Logistic Regression model and cluster analysis based on Portuguese banking institution's marketing campaign

Yuika Cho-1003213186

12/20/2020

Catalog

Introduction

The dataset was obtained from UCI Machine Learning Repository. It is about term deposit campaigns of a Portuguese banking institution. This dataset mainly has two parts. One is about clients' personal information, like age, job, education level, personal loans. Another is about clients' reaction and attitude towards bank's marketing campaign, like previous contacts performed times, outcome of the previous marketing campaign and last contact duration. In general, the dataset has 21 columns and 45211, from May 2008 to November 2010 (figure 1).

The main research question is about which factors lead to the result of bank's marketing campaign, which shows in the 21st column of the dataset. On the one hand, trying to figure out this question can help the bank catch the key factors that could motivate client to choose their finance products. I built a logistic regression model to predict the key factors. On the other hand, Portuguese banking institution should draw a general picture for their main clients. For example, how their personas look like and

what's their basic personal information. I used cluster analysis to segment clients. Thus, the bank could better design their marketing strategy by combining the key factors and personas' information.

Keywords

logistic regression model, cluster analysis, Portuguese banking institution, marketing campaign

Methodology (Data and Model)

For the logistic regression model method, I first rejected all observations with "unknown", like job and education has values with "unknown". Secondly, I substituted all data with "divorced" into "single", because the key point that I'd like to classify is that clients' marital status, which is in marriage or not.

For the cluster analysis, I first selected all variables that directly related to customer information. For example, age, job, marital, education, housing and loan. However, there are still tons of rows in the new dataset, and it is hard to do the cluster analysis directly due to the large capacity. So, I chose May as the month that I analysis. Then, I transferred all Character variables to dummy variables (figure 4). The last step is to standardize this dataset.

Results & Discussion

From the output of the model, it is clearly that "pdays", "poutcome", "duration" and "contacttype" are high correlated. "age" and "education" are significant as well. This information can greatly help the bank to make marketing decisions (figure 3). For example, from "poutcome", we can infer that people with more positive attitude towards marketing campaign are usually more likely to get a term deposit.

From the results of cluster analysis, I first used NbClust formula to calculate the best number of clusters and got the number 3 (figure 2). Then, we can see that there are

five customer segments with different characteristics, and I calculated the marketing campaign success rate based on the 21st column (figure 5):

1) segment: people who work in service industry
   Their education level is usually high school, and work in service industry. They are very young and unmarried, with housing loans but no personal loans.
   Marketing campaign success rate: 12%

2) segment: Technician
   They are middle-aged, married, highly educated, financially sound, and have no homes or personal loans.
   Marketing campaign success rate: 8%

3) Segment: Admin
   They are middle-aged married administrative workers, often without housing or personal loans, and have college degrees.
   Marketing campaign success rate: 18%

As can be seen from the above analysis, technical and managerial personnel have a higher level of education than others. They are the top2 segments among all customers. They are the main target customers. Thus, banks can focus their marketing strategies on technicians and managers or people with higher education levels and marital status.

APPENDIX

Figure 1

```
> summary(data)
      age            job              marital
 Min.   :17.00   Length:41188       Length:41188
 1st Qu.:32.00   Class :character   Class :character
 Median :38.00   Mode  :character   Mode  :character
 Mean   :40.02
 3rd Qu.:47.00
 Max.   :98.00
  education           default
 Length:41188       Length:41188
 Class :character   Class :character
 Mode  :character   Mode  :character


   housing             loan
 Length:41188       Length:41188
 Class :character   Class :character
 Mode  :character   Mode  :character


   contact             duration         campaign
 Length:41188       Min.   :   0.0   Min.   : 1.000
 Class :character   1st Qu.: 102.0   1st Qu.: 1.000
 Mode  :character   Median : 180.0   Median : 2.000
                    Mean   : 258.3   Mean   : 2.568
                    3rd Qu.: 319.0   3rd Qu.: 3.000
                    Max.   :4918.0   Max.   :56.000
     pdays           previous       poutcome
 Min.   :  0.0   Min.   :0.000   Length:41188
 1st Qu.:999.0   1st Qu.:0.000   Class :character
 Median :999.0   Median :0.000   Mode  :character
 Mean   :962.5   Mean   :0.173
 3rd Qu.:999.0   3rd Qu.:0.000
 Max.   :999.0   Max.   :7.000
  emp.var.rate      cons.price.idx  cons.conf.idx
 Min.   :-3.40000   Min.   :92.20   Min.   :-50.8
 1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7
 Median : 1.10000   Median :93.75   Median :-41.8
 Mean   : 0.08189   Mean   :93.58   Mean   :-40.5
 3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4
 Max.   : 1.40000   Max.   :94.77   Max.   :-26.9
   euribor3m          y
 Min.   :0.634   Length:41188
 1st Qu.:1.344   Class :character
 Median :4.857   Mode  :character
 Mean   :3.621
 3rd Qu.:4.961
 Max.   :5.045
```

Figure 2

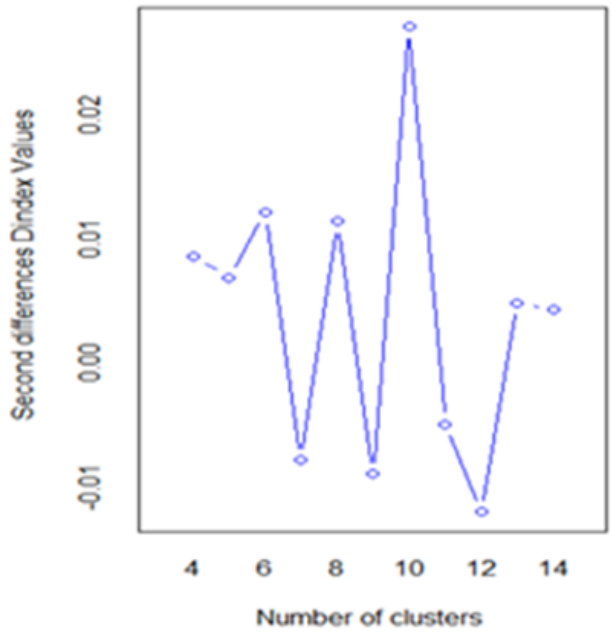

Figure 3

```
> lrm
Generalized Linear Model

30488 samples
   17 predictor
    2 classes: 'no', 'yes'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 24391, 24390, 24390, 24391, 24390, 24392,
 ...
Resampling results:

  Accuracy   Kappa
  0.8996218  0.4558483
```

```
> summary(lrm)

Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.8123  -0.3604  -0.2091  -0.1496   3.3180

Coefficients:
                                Estimate Std. Error
(Intercept)                   -1.197e+02  6.685e+00
age                           -1.564e-04  2.678e-03
`jobblue-collar`              -3.068e-01  8.924e-02
jobentrepreneur               -2.634e-01  1.366e-01
jobhousemaid                   5.954e-03  1.634e-01
jobmanagement                 -7.962e-02  9.023e-02
jobretired                     4.049e-01  1.184e-01
`jobself-employed`            -8.753e-02  1.229e-01
jobservices                   -2.160e-01  9.320e-02
jobstudent                     3.750e-01  1.229e-01
jobtechnician                  2.666e-02  7.468e-02
jobunemployed                  4.532e-02  1.353e-01
maritalmarried                -6.962e-03  7.345e-02
maritalsingle                  7.031e-02  8.295e-02
educationbasic.6y              8.896e-02  1.416e-01
educationbasic.9y             -1.339e-02  1.089e-01
educationhigh.school           5.041e-02  1.040e-01
educationilliterate            1.771e+00  8.360e-01
educationprofessional.course   1.537e-01  1.126e-01
educationuniversity.degree     2.683e-01  1.040e-01
defaultyes                    -7.252e+00  1.135e+02
housingyes                    -1.546e-02  4.418e-02
loanyes                       -8.054e-02  6.132e-02
contacttelephone              -8.129e-01  6.842e-02
duration                       4.407e-03  8.156e-05
campaign                      -4.067e-02  1.289e-02
pdays                         -1.061e-03  2.313e-04
previous                      -3.859e-02  6.320e-02
poutcomenonexistent            5.448e-01  1.003e-01
poutcomesuccess                8.691e-01  2.253e-01
emp.var.rate                  -8.729e-01  7.133e-02
cons.price.idx                 1.267e+00  7.011e-02
cons.conf.idx                  4.624e-02  4.459e-03
euribor3m                     -4.535e-02  5.584e-02
                              z value Pr(>|z|)
(Intercept)                   -17.908  < 2e-16 ***
```

```
`jobblue-collar`               -3.438 0.000587 ***
jobentrepreneur                -1.928 0.053867 .
jobhousemaid                    0.036 0.970940
jobmanagement                  -0.882 0.377536
jobretired                      3.419 0.000628 ***
`jobself-employed`             -0.712 0.476289
jobservices                    -2.318 0.020450 *
jobstudent                      3.051 0.002284 **
jobtechnician                   0.357 0.721067
jobunemployed                   0.335 0.737601
maritalmarried                 -0.095 0.924488
maritalsingle                   0.848 0.396666
educationbasic.6y               0.628 0.529898
educationbasic.9y              -0.123 0.902082
educationhigh.school            0.485 0.627907
educationilliterate             2.118 0.034195 *
educationprofessional.course    1.364 0.172497
educationuniversity.degree      2.579 0.009906 **
defaultyes                     -0.064 0.949060
housingyes                     -0.350 0.726476
loanyes                        -1.314 0.189005
contacttelephone              -11.881  < 2e-16 ***
duration                       54.038  < 2e-16 ***
campaign                       -3.154 0.001609 **
pdays                          -4.589 4.45e-06 ***
previous                       -0.611 0.541422
previous                      -0.611 0.541422
poutcomenonexistent            5.432 5.57e-08 ***
poutcomesuccess                3.857 0.000115 ***
emp.var.rate                 -12.239  < 2e-16 ***
cons.price.idx                18.079  < 2e-16 ***
cons.conf.idx                 10.370  < 2e-16 ***
euribor3m                     -0.812 0.416706
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23160  on 30487  degrees of freedom
Residual deviance: 14421  on 30454  degrees of freedom
AIC: 14489

Number of Fisher Scoring iterations: 10
```

Figure 4

```
# A tibble: 1,593 x 14
     age job   marital education default housing loan  month
   <dbl> <chr> <chr>   <chr>     <chr>   <chr>   <chr> <chr>
 1    35 blue~ single  basic.9y  no      yes     no    may
 2    46 tech~ married high.sch~ no      no      no    may
 3    30 admi~ single  universi~ no      yes     no    may
 4    35 admi~ single  high.sch~ no      yes     yes   may
 5    35 serv~ married basic.6y  no      no      no    may
 6    60 blue~ married basic.9y  no      yes     no    may
 7    36 mana~ married universi~ no      yes     no    may
 8    31 admi~ divorc~ universi~ no      no      no    may
 9    81 reti~ married basic.9y  no      no      no    may
10    30 unem~ married high.sch~ no      no      no    may
# ... with 1,583 more rows, and 6 more variables: duration <dbl>,
#   campaign <dbl>, pdays <dbl>, previous <dbl>, poutcome <chr>,
#   y <chr>
```

Figure 5

```
K-means clustering with 3 clusters of sizes 14001, 18293, 8894

Cluster means:
  bank$age maritaldivorced maritalmarried maritalsingle
1 42.00864      0.13656167      0.6885222    0.17405900
2 30.97939      0.06007763      0.4675559    0.47007052
3 55.50281      0.18000899      0.7572521    0.05981561
  maritalunknown educationbasic.4y educationbasic.6y
1   0.0008570816        0.10549246        0.07799443
2   0.0022959602        0.04329525        0.04411524
3   0.0029233191        0.21441421        0.04418709
  educationbasic.9y educationhigh.school
1         0.1553460            0.2216985
2         0.1511507            0.2662767
3         0.1242411            0.1731504
  educationilliterate educationprofessional.course
1        0.0004285408                    0.1344190
2        0.0001639972                    0.1244192
3        0.0010119181                    0.1219924
  educationuniversity.degree educationunknown jobadmin.
1                  0.2606242       0.04399686 0.2419113
2                  0.3418794       0.02869950 0.2965615
3                  0.2546661       0.06633686 0.1810209
  jobblue.collar jobentrepreneur jobhousemaid
1      0.2632669      0.04399686   0.02892651
2      0.2121030      0.02547423   0.01268245
3      0.1897909      0.04205082   0.04756015
  jobmanagement  jobretired jobself.employed jobservices
1    0.08135133 0.005499607       0.03599743  0.09585030
2    0.05149511 0.001093314       0.03312743  0.11080741
3    0.09478300 0.182482573       0.03496739  0.06746121
   jobstudent jobtechnician jobunemployed  jobunknown
1 0.002428398     0.1657739    0.02671238 0.008285122
2 0.045973870     0.1828568    0.02334226 0.004482589
3 0.000000000     0.1210929    0.02394873 0.014841466
  housingno housingunknown housingyes    loanno
1 0.4605385     0.02371259  0.5157489 0.8267981
2 0.4472202     0.02416225  0.5286175 0.8199311
3 0.4489544     0.02428604  0.5267596 0.8292107
  loanunknown   loanyes
1  0.02371259 0.1494893
2  0.02416225 0.1559066
3  0.02428604 0.1465033
```

```
Clustering vector:
  1   2   3   4   5   6   7   8   9  10  11
  3   3   1   1   3   1   3   1   2   2   1
 12  13  14  15  16  17  18  19  20  21  22
  2   2   3   2   3   2   1   3   1   2   3
 23  24  25  26  27  28  29  30  31  32  33
  3   1   1   2   3   1   3   3   1   3   3
 34  35  36  37  38  39  40  41  42  43  44
  3   3   3   2   3   1   3   3   2   1   3
 45  46  47  48  49  50  51  52  53  54  55
  1   1   3   1   2   1   3   1   3   3   3
 56  57  58  59  60  61  62  63  64  65  66
  3   3   1   3   2   1   3   1   1   1   1
 67  68  69  70  71  72  73  74  75  76  77
  1   2   3   1   1   1   1   3   1   1   3
 78  79  80  81  82  83  84  85  86  87  88
  2   3   1   3   3   1   3   1   2   1   1
 89  90  91  92  93  94  95  96  97  98  99
  3   2   2   3   3   2   1   1   1   3   1
100 101 102 103 104 105 106 107 108 109 110
  1   3   3   3   3   3   1   2   2   1   1
111 112 113 114 115 116 117 118 119 120 121
  3   3   2   3   1   2   3   3   3   1   1
122 123 124 125 126 127 128 129 130 131 132
  3   1   1   2   1   3   2   3   1   1   1
133 134 135 136 137 138 139 140 141 142 143
  2   1   3   1   1   3   2   1   1   3   1
144 145 146 147 148 149 150 151 152 153 154
  3   1   1   1   2   1   3   3   3   1   3
155 156 157 158 159 160 161 162 163 164 165
  1   2   3   3   1   3   2   1   1   1   1
166 167 168 169 170 171 172 173 174 175 176
  1   1   3   1   3   1   2   1   1   1   1
177 178 179 180 181 182 183 184 185 186 187
  1   3   2   1   2   1   1   2   3   3   3
188 189 190 191 192 193 194 195 196 197 198
  3   3   2   1   1   1   1   2   1   1   2
199 200 201 202 203 204 205 206 207 208 209
  1   1   2   2   1   1   1   2   2   2   2
210 211 212 213 214 215 216 217 218 219 220
  2   2   3   3   3   1   1   2   2   3   1
221 222 223 224 225 226 227 228 229 230 231
  2   1   2   1   1   3   1   2   1   1   1
232 233 234 235 236 237 238 239 240 241 242
  3   2   1   3   1   3   1   2   1   3   2
243 244 245 246 247 248 249 250 251 252 253
  2   2   2   2   2   2   2   2   1   1   3
254 255 256 257 258 259 260 261 262 263 264
  2   2   3   3   1   1   2   2   2   1   1
265 266 267 268 269 270 271 272 273 274 275
```

```
   2    2    1    2    3    1    3    3    2    2    1
 793  794  795  796  797  798  799  800  801  802  803
   2    1    2    2    1    1    2    1    1    1    2
 804  805  806  807  808  809  810  811  812  813  814
   1    2    2    2    3    2    2    3    1    1    3
 815  816  817  818  819  820  821  822  823  824  825
   3    1    1    2    3    2    2    2    2    2    2
 826  827  828  829  830  831  832  833  834  835  836
   3    1    2    1    3    3    1    2    3    1    1
 837  838  839  840  841  842  843  844  845  846  847
   3    1    1    2    1    1    3    3    1    3    2
 848  849  850  851  852  853  854  855  856  857  858
   2    1    3    1    2    2    2    1    1    1    1
 859  860  861  862  863  864  865  866  867  868  869
   1    2    1    1    2    1    2    2    1    1    1
 870  871  872  873  874  875  876  877  878  879  880
   1    1    1    1    3    1    1    1    1    2    1
 881  882  883  884  885  886  887  888  889  890  891
   2    2    1    1    1    1    2    3    2    2    1
 892  893  894  895  896  897  898  899  900  901  902
   1    2    1    1    1    2    1    2    1    2    3
 903  904  905  906  907  908  909  910  911  912  913
   1    3    3    2    2    1    3    1    1    1    1
 914  915  916  917  918  919  920  921  922  923  924
   2    2    1    2    1    1    2    3    1    1    3
 925  926  927  928  929  930  931  932  933  934  935
   3    1    3    3    2    2    3    2    2    2    1
 936  937  938  939  940  941  942  943  944  945  946
   3    1    1    2    2    3    1    1    3    1    1
 947  948  949  950  951  952  953  954  955  956  957
   3    1    3    1    1    2    1    1    2    2    1
 958  959  960  961  962  963  964  965  966  967  968
   1    1    1    2    3    1    1    2    1    1    2
 969  970  971  972  973  974  975  976  977  978  979
   1    2    3    2    2    1    3    1    1    2    1
 980  981  982  983  984  985  986  987  988  989  990
   2    2    3    1    1    2    1    2    3    1    3
 991  992  993  994  995  996  997  998  999 1000
   1    1    1    1    1    2    1    3    3    2
 [ reached getOption("max.print") -- omitted 40188 entries ]

Within cluster sum of squares by cluster:
[1] 209272.6 298922.7 403272.8
 (between_SS / total_SS =  80.2 %)

Available components:

[1] "cluster"      "centers"      "totss"
[4] "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"

          1     2     3
  no  12828 16050  7670
  yes  1173  2243  1224
```

References

UCI Machine Learning Repository: Bank Marketing Data Set. (2012). UCI. http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#