Untitled

September 20, 2022

[1]:
```
import matplotlib.pyplot as plt

import pandas as pd

import numpy as np

from sklearn import (
    ensemble,
    model_selection,
    preprocessing,
    tree,
)

from sklearn.metrics import (
    auc,
    confusion_matrix,
    roc_auc_score,
    roc_curve,
)

from sklearn.model_selection import(
    train_test_split,
    StratifiedKFold,
)

from yellowbrick.classifier import (
    ConfusionMatrix,
    ROCAUC,
)

from yellowbrick.model_selection import (
    LearningCurve,
)
```

[2]:
```
df = pd.read_excel('titanic3.xls')
```

[3]:
```
from io import StringIO
import sys
```

[4]:
```
csv_data = \
'''A,B,C,D
1.0,2.0,3.0,4.0
5.0,6.0,,8.0
10.0,11.0,12.0,'''
```

[5]:
```
if (sys.version_info < (3, 0)):
    csv_data=unicode(csv_data)
```

[6]:
```
df2 = pd.read_csv(StringIO(csv_data))
```

[7]:
```
df2
```

[7]:

	A	B	C	D
0	1.0	2.0	3.0	4.0
1	5.0	6.0	NaN	8.0
2	10.0	11.0	12.0	NaN

[8]:
```
org_df=df
```

[9]:
```
df.dtypes
```

[9]:
```
pclass          int64
survived        int64
name           object
sex            object
age           float64
sibsp           int64
parch           int64
ticket         object
fare          float64
cabin          object
embarked       object
boat           object
body          float64
home.dest      object
dtype: object
```

[10]:
```
!pip install ipywidgets pandas_profiling
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: ipywidgets in
/home/hefesto/.local/lib/python3.10/site-packages (8.0.2)
Requirement already satisfied: pandas_profiling in
/home/hefesto/.local/lib/python3.10/site-packages (3.3.0)
Requirement already satisfied: traitlets>=4.3.1 in /usr/lib/python3/dist-
packages (from ipywidgets) (5.3.0)
Requirement already satisfied: ipykernel>=4.5.1 in /usr/lib/python3/dist-
```

```
packages (from ipywidgets) (6.15.1)
Requirement already satisfied: ipython>=6.1.0 in /usr/lib/python3/dist-packages
(from ipywidgets) (7.31.1)
Requirement already satisfied: jupyterlab-widgets~=3.0 in
/home/hefesto/.local/lib/python3.10/site-packages (from ipywidgets) (3.0.3)
Requirement already satisfied: widgetsnbextension~=4.0 in
/home/hefesto/.local/lib/python3.10/site-packages (from ipywidgets) (4.0.3)
Requirement already satisfied: pydantic<1.10,>=1.8.1 in
/home/hefesto/.local/lib/python3.10/site-packages (from pandas_profiling)
(1.9.2)
Requirement already satisfied: jinja2<3.2,>=2.11.1 in /usr/lib/python3/dist-
packages (from pandas_profiling) (3.0.3)
Requirement already satisfied: htmlmin==0.1.12 in
/home/hefesto/.local/lib/python3.10/site-packages (from pandas_profiling)
(0.1.12)
Requirement already satisfied: multimethod<1.9,>=1.4 in
/home/hefesto/.local/lib/python3.10/site-packages (from pandas_profiling) (1.8)
Requirement already satisfied: numpy<1.24,>=1.16.0 in /usr/lib/python3/dist-
packages (from pandas_profiling) (1.21.5)
Requirement already satisfied: statsmodels<0.14,>=0.13.2 in
/home/hefesto/.local/lib/python3.10/site-packages (from pandas_profiling)
(0.13.2)
Requirement already satisfied: tangled-up-in-unicode==0.2.0 in
/home/hefesto/.local/lib/python3.10/site-packages (from pandas_profiling)
(0.2.0)
Requirement already satisfied: phik<0.13,>=0.11.1 in
/home/hefesto/.local/lib/python3.10/site-packages (from pandas_profiling)
(0.12.2)
Requirement already satisfied: tqdm<4.65,>=4.48.2 in
/home/hefesto/.local/lib/python3.10/site-packages (from pandas_profiling)
(4.64.1)
Requirement already satisfied: matplotlib<3.6,>=3.2 in /usr/lib/python3/dist-
packages (from pandas_profiling) (3.5.2)
Requirement already satisfied: missingno<0.6,>=0.4.2 in
/home/hefesto/.local/lib/python3.10/site-packages (from pandas_profiling)
(0.5.1)
Requirement already satisfied: seaborn<0.12,>=0.10.1 in
/home/hefesto/.local/lib/python3.10/site-packages (from pandas_profiling)
(0.11.2)
Requirement already satisfied: pandas!=1.4.0,<1.5,>1.1 in
/home/hefesto/.local/lib/python3.10/site-packages (from pandas_profiling)
(1.4.3)
Requirement already satisfied: visions[type_image_path]==0.7.5 in
/home/hefesto/.local/lib/python3.10/site-packages (from pandas_profiling)
(0.7.5)
Requirement already satisfied: joblib~=1.1.0 in
/home/hefesto/.local/lib/python3.10/site-packages (from pandas_profiling)
(1.1.0)
```

```
Requirement already satisfied: PyYAML<6.1,>=5.0.0 in /usr/lib/python3/dist-
packages (from pandas_profiling) (5.4.1)
Requirement already satisfied: scipy<1.10,>=1.4.1 in /usr/lib/python3/dist-
packages (from pandas_profiling) (1.7.3)
Requirement already satisfied: requests<2.29,>=2.24.0 in /usr/lib/python3/dist-
packages (from pandas_profiling) (2.27.1)
Requirement already satisfied: networkx>=2.4 in
/home/hefesto/.local/lib/python3.10/site-packages (from
visions[type_image_path]==0.7.5->pandas_profiling) (2.8.6)
Requirement already satisfied: attrs>=19.3.0 in /usr/lib/python3/dist-packages
(from visions[type_image_path]==0.7.5->pandas_profiling) (22.1.0)
Requirement already satisfied: imagehash in
/home/hefesto/.local/lib/python3.10/site-packages (from
visions[type_image_path]==0.7.5->pandas_profiling) (4.3.0)
Requirement already satisfied: Pillow in /usr/lib/python3/dist-packages (from
visions[type_image_path]==0.7.5->pandas_profiling) (9.2.0)
Requirement already satisfied: tornado>=6.1 in /usr/lib/python3/dist-packages
(from ipykernel>=4.5.1->ipywidgets) (6.2)
Requirement already satisfied: packaging in /usr/lib/python3/dist-packages (from
ipykernel>=4.5.1->ipywidgets) (21.3)
Requirement already satisfied: psutil in /usr/lib/python3/dist-packages (from
ipykernel>=4.5.1->ipywidgets) (5.9.0)
Requirement already satisfied: pyzmq>=17 in /usr/local/lib/python3.10/dist-
packages (from ipykernel>=4.5.1->ipywidgets) (23.2.0)
Requirement already satisfied: nest-asyncio in /usr/lib/python3/dist-packages
(from ipykernel>=4.5.1->ipywidgets) (1.5.4)
Requirement already satisfied: matplotlib-inline>=0.1 in /usr/lib/python3/dist-
packages (from ipykernel>=4.5.1->ipywidgets) (0.1.3)
Requirement already satisfied: debugpy>=1.0 in /usr/local/lib/python3.10/dist-
packages (from ipykernel>=4.5.1->ipywidgets) (1.6.2)
Requirement already satisfied: jupyter-client>=6.1.12 in /usr/lib/python3/dist-
packages (from ipykernel>=4.5.1->ipywidgets) (7.3.4)
Requirement already satisfied: python-dateutil>=2.8.1 in
/usr/local/lib/python3.10/dist-packages (from
pandas!=1.4.0,<1.5,>1.1->pandas_profiling) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/lib/python3/dist-packages
(from pandas!=1.4.0,<1.5,>1.1->pandas_profiling) (2022.2.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/home/hefesto/.local/lib/python3.10/site-packages (from
pydantic<1.10,>=1.8.1->pandas_profiling) (4.3.0)
Requirement already satisfied: patsy>=0.5.2 in
/home/hefesto/.local/lib/python3.10/site-packages (from
statsmodels<0.14,>=0.13.2->pandas_profiling) (0.5.2)
Requirement already satisfied: entrypoints in /usr/lib/python3/dist-packages
(from jupyter-client>=6.1.12->ipykernel>=4.5.1->ipywidgets) (0.4)
Requirement already satisfied: jupyter-core>=4.9.2 in /usr/lib/python3/dist-
packages (from jupyter-client>=6.1.12->ipykernel>=4.5.1->ipywidgets) (4.11.1)
Requirement already satisfied: six in /usr/lib/python3/dist-packages (from
```

```
patsy>=0.5.2->statsmodels<0.14,>=0.13.2->pandas_profiling) (1.16.0)
Requirement already satisfied: PyWavelets in
/home/hefesto/.local/lib/python3.10/site-packages (from
imagehash->visions[type_image_path]==0.7.5->pandas_profiling) (1.4.1)
```

[11]:
```
import pandas_profiling
```

[12]:
```
pandas_profiling.ProfileReport(df)
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

<IPython.core.display.HTML object>
```

[12]:

[13]:
```
from pandas_profiling import ProfileReport
```

[14]:
```
profile = ProfileReport(df, title="Pandas Profiling Report")
profile.to_file("your_report.html")
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

Export report to file:   0%|          | 0/1 [00:00<?, ?it/s]
```

[15]:
```
df.shape
```

[15]:
```
(1309, 14)
```

[16]:
```
df.describe()
```

[16]:

	pclass	survived	age	sibsp	parch	fare	body
count	1309.000000	1309.000000	1046.000000	1309.000000	1309.000000	1308.000000	121.000000
mean	2.294882	0.381971	29.881135	0.498854	0.385027	33.295479	160.809917
std	0.837836	0.486055	14.413500	1.041658	0.865560	51.758668	97.696922
min	1.000000	0.000000	0.166700	0.000000	0.000000		
25%	2.000000	0.000000	21.000000	0.000000	0.000000		
50%	3.000000	0.000000	28.000000	0.000000	0.000000		
75%	3.000000	1.000000	39.000000	1.000000	0.000000		
max	3.000000	1.000000	80.000000	8.000000	9.000000		

```
min       0.000000    1.000000
25%       7.895800   72.000000
50%      14.454200  155.000000
75%      31.275000  256.000000
max     512.329200  328.000000
```

[17]:
```
df.describe().iloc[:,:2]
```

[17]:
```
             pclass     survived
count  1309.000000  1309.000000
mean      2.294882     0.381971
std       0.837836     0.486055
min       1.000000     0.000000
25%       2.000000     0.000000
50%       3.000000     0.000000
75%       3.000000     1.000000
max       3.000000     1.000000
```

[18]:
```
df.isnull().sum()
```

[18]:
```
pclass          0
survived        0
name            0
sex             0
age           263
sibsp           0
parch           0
ticket          0
fare            1
cabin        1014
embarked        2
boat          823
body         1188
home.dest     564
dtype: int64
```

[19]:
```
df.isnull().mean()*100
```

[19]:
```
pclass        0.000000
survived      0.000000
name          0.000000
sex           0.000000
age          20.091673
sibsp         0.000000
parch         0.000000
ticket        0.000000
fare          0.076394
```

```
cabin        77.463713
embarked      0.152788
boat         62.872422
body         90.756303
home.dest    43.086325
dtype: float64
```

[20]:
```
mascara=df.isnull().any(axis=1)
```

[21]:
```
mascara.head()
```

[21]:
```
0    True
1    True
2    True
3    True
4    True
dtype: bool
```

[22]:
```
df.sex.value_counts(dropna=False)
```

[22]:
```
male      843
female    466
Name: sex, dtype: int64
```

[23]:
```
df.embarked.value_counts(dropna=False)
```

[23]:
```
S      914
C      270
Q      123
NaN      2
Name: embarked, dtype: int64
```

[24]:
```
df.body.value_counts(dropna=False)
```

[24]:
```
NaN      1188
58.0        1
285.0       1
156.0       1
143.0       1
           …
174.0       1
169.0       1
245.0       1
46.0        1
304.0       1
Name: body, Length: 122, dtype: int64
```

[25]:
```
df.name.head(10)
```

[25]:
```
0                      Allen, Miss. Elisabeth Walton
1                     Allison, Master. Hudson Trevor
2                       Allison, Miss. Helen Loraine
3               Allison, Mr. Hudson Joshua Creighton
4    Allison, Mrs. Hudson J C (Bessie Waldo Daniels)
5                                 Anderson, Mr. Harry
6                   Andrews, Miss. Kornelia Theodosia
7                             Andrews, Mr. Thomas Jr
8      Appleton, Mrs. Edward Dale (Charlotte Lamson)
9                            Artagaveytia, Mr. Ramon
Name: name, dtype: object
```

[26]:
```
df = df.drop( columns=['name','ticket','home.dest','boat','body','cabin'])
```

[27]:
```
df
```

[27]:

	pclass	survived	sex	age	sibsp	parch	fare	embarked
0	1	1	female	29.0000	0	0	211.3375	S
1	1	1	male	0.9167	1	2	151.5500	S
2	1	0	female	2.0000	1	2	151.5500	S
3	1	0	male	30.0000	1	2	151.5500	S
4	1	0	female	25.0000	1	2	151.5500	S
…	…	…	…	…	…	…	…	…
1304	3	0	female	14.5000	1	0	14.4542	C
1305	3	0	female	NaN	1	0	14.4542	C
1306	3	0	male	26.5000	0	0	7.2250	C
1307	3	0	male	27.0000	0	0	7.2250	C
1308	3	0	male	29.0000	0	0	7.8750	S

[1309 rows x 8 columns]

[28]:
```
df.columns
```

[28]:
```
Index(['pclass', 'survived', 'sex', 'age', 'sibsp', 'parch', 'fare',
       'embarked'],
      dtype='object')
```

[29]:
```
df
```

[29]:

	pclass	survived	sex	age	sibsp	parch	fare	embarked
0	1	1	female	29.0000	0	0	211.3375	S
1	1	1	male	0.9167	1	2	151.5500	S
2	1	0	female	2.0000	1	2	151.5500	S
3	1	0	male	30.0000	1	2	151.5500	S
4	1	0	female	25.0000	1	2	151.5500	S
…	…	…	…	…	…	…	…	…
1304	3	0	female	14.5000	1	0	14.4542	C
1305	3	0	female	NaN	1	0	14.4542	C

```
1306       3         0    male  26.5000      0      0    7.2250        C
1307       3         0    male  27.0000      0      0    7.2250        C
1308       3         0    male  29.0000      0      0    7.8750        S

[1309 rows x 8 columns]
```

[30]:
```
df=pd.get_dummies(df)
```

[31]:
```
df
```

[31]:
```
      pclass  survived      age  sibsp  parch      fare  sex_female  sex_male  \
0          1         1  29.0000      0      0  211.3375           1         0
1          1         1   0.9167      1      2  151.5500           0         1
2          1         0   2.0000      1      2  151.5500           1         0
3          1         0  30.0000      1      2  151.5500           0         1
4          1         0  25.0000      1      2  151.5500           1         0
…        …         …       …    …      …         …          …         …
1304       3         0  14.5000      1      0   14.4542           1         0
1305       3         0     NaN      1      0   14.4542           1         0
1306       3         0  26.5000      0      0    7.2250           0         1
1307       3         0  27.0000      0      0    7.2250           0         1
1308       3         0  29.0000      0      0    7.8750           0         1

      embarked_C  embarked_Q  embarked_S
0              0           0           1
1              0           0           1
2              0           0           1
3              0           0           1
4              0           0           1
…            …           …           …
1304           1           0           0
1305           1           0           0
1306           1           0           0
1307           1           0           0
1308           0           0           1

[1309 rows x 11 columns]
```

[32]:
```
df.columns
```

[32]:
```
Index(['pclass', 'survived', 'age', 'sibsp', 'parch', 'fare', 'sex_female',
       'sex_male', 'embarked_C', 'embarked_Q', 'embarked_S'],
      dtype='object')
```

[33]:
```
df = df.drop( columns=['sex_male'])
```

df.columns

[34]:
```
df.columns
```

[34]:
```
Index(['pclass', 'survived', 'age', 'sibsp', 'parch', 'fare', 'sex_female',
       'embarked_C', 'embarked_Q', 'embarked_S'],
      dtype='object')
```

[35]:
```
X = df.drop( columns=['survived'])
```

[36]:
```
X
```

[36]:

	pclass	age	sibsp	parch	fare	sex_female	embarked_C	embarked_Q	embarked_S
0	1	29.0000	0	0	211.3375	1	0	0	1
1	1	0.9167	1	2	151.5500	0	0	0	1
2	1	2.0000	1	2	151.5500	1	0	0	1
3	1	30.0000	1	2	151.5500	0	0	0	1
4	1	25.0000	1	2	151.5500	1	0	0	1
…	…	…	…	…	…	…	…	…	…
1304	3	14.5000	1	0	14.4542	1	1	0	0
1305	3	NaN	1	0	14.4542	1	1	0	0
1306	3	26.5000	0	0	7.2250	0	1	0	0
1307	3	27.0000	0	0	7.2250	0	1	0	0
1308	3	29.0000	0	0	7.8750	0	0	0	1

[1309 rows x 9 columns]

[37]:
```
y=df.survived
```

[38]:
```
y
```

[38]:
```
0       1
1       1
2       0
3       0
4       0
       ..
```

```
1304    0
1305    0
1306    0
1307    0
1308    0
Name: survived, Length: 1309, dtype: int64
```

[39]:
```
df2.isnull().sum()
```

[39]:
```
A    0
B    0
C    1
D    1
dtype: int64
```

[40]:
```
df2
```

[40]:

	A	B	C	D
0	1.0	2.0	3.0	4.0
1	5.0	6.0	NaN	8.0
2	10.0	11.0	12.0	NaN

[41]:
```
df2.values
```

[41]:
```
array([[ 1.,  2.,  3.,  4.],
       [ 5.,  6., nan,  8.],
       [10., 11., 12., nan]])
```

[42]:
```
df2.dropna(axis=0)
```

[42]:

	A	B	C	D
0	1.0	2.0	3.0	4.0

[43]:
```
df2.dropna(axis=1)
```

[43]:

	A	B
0	1.0	2.0
1	5.0	6.0
2	10.0	11.0

[44]:
```
df2.dropna(subset=['C'])
```

[44]:

	A	B	C	D
0	1.0	2.0	3.0	4.0
2	10.0	11.0	12.0	NaN

[45]:
```
from sklearn.impute import SimpleImputer
import numpy as np
```

```
imr=SimpleImputer(missing_values=np.nan, strategy='mean')
imr=imr.fit(df2.values)
imputed_data=imr.transform(df2.values)
```

[46]:
```
imputed_data
```

[46]:
```
array([[ 1. ,  2. ,  3. ,  4. ],
       [ 5. ,  6. ,  7.5,  8. ],
       [10. , 11. , 12. ,  6. ]])
```

[47]:
```
df2
```

[47]:

	A	B	C	D
0	1.0	2.0	3.0	4.0
1	5.0	6.0	NaN	8.0
2	10.0	11.0	12.0	NaN

[48]:
```
df2.fillna(df2.mean())
```

[48]:

	A	B	C	D
0	1.0	2.0	3.0	4.0
1	5.0	6.0	7.5	8.0
2	10.0	11.0	12.0	6.0

[50]:
```
import janitor as jn
```

[51]:
```
X, y = jn.get_features_targets(df, target_columns='survived')
```

```
/usr/local/lib/python3.10/dist-packages/janitor/utils.py:290: FutureWarning:
get_features_targets() has moved. Please use ml.get_features_targets().
  warnings.warn(message, FutureWarning)
```

[52]:
```
X_treino, X_test, Y_treino, Y_test= model_selection.train_test_split(X, y,␣
↪test_size=0.3, random_state=42)
```

[53]:
```
X_treino
```

[53]:

	pclass	age	sibsp	parch	fare	sex_female	embarked_C	embarked_Q	\
1214	3	NaN	0	0	8.6625	0	0	0	
677	3	26.0	0	0	7.8958	0	0	0	
534	2	19.0	0	0	26.0000	1	0	0	
1174	3	NaN	8	2	69.5500	1	0	0	
864	3	28.0	0	0	7.7750	1	0	0	
…	…	…	…	…	…	…	…	…	
1095	3	NaN	0	0	7.6292	1	0	1	
1130	3	18.0	0	0	7.7750	1	0	0	
1294	3	28.5	0	0	16.1000	0	0	0	

```
860        3  26.0      0      0   7.9250           1           0           0
1126       3  28.0      0      0   7.8958           1           0           0

      embarked_S
1214           1
677            1
534            1
1174           1
864            1
…            …
1095           0
1130           1
1294           1
860            1
1126           1

[916 rows x 9 columns]
```

[54]:
```
Y_treino
```

[54]:
```
1214    0
677     0
534     1
1174    0
864     0
       ..
1095    0
1130    0
1294    0
860     1
1126    0
Name: survived, Length: 916, dtype: int64
```

[55]:
```
X_treino.columns
```

[55]:
```
Index(['pclass', 'age', 'sibsp', 'parch', 'fare', 'sex_female', 'embarked_C',
       'embarked_Q', 'embarked_S'],
      dtype='object')
```

[ ]:
```

```