21_PCA

August 1, 2025

imamo multikolinearnost…. kada ručno micati varijable, a kada koristiti metodu PCA?

PCA metoda je objedinjavanje nezavisnih varijabli u jednu. Te nezavisne varijable su medjusobno korelirane

0.1 Primjer 1

- imamo nezavisne varijable: povrsina stana, broj soba, broj kupaonica
- imamo zavisnu varijablu: cijena stana
- sve nezavisne varijable utječu na cijenu stana
- takodjer, sve nezavisne varijable su medjusobno korelirane
- Vaš investitor želi znati koliko svaki novi metar stana utječe na cijenu, važna mu je interpretabilnost
- vaš investitor takodjer želi znati koliko broj soba ili broj kupaona utječe na cijenu stana
- radi bolje točnosti modela (bolje predikcije cijene stana), mičete varijablu broj kupaona (mičete onu koja nosi najmanje informacija)

0.2 Primjer 2 - angažman na web stranici, zanima vas koji od vaših klijenta sljedeći mjesec odustaje od kupovine (da/ne), to je zavisna varijabla

- nezavisne varijable su broj posjeta web stranici, prosječno vrijeme na stranici, broj pregledanih proizvoda
- nezavisne varijable su medjusobno korelirane
- vas ne zanima INTERPRETABILNOST
- zanima vas samo hoće li kupac odustati
- korištenjem metode PCA stvarate novu varijablu (angažiranost), koja će objediniti sve tri ulazne varijable, zadržati njhovu informativnost, a "sakriti" njihove medjusobne korelacije i poboljšati točnost predikcije
- što je veći "angažman" kupca, on i dalje ostaje naš klijent

[1]:
```
import pandas as pd
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
```

[2]:
```
#angazman = pd.read_csv ("angazman.csv") - ovaj dataset nedostaje
d = {
    'Broj_Posjeta_Web_Stranici_u_Tjednu': [47,41,53,63,60,42,20,48,40,32],
    'Prosjecno_Vrijeme_na_Stranici_Min': [52,51,71,87,51,61,41,52,56,43],
    'Broj_Pregledanih_Proizvoda': [95,60,100,125,74,96,54,86,76,49],
```

```
    'Odustajanje_Kupca': [1,0,1,0,1,0,1,0,1,1],
    'Angaziranost': [0.655226, -0.464907, 1.645457, 3.189451, -0.644420, 0.
↪732590, -1.883173, 0.491993, 0.018530, -1.385621]
}
angazman = pd.DataFrame(d)
```

[3]:
```
angazman.head()
```

[3]:
```
   Broj_Posjeta_Web_Stranici_u_Tjednu  Prosjecno_Vrijeme_na_Stranici_Min  \
0                                  47                                 52
1                                  41                                 51
2                                  53                                 71
3                                  63                                 87
4                                  60                                 51

   Broj_Pregledanih_Proizvoda  Odustajanje_Kupca  Angaziranost
0                          95                  1      0.655226
1                          60                  0     -0.464907
2                         100                  1      1.645457
3                         125                  0      3.189451
4                          74                  1     -0.644420
```

[4]:
```
angazman.corr()
```

[4]:
```
                                    Broj_Posjeta_Web_Stranici_u_Tjednu  \
Broj_Posjeta_Web_Stranici_u_Tjednu                            1.000000
Prosjecno_Vrijeme_na_Stranici_Min                             0.721033
Broj_Pregledanih_Proizvoda                                    0.751867
Odustajanje_Kupca                                            -0.262783
Angaziranost                                                  0.761536

                                    Prosjecno_Vrijeme_na_Stranici_Min  \
Broj_Posjeta_Web_Stranici_u_Tjednu                           0.721033
Prosjecno_Vrijeme_na_Stranici_Min                            1.000000
Broj_Pregledanih_Proizvoda                                   0.897363
Odustajanje_Kupca                                           -0.393187
Angaziranost                                                 0.952756

                                    Broj_Pregledanih_Proizvoda  \
Broj_Posjeta_Web_Stranici_u_Tjednu                    0.751867
Prosjecno_Vrijeme_na_Stranici_Min                     0.897363
Broj_Pregledanih_Proizvoda                            1.000000
Odustajanje_Kupca                                    -0.373959
Angaziranost                                          0.964419

                                    Odustajanje_Kupca  Angaziranost
Broj_Posjeta_Web_Stranici_u_Tjednu          -0.262783      0.761536
```

```
Prosjecno_Vrijeme_na_Stranici_Min          -0.393187      0.952756
Broj_Pregledanih_Proizvoda                 -0.373959      0.964419
Odustajanje_Kupca                           1.000000     -0.438088
Angaziranost                               -0.438088      1.000000
```

[5]:
```
# PCA je osjetljiv na skalu varijabli, pa ih moramo standardizirati.
# Stvaramo novi DataFrame samo s nezavisnim varijablama za PCA.
X = angazman[['Broj_Posjeta_Web_Stranici_u_Tjednu',
        'Prosjecno_Vrijeme_na_Stranici_Min',
        'Broj_Pregledanih_Proizvoda']]
```

[6]:
```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

[7]:
```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

[8]:
```
X_scaled
```

[8]:
```
array([[ 0.19805648, -0.34671849,  0.60322583],
       [-0.29708472, -0.42376705, -0.96069299],
       [ 0.69319768,  1.11720403,  0.8266428 ],
       [ 1.518433  ,  2.3499809 ,  1.94372767],
       [ 1.27086241, -0.42376705, -0.33512546],
       [-0.21456119,  0.34671849,  0.64790922],
       [-2.03007891, -1.19425259, -1.22879336],
       [ 0.28058001, -0.34671849,  0.20107528],
       [-0.37960825, -0.03852428, -0.24575867],
       [-1.03979651, -1.04015548, -1.45221033]])
```

objedinit ćemo tri ulazne varijable u jednu

[9]:
```
pca = PCA(n_components=1)
principal_components = pca.fit_transform(X_scaled)
```

dodajemo novu kolonu angažiranost na naš dataset

[10]:
```
angazman['Angaziranost'] = principal_components
```

[11]:
```
angazman
```

[11]:
```
   Broj_Posjeta_Web_Stranici_u_Tjednu  Prosjecno_Vrijeme_na_Stranici_Min  \
0                                  47                                 52
1                                  41                                 51
2                                  53                                 71
3                                  63                                 87
4                                  60                                 51
5                                  42                                 61
6                                  20                                 41
```

```
7                                  48                                  52
8                                  40                                  56
9                                  32                                  43
```

```
   Broj_Pregledanih_Proizvoda  Odustajanje_Kupca  Angaziranost
0                          95                  1      0.263634
1                          60                  0     -0.982905
2                         100                  1      1.528244
3                         125                  0      3.369489
4                          74                  1      0.250682
5                          96                  0      0.470572
6                          54                  1     -2.547325
7                          86                  0      0.070089
8                          76                  1     -0.377307
9                          49                  1     -2.045172
```

[12]:
```
angazman.corr()
```

[12]:
```
                                Broj_Posjeta_Web_Stranici_u_Tjednu  \
Broj_Posjeta_Web_Stranici_u_Tjednu                        1.000000
Prosjecno_Vrijeme_na_Stranici_Min                         0.721033
Broj_Pregledanih_Proizvoda                                0.751867
Odustajanje_Kupca                                        -0.262783
Angaziranost                                              0.883472

                                Prosjecno_Vrijeme_na_Stranici_Min  \
Broj_Posjeta_Web_Stranici_u_Tjednu                       0.721033
Prosjecno_Vrijeme_na_Stranici_Min                        1.000000
Broj_Pregledanih_Proizvoda                               0.897363
Odustajanje_Kupca                                       -0.393187
Angaziranost                                             0.943769

                                Broj_Pregledanih_Proizvoda  \
Broj_Posjeta_Web_Stranici_u_Tjednu                0.751867
Prosjecno_Vrijeme_na_Stranici_Min                 0.897363
Broj_Pregledanih_Proizvoda                        1.000000
Odustajanje_Kupca                                -0.373959
Angaziranost                                      0.954752

                                Odustajanje_Kupca  Angaziranost
Broj_Posjeta_Web_Stranici_u_Tjednu      -0.262783     0.883472
Prosjecno_Vrijeme_na_Stranici_Min       -0.393187     0.943769
Broj_Pregledanih_Proizvoda              -0.373959     0.954752
Odustajanje_Kupca                        1.000000    -0.371801
Angaziranost                            -0.371801     1.000000
```

u budućim analzama gdje želimo predvidjeti odustajanje kupca, koristit će se varijabla angaziranost. Jako teško ju je interpretirati, ali nas to i ne zanima.

[ ]: