Endprojekt_Hoermann

Paul Hörmann

1/12/2020

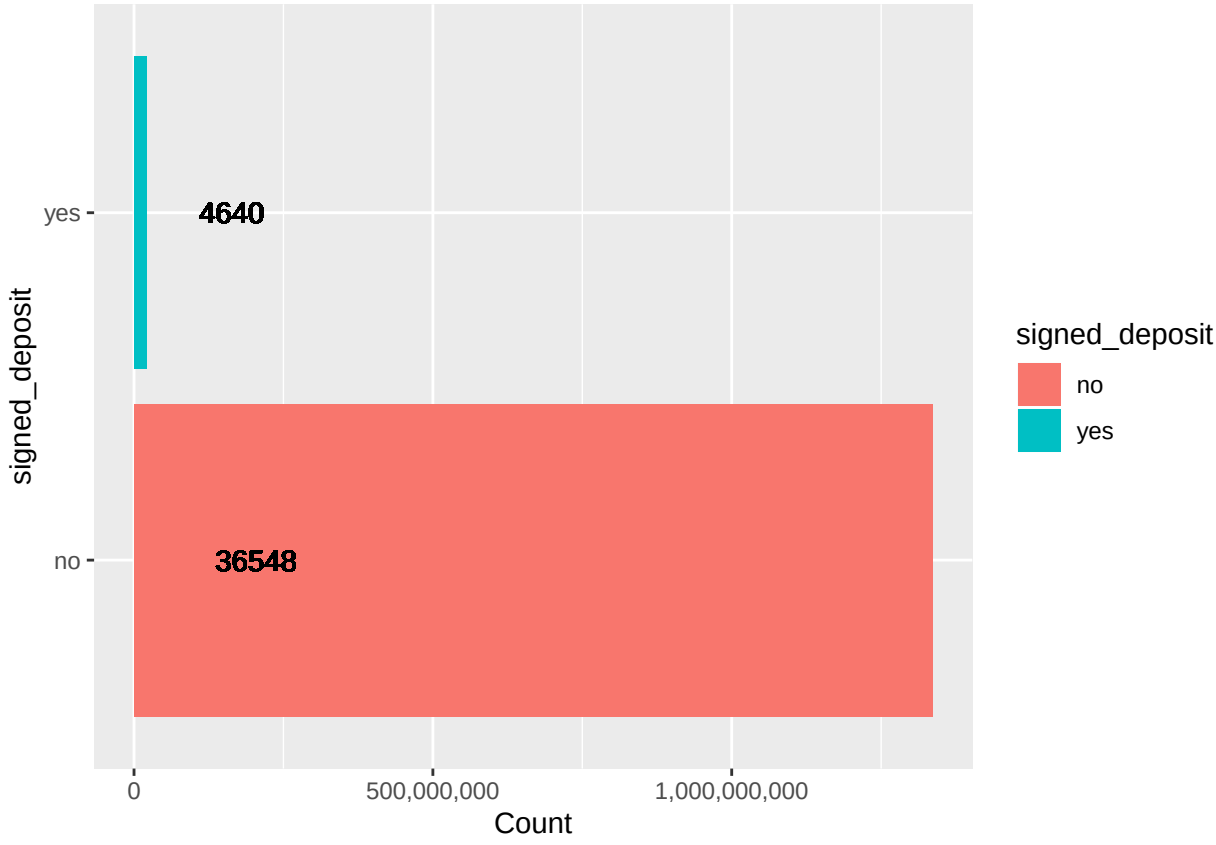
Acknowledgement

If not otherwise stated a error of 5% is used. For all Tests / Overviews data was removed if the attribute of interest had an unknown value. As we are taking a look at if the loan was given or not, data where it's unknown has been removed right away.

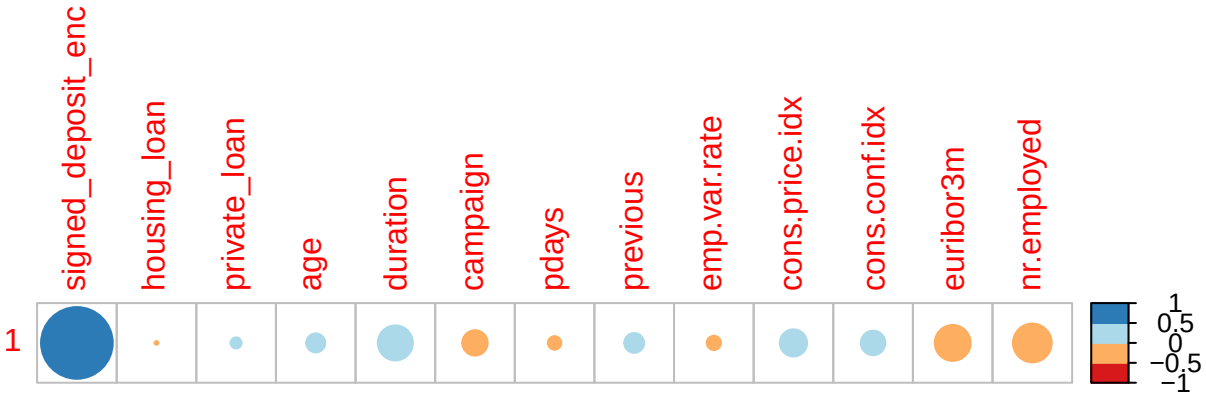
Overview

```
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.

##

## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.

##
## Attaching package: 'gdata'

## The following object is masked from 'package:stats':
##
##     nobs

## The following object is masked from 'package:utils':
##
##     object.size

## The following object is masked from 'package:base':
##
##     startsWith

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:gdata':
##
##     first, last

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:gdata':
##
##     combine, first, last

## The following objects are masked from 'package:stats':
##
```

```
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## corrplot 0.84 loaded

##    age       job marital   education default housing loan   contact month
## 1:  56 housemaid married    basic.4y      no      no   no telephone   may
## 2:  57  services married high.school unknown      no   no telephone   may
## 3:  37  services married high.school      no     yes   no telephone   may
## 4:  40    admin. married    basic.6y      no      no   no telephone   may
## 5:  56  services married high.school      no      no  yes telephone   may
## 6:  45  services married    basic.9y unknown      no   no telephone   may
##    day_of_week duration campaign pdays previous    poutcome emp.var.rate
## 1:         mon      261        1   999        0 nonexistent          1.1
## 2:         mon      149        1   999        0 nonexistent          1.1
## 3:         mon      226        1   999        0 nonexistent          1.1
## 4:         mon      151        1   999        0 nonexistent          1.1
## 5:         mon      307        1   999        0 nonexistent          1.1
## 6:         mon      198        1   999        0 nonexistent          1.1
##    cons.price.idx cons.conf.idx euribor3m nr.employed  y
## 1:         93.994         -36.4     4.857        5191 no
## 2:         93.994         -36.4     4.857        5191 no
## 3:         93.994         -36.4     4.857        5191 no
## 4:         93.994         -36.4     4.857        5191 no
## 5:         93.994         -36.4     4.857        5191 no
## 6:         93.994         -36.4     4.857        5191 no
```



Correlation Analysis



As Visible above there are 5 main factors correlating with the signage of a deposit. Positively correlating is the duration of the call, the consumer price index and the consumer confidence index. Negatively correlating is the 3 month euribor and the number of employed.

Below you can see the values ordered by their correlation.

```
##    signed_deposit_enc nr.employed euribor3m  duration cons.price.idx
## 1:                  1  -0.2924534 -0.250981 0.2393674      0.1441281
##      campaign cons.conf.idx   previous        age emp.var.rate       pdays
## 1: -0.1254179     0.1168678 0.07659563 0.07028341  -0.03948844 -0.03522348
##    private_loan housing_loan
## 1:   0.02371457  -0.00267077
```

Now we take a look at the coefficient of determination, to find out how much influence those values really have.

```
## [1] 0.1257967
```

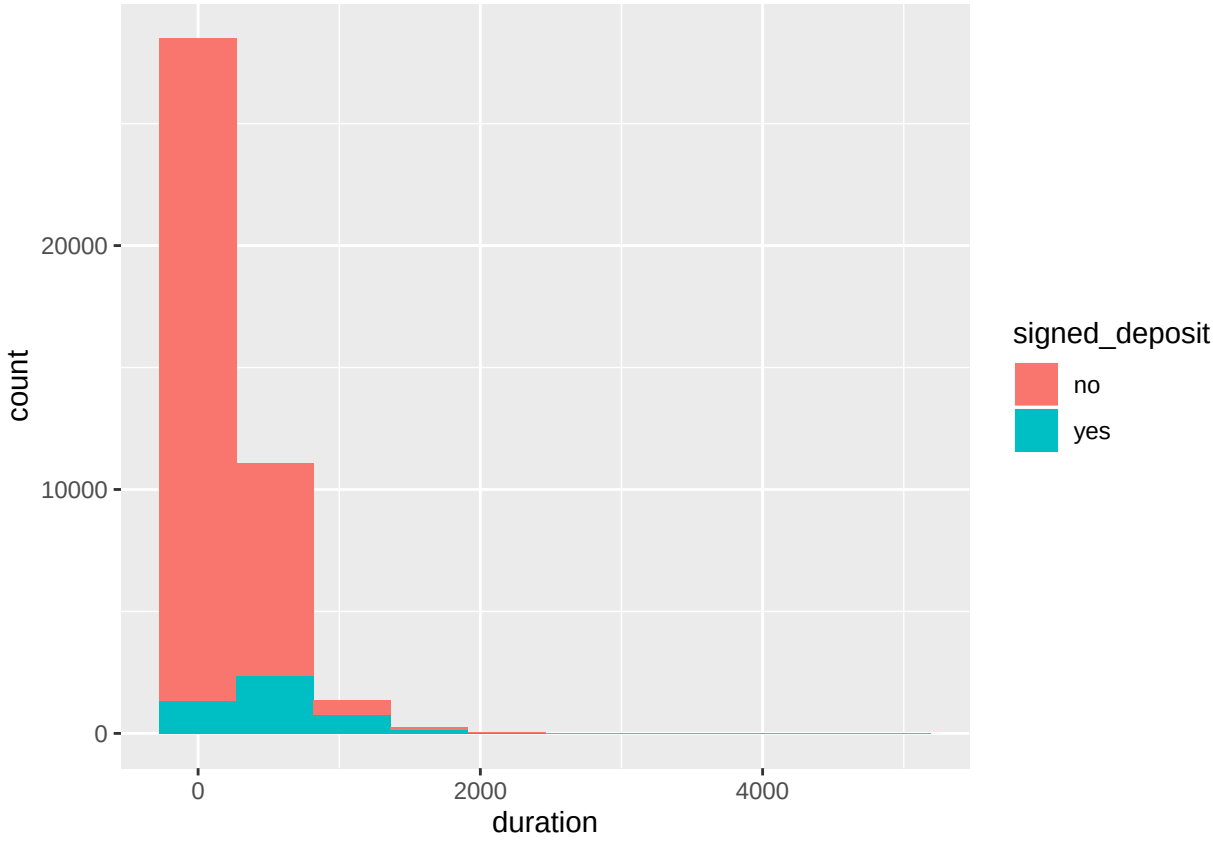
```
## [1] 0.09472324
```

```
## [1] 0.1642469
```

```
## [1] 0.01855349
```

```
## [1] 0.003011589
```

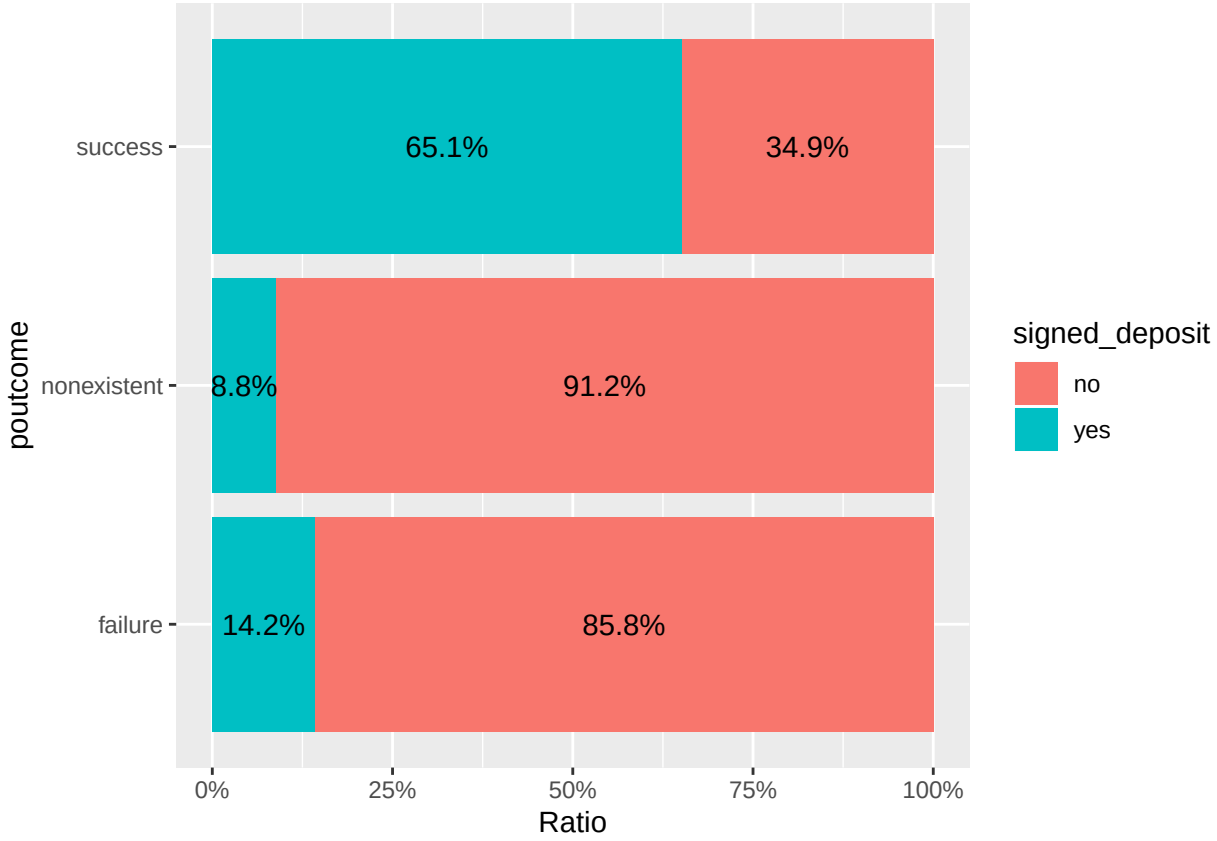
Taking a look at the most important factors



As you can see, the ratio of yes to no increases the longer the duration of the call was.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  source$signed_deposit by source$signed_deposit_enc
## Kruskal-Wallis chi-squared = 41187, df = 1, p-value < 2.2e-16
```

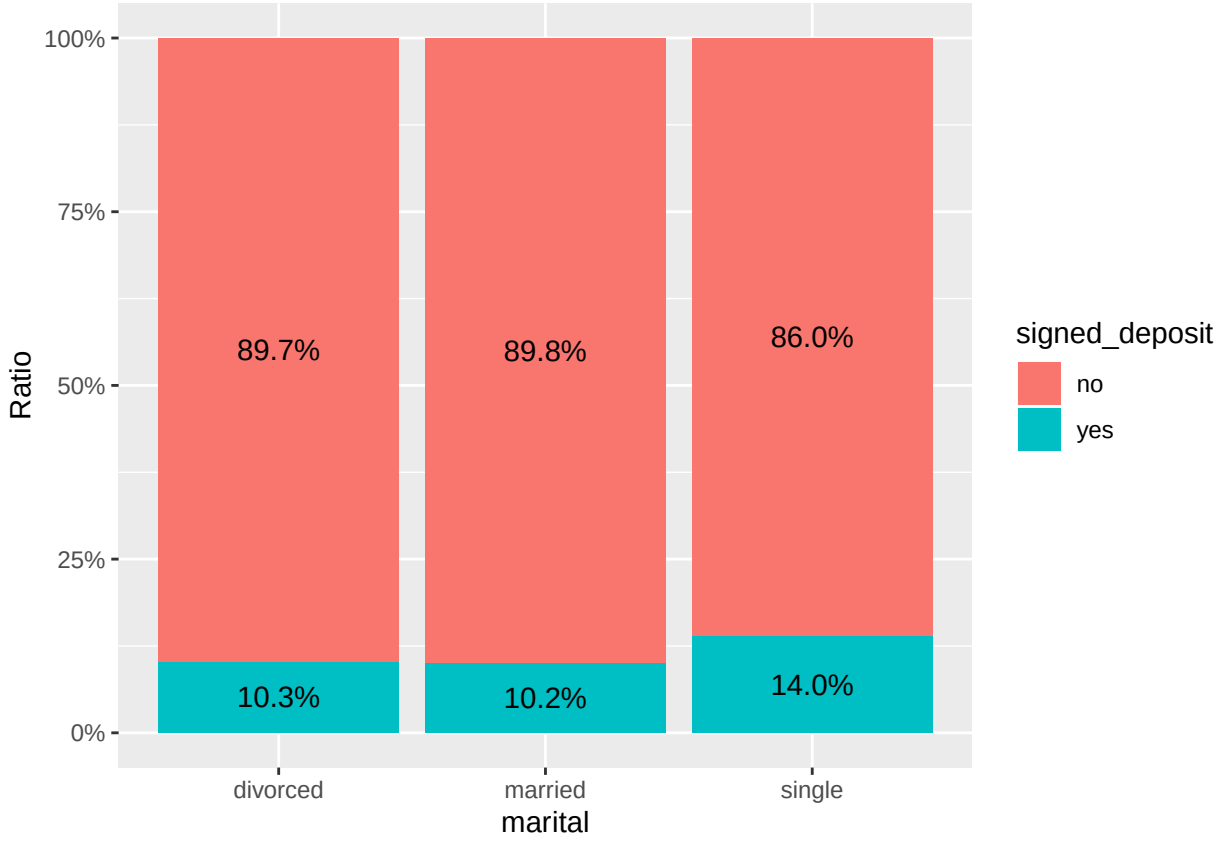
And the test shows, that the influence is statistifically significant.

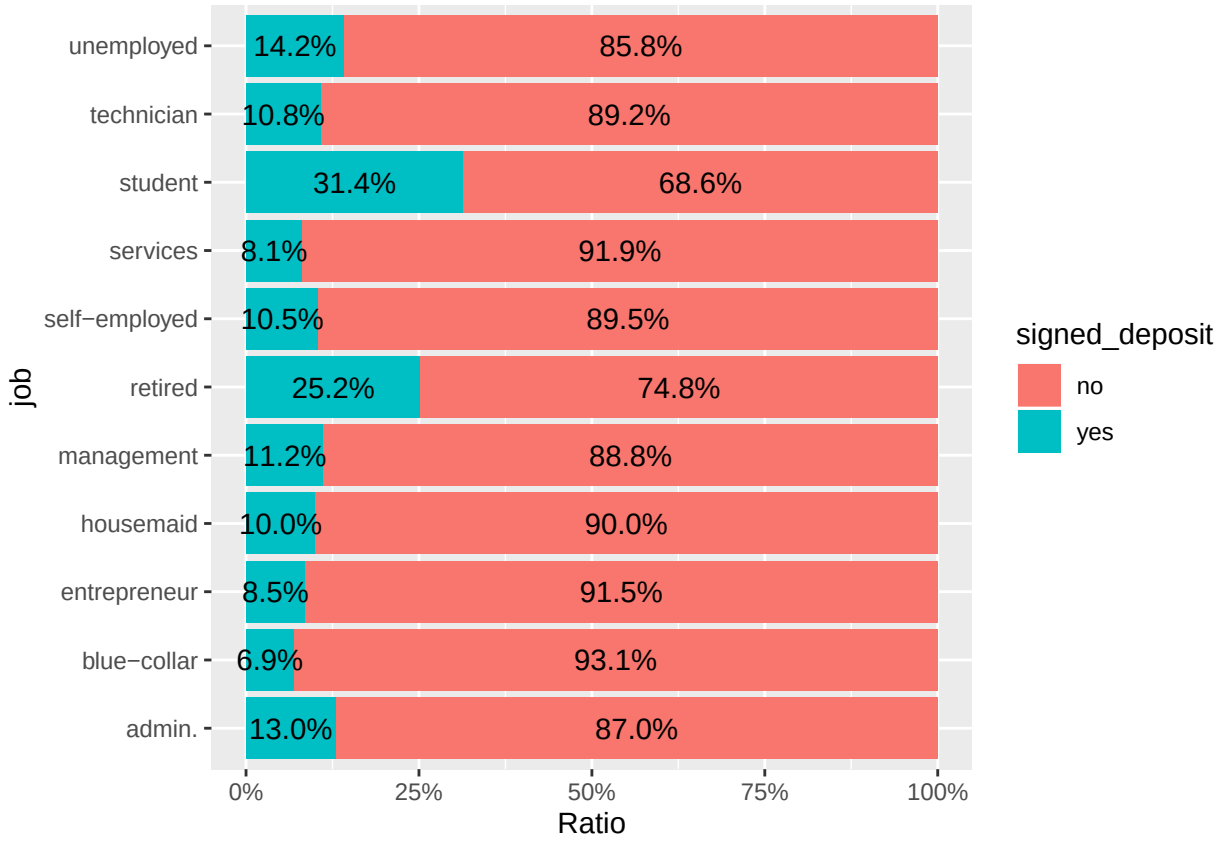


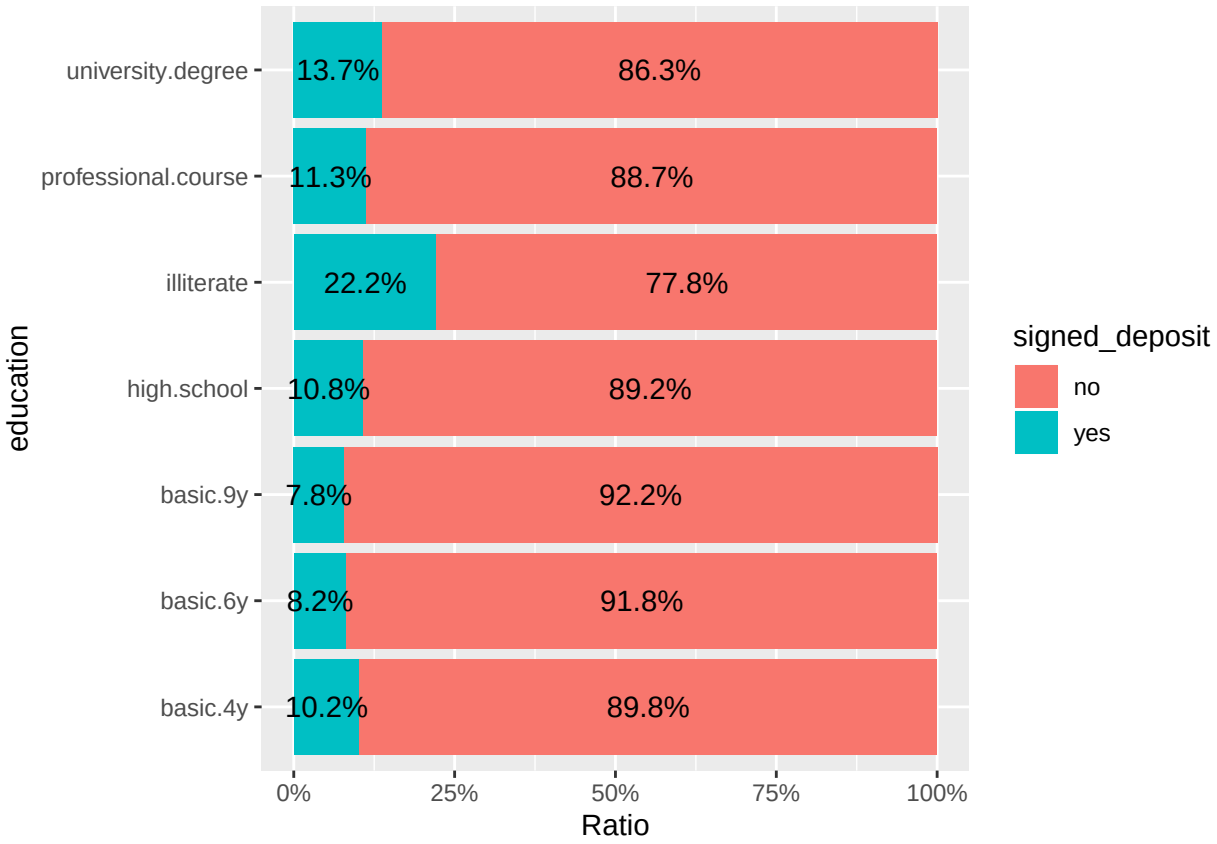
Especially people previously signed tend to do the same again.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  cleaned_poutcome$signed_deposit by cleaned_poutcome$poutcome
## Kruskal-Wallis chi-squared = 4230.4, df = 2, p-value < 2.2e-16
```

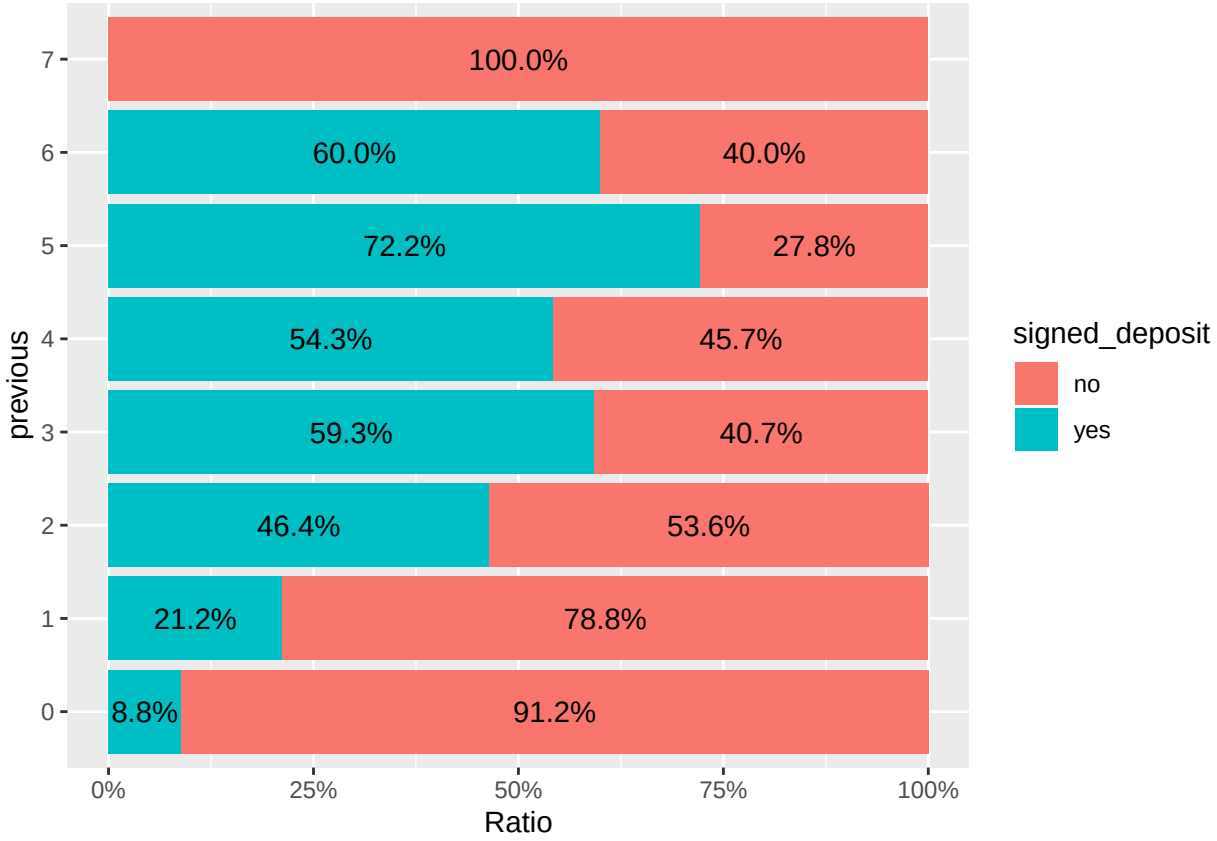
Which is also statistically significant.

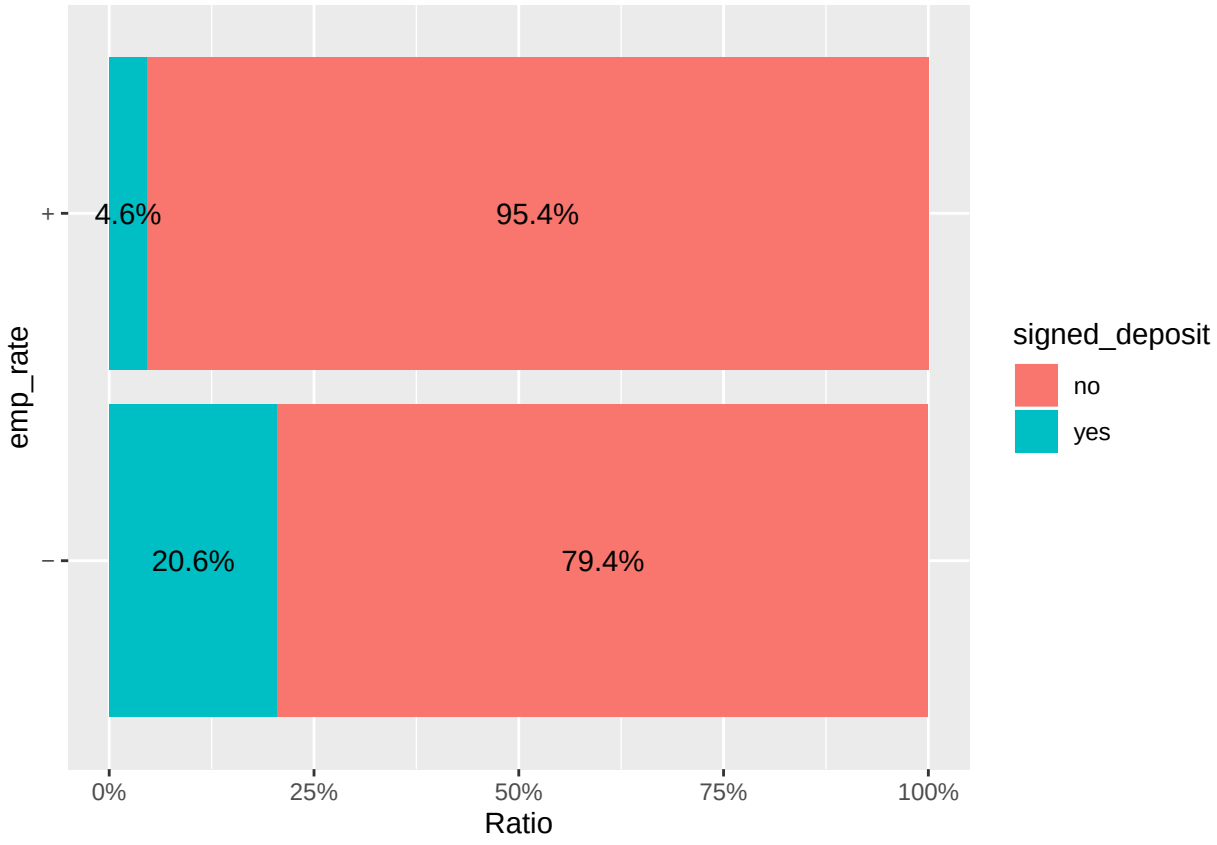

100%
75%
50%
25%
0%
Ratio
89.7%
89.8%
86.0%
10.3%
10.2%
14.0%
divorced
married
single
marital
signed_deposit
no
yes


unemployed
technician
student
services
self−employed
retired
management
housemaid
entrepreneur
blue−collar
admin.
job
14.2%
85.8%
10.8%
89.2%
31.4%
68.6%
8.1%
91.9%
10.5%
89.5%
25.2%
74.8%
11.2%
88.8%
10.0%
90.0%
8.5%
91.5%
6.9%
93.1%
13.0%
87.0%
0%
25%
50%
75%
100%
Ratio
signed_deposit
no
yes

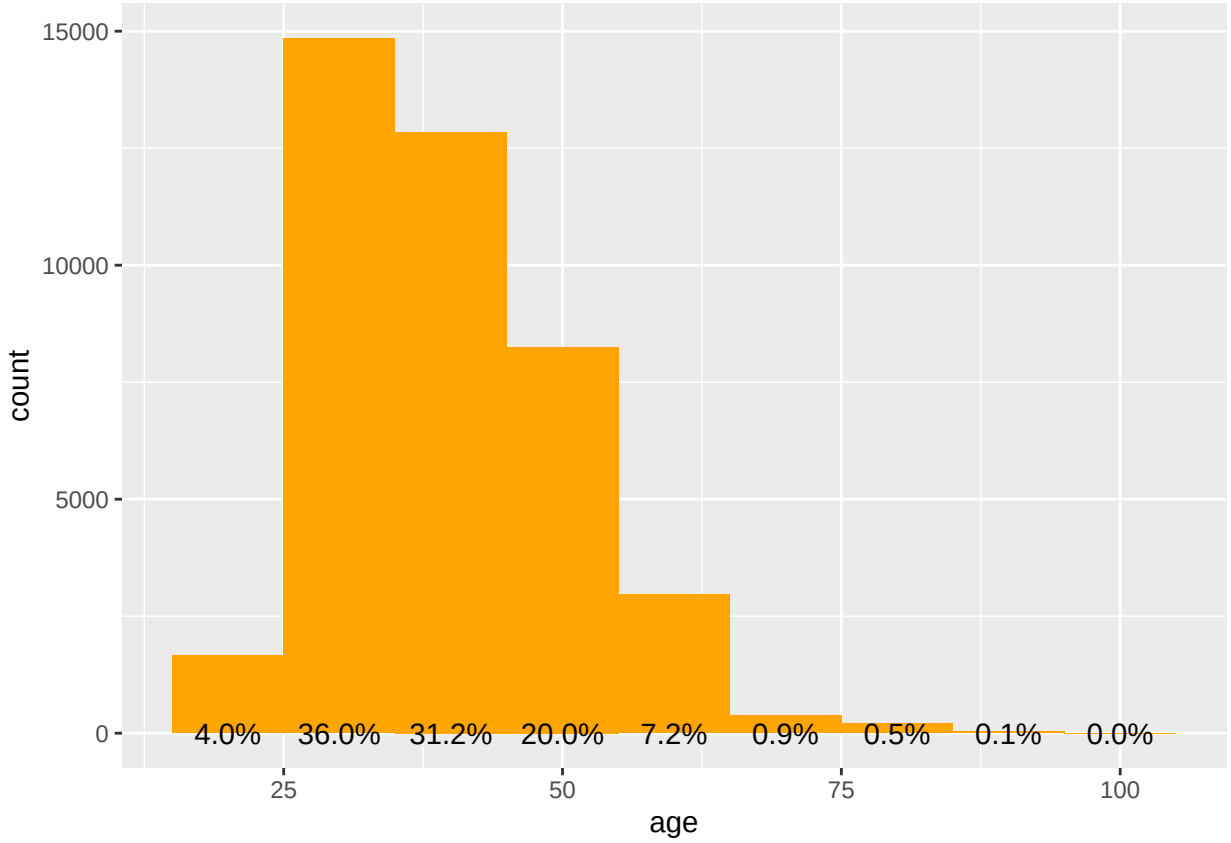


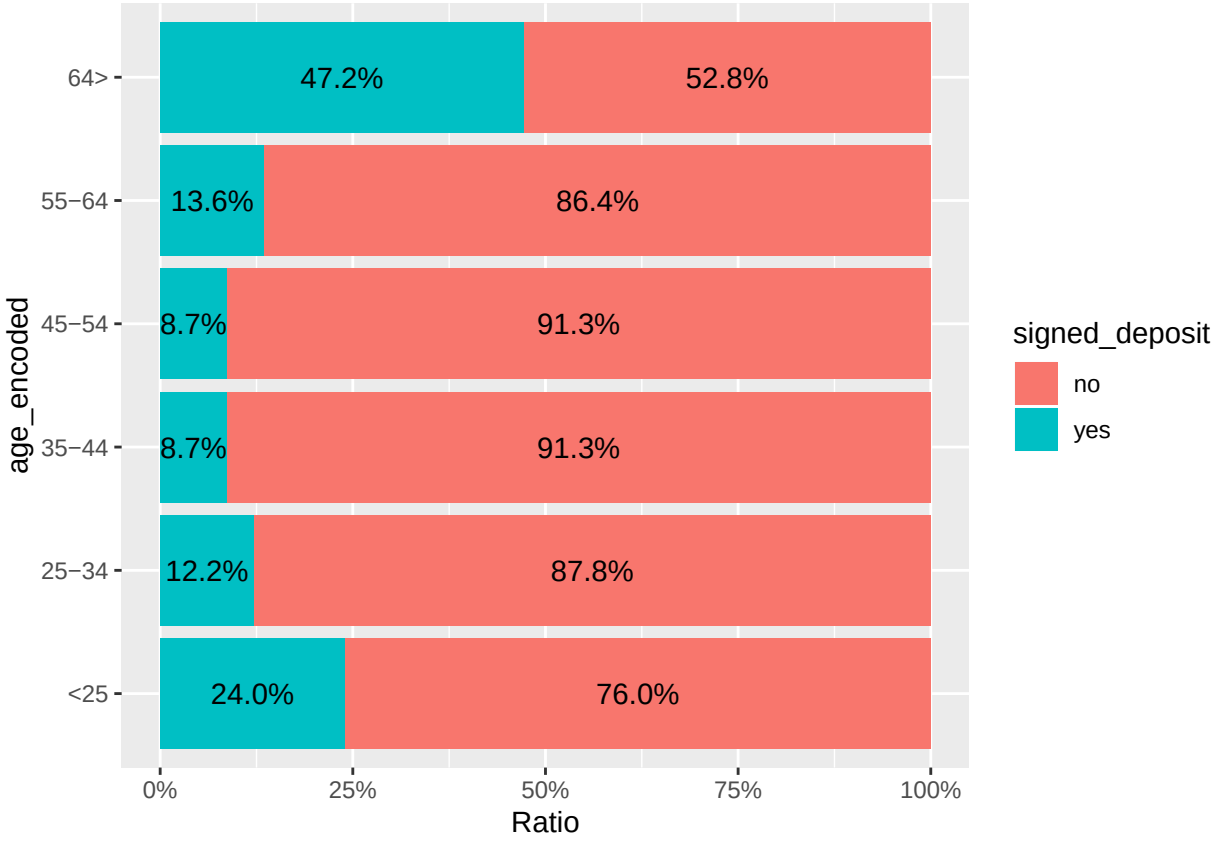
As you can see from the three previous graphs students, retired people and also illiterate people tend more to sign a deposit than others, whereas marital status has no big influence.


previous
7
6
5
4
3
2
1
0
100.0%
60.0%
40.0%
72.2%
27.8%
54.3%
45.7%
59.3%
40.7%
46.4%
53.6%
21.2%
78.8%
8.8%
91.2%
0%
25%
50%
75%
100%
Ratio
signed_deposit
no
yes


emp_rate
+
−
4.6%
95.4%
20.6%
79.4%
0%
25%
50%
75%
100%
Ratio
signed_deposit
no
yes

As already seen before, the employment rate has big influence on wether people tend to sign or not.






```
##
##  Kruskal-Wallis rank sum test
##
## data:  age_enc$signed_deposit_enc by age_enc$age_encoded
## Kruskal-Wallis chi-squared = 1211.3, df = 5, p-value < 2.2e-16
```

The age more or less matches the job, meaning younger students and older retired people are more likely to sign a deposit.