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	does he/she has a personal loan? (No, yes, unknown)	Campaign
pdays	does he/she has a personal loan? (No, yes, unknown)	Campaign
previous	does he/she has a personal loan? (No, yes, unknown)	Campaign
poutcome	does he/she has a personal loan? (No, yes, unknown)	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 we have that all our variables are integers or factors. 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.

```
## Observations: 28,645
## Variables: 21
## $ id             <int> 1, 2, 5, 6, 8, 10, 11, 12, 14, 15, 16, 18, 19, ...
## $ age            <int> 52, 33, 54, 53, 42, 36, 40, 44, 36, 48, 48, 27,...
## $ job            <fct> technician, admin., admin., housemaid, self-emp...
## $ marital        <fct> married, single, single, married, married, marr...
## $ education      <fct> high.school, university.degree, university.degr...
## $ default        <fct> no, no, no, no, unknown, no, no, no, no, no, no...
## $ housing        <fct> yes, yes, yes, no, yes, no, yes, yes, no, no, y...
## $ loan           <fct> no, no, no, yes, no, yes, no, no, no, no, no, n...
## $ contact        <fct> cellular, cellular, cellular, cellular, cellula...
```

```
## $ month          <fct> nov, nov, may, jun, aug, jul, apr, mar, may, ju...
## $ day_of_week    <fct> tue, thu, mon, thu, tue, wed, thu, tue, wed, th...
## $ campaign       <int> 1, 1, 1, 1, 2, 4, 1, 2, 1, 4, 1, 2, 2, 2, 1, 2,...
## $ pdays          <int> 999, 999, 999, 999, 999, 999, 999, 999, 999, 99...
## $ previous       <int> 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,...
## $ poutcome       <fct> nonexistent, nonexistent, nonexistent, failure,...
## $ emp.var.rate   <dbl> -0.1, -0.1, -1.8, -2.9, 1.4, 1.4, -1.8, -1.8, 1...
## $ cons.price.idx <dbl> 93.200, 93.200, 92.893, 92.963, 93.444, 93.918,...
## $ cons.conf.idx  <dbl> -42.0, -42.0, -46.2, -40.8, -36.1, -42.7, -47.1...
## $ euribor3m      <dbl> 4.153, 4.076, 1.264, 1.260, 4.966, 4.963, 1.365...
## $ nr.employed    <dbl> 5195.8, 5195.8, 5099.1, 5076.2, 5228.1, 5228.1,...
## $ y              <fct> no, no, no, yes, no, no, yes, yes, no, no, no, ...
```

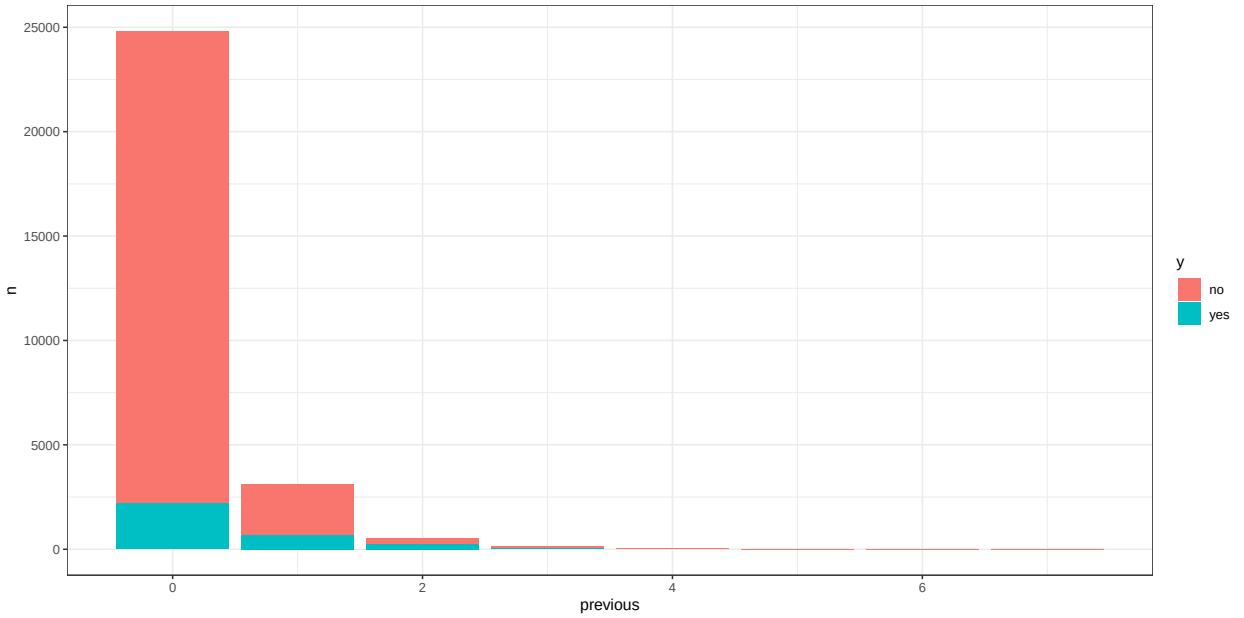
```
## # A tibble: 11 x 5
##    col_name      cnt common            common_pcnt levels
##    <chr>       <int> <chr>                   <dbl> <list>
##  1 contact         2 cellular                 63.5 <tibble [2 x 3]>
##  2 day_of_week     5 thu                      21.2 <tibble [5 x 3]>
##  3 default         3 no                       79.2 <tibble [3 x 3]>
##  4 education       8 university.degree        29.4 <tibble [8 x 3]>
##  5 housing         3 yes                      52.4 <tibble [3 x 3]>
##  6 job            12 admin.                   25.2 <tibble [12 x 3]>
##  7 loan            3 no                       82.4 <tibble [3 x 3]>
##  8 marital         4 married                  60.6 <tibble [4 x 3]>
##  9 month          10 may                      33.3 <tibble [10 x 3]>
## 10 poutcome        3 nonexistent              86.7 <tibble [3 x 3]>
## 11 y               2 no                       88.5 <tibble [2 x 3]>
```

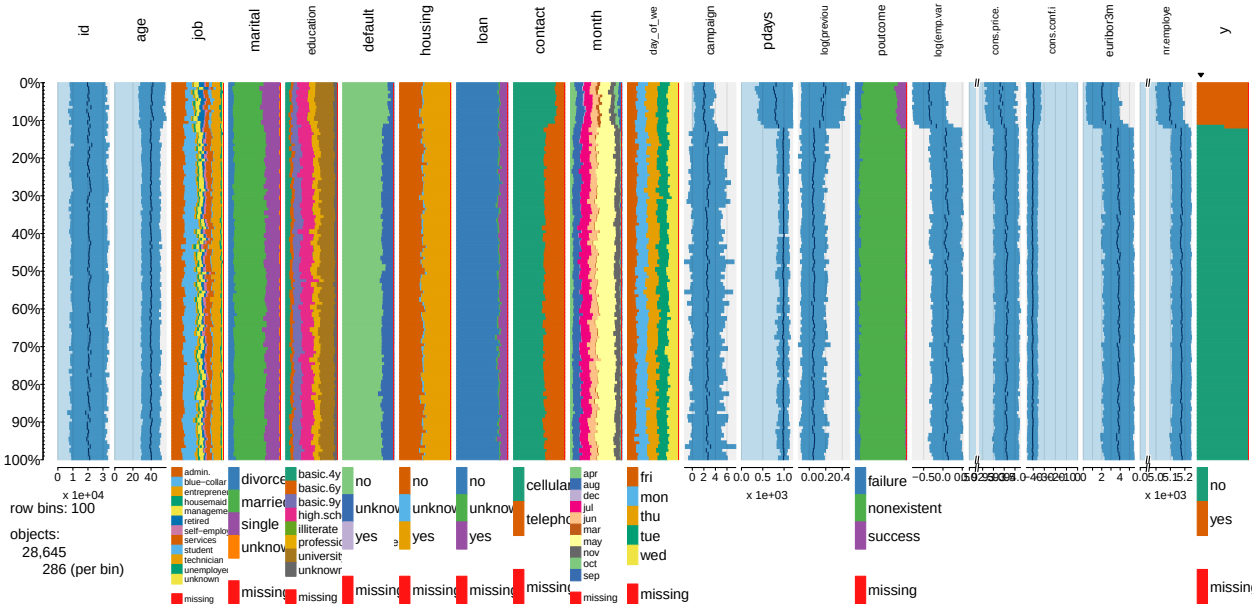
```
## # A tibble: 10 x 10
##    col_name      min      q1  median     mean      q3     max      sd
##    <chr>       <dbl>   <dbl>   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>
##  1 id        1.00e+0  1.04e4  2.05e4  2.06e+4  3.09e4  4.12e4 1.19e+4
##  2 age       1.70e+1  3.20e1  3.80e1  4.00e+1  4.70e1  9.80e1 1.04e+1
##  3 campaign  1.00e+0  1.00e0  2.00e0  2.56e+0  3.00e0  4.30e1 2.76e+0
##  4 pdays     0.       9.99e2  9.99e2  9.63e+2  9.99e2  9.99e2 1.87e+2
##  5 previous  0.       0.      0.      1.69e-1  0.      7.00e0 4.87e-1
##  6 emp.var~ -3.40e+0 -1.80e0  1.10e0  8.15e-2  1.40e0  1.40e0 1.57e+0
##  7 cons.pr~  9.22e+1  9.31e1  9.38e1  9.36e+1  9.40e1  9.48e1 5.80e-1
##  8 cons.co~ -5.08e+1 -4.27e1 -4.18e1 -4.05e+1 -3.64e1 -2.69e1 4.64e+0
##  9 euribor~  6.34e-1  1.34e0  4.86e0  3.62e+0  4.96e0  5.04e0 1.74e+0
## 10 nr.empl~  4.96e+3  5.10e3  5.19e3  5.17e+3  5.23e3  5.23e3 7.23e+1
## # ... with 2 more variables: pcnt_na <dbl>, hist <list>
```

```
##
##        no       yes
## 0.8853901 0.1146099
```

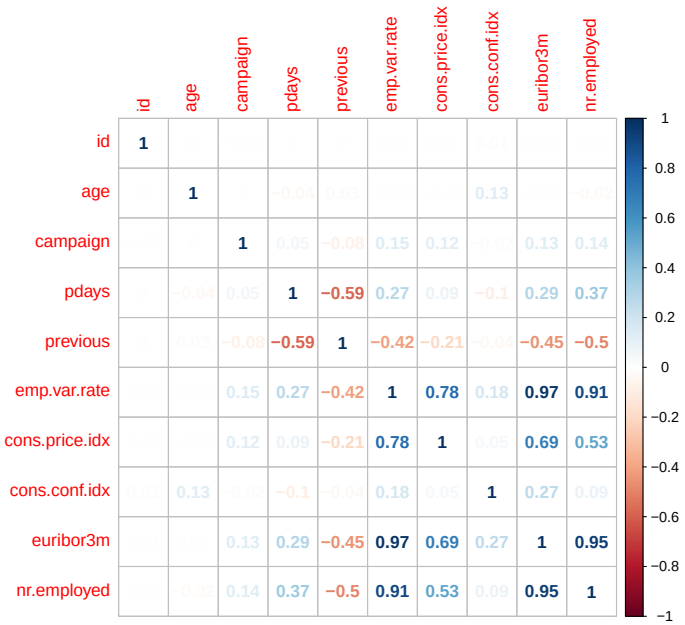
```
## [1] -0.5884877
```

Dataset is UNbalanced. Resampling is required






```
## # A tibble: 45 x 6
##    col_1          col_2           corr p_value  lower  upper
##    <chr>          <chr>          <dbl>   <dbl>  <dbl>  <dbl>
##  1 euribor3m      emp.var.rate   0.972       0  0.972  0.973
##  2 nr.employed    euribor3m      0.945       0  0.944  0.946
##  3 nr.employed    emp.var.rate   0.907       0  0.905  0.909
##  4 cons.price.idx emp.var.rate   0.778       0  0.773  0.782
##  5 euribor3m      cons.price.idx  0.691       0  0.685  0.697
##  6 previous       pdays          -0.588       0 -0.596 -0.581
##  7 nr.employed    cons.price.idx  0.525       0  0.517  0.533
##  8 nr.employed    previous       -0.495       0 -0.504 -0.486
##  9 euribor3m      previous       -0.450       0 -0.459 -0.441
## 10 emp.var.rate   previous       -0.419       0 -0.429 -0.410
## # ... with 35 more rows
```



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

aaaaa

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)

aaaaa

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

aaaaa

Validate the model y checking the assumptions

aaaaa

Interpret the final model

aaaaa