8d0szfrko

August 2, 2023

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
```

[2]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[3]:
```
df=pd.read_csv("/content/drive/MyDrive/mydatasets/C3_bot_detection_data.csv")
df
```

[3]:

	User ID	Username	\
0	132131	flong	
1	289683	hinesstephanie	
2	779715	roberttran	
3	696168	pmason	
4	704441	noah87	
…	…	…	
49995	491196	uberg	
49996	739297	jessicamunoz	
49997	674475	lynncunningham	
49998	167081	richardthompson	
49999	311204	daniel29	

	Tweet	Retweet Count	\
0	Station activity person against natural majori…	85	
1	Authority research natural life material staff…	55	
2	Manage whose quickly especially foot none to g…	6	
3	Just cover eight opportunity strong policy which.	54	
4	Animal sign six data good or.	26	
…	…	…	
49995	Want but put card direction know miss former h…	64	
49996	Provide whole maybe agree church respond most …	18	
49997	Bring different everyone international capital…	43	

```
49998  Than about single generation itself seek sell …             45
49999  Here morning class various room human true bec…             91

       Mention Count  Follower Count  Verified  Bot Label           Location  \
0                  1            2353     False          1          Adkinston
1                  5            9617      True          0          Sanderston
2                  2            4363      True          0        Harrisonfurt
3                  5            2242      True          1        Martinezberg
4                  3            8438     False          1        Camachoville
…                …               …         …         …                …
49995              0            9911      True          1  Lake Kimberlyburgh
49996              5            9900     False          1           Greenbury
49997              3            6313      True          1         Deborahfort
49998              1            6343     False          0         Stephenside
49999              4            4006     False          0           Novakberg

                Created At                         Hashtags
0      2020-05-11 15:29:50                              NaN
1      2022-11-26 05:18:10                        both live
2      2022-08-08 03:16:54                      phone ahead
3      2021-08-14 22:27:05                ever quickly new I
4      2020-04-13 21:24:21                   foreign mention
…                      …                                …
49995  2023-04-20 11:06:26  teach quality ten education any
49996  2022-10-18 03:57:35            add walk among believe
49997  2020-07-08 03:54:08          onto admit artist first
49998  2022-03-22 12:13:44                             star
49999  2022-12-03 06:11:07                             home

[50000 rows x 11 columns]
```

[4]:
```
df.head()
```

[4]:
```
   User ID        Username                                              Tweet  \
0   132131           flong  Station activity person against natural majori…
1   289683  hinesstephanie  Authority research natural life material staff…
2   779715      roberttran  Manage whose quickly especially foot none to g…
3   696168          pmason  Just cover eight opportunity strong policy which.
4   704441          noah87                      Animal sign six data good or.

   Retweet Count  Mention Count  Follower Count  Verified  Bot Label  \
0             85              1            2353     False          1
1             55              5            9617      True          0
2              6              2            4363      True          0
3             54              5            2242      True          1
4             26              3            8438     False          1
```

```
       Location           Created At            Hashtags
0     Adkinston  2020-05-11 15:29:50                 NaN
1    Sanderston  2022-11-26 05:18:10           both live
2  Harrisonfurt  2022-08-08 03:16:54         phone ahead
3  Martinezberg  2021-08-14 22:27:05  ever quickly new I
4  Camachoville  2020-04-13 21:24:21     foreign mention
```

1 Data Cleaning and Data Preprocessing

[5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   User ID         50000 non-null  int64
 1   Username        50000 non-null  object
 2   Tweet           50000 non-null  object
 3   Retweet Count   50000 non-null  int64
 4   Mention Count   50000 non-null  int64
 5   Follower Count  50000 non-null  int64
 6   Verified        50000 non-null  bool
 7   Bot Label       50000 non-null  int64
 8   Location        50000 non-null  object
 9   Created At      50000 non-null  object
 10  Hashtags        41659 non-null  object
dtypes: bool(1), int64(5), object(5)
memory usage: 3.9+ MB
```

[6]:
```
df.describe()
```

[6]:
```
             User ID  Retweet Count  Mention Count  Follower Count  \
count   50000.000000    50000.00000   50000.000000    50000.000000
mean   548890.680540       50.00560       2.513760     4988.602380
std    259756.681425       29.18116       1.708563     2878.742898
min    100025.000000        0.00000       0.000000        0.000000
25%    323524.250000       25.00000       1.000000     2487.750000
50%    548147.000000       50.00000       3.000000     4991.500000
75%    772983.000000       75.00000       4.000000     7471.000000
max    999995.000000      100.00000       5.000000    10000.000000

          Bot Label
count  50000.000000
mean       0.500360
std        0.500005
```

```
min         0.000000
25%         0.000000
50%         1.000000
75%         1.000000
max         1.000000
```

[7]:
```
df.columns
```

[7]:
```
Index(['User ID', 'Username', 'Tweet', 'Retweet Count', 'Mention Count',
       'Follower Count', 'Verified', 'Bot Label', 'Location', 'Created At',
       'Hashtags'],
      dtype='object')
```

[8]:
```
feature_matrix = df[['User ID','Retweet Count', 'Mention Count',
       'Follower Count', 'Bot Label']]
target_vector = df[["Verified"]]
```

[9]:
```
fs = StandardScaler().fit_transform(feature_matrix)
logr = LogisticRegression()
logr.fit(fs,target_vector)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1143:
DataConversionWarning: A column-vector y was passed when a 1d array was
expected. Please change the shape of y to (n_samples, ), for example using
ravel().
  y = column_or_1d(y, warn=True)
```

[9]:
```
LogisticRegression()
```

[10]:
```
observation=[[1,2,3,4,5]]
prediction = logr.predict(observation)
print(prediction)
```

```
[ True]
```

[11]:
```
logr.classes_
```

[11]:
```
array([False,  True])
```

[12]:
```
logr.predict_proba(observation)
```

[12]:
```
array([[0.48759575, 0.51240425]])
```

Random Forest

[13]:
```
df['Verified'].value_counts()
```

[13]:
```
True     25004
False    24996
Name: Verified, dtype: int64
```

[14]:
```
x=df[['User ID','Retweet Count', 'Mention Count',
       'Follower Count', 'Bot Label']]
y=df['Verified']
```

[17]:
```
g1={"Verified":{True:1, False:2}}
df=df.replace(g1)
df
```

[17]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Location
0	132131	flong	Station activity person against natural majori…	85	1	2353	2	1	Adkinston
1	289683	hinesstephanie	Authority research natural life material staff…	55	5	9617	1	0	Sanderston
2	779715	roberttran	Manage whose quickly especially foot none to g…	6	2	4363	1	0	Harrisonfurt
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	1	1	Martinezberg
4	704441	noah87	Animal sign six data good or.	26	3	8438	2	1	Camachoville
…	…	…	…	…	…	…	…	…	…
49995	491196	uberg	Want but put card direction know miss former h…	64	0	9911	1	1	Lake Kimberlyburgh
49996	739297	jessicamunoz	Provide whole maybe agree church respond most …	18	5	9900	2	1	Greenbury
49997	674475	lynncunningham	Bring different everyone international capital…	43					
49998	167081	richardthompson	Than about single generation itself seek sell …	45					
49999	311204	daniel29	Here morning class various room human true bec…	91					

```
49997              3           6313         1          1         Deborahfort
49998              1           6343         2          0         Stephenside
49999              4           4006         2          0           Novakberg

                Created At                          Hashtags
0      2020-05-11 15:29:50                               NaN
1      2022-11-26 05:18:10                         both live
2      2022-08-08 03:16:54                      phone ahead
3      2021-08-14 22:27:05               ever quickly new I
4      2020-04-13 21:24:21                  foreign mention
…                      …                                 …
49995  2023-04-20 11:06:26  teach quality ten education any
49996  2022-10-18 03:57:35           add walk among believe
49997  2020-07-08 03:54:08         onto admit artist first
49998  2022-03-22 12:13:44                            star
49999  2022-12-03 06:11:07                            home

[50000 rows x 11 columns]
```

[18]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,train_size=0.70)
```

[19]:
```
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier()
rfc.fit(x_train,y_train)
```

[19]:
```
RandomForestClassifier()
```

[20]:
```
parameters = {'max_depth':[1,2,3,4,5],'min_samples_leaf':[5,10,15,20,25],
              'n_estimators': [10,20,30,40,50]
              }
```

[21]:
```
from sklearn.model_selection import GridSearchCV
grid_search =␣
 ↪GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring="accuracy")
grid_search.fit(x_train,y_train)
```

[21]:
```
GridSearchCV(cv=2, estimator=RandomForestClassifier(),
             param_grid={'max_depth': [1, 2, 3, 4, 5],
                         'min_samples_leaf': [5, 10, 15, 20, 25],
                         'n_estimators': [10, 20, 30, 40, 50]},
             scoring='accuracy')
```

[22]:
```
grid_search.best_score_
```

[22]:
```
0.5062571428571429
```

[23]:
```
rfc_best = grid_search.best_estimator_
```

[24]:
```
from sklearn.tree import plot_tree
plt.figure(figsize=(89,40))
plot_tree(rfc_best.estimators_[5], feature_names=x.columns, class_names=['Yes',␣
↪'No'], filled=True)
```

[24]:
```
[Text(0.5, 0.75, 'Retweet Count <= 1.5\ngini = 0.5\nsamples = 22157\nvalue =
[17525, 17475]\nclass = Yes'),
 Text(0.25, 0.25, 'gini = 0.497\nsamples = 438\nvalue = [374, 324]\nclass =
Yes'),
 Text(0.75, 0.25, 'gini = 0.5\nsamples = 21719\nvalue = [17151, 17151]\nclass =
Yes')]
```