SAP projekt Porez123 - prva točka

Matija Kukić

11.12.2024

Provjera povezanosti zanimanja i bračnog statusa

U ovoj bilježnici ćemo pokušati jednostavnom analizom vidjeti jesu li povezani zanimanje klijenta i njihovog bračnog statusa.

```
library('dplyr')
```

Prvo učitavamo podatke.

```
marketing = read.csv("../data/data.csv")
head(marketing)
```

```
##   age          job marital_status education default balance housing_loan
## 1  58   management        married  tertiary      no    2143          yes
## 2  44   technician         single secondary      no      29          yes
## 3  33 entrepreneur        married secondary      no       2          yes
## 4  47  blue-collar        married   unknown      no    1506          yes
## 5  33      unknown         single   unknown      no       1           no
## 6  35   management        married  tertiary      no     231          yes
##   personal_loan contact last_contact_day last_contact_month
## 1            no unknown               5               may
## 2            no unknown               5               may
## 3           yes unknown               5               may
## 4            no unknown               5               may
## 5            no unknown               5               may
## 6            no unknown               5               may
##   last_contact_duration campaign_contacts_count
## 1                   261                       1
## 2                   151                       1
## 3                    76                       1
## 4                    92                       1
## 5                   198                       1
## 6                   139                       1
##   days_from_previous_campaign_contact previous_contacts_count
## 1                                  -1                       0
## 2                                  -1                       0
## 3                                  -1                       0
## 4                                  -1                       0
## 5                                  -1                       0
## 6                                  -1                       0
##   previous_campaign_outcome term_deposit_accepted
```

```
## 1                   unknown                   no
## 2                   unknown                   no
## 3                   unknown                   no
## 4                   unknown                   no
## 5                   unknown                   no
## 6                   unknown                   no
```

```
dim(marketing)
```

```
## [1] 45211    17
```

Pregled jednostavnih statistika.

```
summary(marketing)
```

```
##       age            job            marital_status      education
##  Min.   :18.00   Length:45211       Length:45211       Length:45211
##  1st Qu.:33.00   Class :character   Class :character   Class :character
##  Median :39.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :40.94
##  3rd Qu.:48.00
##  Max.   :95.00
##    default             balance       housing_loan       personal_loan
##  Length:45211       Min.   : -8019   Length:45211       Length:45211
##  Class :character   1st Qu.:    72   Class :character   Class :character
##  Mode  :character   Median :   448   Mode  :character   Mode  :character
##                     Mean   :  1362
##                     3rd Qu.:  1428
##                     Max.   :102127
##    contact          last_contact_day last_contact_month last_contact_duration
##  Length:45211       Min.   : 1.00    Length:45211       Min.   :   0.0
##  Class :character   1st Qu.: 8.00    Class :character   1st Qu.: 103.0
##  Mode  :character   Median :16.00    Mode  :character   Median : 180.0
##                     Mean   :15.81                       Mean   : 258.2
##                     3rd Qu.:21.00                       3rd Qu.: 319.0
##                     Max.   :31.00                       Max.   :4918.0
##  campaign_contacts_count days_from_previous_campaign_contact
##  Min.   : 1.000          Min.   : -1.0
##  1st Qu.: 1.000          1st Qu.: -1.0
##  Median : 2.000          Median : -1.0
##  Mean   : 2.764          Mean   : 40.2
##  3rd Qu.: 3.000          3rd Qu.: -1.0
##  Max.   :63.000          Max.   :871.0
##  previous_contacts_count previous_campaign_outcome term_deposit_accepted
##  Min.   :  0.0000        Length:45211              Length:45211
##  1st Qu.:  0.0000        Class :character          Class :character
##  Median :  0.0000        Mode  :character          Mode  :character
##  Mean   :  0.5803
##  3rd Qu.:  0.0000
##  Max.   :275.0000
```

Pregledavamo kakve vrijednosti imamo u datasetu, nakon toga još gledamo imamo li neke nepoznate podatke.

```
levels(factor(marketing$job))
```

```
##  [1] "admin."        "blue-collar"   "entrepreneur"  "housemaid"
##  [5] "management"    "retired"       "self-employed" "services"
##  [9] "student"       "technician"    "unemployed"    "unknown"
```

```
levels(factor(marketing$marital_status))
```

```
## [1] "divorced" "married"  "single"
```

```
s <- c('job', 'marital_status')
for (col_name in s){
  if (sum(is.na(marketing[,col_name])) > 0){
    cat('Ukupno nedostajućih vrijednosti za varijablu ',col_name, ': ', sum(is.na(marketing[,col_name]))
  }
  else {
    cat('Nema nedostajućih vrijednosti.\n')
  }
}
```

```
## Nema nedostajućih vrijednosti.
## Nema nedostajućih vrijednosti.
```

Nemamo nedostajućih vrijednosti. Logično možemo u iz podataka uzeti samo stupce koji nam trebaju odnosno posao i bračni status.

```
rel = marketing[names(marketing) %in% c('job','marital_status')]
head(rel)
```

```
##            job marital_status
## 1   management        married
## 2   technician         single
## 3 entrepreneur        married
## 4  blue-collar        married
## 5      unknown         single
## 6   management        married
```

```
dim(rel)
```

```
## [1] 45211     2
```

```
tracemem(rel)==tracemem(marketing)
```

```
## [1] FALSE
```

Možemo i maknuti nepoznati posao.

```
rel = rel[rel$job != 'unknown', ]
dim(rel)
```

```
## [1] 44923     2
```

Primjećujemo da nismo izgubili mnogo podataka.

Budući da imamo kategorijske podatke, radimo kontigencijsku tablicu i provodimo χ^2 test za

```
tab = table(rel$job,rel$marital_status)
tab
```

```
##
##                 divorced married single
##   admin.             750    2693   1728
##   blue-collar        750    6968   2014
##   entrepreneur       179    1070    238
##   housemaid          184     912    144
##   management        1111    5400   2947
##   retired            425    1731    108
##   self-employed      140     993    446
##   services           549    2407   1198
##   student              6      54    878
##   technician         925    4052   2620
##   unemployed         171     731    401
```

```
kontab = addmargins(tab)
kontab
```

```
##
##                 divorced married single   Sum
##   admin.             750    2693   1728  5171
##   blue-collar        750    6968   2014  9732
##   entrepreneur       179    1070    238  1487
##   housemaid          184     912    144  1240
##   management        1111    5400   2947  9458
##   retired            425    1731    108  2264
##   self-employed      140     993    446  1579
##   services           549    2407   1198  4154
##   student              6      54    878   938
##   technician         925    4052   2620  7597
##   unemployed         171     731    401  1303
##   Sum               5190   27011  12722 44923
```

Sada moramo pogledati očekivane vrijednosti svakog stupca i retka.

```
for (col_names in colnames(kontab)){
  for (row_names in rownames(kontab)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Očekivane frekvencije za razred ',col_names,'-',row_names,': ',(kontab[row_names,'Sum'] * kon
    }
  }
}
```

```
## Očekivane frekvencije za razred  divorced - admin. :  597.4109
## Očekivane frekvencije za razred  divorced - blue-collar :  1124.348
## Očekivane frekvencije za razred  divorced - entrepreneur :  171.7946
## Očekivane frekvencije za razred  divorced - housemaid :  143.2585
## Očekivane frekvencije za razred  divorced - management :  1092.692
## Očekivane frekvencije za razred  divorced - retired :  261.5622
## Očekivane frekvencije za razred  divorced - self-employed :  182.4235
## Očekivane frekvencije za razred  divorced - services :  479.9159
## Očekivane frekvencije za razred  divorced - student :  108.3681
## Očekivane frekvencije za razred  divorced - technician :  877.6892
## Očekivane frekvencije za razred  divorced - unemployed :  150.5369
## Očekivane frekvencije za razred  married - admin. :  3109.184
## Očekivane frekvencije za razred  married - blue-collar :  5851.592
## Očekivane frekvencije za razred  married - entrepreneur :  894.0934
## Očekivane frekvencije za razred  married - housemaid :  745.5789
## Očekivane frekvencije za razred  married - management :  5686.843
## Očekivane frekvencije za razred  married - retired :  1361.283
## Očekivane frekvencije za razred  married - self-employed :  949.4105
## Očekivane frekvencije za razred  married - services :  2497.689
## Očekivane frekvencije za razred  married - student :  563.9943
## Očekivane frekvencije za razred  married - technician :  4567.873
## Očekivane frekvencije za razred  married - unemployed :  783.4591
## Očekivane frekvencije za razred  single - admin. :  1464.405
## Očekivane frekvencije za razred  single - blue-collar :  2756.06
## Očekivane frekvencije za razred  single - entrepreneur :  421.112
## Očekivane frekvencije za razred  single - housemaid :  351.1627
## Očekivane frekvencije za razred  single - management :  2678.465
## Očekivane frekvencije za razred  single - retired :  641.155
## Očekivane frekvencije za razred  single - self-employed :  447.166
## Očekivane frekvencije za razred  single - services :  1176.395
## Očekivane frekvencije za razred  single - student :  265.6376
## Očekivane frekvencije za razred  single - technician :  2151.438
## Očekivane frekvencije za razred  single - unemployed :  369.004
```

Sve su veće od 5 pa možemo provesti test.

```
chisq.test(kontab,correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  kontab
## X-squared = 3819.6, df = 33, p-value < 2.2e-16
```

Možemo odbaciti hipotezu H0 koja je da su tablični podatci nezavisnu u korist hipoteze H1 koja je da postoji zavisnost među kategorijama.