# QDA_feature_selection

December 22, 2022

[3]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

import sklearn.linear_model as skl_lm
import sklearn.discriminant_analysis as skl_da
from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler
```

[156]:
```
# creating the data set (initial columns + new ones that we used in the data␣
 ↪analysis)

d = pd.read_csv(r"C:\Users\billt\OneDrive\Desktop\SML_project\train.csv")
d["Number of words co-lead"] = d["Number of words lead"] - d["Difference in␣
 ↪words lead and co-lead"]
d = d.drop( ["Difference in words lead and co-lead"], axis="columns")
d2 = pd.DataFrame()

lead=list()
colead=list()
femrest=list()
malerest=list()


for i in range(1039):
    lead.append(d.iloc[i,2] / d.iloc[i,1])
    colead.append(d.iloc[i,13] / d.iloc[i,1])
    femrest.append( (d.iloc[i,0] / d.iloc[i,1]))
    malerest.append( (d.iloc[i,6] / d.iloc[i,1]))

d2 = pd.DataFrame( {"lead perc":lead, "colead perc":colead, "fem rest perc":
 ↪femrest, "male rest perc":malerest, "Lead":d["Lead"], "year":d["Year"],␣
 ↪"gross":d["Gross"] } )

d["lead perc"]=d2["lead perc"]
d["colead perc"]=d2["colead perc"]
```

```
d["fem rest perc"]=d2["fem rest perc"]
d["male rest perc"]=d2["male rest perc"]
d
```

[156]:

| | Number words female | Total words | Number of words lead | \ |
|---|---|---|---|---|
| 0 | 1512 | 6394 | 2251.0 | |
| 1 | 1524 | 8780 | 2020.0 | |
| 2 | 155 | 4176 | 942.0 | |
| 3 | 1073 | 9855 | 3440.0 | |
| 4 | 1317 | 7688 | 3835.0 | |
| … | … | … | … | |
| 1034 | 303 | 2398 | 1334.0 | |
| 1035 | 632 | 8404 | 1952.0 | |
| 1036 | 1326 | 2750 | 877.0 | |
| 1037 | 462 | 3994 | 775.0 | |
| 1038 | 2735 | 11946 | 3410.0 | |

| | Number of male actors | Year | Number of female actors | Number words male | \ |
|---|---|---|---|---|---|
| 0 | 2 | 1995 | 5 | 2631 | |
| 1 | 9 | 2001 | 4 | 5236 | |
| 2 | 7 | 1968 | 1 | 3079 | |
| 3 | 12 | 2002 | 2 | 5342 | |
| 4 | 8 | 1988 | 4 | 2536 | |
| … | … | … | … | … | |
| 1034 | 5 | 1973 | 2 | 761 | |
| 1035 | 6 | 1992 | 2 | 5820 | |
| 1036 | 2 | 2000 | 3 | 547 | |
| 1037 | 8 | 1996 | 3 | 2757 | |
| 1038 | 13 | 2007 | 4 | 5801 | |

| | Gross | Mean Age Male | Mean Age Female | Age Lead | Age Co-Lead | Lead | \ |
|---|---|---|---|---|---|---|---|
| 0 | 142.0 | 51.500000 | 42.333333 | 46.0 | 65.0 | Female | |
| 1 | 37.0 | 39.125000 | 29.333333 | 58.0 | 34.0 | Male | |
| 2 | 376.0 | 42.500000 | 37.000000 | 46.0 | 37.0 | Male | |
| 3 | 19.0 | 35.222222 | 21.500000 | 33.0 | 23.0 | Male | |
| 4 | 40.0 | 45.250000 | 45.000000 | 36.0 | 39.0 | Male | |
| … | … | … | … | … | … | | |
| 1034 | 174.0 | 43.200000 | 31.000000 | 46.0 | 24.0 | Male | |
| 1035 | 172.0 | 37.166667 | 24.000000 | 21.0 | 34.0 | Female | |
| 1036 | 53.0 | 27.500000 | 27.666667 | 28.0 | 25.0 | Male | |
| 1037 | 32.0 | 42.857143 | 38.500000 | 29.0 | 32.0 | Female | |
| 1038 | 32.0 | 44.090909 | 50.000000 | 38.0 | 48.0 | Male | |

| | Number of words co-lead | lead perc | colead perc | fem rest perc | \ |
|---|---|---|---|---|---|
| 0 | 1908.0 | 0.352049 | 0.298405 | 0.236472 | |
| 1 | 801.0 | 0.230068 | 0.091230 | 0.173576 | |
| 2 | 155.0 | 0.225575 | 0.037117 | 0.037117 | |

```
3                       817.0   0.349061     0.082902       0.108879
4                       686.0   0.498829     0.089230       0.171306
…                         …       …            …              …
1034                    168.0   0.556297     0.070058       0.126355
1035                   1765.0   0.232270     0.210019       0.075202
1036                    521.0   0.318909     0.189455       0.482182
1037                    723.0   0.194041     0.181022       0.115674
1038                   1874.0   0.285451     0.156873       0.228947

      male rest perc
0           0.411480
1           0.596355
2           0.737308
3           0.542060
4           0.329865
…                ...
1034        0.317348
1035        0.692527
1036        0.198909
1037        0.690285
1038        0.485602

[1039 rows x 18 columns]
```

[ ]:
```
#FEATURE COMBINATIONS
results = pd.DataFrame({"columns":[] , "train accuracy":[] , "test accuracy":[]
 ↪})
y = d["Lead"]
import random

import itertools
col = d.columns.tolist()
col.remove("Lead")
combinations = []
for r in range(len(col)+1):
    for combination in itertools.combinations(col, r):
        combinations.append(combination)
combinations = combinations[1:]

for comb in combinations:
    x = d.drop("Lead", axis="columns")
    x = x.drop(list(set(x) - set(comb)), axis="columns")

    x_train, x_test, y_train, y_test = train_test_split(x, y, random_state=4045)

    scaler1 = StandardScaler()
    scaler1.fit(x_train)
```

```
    x_train=scaler1.transform(x_train)
    x_train = pd.DataFrame(x_train)

    x_test=scaler1.transform(x_test)
    x_test = pd.DataFrame(x_test)

    qda = skl_da.QuadraticDiscriminantAnalysis()
    qda.fit(x_train, y_train)

    l=list()
    l.append(comb)
    l.append(np.mean(qda.predict(x_train.iloc[:,])==y_train))
    l.append(np.mean(qda.predict(x_test.iloc[:,])==y_test) )

    results.loc[len(results.index)] = l
```

[159]:
```
results.loc[results["test accuracy"]==max(results["test accuracy"])]
```

[159]:

| | columns | train accuracy | test accuracy |
|---|---|---|---|
| 126693 | (Total words, Number of words lead, Number of … | 0.913992 | 0.95 |

[167]:
```
r2 = results.sort_values("test accuracy",ascending=False)
r2.head(20)
```

[167]:

| | columns | train accuracy | test accuracy |
|---|---|---|---|
| 126693 | (Total words, Number of words lead, Number of … | 0.913992 | 0.95 |
| 121420 | (Number of male actors, Year, Number of female… | 0.925546 | 0.95 |
| 124173 | (Number words female, Total words, Number of m… | 0.939666 | 0.95 |
| 122678 | (Number words female, Total words, Number of w… | 0.939666 | 0.95 |
| 125174 | (Number words female, Number of words lead, Nu… | 0.939666 | 0.95 |
| 121418 | (Number of male actors, Year, Number of female… | 0.925546 | 0.95 |
| 121417 | (Number of male actors, Year, Number of female… | 0.925546 | 0.95 |
| 126539 | (Total words, Number of words lead, Number of … | 0.939666 | 0.95 |
| 110602 | (Number words female, Total words, Number of w… | 0.935815 | 0.95 |
| 94369 | (Number words female, Total words, Number of m… | 0.934531 | 0.95 |
| 116921 | (Number words female, Number of male actors, N… | 0.938383 | 0.95 |
| 116922 | (Number words female, Number of male actors, N… | 0.938383 | 0.95 |
| 116924 | (Number words female, Number of male actors, N… | 0.938383 | 0.95 |
| 113130 | (Number words female, Total words, Number of m… | 0.937099 | 0.95 |
| 115223 | (Number words female, Number of words lead, Nu… | 0.935815 | 0.95 |
| 115132 | (Number words female, Number of words lead, Nu… | 0.937099 | 0.95 |
| 124805 | (Number words female, Number of words lead, Nu… | 0.934531 | 0.95 |
| 97372 | (Number words female, Number of words lead, Nu… | 0.934531 | 0.95 |

```
113221  (Number words female, Total words, Number of m…        0.935815
127797  (Number of male actors, Year, Number of female…        0.929397

        test accuracy
126693       0.950000
121420       0.946154
124173       0.946154
122678       0.946154
125174       0.946154
121418       0.946154
121417       0.946154
126539       0.946154
110602       0.942308
94369        0.942308
116921       0.942308
116922       0.942308
116924       0.942308
113130       0.942308
115223       0.942308
115132       0.942308
124805       0.942308
97372        0.942308
113221       0.942308
127797       0.942308
```

[ ]:
```
w=d[list(('Total words',
 'Number of male actors',
 'Number of female actors',
 'Mean Age Male',
 'Mean Age Female',
 'Age Lead',
 'Age Co-Lead',
 'lead perc',
 'colead perc',
 'fem rest perc'))]
y=d["Lead"]
```

[165]:
```
results
```

[165]:
```
                                                columns  train accuracy  \
0                                  (Number words female,)        0.754814
1                                          (Total words,)        0.757381
2                                (Number of words lead,)        0.757381
3                               (Number of male actors,)        0.757381
4                                                (Year,)        0.757381
…                                                      …             …
131066  (Number words female, Total words, Number of w…        0.876765
```

```
131067  (Number words female, Total words, Number of m…        0.921694
131068  (Number words female, Number of words lead, Nu…        0.915276
131069  (Total words, Number of words lead, Number of …        0.899872
131070  (Number words female, Total words, Number of w…        0.835687

        test accuracy
0            0.746154
1            0.750000
2            0.750000
3            0.750000
4            0.750000
…               …
131066       0.811538
131067       0.907692
131068       0.896154
131069       0.876923
131070       0.803846

[131071 rows x 3 columns]
```