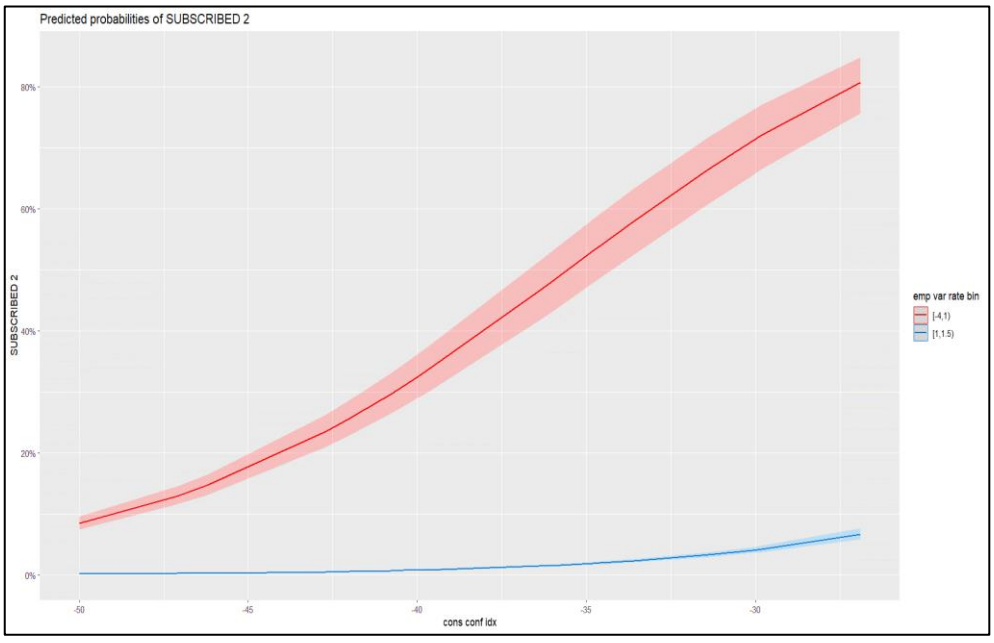

Figure 3-5 How the consumer confidence index affect the probability of subscription in the two employment groups (-4,1) , (1,1.5).

The next Figure, makes it clear how the consumer confidence index, affect the probability of subscription, depending on the type of job (admin, blue-colar, services, technician and others) per employment group (-4,1) and (1,1.5). The effect of the job type is stronger in the employment group (-4,1) compared to the second group (1,1.5). Also, the probability is always higher for admin and blue collar, and services compared to the other jobs.

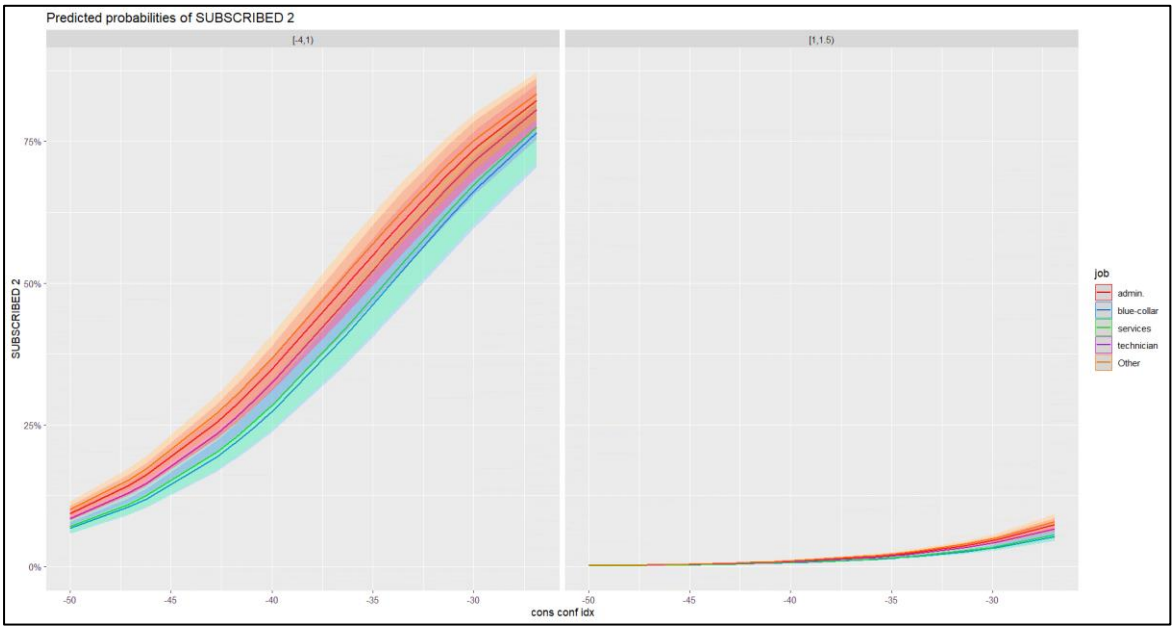

Figure 3-6 How the consumer confidence index, affects the probability of subscription, depending on the type of job per employment group (-4,1) and (1,1.5).

The next graph makes it clear how the consumer confidence index affect the probability of subscription in the three duration groups: [0,200), [200,400), [400,5.1e+03). The effect of the consumer confidence index is stronger in the duration group [400,5.1e+03) followed by the group [200,400). Thus, duration and consumer confidence index both have a positive effect.

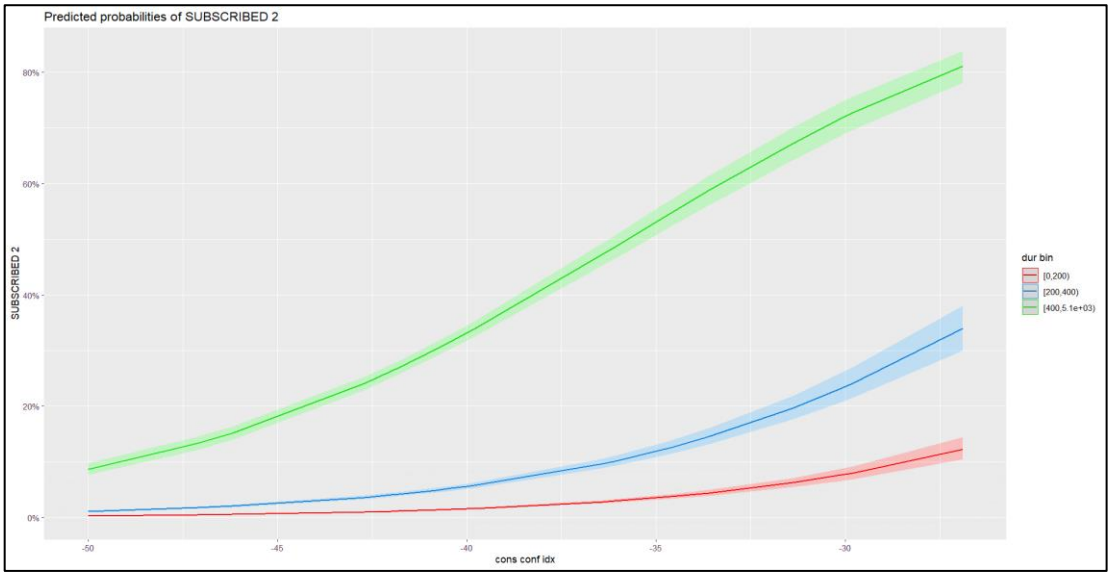

Figure 3-7 How the consumer confidence index affect the probability of subscription in the three duration groups: [0,200), [200,400), [400,5.1e+03).

The next graph makes it clear how the type of job (admin, blue-colar, services, technician and others) affect the probability of subscription in the three duration groups: [0,200) [200,400) [400,5.1e+03). The effect of job type is stronger in the duration groups [400,5.1e+03) followed by the group [200,400) .

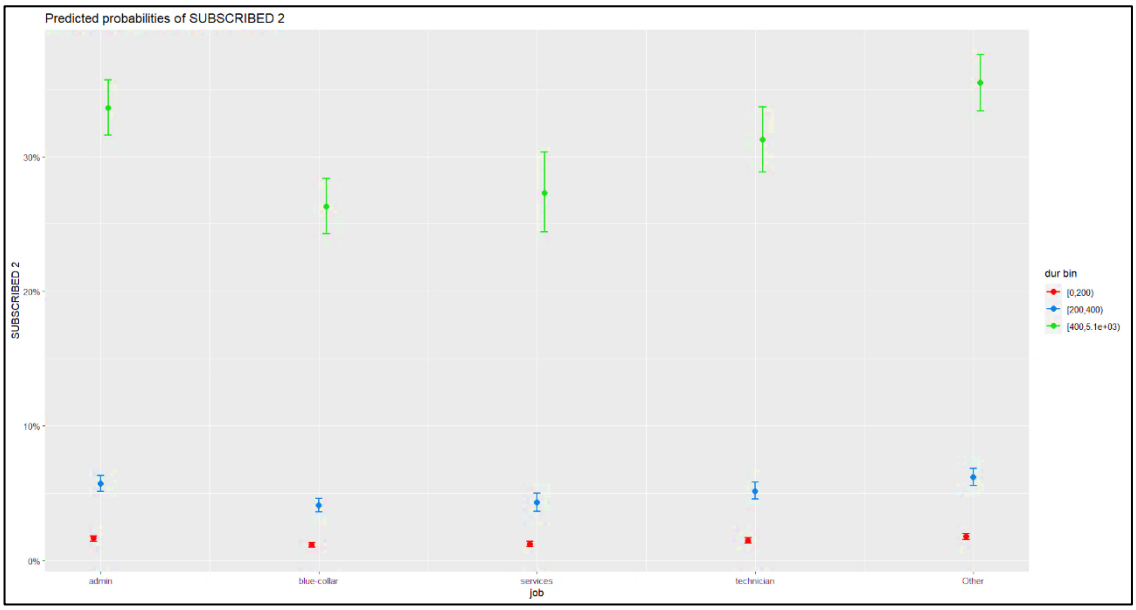

Figure 3-8 How the type of job (admin, blue-colar, services, technician and others) affect the probability of subscription in the three duration groups: [0,200) [200,400) [400,5.1e+03).

The next graph makes it clear how the outcome of the previous marketing campaign (categorical: 'failure',' nonexistent', 'success') affect the probability of subscription in the three duration groups: [0,200), [200,400), [400,5.1e+03). Longer duration has a key role in subscription probability, even when the previous outcome was failure, and even more so if the previous outcome was successful.

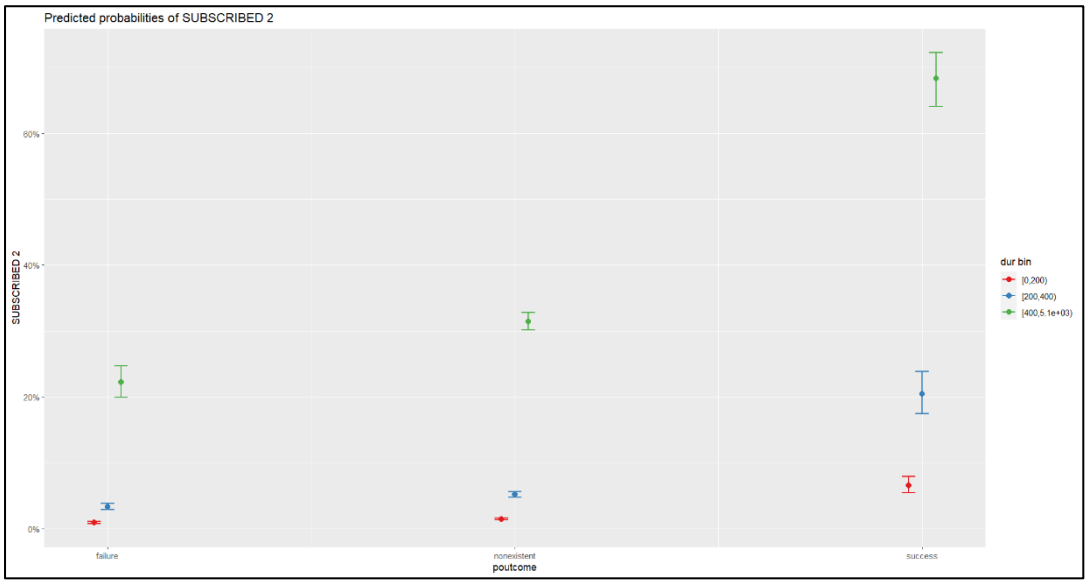

Figure 3-9 How the outcome of the previous marketing campaign (categorical: 'failure',' nonexistent', 'success') affect the probability of subscription in the three duration groups: [0,200), [200,400), [400,5.1e+03).

4. Linearity Assumption

The logistic regression assumes that there is a linear relationship between the logit of the outcome and each predictor variable. This is done visually in the next Figure, inspecting the scatter plot between the standardized (Pearson's) residuals and the numerical predictors “cons.price.idx” and “cons.conf.idx”. If the assumption is fulfilled, we expect to see an approximately straight horizontal line, indicative of no non-linear patterns (as is the case in the next Figure 4-1).

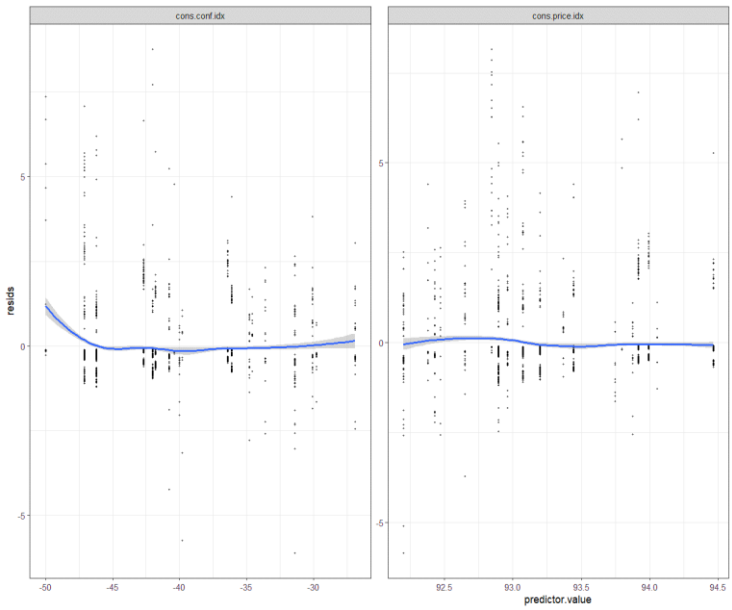

Figure 4-1 Describes how the residuals are quite linearly associated with the outcome “SUBSCRIBED2”.

5. Conclusions

Conducting the above analysis with different tests and plots we ended up with a final model that fits well with our data set. The methodology followed was to first undergo data cleaning and transformation in order to achieve the highest possible accuracy later on in our analysis. Beyond that we created plots for a deeper look into the shapes of the relationships and patterns, without the need for complex summary statistics. Third, implemented Stepwise methods AIC and BIC, in order to aid in the examination of the statistical significance of each predictor variable and to remove the insignificant ones from our model. Fourth, we checked the linearity assumption of our model. Interpreting the model parameters, was key to understand deeper some of our findings. Taking into account the effect of each parameter on odds ratio, the most important parameters were the duration, the employment variation rate and the consumer confidence index.

6. Appendix

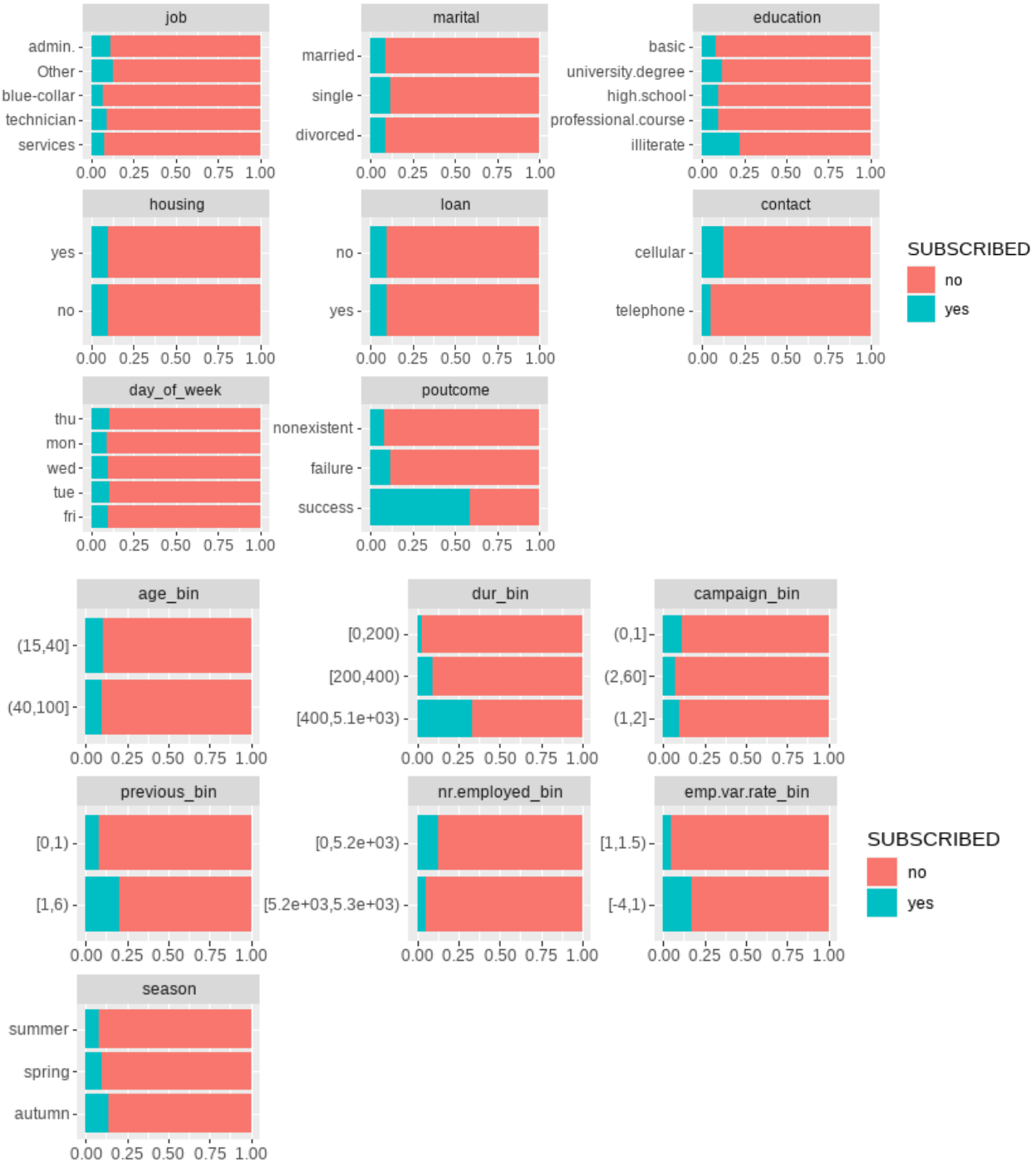

Figure 6-1 Subscription per Categorical Variable

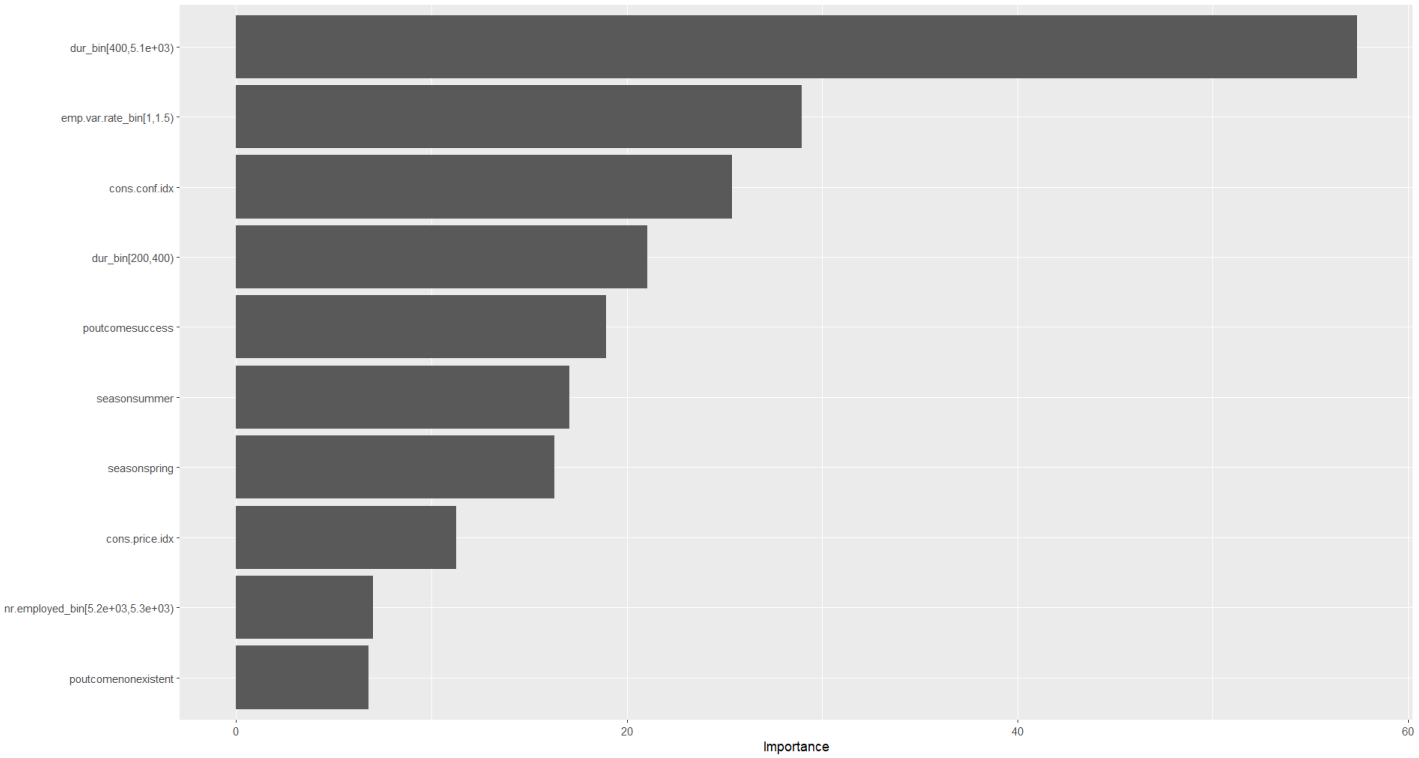

Figure 6-2 Model m1 Variable Importance Plot