Advanced Statistical Modelling: Logistic Regression

Ricard Meyerhofer & Joel Cantero

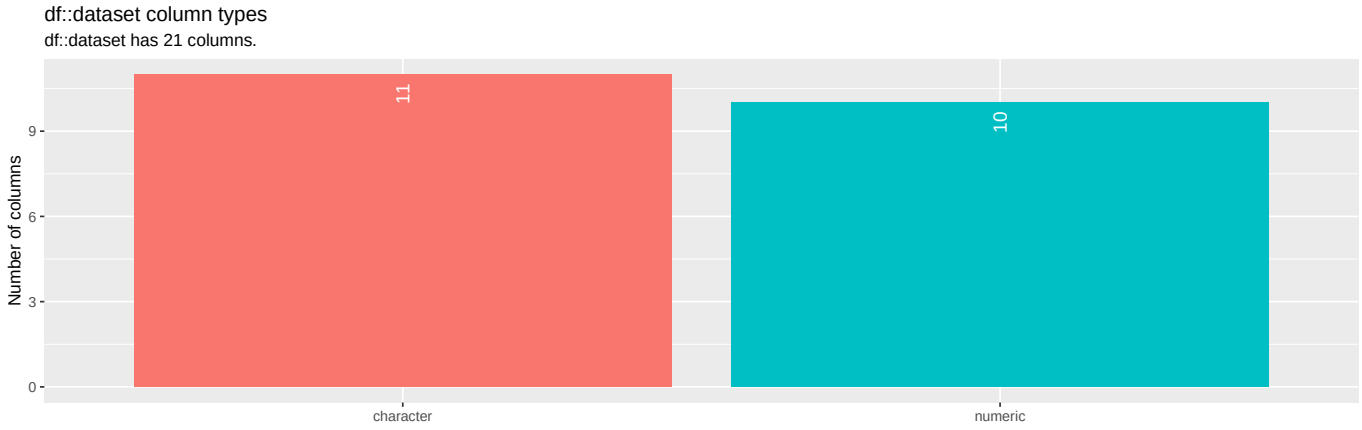
4/11/2019

Exploratory data analysis

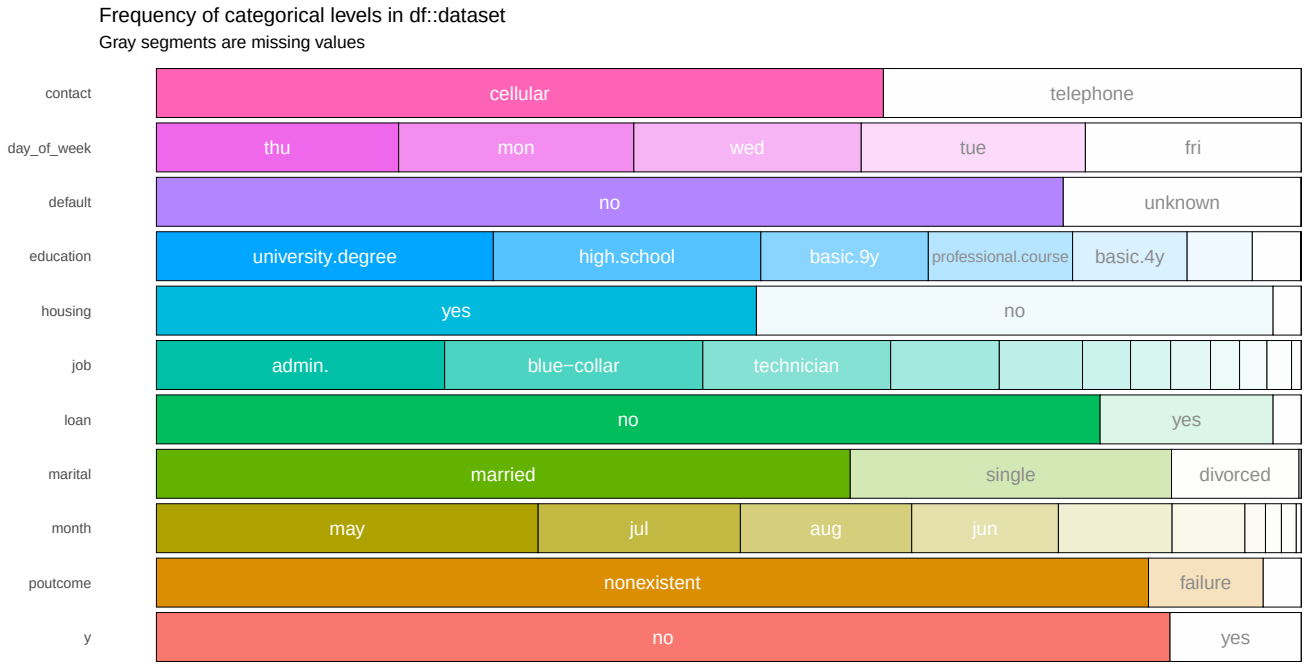
As explained in the problem statement, our dataset is composed by 28645 calls from JYB. JYB has the purpose of reducing the telemarketing costs by decreasing the number of calls to clients not likely to buy the product. This is the list of the available variables:

Variable	Description	Attribute type
id	Customer ID	Client
age	age in years	Client
job	(admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown)	Client
marital	Marital status (Divorced, married, single, unknown)	Client
education	Education level (basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown)	Client
default	is he/she a defaulter? (No, yes, unknown)	Client
housing	does he/she has a mortgage? (No, yes, unknown)	Client
loan	does he/she has a personal loan? (No, yes, unknown)	Client
contact	phone type (cellular, telephone)	Call
month	month of the call	Call
day_of_week	day of the call (mon, tue, wed, thu, fri)	Call
campaign	Number of contacts made this campaign for this client (including the current one)	Campaign
pdays	number of days that have passed since the customer was contacted for the last time for a previous campaign (999 means that it was not previously contacted)	Campaign
previous	number of calls made to this client before this campaign	Campaign
poutcome	previous campaign result (failure, nonexistent, success)	Campaign
emp.var.rate	employment variation rate (quarterly)	Indicators
cons.price.idx	Consumer Price Index (monthly)	Indicators
cons.conf.idx	Consumer confidence index (monthly)	Indicators
euribor3m	euribor a 3 mesos (daily)	Indicators
nr.employed	number of employed (quarterly)	Indicators
Y	The customer subscribed the deposit? (yes,no)	Response

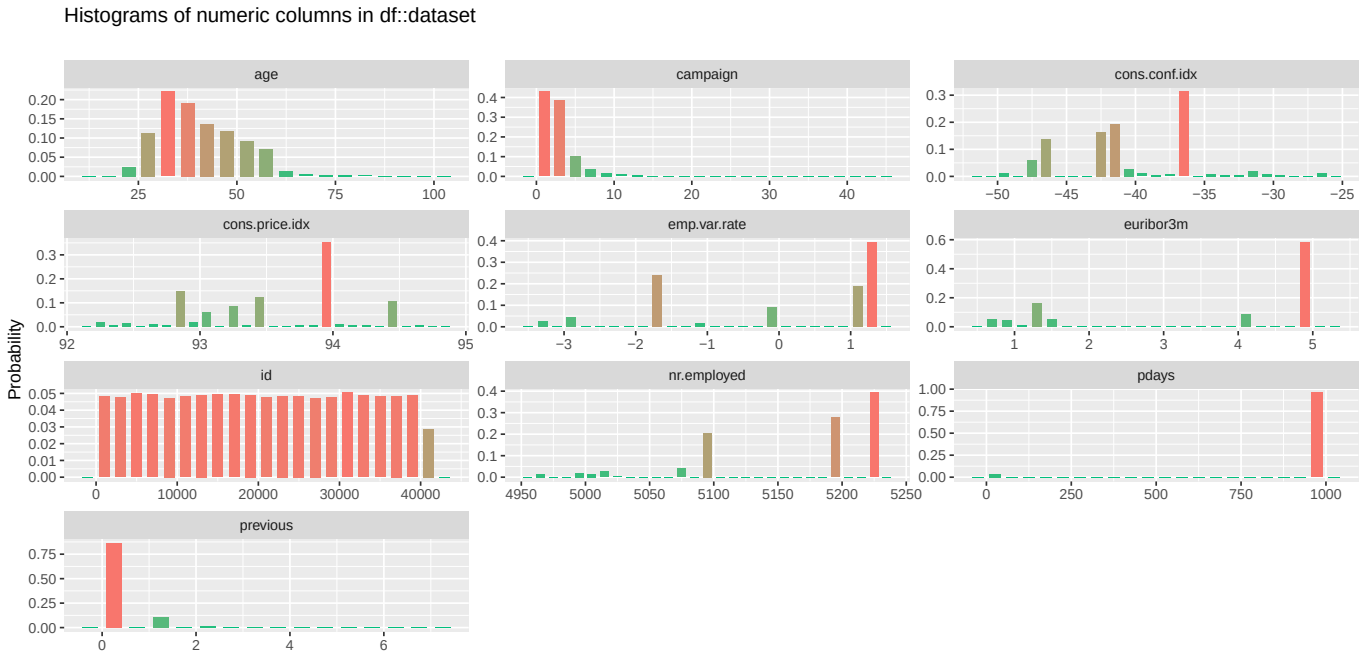
As we can see in the plot, we have 11 factors and 10 numeric values. Furthermore, we have seen that there our dataset is complete which means that it has no missing values. However, this does not imply that there are no outliers.



Exploration of the categorical variables where we can see a map with all their possible values and their representation. We can see that there is a clear unbalance in some variables such as *poutcome*, *load*, *contact* or the response variable y. So we might need to resample our dataset.



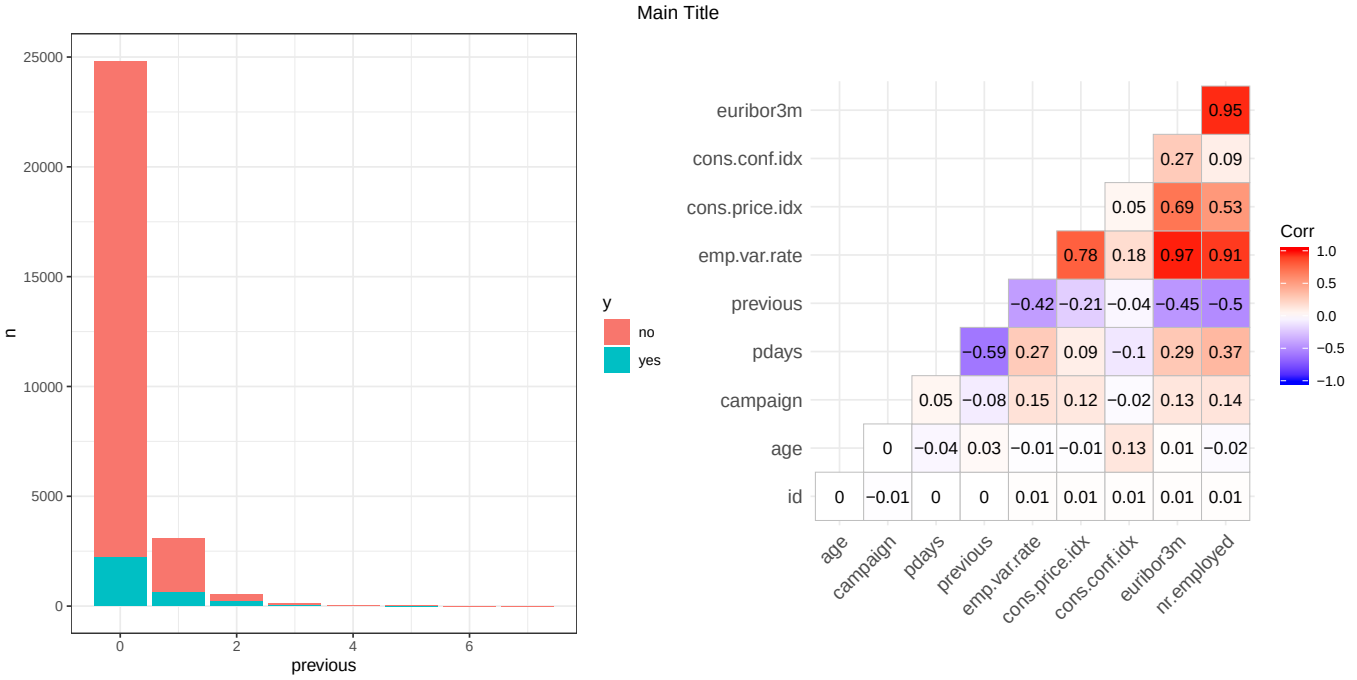
If we now explore of the numerical variables, we see that most of variables are not following a normal. This might be because they are ratios and also are taken most of them in a monthly basis.



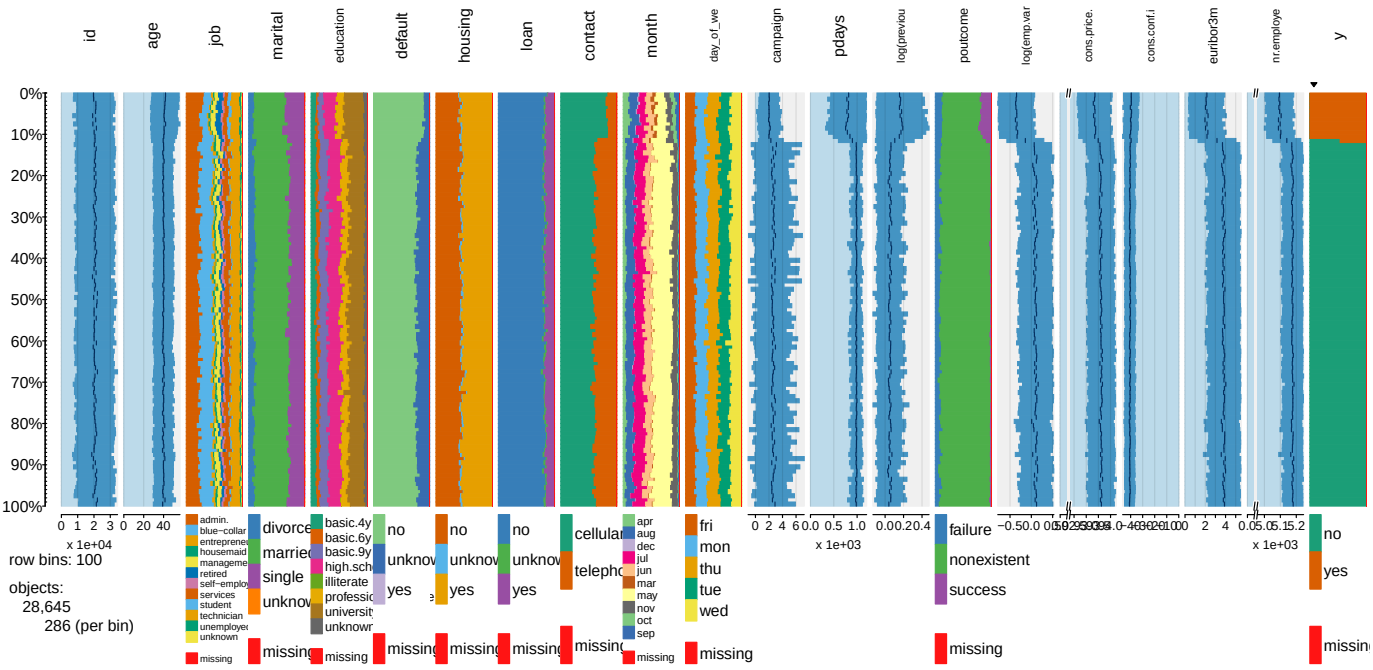
We can also see that the correlation between the variable *previous* and *pdays* is of 0.50 which means that they are moderately correlated. But if we pay attention to number of 0's and 999 and their description, we can see that they are describing the same thing. So we are going to see how these variables are related to the response variable and as we can see below, *pdays* is not necessary. Furthermore, if we take a look at the correlation matrix, we can see that there is a very high correlation between *euribor*3*m* and *emp.var.rate*, *emp.var.rate* with *nr.employed* and *nr.employed* with *euribor*3*m*.

This variables could be dropped but since in the next statement says that we should use the original variables, we are

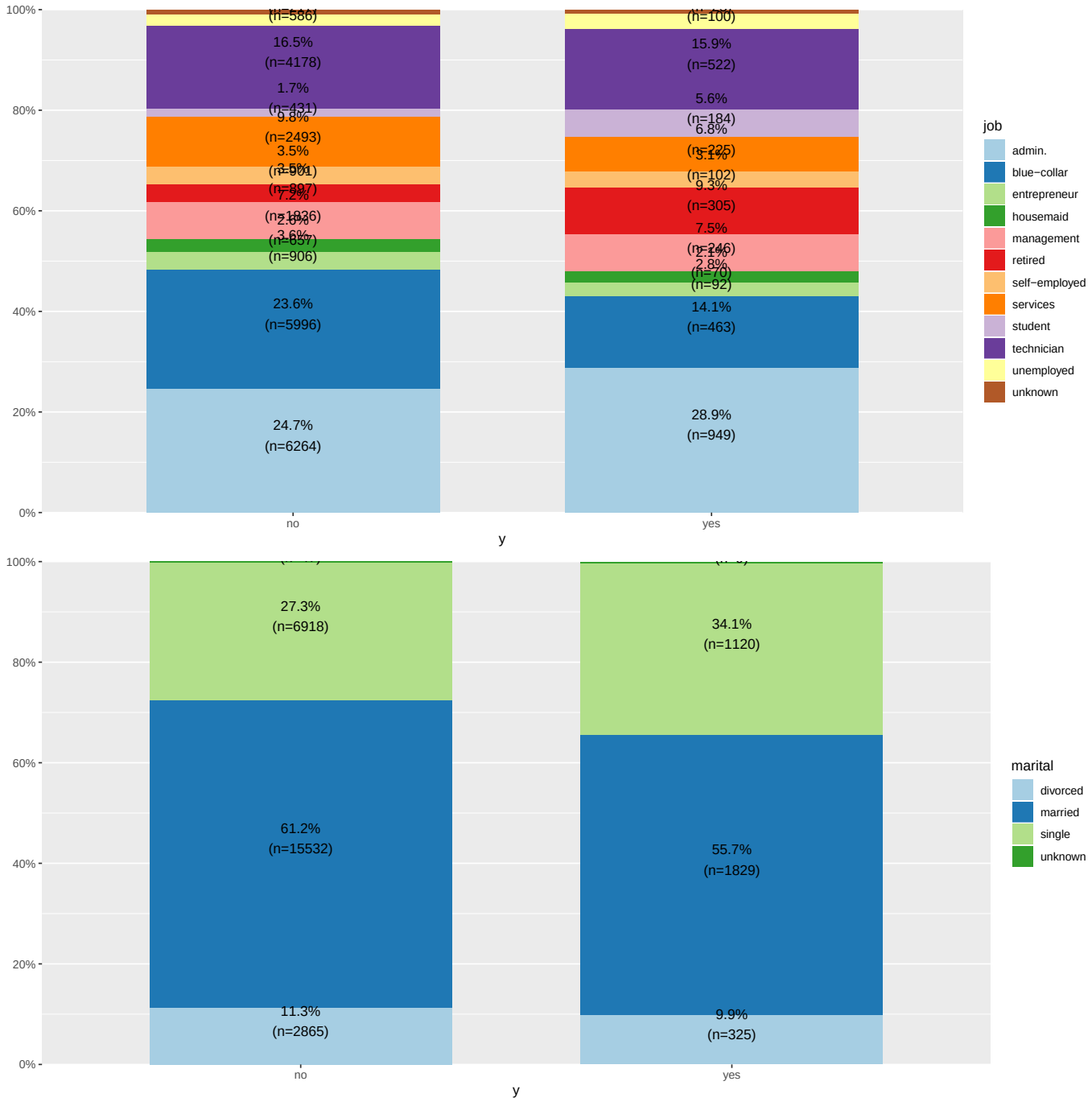
not dropping them even that at first glance, looks more than obvious that this variables will not be relevant with the AIC tests.

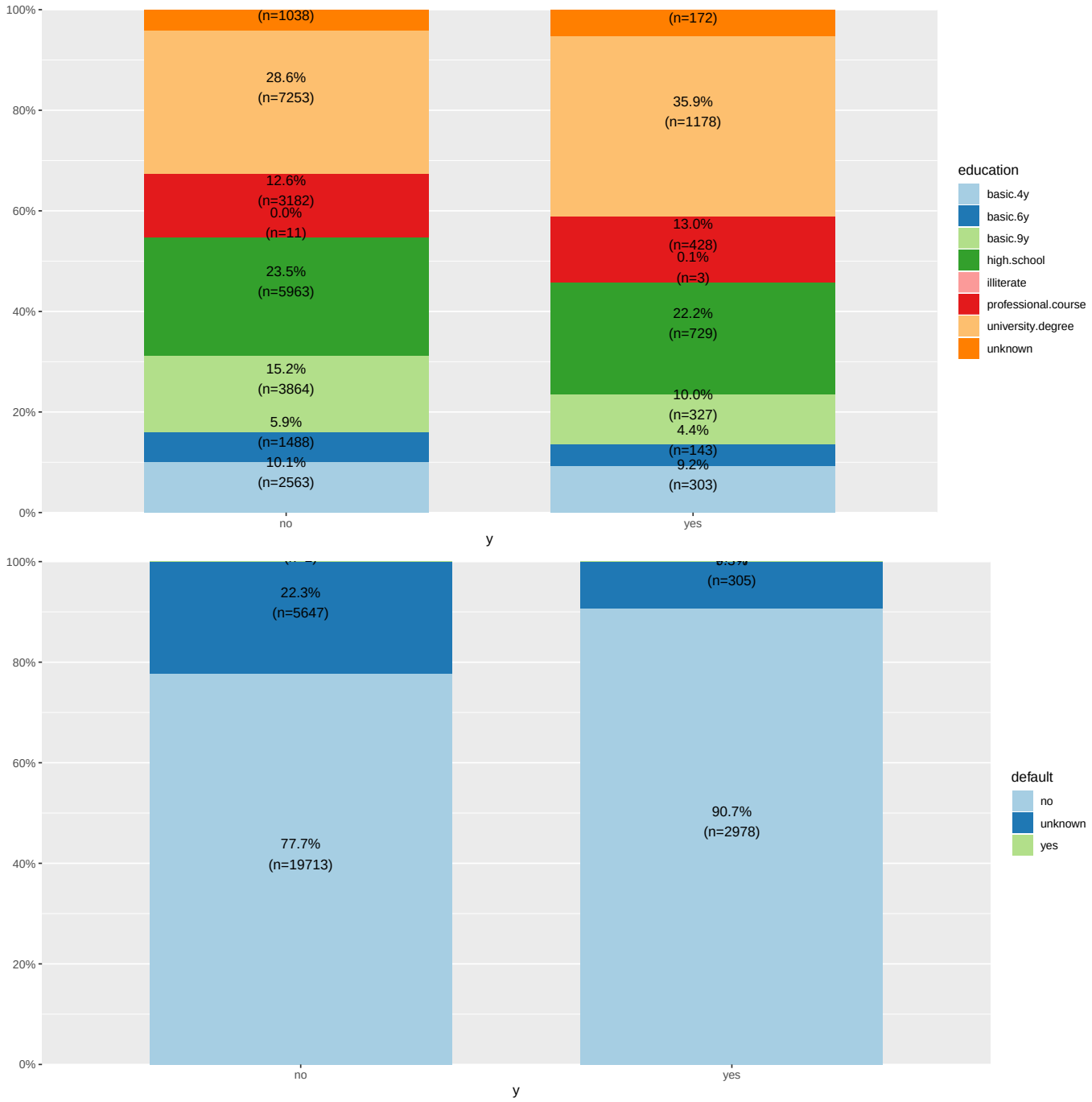


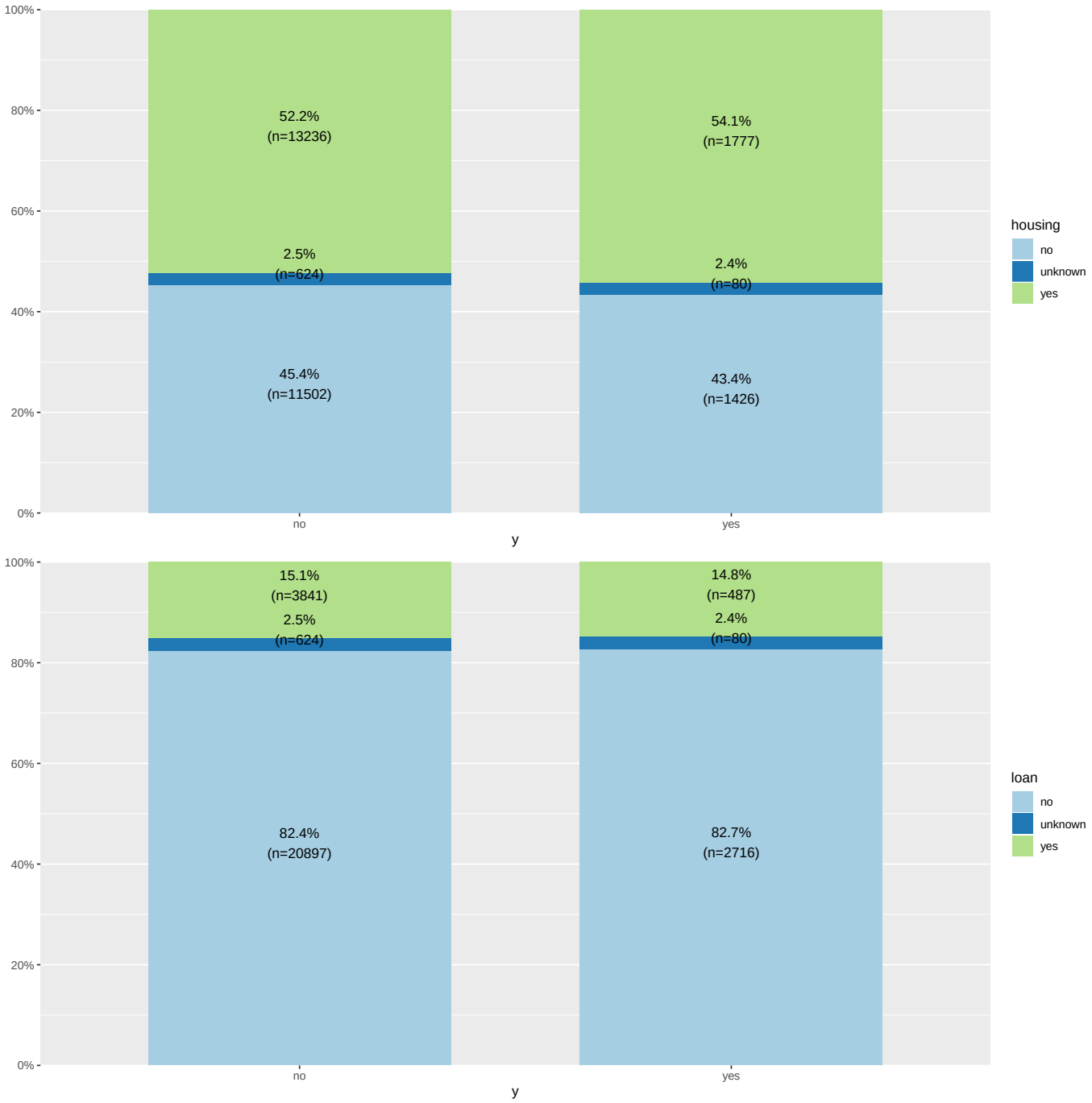
Also we can do a general plot to have a first idea of which variables help us to distinguish between the two groups. As we can see below, it looks like *housing*, *contact* and *potcome* distinguish between the two groups.

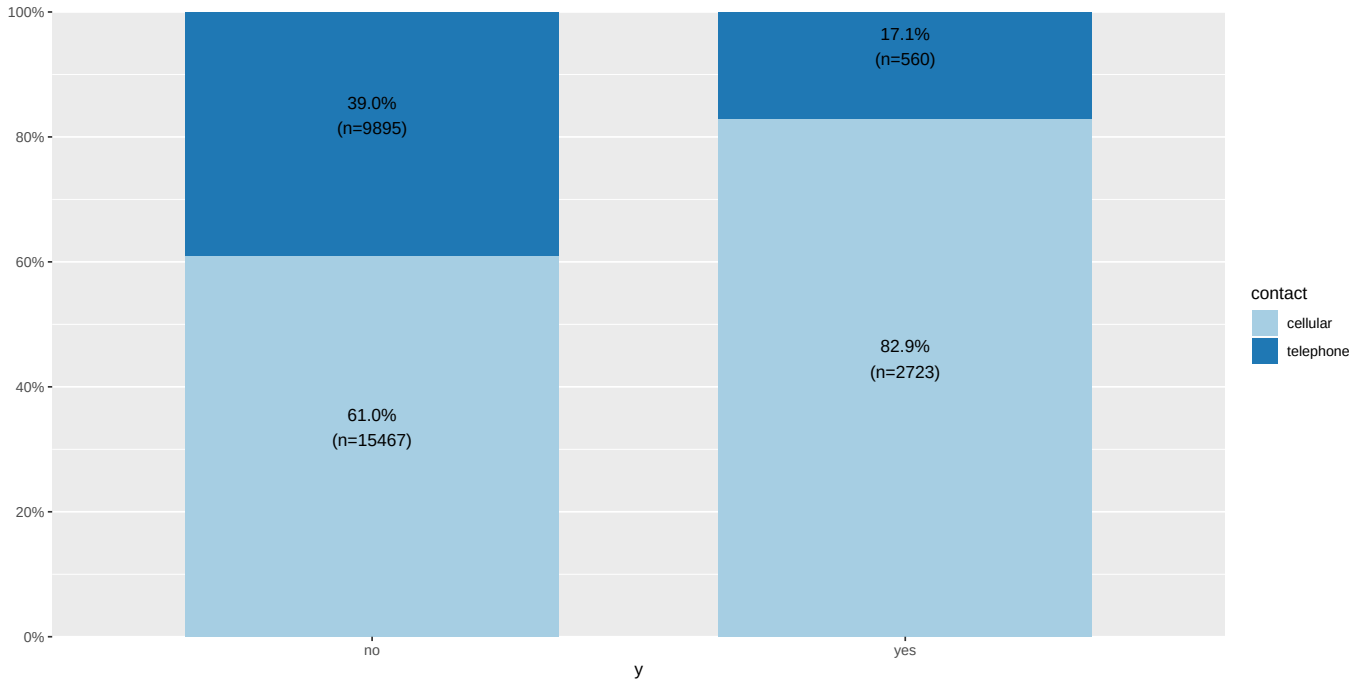


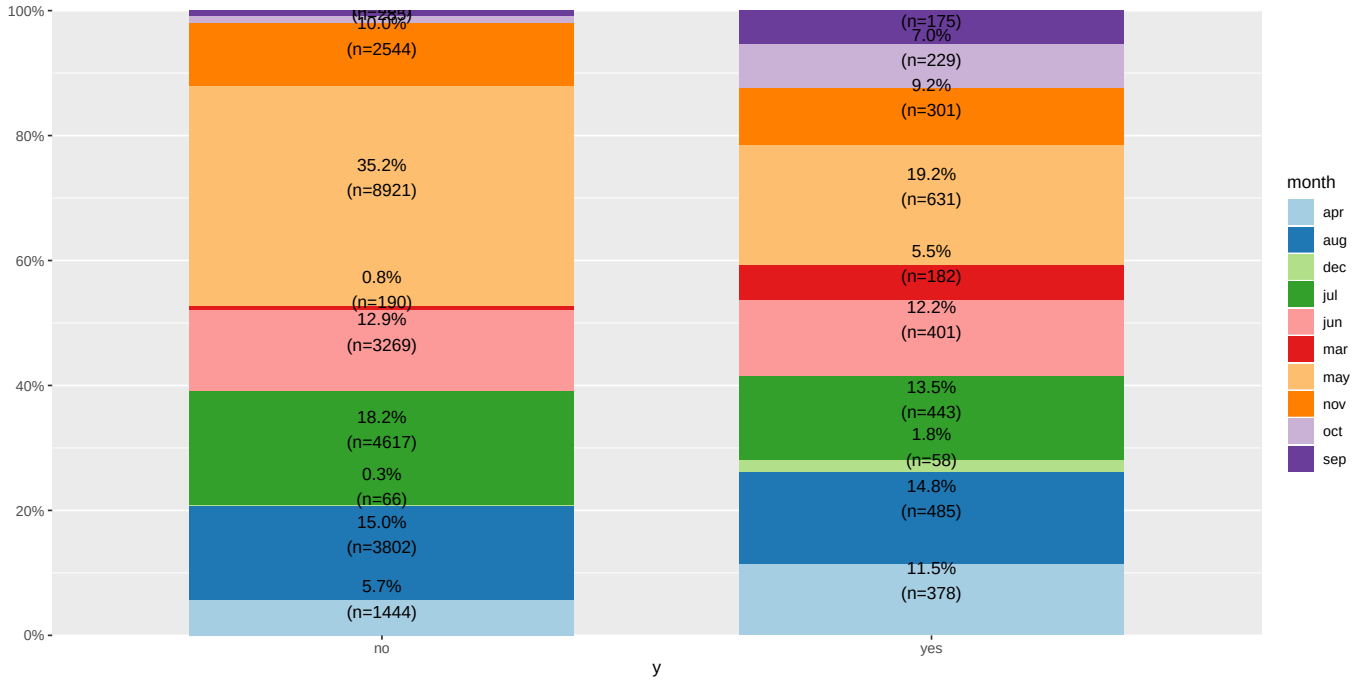
Finally, we can see all the categorical variables by the response variable. We can see that most of them behave in a very similar manner even that there are some differences in the *default*, *contact* and *poutcome* distributions.

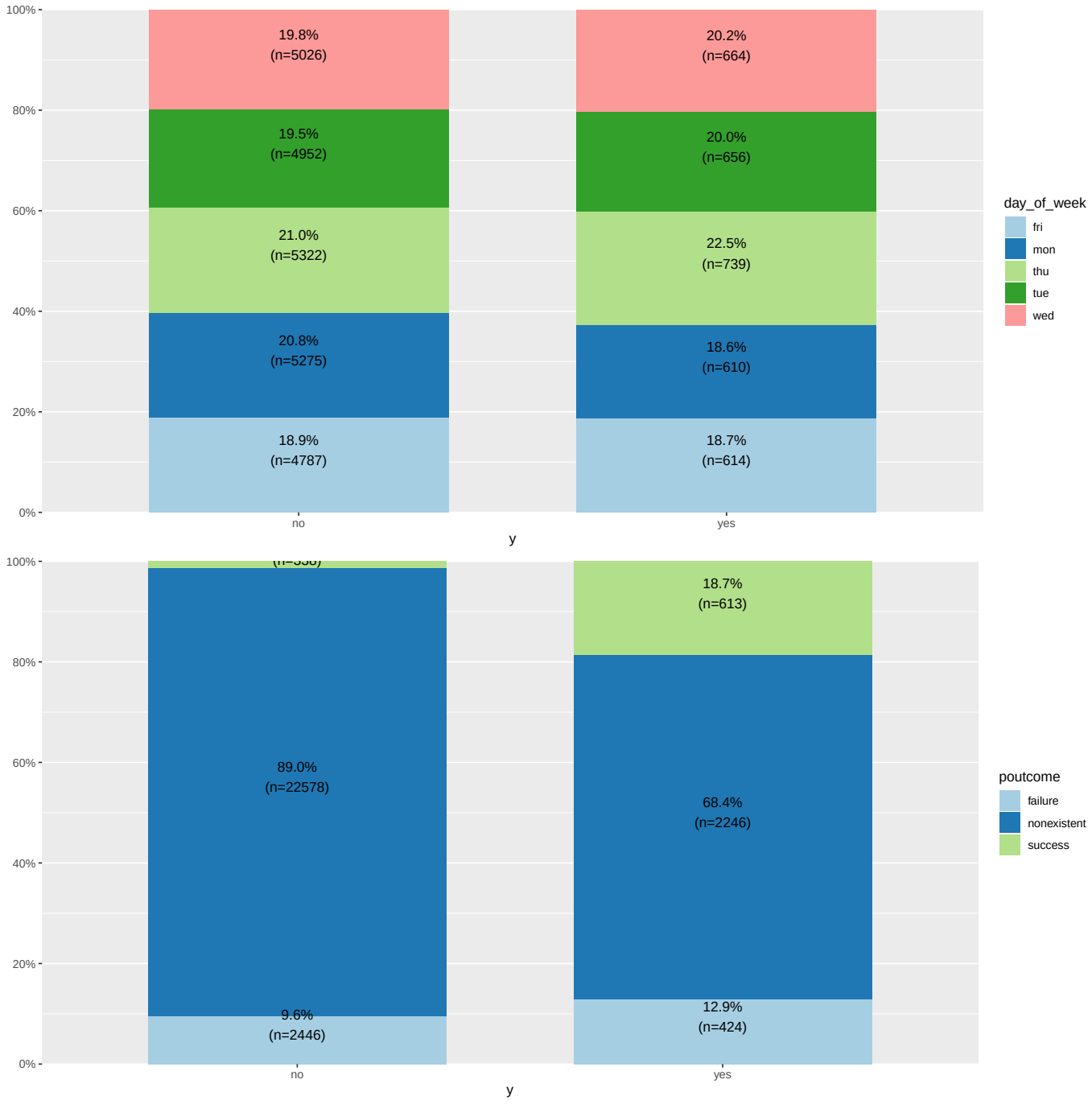

100%
80%
60%
40%
20%
0%
16.5%
(n=4178)
1.7%
(n=431)
9.8%
(n=2493)
3.5%
(n=901)
7.2%
(n=1836)
3.6%
(n=906)
23.6%
(n=5996)
24.7%
(n=6264)
15.9%
(n=522)
5.6%
(n=184)
6.8%
(n=225)
3.1%
(n=102)
9.3%
(n=305)
7.5%
(n=246)
2.8%
(n=92)
14.1%
(n=463)
28.9%
(n=949)
no
yes
y
job
admin.
blue−collar
entrepreneur
housemaid
management
retired
self−employed
services
student
technician
unemployed
unknown
27.3%
(n=6918)
61.2%
(n=15532)
11.3%
(n=2865)
34.1%
(n=1120)
55.7%
(n=1829)
9.9%
(n=325)
no
yes
y
marital
divorced
married
single
unknown


100%
80%
60%
40%
20%
0%
(n=1038)
28.6%
(n=7253)
12.6%
(n=3182)
0.0%
(n=11)
23.5%
(n=5963)
15.2%
(n=3864)
5.9%
(n=1488)
10.1%
(n=2563)
(n=172)
35.9%
(n=1178)
13.0%
(n=428)
0.1%
(n=3)
22.2%
(n=729)
10.0%
(n=327)
4.4%
(n=143)
9.2%
(n=303)
no
yes
y
education
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown
22.3%
(n=5647)
77.7%
(n=19713)
(n=305)
90.7%
(n=2978)
default
no
unknown
yes

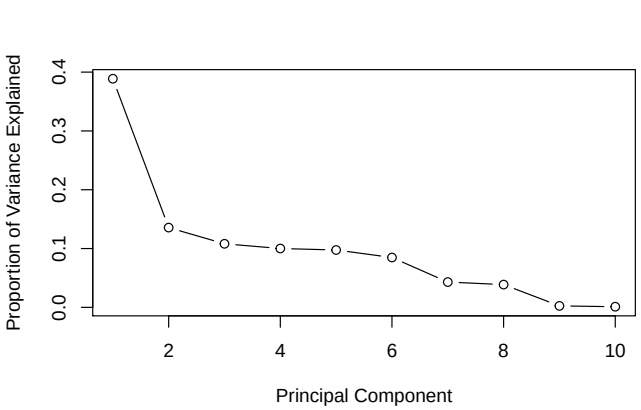

100%
80%
60%
40%
20%
0%
52.2%
(n=13236)
2.5%
(n=624)
45.4%
(n=11502)
54.1%
(n=1777)
2.4%
(n=80)
43.4%
(n=1426)
no
yes
y
housing
no
unknown
yes
15.1%
(n=3841)
2.5%
(n=624)
82.4%
(n=20897)
14.8%
(n=487)
2.4%
(n=80)
82.7%
(n=2716)
loan

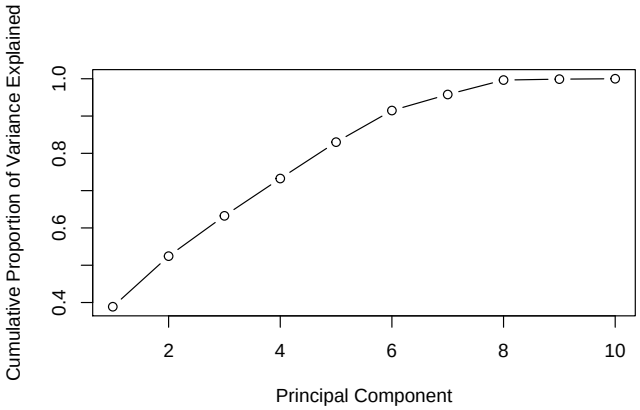

100%
80%
60%
40%
20%
0%
39.0%
(n=9895)
61.0%
(n=15467)
17.1%
(n=560)
82.9%
(n=2723)
no
yes
y
contact
cellular
telephone


100%
80%
60%
40%
20%
0%
10.0%
(n=2544)
35.2%
(n=8921)
0.8%
(n=190)
12.9%
(n=3269)
18.2%
(n=4617)
0.3%
(n=66)
15.0%
(n=3802)
5.7%
(n=1444)
(n=175)
7.0%
(n=229)
9.2%
(n=301)
19.2%
(n=631)
5.5%
(n=182)
12.2%
(n=401)
13.5%
(n=443)
1.8%
(n=58)
14.8%
(n=485)
11.5%
(n=378)
no
yes
y
month
apr
aug
dec
jul
jun
mar
may
nov
oct
sep



Finally we have performed the screeplot to see the significant variables and a PCA to see which variables are related with which others. But as said previously, this is just an indicative of we should see when doing the AIC and BIC steps which later we will see if corresponds to our final result.





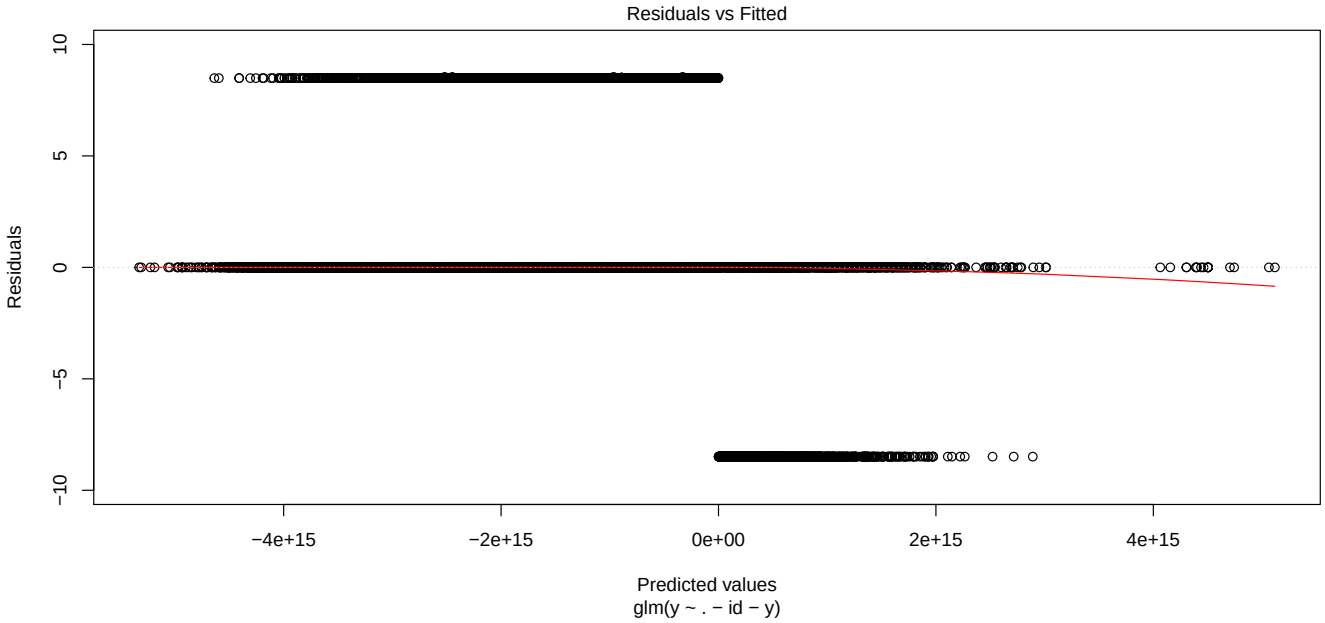
With the original variables, fit the complete model without interactions and using the logit link function.

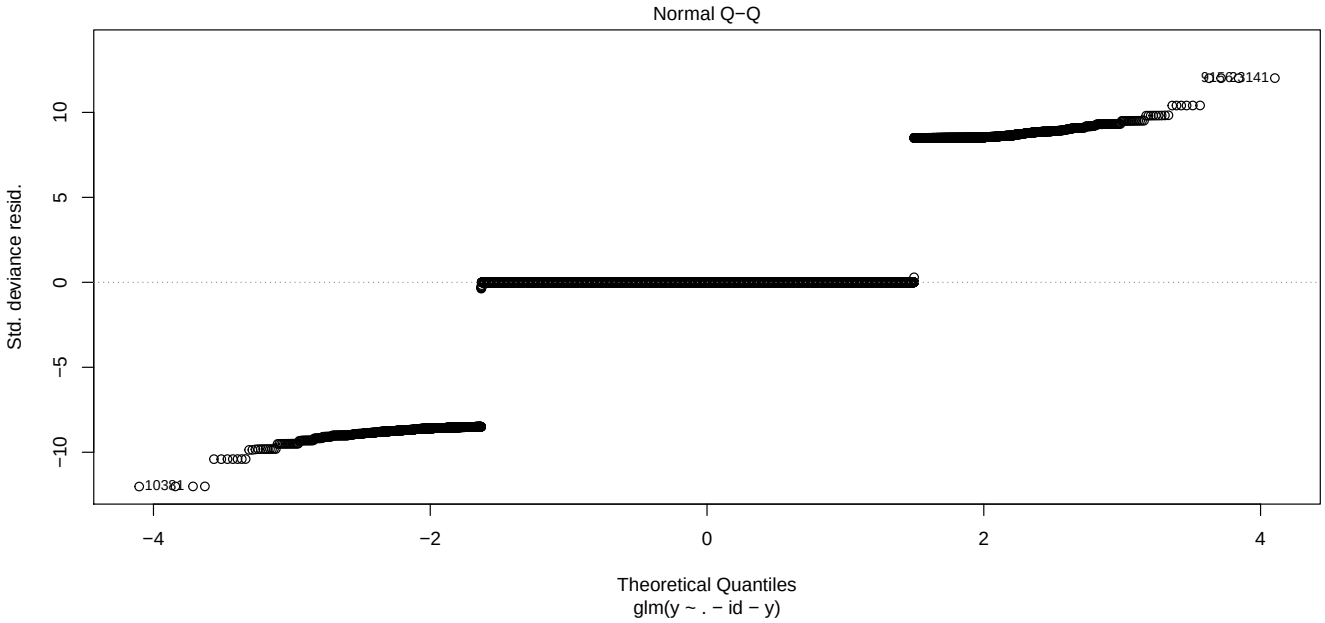
In this section we are going to perform a logistic regression model with all the variables without interactions. We are going to exclude id because it is a identificator number. The logit link is the default for the binomial family so doesn't need to be specified. To evaluate the model, we have splitten our dataset into training sample (90%) and test sample (10%). We will calculate accuracy model for each model that we build, it is necessary to compare models. Then, we have passed Anova test comparing it to a null model.

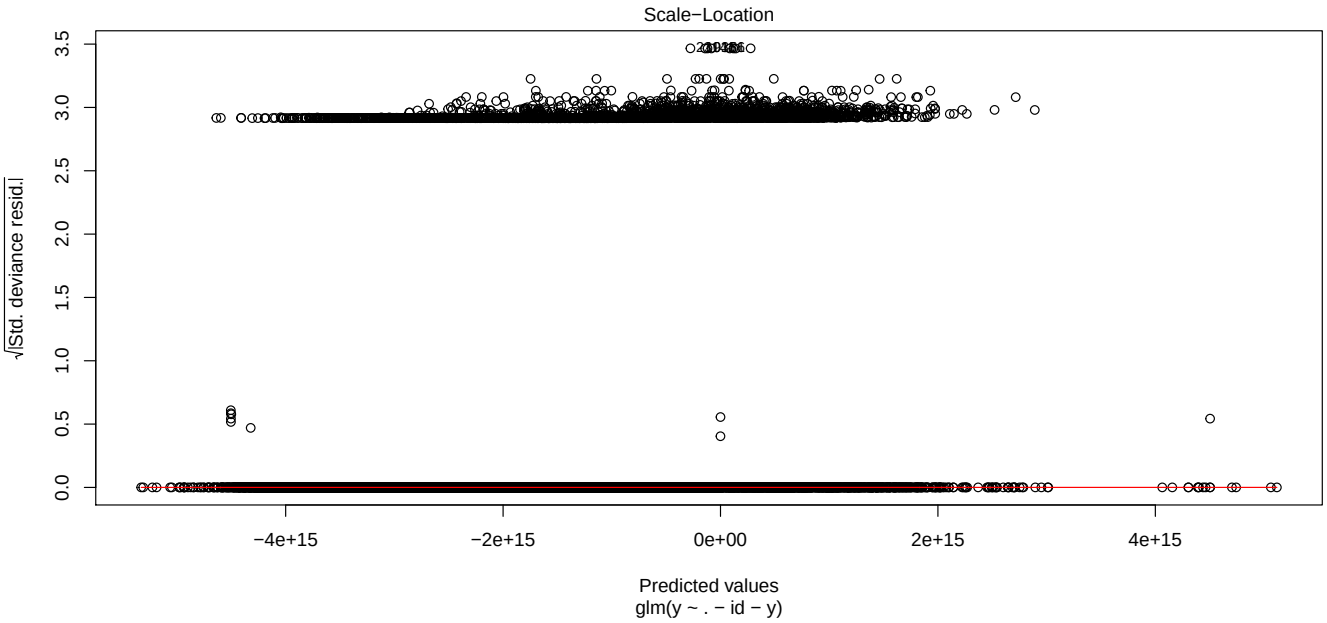
```
levels(dataset$y) <- c(FALSE,TRUE)
dataset$y <- as.logical.factor(dataset$y)

# Split train and test
require(caTools)
library(caTools)
sample <- sample.split(dataset,SplitRatio = 0.90)
train <- subset(dataset,sample==TRUE)
test <- subset(dataset, sample==FALSE)

summary(completeModel<-glm(y ~ . - id - y, train, family = binomial))

# Null Model
nullModel <- lm(y ~ 1, train)

# Predictions
completeModel$xlevels[["euribor3m"]] <- union(completeModel$xlevels[["euribor3m"]], levels(test$euribor3m))
preds <- predict(completeModel, test, type = "response")
preds[preds > .5] = TRUE
preds[preds < .5] = FALSE
preds <- as.logical(preds)
mean(preds == test$y)

# Anova test
anova(nullModel, completeModel, test = "Chisq")
```

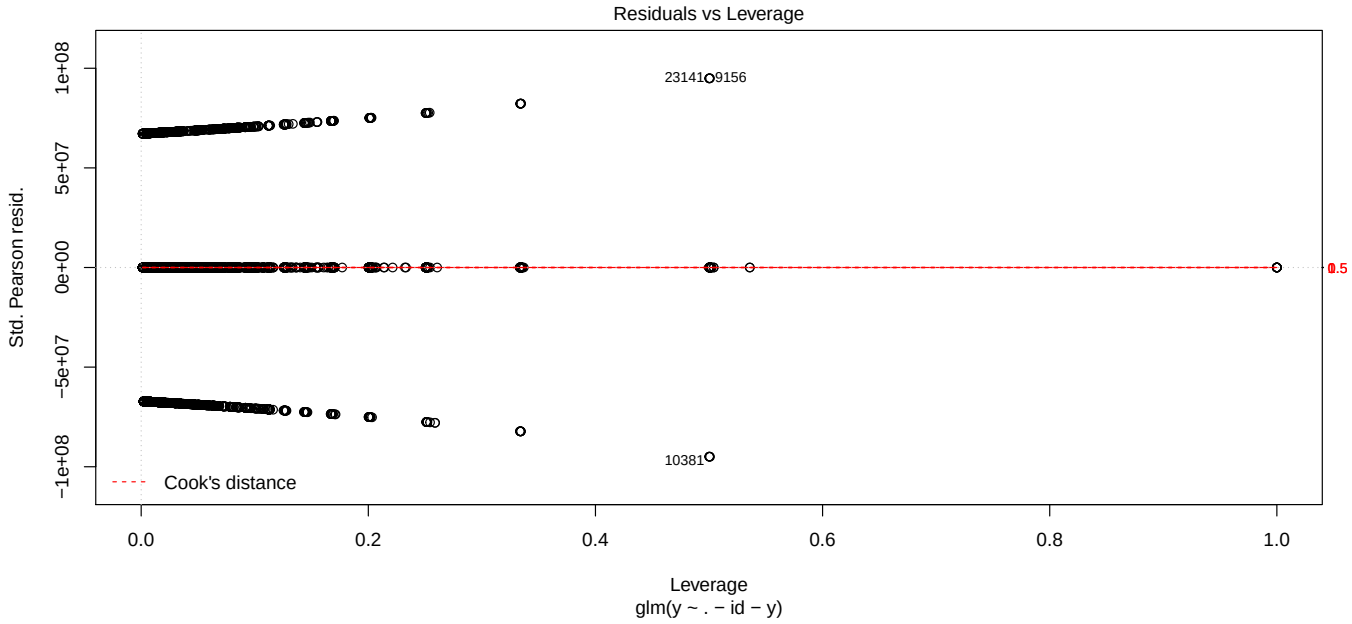
As we can see, we obtain a 86.65% accuracy in our test set and 197887 AIC.

```
plot(completeModel)
```










```
op<-par(mfrow=c(2,2))
```

Evaluate possible first order interactions (between two factors or between a factor and a covariable) and include them in the model (if there were any).

As we cannot build a model with all interactions for computational reasons, we have tried some interactions and we think that the formula obtained is a right one.

```
summary(interactionsModel<-glm(y ~ . + euribor3m*previous, family = binomial, data = train))
```

With InteractionsModel we obtain a better AIC (192265) than completeModel (197887)

Perform an automatic variables selection basen on the AIC & BIC. Make a comparison of the models and argue which one is chosen.

In order to create a model with the most significant variables, we decided to choose both criterions: the BIC and AIC criterion. Then, we will make a comparision of the models to select which one is better.

There are two heuristic strategies when modelling:

- **Forward:** In this strategy we start the case with none available predictor variables and add one at a time
- **Backward:** In this strategy we start with all available predictor variables and delete one at a time

In this case, due the amount of variables we have on our dataset, we rather follow forward heuristic strategy. Just because if we try to follow backward model it lasts very longer.

```
# BIC criterion
BICModel <- step(nullModel, scope = list(lower=nullModel, upper=completeModel), direction="both", criterion

preds <- predict(BICModel, test, type = "response")
preds[preds > .5] = TRUE
preds[preds < .5] = FALSE
preds <- as.logical(preds)
mean(preds == test$y)



# BIC Model = -56410.7
# Accuracy: 89,86%
# formula = y ~ cons.conf.idx + poutcome + contact + pdays

require(MASS)
library(MASS)
AICModel <- step(nullModel, scope = list(lower=nullModel, upper=completeModel),
        direction = "both",
        trace = 0,
        k = 2)
summary(AICModel)

AICModel$xlevels[["euribor3m"]] <- union(AICModel$xlevels[["euribor3m"]], levels(test$euribor3m))

preds <- predict(AICModel, test, type = "response")
preds[preds > .5] = TRUE
preds[preds < .5] = FALSE
preds <- as.logical(preds)

mean(preds == test$y)
```

As we can see, BIC Criterion gives us the formula (y ~ cons.conf.idx + poutcome + contact + pdays) and an accuracy about 89,87%, and R-squared around 0.206. AIC Model gets 88,51% of accuracy and a formula more complex than BIC Criterion (y ~ euribor3m + poutcome + cons.conf.idx + contact + pdays + day_of_week + campaign + previous + default + age) and a R-squared around 0.2331.

We are going to select BIC Criterion because it perfoms better than AIC.

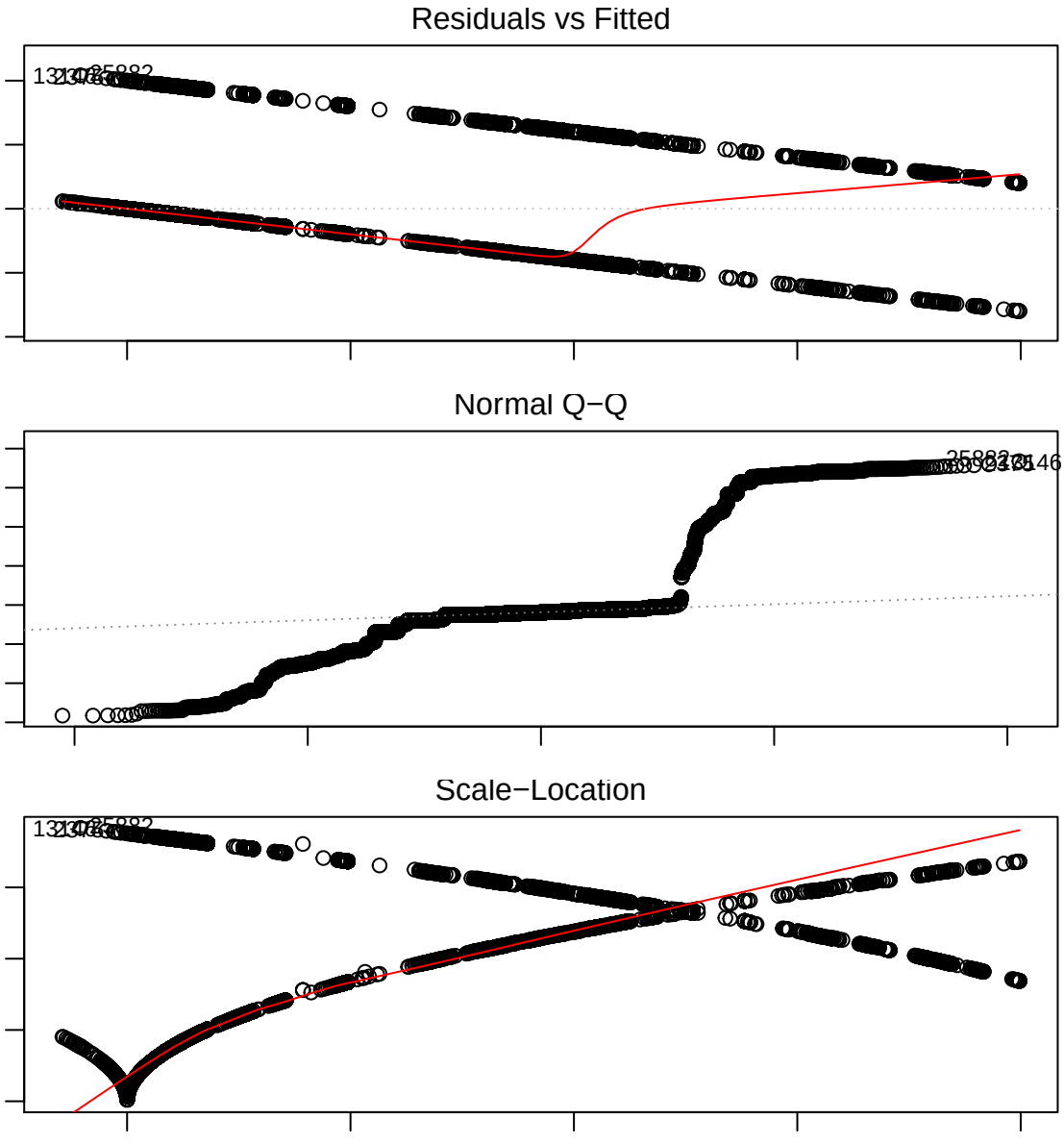
Validate the model by checking the assumptions

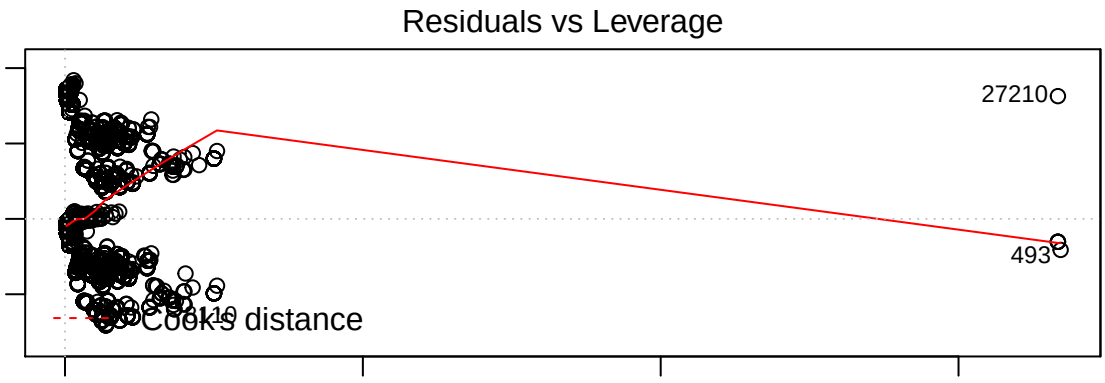
Once we have our model selected, we are going to validate it with anova comparing the null model with our final model. Then we can see the plots obtained.

```
anova(nullModel, BICModel)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ cons.price.idx + poutcome + contact + pdays + campaign
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  24552 2492.3
## 2  24522 1970.0 30    522.22 216.67 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
op<-par(mar=c(1,1,1,1))
plot(BICModel)
```





In this case we can see that our model, passes the test since our p-value is lower than 0.05. Normal Q-Q shows us that the distribution is not symetry. From the Scale-Location plot, it seems that the assumption of homoskedasticity is not validated.

Interpret the final model

Even our final model has an adjusted r-squared of 0.206 which is a low value, we can assume that has a good AIC and it performs good with an accuracy around 89,86%, one of our purposes in this project.

As we have previously said, we would rather a simple model with less predictor variables than a complex model with many of them. It has been rough working with this dataset due its size (28645 observations). Probably it would be much wiser to perform an imputation with unknown instances. Regardless, we do not have this many variables that can be relevant as we have seen in the exploratory analysis.

We can say that our final model explains that if we increase the consumer price index for a client, it means a high probability of subscribing to the deposit.

On the other hand, we can interpret that is better to contact clients via by cellular than by phone.