```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import warnings
warnings.simplefilter('ignore')
```

```
columns = [ 'UsingIP', 'LongURL', 'ShortURL', 'Symbol@',
'Redirecting//','PrefixSuffix-', 'SubDomains',
           'HTTPS', 'DomainRegLen', 'Favicon', 'NonStdPort',
'HTTPSDomainURL', 'RequestURL', 'AnchorURL',
           'LinksInScriptTags', 'ServerFormHandler', 'InfoEmail',
'AbnormalURL', 'WebsiteForwarding',
           'StatusBarCust', 'DisableRightClick','UsingPopupWindow',
'IframeRedirection', 'AgeofDomain',
           'DNSRecording', 'WebsiteTraffic', 'PageRank',
'GoogleIndex','LinksPointingToPage', 'StatsReport', 'class']
```

```
df_phishing = pd.read_csv('phishing.txt',sep = '\t',delimiter =
',',header=None)
df_phishing
```

```
       0   1   2   3   4   5   6   7   8   9   10  11  12  13  14  15
16  17  \
0      -1   1   1   1  -1  -1  -1  -1  -1   1   1  -1   1  -1   1  -1
-1  -1
1       1   1   1   1   1  -1   0   1  -1   1   1  -1   1   0  -1  -1
1   1
2       1   0   1   1   1  -1  -1  -1  -1   1   1  -1   1   0  -1  -1
-1  -1
3       1   0   1   1   1  -1  -1  -1   1   1   1  -1  -1   0   0  -1
1   1
4       1   0  -1   1   1  -1   1   1  -1   1   1   1   1   0   0  -1
1   1
...    ..  ..  ..  ..  ..  ..  ..  ..  ..  ..  ..  ..  ..  ..  ..  ..
..  ..
11050   1  -1   1  -1   1   1   1   1  -1  -1  -1   1   1   1   1  -1
-1   1
11051  -1   1   1  -1  -1  -1   1  -1  -1  -1  -1   1   1  -1  -1   0
-1  -1
11052   1  -1   1   1   1  -1   1  -1  -1   1   1   1   1   0  -1  -1
1   1
11053  -1  -1   1   1   1  -1  -1  -1   1  -1   1   1  -1  -1   1  -1
1   1
11054  -1  -1   1   1   1  -1  -1  -1   1   1   1   1  -1  -1   0  -1
1   1

       18  19  20  21  22  23  24  25  26  27  28  29  30
0       0   1   1   1   1  -1  -1  -1  -1   1   1  -1  -1
1       0   1   1   1   1  -1  -1   0  -1   1   1   1  -1
```

```
2       0   1   1   1   1   1  -1   1  -1   1   0  -1  -1
3       0   1   1   1   1  -1  -1   1  -1   1  -1   1  -1
4       0  -1   1  -1   1  -1  -1   0  -1   1   1   1   1
...    ..  ..  ..  ..  ..  ..  ..  ..  ..  ..  ..  ..  ..
11050   0  -1  -1  -1  -1   1   1  -1  -1   1   1   1   1
11051   1  -1   1  -1   1   1   1   1   1   1  -1   1  -1
11052   0   1   1   1   1   1   1   1  -1   1   0   1  -1
11053   0  -1   1  -1   1   1   1   1  -1   1   1   1  -1
11054   0   1   1   1   1  -1   1  -1  -1  -1   1  -1  -1

[11055 rows x 31 columns]
```

```
df_phishing.columns = columns
```

```
pd.set_option('Display.max_columns',None)
```

```
df_phishing.head()
```

```
   UsingIP  LongURL  ShortURL  Symbol@  Redirecting//
PrefixSuffix-  \
0       -1        1         1        1             -1             -1

1        1        1         1        1              1             -1

2        1        0         1        1              1             -1

3        1        0         1        1              1             -1

4        1        0        -1        1              1             -1



   SubDomains  HTTPS  DomainRegLen  Favicon  NonStdPort
HTTPSDomainURL  \
0          -1     -1            -1        1           1              -
1
1           0      1            -1        1           1              -
1
2          -1     -1            -1        1           1              -
1
3          -1     -1             1        1           1              -
1
4           1      1            -1        1           1
1

   RequestURL  AnchorURL  LinksInScriptTags  ServerFormHandler
InfoEmail  \
0           1         -1                  1                 -1
-1
1           1          0                 -1                 -1
1
```

```
2           1          0                 -1                 -1
-1
3          -1          0                  0                 -1
1
4           1          0                  0                 -1
1

   AbnormalURL  WebsiteForwarding  StatusBarCust  DisableRightClick  \
0           -1                  0              1                  1
1            1                  0              1                  1
2           -1                  0              1                  1
3            1                  0              1                  1
4            1                  0             -1                  1

   UsingPopupWindow  IframeRedirection  AgeofDomain  DNSRecording  \
0                 1                  1           -1            -1
1                 1                  1           -1            -1
2                 1                  1            1            -1
3                 1                  1           -1            -1
4                -1                  1           -1            -1

   WebsiteTraffic  PageRank  GoogleIndex  LinksPointingToPage
StatsReport  \
0              -1        -1            1                    1
-1
1               0        -1            1                    1
1
2               1        -1            1                    0
-1
3               1        -1            1                   -1
1
4               0        -1            1                    1
1

   class
0     -1
1     -1
2     -1
3     -1
4      1
```

```
df_phishing.tail()
```

```
       UsingIP  LongURL  ShortURL  Symbol@  Redirecting//
PrefixSuffix-  \
11050        1       -1         1       -1              1
1
11051       -1        1         1       -1             -1
-1
11052        1       -1         1        1              1
```

```
-1
11053       -1       -1         1        1              1
-1
11054       -1       -1         1        1              1
-1

       SubDomains  HTTPS  DomainRegLen  Favicon  NonStdPort
HTTPSDomainURL  \
11050           1      1            -1       -1          -1
1
11051           1     -1            -1       -1          -1
1
11052           1     -1            -1        1           1
1
11053          -1     -1             1       -1           1
1
11054          -1     -1             1        1           1
1

       RequestURL  AnchorURL  LinksInScriptTags  ServerFormHandler
InfoEmail  \
11050           1          1                  1                 -1
-1
11051           1         -1                 -1                  0
-1
11052           1          0                 -1                 -1
1
11053          -1         -1                  1                 -1
1
11054          -1         -1                  0                 -1
1

       AbnormalURL  WebsiteForwarding  StatusBarCust
DisableRightClick  \
11050            1                  0             -1                 -
1
11051           -1                  1             -1
1
11052            1                  0              1
1
11053            1                  0             -1
1
11054            1                  0              1
1

       UsingPopupWindow  IframeRedirection  AgeofDomain  DNSRecording
\
11050                -1                 -1            1             1

11051                -1                  1            1             1
```

```
11052                 1                  1            1             1

11053                -1                  1            1             1

11054                 1                  1           -1             1


       WebsiteTraffic  PageRank  GoogleIndex  LinksPointingToPage  \
11050              -1        -1            1                    1
11051               1         1            1                   -1
11052               1        -1            1                    0
11053               1        -1            1                    1
11054              -1        -1           -1                    1

       StatsReport  class
11050            1      1
11051            1     -1
11052            1     -1
11053            1     -1
11054           -1     -1
```

```
df_phishing.shape
```

```
(11055, 31)
```

```
df_phishing.keys()
```

```
Index(['UsingIP', 'LongURL', 'ShortURL', 'Symbol@', 'Redirecting//',
       'PrefixSuffix-', 'SubDomains', 'HTTPS', 'DomainRegLen',
'Favicon',
       'NonStdPort', 'HTTPSDomainURL', 'RequestURL', 'AnchorURL',
       'LinksInScriptTags', 'ServerFormHandler', 'InfoEmail',
'AbnormalURL',
       'WebsiteForwarding', 'StatusBarCust', 'DisableRightClick',
       'UsingPopupWindow', 'IframeRedirection', 'AgeofDomain',
'DNSRecording',
       'WebsiteTraffic', 'PageRank', 'GoogleIndex',
'LinksPointingToPage',
       'StatsReport', 'class'],
      dtype='object')
```

```
df_phishing.size
```

```
342705
```

```
df_phishing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11055 entries, 0 to 11054
Data columns (total 31 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   UsingIP             11055 non-null  int64
 1   LongURL             11055 non-null  int64
 2   ShortURL            11055 non-null  int64
 3   Symbol@             11055 non-null  int64
 4   Redirecting//       11055 non-null  int64
 5   PrefixSuffix-       11055 non-null  int64
 6   SubDomains          11055 non-null  int64
 7   HTTPS               11055 non-null  int64
 8   DomainRegLen        11055 non-null  int64
 9   Favicon             11055 non-null  int64
 10  NonStdPort          11055 non-null  int64
 11  HTTPSDomainURL      11055 non-null  int64
 12  RequestURL          11055 non-null  int64
 13  AnchorURL           11055 non-null  int64
 14  LinksInScriptTags   11055 non-null  int64
 15  ServerFormHandler   11055 non-null  int64
 16  InfoEmail           11055 non-null  int64
 17  AbnormalURL         11055 non-null  int64
 18  WebsiteForwarding   11055 non-null  int64
 19  StatusBarCust       11055 non-null  int64
 20  DisableRightClick   11055 non-null  int64
 21  UsingPopupWindow    11055 non-null  int64
 22  IframeRedirection   11055 non-null  int64
 23  AgeofDomain         11055 non-null  int64
 24  DNSRecording        11055 non-null  int64
 25  WebsiteTraffic      11055 non-null  int64
 26  PageRank            11055 non-null  int64
 27  GoogleIndex         11055 non-null  int64
 28  LinksPointingToPage 11055 non-null  int64
 29  StatsReport         11055 non-null  int64
 30  class               11055 non-null  int64
dtypes: int64(31)
memory usage: 2.6 MB
```

All are integer datatypes

```
df_phishing.describe()
```

```
            UsingIP       LongURL      ShortURL       Symbol@
Redirecting//  \
count  11055.000000  11055.000000  11055.000000  11055.000000
11055.000000
mean       0.313795     -0.633198      0.738761      0.700588
0.741474
std        0.949534      0.766095      0.673998      0.713598
0.671011
min       -1.000000     -1.000000     -1.000000     -1.000000      -
1.000000
25%       -1.000000     -1.000000      1.000000      1.000000
```

```
1.000000
50%        1.000000     -1.000000      1.000000      1.000000
1.000000
75%        1.000000     -1.000000      1.000000      1.000000
1.000000
max        1.000000      1.000000      1.000000      1.000000
1.000000

       PrefixSuffix-    SubDomains         HTTPS  DomainRegLen
Favicon  \
count   11055.000000  11055.000000  11055.000000  11055.000000
11055.000000
mean       -0.734962      0.063953      0.250927     -0.336771
0.628584
std         0.678139      0.817518      0.911892      0.941629
0.777777
min        -1.000000     -1.000000     -1.000000     -1.000000     -
1.000000
25%        -1.000000     -1.000000     -1.000000     -1.000000
1.000000
50%        -1.000000      0.000000      1.000000     -1.000000
1.000000
75%        -1.000000      1.000000      1.000000      1.000000
1.000000
max         1.000000      1.000000      1.000000      1.000000
1.000000

         NonStdPort  HTTPSDomainURL    RequestURL     AnchorURL  \
count  11055.000000    11055.000000  11055.000000  11055.000000
mean       0.728268        0.675079      0.186793     -0.076526
std        0.685324        0.737779      0.982444      0.715138
min       -1.000000       -1.000000     -1.000000     -1.000000
25%        1.000000        1.000000     -1.000000     -1.000000
50%        1.000000        1.000000      1.000000      0.000000
75%        1.000000        1.000000      1.000000      0.000000
max        1.000000        1.000000      1.000000      1.000000

       LinksInScriptTags  ServerFormHandler     InfoEmail
AbnormalURL  \
count       11055.000000       11055.000000  11055.000000
11055.000000
mean           -0.118137          -0.595749      0.635640
0.705292
std             0.763973           0.759143      0.772021
0.708949
min            -1.000000          -1.000000     -1.000000     -
1.000000
25%            -1.000000          -1.000000      1.000000
1.000000
50%             0.000000          -1.000000      1.000000
```

```
1.000000
75%             0.000000          -1.000000      1.000000
1.000000
max             1.000000           1.000000      1.000000
1.000000

       WebsiteForwarding  StatusBarCust  DisableRightClick
UsingPopupWindow  \
count       11055.000000   11055.000000       11055.000000
11055.000000
mean            0.115694       0.762099           0.913885
0.613388
std             0.319872       0.647490           0.405991
0.789818
min             0.000000      -1.000000          -1.000000         -
1.000000
25%             0.000000       1.000000           1.000000
1.000000
50%             0.000000       1.000000           1.000000
1.000000
75%             0.000000       1.000000           1.000000
1.000000
max             1.000000       1.000000           1.000000
1.000000

       IframeRedirection   AgeofDomain  DNSRecording
WebsiteTraffic  \
count       11055.000000  11055.000000  11055.000000    11055.000000

mean            0.816915      0.061239      0.377114        0.287291

std             0.576784      0.998168      0.926209        0.827733

min            -1.000000     -1.000000     -1.000000       -1.000000

25%             1.000000     -1.000000     -1.000000        0.000000

50%             1.000000      1.000000      1.000000        1.000000

75%             1.000000      1.000000      1.000000        1.000000

max             1.000000      1.000000      1.000000        1.000000


           PageRank   GoogleIndex  LinksPointingToPage
StatsReport  \
count  11055.000000  11055.000000         11055.000000  11055.000000

mean      -0.483673      0.721574             0.344007      0.719584
```

```
std        0.875289      0.692369             0.569944      0.694437

min       -1.000000     -1.000000            -1.000000     -1.000000

25%       -1.000000      1.000000             0.000000      1.000000

50%       -1.000000      1.000000             0.000000      1.000000

75%        1.000000      1.000000             1.000000      1.000000

max        1.000000      1.000000             1.000000      1.000000
```

```
              class
count  11055.000000
mean       0.113885
std        0.993539
min       -1.000000
25%       -1.000000
50%        1.000000
75%        1.000000
max        1.000000
```

```
df_phishing.var().sort_values(ascending=False)
```

```
AgeofDomain            0.996340
class                  0.987119
RequestURL             0.965196
UsingIP                0.901614
DomainRegLen           0.886666
DNSRecording           0.857862
HTTPS                  0.831548
PageRank               0.766130
WebsiteTraffic         0.685142
SubDomains             0.668336
UsingPopupWindow       0.623812
Favicon                0.604936
InfoEmail              0.596016
LongURL                0.586902
LinksInScriptTags      0.583654
ServerFormHandler      0.576298
HTTPSDomainURL         0.544317
AnchorURL              0.511422
Symbol@                0.509223
AbnormalURL            0.502609
StatsReport            0.482243
GoogleIndex            0.479374
NonStdPort             0.469669
PrefixSuffix-          0.459873
```

```
ShortURL               0.454274
Redirecting//          0.450256
StatusBarCust          0.419244
IframeRedirection      0.332679
LinksPointingToPage    0.324836
DisableRightClick      0.164829
WebsiteForwarding      0.102318
dtype: float64
```

Checking the null values.

```
df_phishing.isna().sum()
```

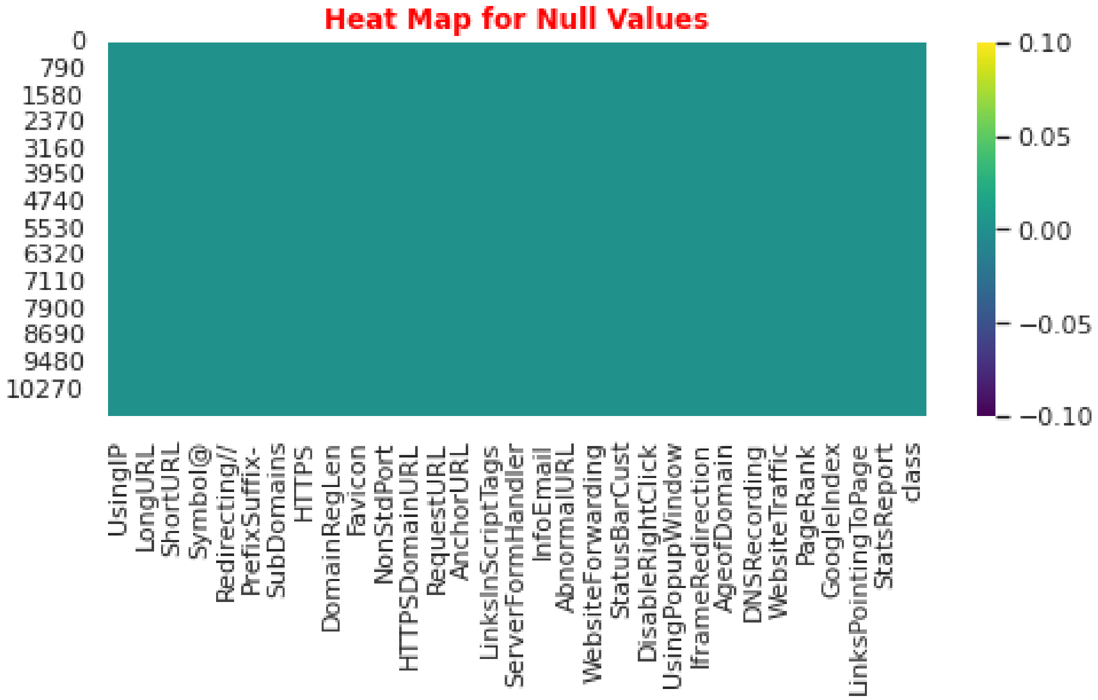
```
UsingIP                0
LongURL                0
ShortURL               0
Symbol@                0
Redirecting//          0
PrefixSuffix-          0
SubDomains             0
HTTPS                  0
DomainRegLen           0
Favicon                0
NonStdPort             0
HTTPSDomainURL         0
RequestURL             0
AnchorURL              0
LinksInScriptTags      0
ServerFormHandler      0
InfoEmail              0
AbnormalURL            0
WebsiteForwarding      0
StatusBarCust          0
DisableRightClick      0
UsingPopupWindow       0
IframeRedirection      0
AgeofDomain            0
DNSRecording           0
WebsiteTraffic         0
PageRank               0
GoogleIndex            0
LinksPointingToPage    0
StatsReport            0
class                  0
dtype: int64
```

```
def null(values):
    if df_phishing.isna().any().any():
        print('Yes')
    else:
        print('No')
```

```
null(df_phishing)
```

```
No
```

```
## Heat map for null values.
plt.figure(figsize=(8,3))
sns.heatmap(df_phishing.isna(),cmap = 'viridis')
plt.title('Heat Map for Null Values',weight = 'bold' , color = 'red')
plt.show()
```



since there is no null values in the dataset data is already cleaned

```
#### Check the output / target vaeiable.
```

```
df_phishing['class'].nunique()
```

```
2
```

```
print(df_phishing['class'].value_counts())
df_phishing['class'].value_counts(normalize=True)*100
```

```
 1    6157
-1    4898
Name: class, dtype: int64
```

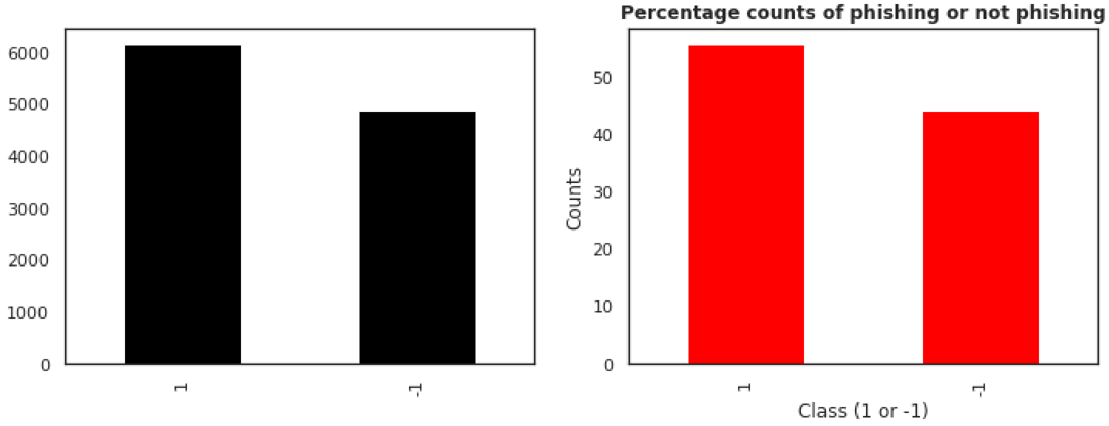
```
 1    55.694256
-1    44.305744
Name: class, dtype: float64
```

```
fig,ax = plt.subplots(1,2,figsize=(12,4))
```

```
df_phishing['class'].value_counts().plot(kind='bar',color='black',ax=a
x[0])
plt.xlabel('Class (1 or -1)')
plt.ylabel('Counts')
plt.title('No of counts of phishing or not phishing',weight = 'bold')

(df_phishing['class'].value_counts(normalize=True)*100).plot(kind='bar
',color='red',ax=ax[1])
plt.title('Percentage counts of phishing or not phishing',weight =
'bold')
plt.xlabel('Class (1 or -1)')
```

```
Text(0.5, 0, 'Class (1 or -1)')
```



Checking the correlation

```
df_phishing.corr().sort_values(by='class',ascending=False)
```

```
                     UsingIP   LongURL  ShortURL   Symbol@
Redirecting//  \
class               0.094160  0.057430 -0.067966  0.052948      -
0.038608
HTTPS               0.071414  0.048754 -0.061426  0.031220      -
0.036200
AnchorURL           0.099847 -0.023396  0.000561  0.057914      -
0.005036
PrefixSuffix-      -0.005257  0.055247 -0.080471 -0.011726      -
0.085590
WebsiteTraffic      0.002922  0.008993 -0.047074  0.032918      -
0.062369
SubDomains         -0.080745  0.003997 -0.041916 -0.058976      -
0.043079
RequestURL          0.029773  0.246348 -0.037235  0.027909      -
0.026368
LinksInScriptTags   0.006212  0.052869 -0.133379 -0.070861      -
0.125583
ServerFormHandler  -0.010962  0.414196 -0.022723 -0.008672      -
0.041672
```

```
GoogleIndex          0.029153  0.002902  0.155844  0.037061
0.178415
AgeofDomain         -0.010446  0.179426 -0.052596 -0.005499      -
0.050107
PageRank            -0.091774  0.183518  0.014591 -0.064735      -
0.003132
UsingIP              1.000000 -0.052411  0.403461  0.158699
0.397389
StatsReport         -0.019103 -0.067153  0.085461 -0.080357
0.070390
DNSRecording        -0.050733 -0.040823  0.436064 -0.047872
0.431409
LongURL             -0.052411  1.000000 -0.097881 -0.075108      -
0.081247
Symbol@              0.158699 -0.075108  0.104447  1.000000
0.086960
StatusBarCust        0.084059 -0.045103  0.062383  0.279697
0.086635
NonStdPort           0.060979  0.000323  0.002201  0.364891
0.025060
LinksPointingToPage -0.339065 -0.022987 -0.198410 -0.006080      -
0.194165
InfoEmail            0.077989 -0.014457  0.049328  0.370123
0.031898
DisableRightClick    0.042881 -0.013613  0.038118  0.219503
0.025863
UsingPopupWindow     0.096882 -0.049381  0.036616  0.290893
0.054463
Favicon              0.087025 -0.042497  0.006101  0.304899
0.035100
IframeRedirection    0.054694 -0.013838  0.016581  0.284410
0.010459
WebsiteForwarding   -0.321181  0.046832 -0.534530 -0.028160      -
0.591478
Redirecting//        0.397389 -0.081247  0.842796  0.086960
1.000000
HTTPSDomainURL       0.363534 -0.089383  0.757838  0.104561
0.760799
AbnormalURL          0.336549 -0.106761  0.739290  0.203945
0.723724
ShortURL             0.403461 -0.097881  1.000000  0.104447
0.842796
DomainRegLen        -0.022739 -0.221892  0.060923  0.015522
0.047464

                     PrefixSuffix-  SubDomains     HTTPS  DomainRegLen
\
class                     0.348606    0.298323  0.714741     -0.225789

HTTPS                     0.261391    0.267649  1.000000     -0.193622
```

```
AnchorURL                 0.348871    0.229491  0.535786     -0.160257
PrefixSuffix-             1.000000    0.087891  0.261391     -0.096799
WebsiteTraffic            0.110598   -0.005764  0.258768     -0.134454
SubDomains                0.087891    1.000000  0.267649     -0.082839
RequestURL                0.098675    0.104857  0.193054     -0.609970
LinksInScriptTags         0.100254    0.093646  0.176825     -0.101084
ServerFormHandler         0.001326    0.096089  0.171402     -0.136422
GoogleIndex               0.067781    0.057673  0.096051     -0.039766
AgeofDomain               0.074116    0.119254  0.162809     -0.062851
PageRank                 -0.006834    0.120730  0.074545     -0.059898
UsingIP                  -0.005257   -0.080745  0.071414     -0.022739
StatsReport              -0.002763    0.081627  0.063411     -0.002212
DNSRecording             -0.016556    0.125493  0.050972     -0.010477
LongURL                   0.055247    0.003997  0.048754     -0.221892
Symbol@                  -0.011726   -0.058976  0.031220      0.015522
StatusBarCust             0.012578   -0.018082  0.023586      0.023784
NonStdPort               -0.022546    0.004863  0.027473      0.022478
LinksPointingToPage       0.067423   -0.010526 -0.011710      0.122672
InfoEmail                -0.045000    0.008830  0.008062      0.039260
DisableRightClick        -0.024868    0.018230  0.015854      0.023520
UsingPopupWindow         -0.014733   -0.025312 -0.013005      0.051410
Favicon                  -0.007504   -0.016704 -0.014757      0.054253
IframeRedirection        -0.036904    0.010637 -0.002773      0.004393
WebsiteForwarding         0.016271    0.031206 -0.021070     -0.016300
```

```
Redirecting//            -0.085590   -0.043079 -0.036200      0.047464

HTTPSDomainURL           -0.070153   -0.037239 -0.029941      0.059161

AbnormalURL              -0.077620   -0.034908 -0.046245      0.058109

ShortURL                 -0.080471   -0.041916 -0.061426      0.060923

DomainRegLen             -0.096799   -0.082839 -0.193622      1.000000

                       Favicon  NonStdPort  HTTPSDomainURL  RequestURL
\
class                -0.000280    0.036419       -0.039854    0.253372

HTTPS                -0.014757    0.027473       -0.029941    0.193054

AnchorURL             0.037698    0.039891        0.011851    0.177693

PrefixSuffix-        -0.007504   -0.022546       -0.070153    0.098675

WebsiteTraffic       -0.050922   -0.028543       -0.039708    0.161166

SubDomains           -0.016704    0.004863       -0.037239    0.104857

RequestURL           -0.004620    0.027561       -0.006620    1.000000

LinksInScriptTags    -0.100341   -0.066502       -0.104381    0.067491

ServerFormHandler    -0.012279    0.006672       -0.009680    0.126661

GoogleIndex          -0.016668   -0.005413        0.115450    0.046409

AgeofDomain          -0.002628    0.008459       -0.049632    0.090455

PageRank              0.011699    0.017954        0.021104    0.055734

UsingIP               0.087025    0.060979        0.363534    0.029773

StatsReport           0.300917    0.343987        0.096187    0.035412

DNSRecording          0.088211    0.054849        0.395387    0.015933

LongURL              -0.042497    0.000323       -0.089383    0.246348

Symbol@               0.304899    0.364891        0.104561    0.027909
```

```
StatusBarCust        0.706179    0.623298        0.110113    0.008144
NonStdPort           0.803834    1.000000        0.004999    0.027561
LinksPointingToPage -0.127243   -0.139104       -0.128724   -0.067109
InfoEmail            0.668317    0.799088        0.075478    0.018178
DisableRightClick    0.414382    0.481631        0.009265   -0.020452
UsingPopupWindow     0.939633    0.748517        0.066957   -0.004622
Favicon              1.000000    0.803834        0.049483   -0.004620
IframeRedirection    0.627607    0.687044        0.017509    0.016934
WebsiteForwarding   -0.015621   -0.022472       -0.460165    0.002329
Redirecting//        0.035100    0.025060        0.760799   -0.026368
HTTPSDomainURL       0.049483    0.004999        1.000000   -0.006620
AbnormalURL          0.071848    0.054126        0.716287   -0.036034
ShortURL             0.006101    0.002201        0.757838   -0.037235
DomainRegLen         0.054253    0.022478        0.059161   -0.609970

                     AnchorURL  LinksInScriptTags
ServerFormHandler  \
class                 0.692935           0.248229           0.221419
HTTPS                 0.535786           0.176825           0.171402
AnchorURL             1.000000           0.136283           0.114311
PrefixSuffix-         0.348871           0.100254           0.001326
WebsiteTraffic        0.326293           0.064548           0.052706
SubDomains            0.229491           0.093646           0.096089
RequestURL            0.177693           0.067491           0.126661
LinksInScriptTags     0.136283           1.000000           0.066598
ServerFormHandler     0.114311           0.066598           1.000000
```

```
GoogleIndex           0.038816           0.045557           0.027588
AgeofDomain           0.075508           0.078057          -0.015840
PageRank              0.099261          -0.006450           0.001979
UsingIP               0.099847           0.006212          -0.010962
StatsReport           0.077377          -0.087343          -0.005289
DNSRecording          0.093288          -0.038545           0.034440
LongURL              -0.023396           0.052869           0.414196
Symbol@               0.057914          -0.070861          -0.008672
StatusBarCust         0.067742          -0.077670           0.007579
NonStdPort            0.039891          -0.066502           0.006672
LinksPointingToPage   0.018651           0.013561          -0.009068
InfoEmail             0.033386          -0.043231           0.011473
DisableRightClick     0.022168          -0.037469           0.008467
UsingPopupWindow      0.041150          -0.112282          -0.004863
Favicon               0.037698          -0.100341          -0.012279
IframeRedirection     0.013403          -0.070030           0.007067
WebsiteForwarding    -0.000839           0.041497           0.049907
Redirecting//        -0.005036          -0.125583          -0.041672
HTTPSDomainURL        0.011851          -0.104381          -0.009680
AbnormalURL          -0.010585          -0.116065          -0.030752
ShortURL              0.000561          -0.133379          -0.022723
DomainRegLen         -0.160257          -0.101084          -0.136422

                     InfoEmail  AbnormalURL  WebsiteForwarding
StatusBarCust  \
class                 0.018249    -0.060488          -0.020113
0.041838
HTTPS                 0.008062    -0.046245          -0.021070
0.023586
AnchorURL             0.033386    -0.010585          -0.000839
0.067742
PrefixSuffix-        -0.045000    -0.077620           0.016271
0.012578
WebsiteTraffic       -0.015685    -0.052416           0.004631      -
0.036531
SubDomains            0.008830    -0.034908           0.031206      -
0.018082
RequestURL            0.018178    -0.036034           0.002329
0.008144
LinksInScriptTags    -0.043231    -0.116065           0.041497      -
0.077670
ServerFormHandler     0.011473    -0.030752           0.049907
0.007579
GoogleIndex          -0.008378     0.124751           0.057230      -
0.006510
AgeofDomain           0.007357    -0.032533          -0.022476
0.013306
PageRank              0.026208     0.007318           0.052867
0.015634
UsingIP               0.077989     0.336549          -0.321181
0.084059
StatsReport           0.352074     0.186399          -0.059194
0.277347
DNSRecording          0.064145     0.366833          -0.211096
0.087161
LongURL              -0.014457    -0.106761           0.046832      -
0.045103
Symbol@               0.370123     0.203945          -0.028160
0.279697
StatusBarCust         0.531656     0.117638          -0.034823
1.000000
NonStdPort            0.799088     0.054126          -0.022472
0.623298
LinksPointingToPage  -0.039956    -0.161027           0.161278      -
0.038551
InfoEmail             1.000000     0.195850          -0.007321
0.531656
DisableRightClick     0.398629     0.023710          -0.023586
0.474054
UsingPopupWindow      0.629462     0.091188          -0.026327
0.733629
Favicon               0.668317     0.071848          -0.015621
0.706179
IframeRedirection     0.577490     0.017590          -0.012668
0.659478
```

```
WebsiteForwarding    -0.007321    -0.459187           1.000000      -
0.034823
Redirecting//         0.031898     0.723724          -0.591478
0.086635
HTTPSDomainURL        0.075478     0.716287          -0.460165
0.110113
AbnormalURL           0.195850     1.000000          -0.459187
0.117638
ShortURL              0.049328     0.739290          -0.534530
0.062383
DomainRegLen          0.039260     0.058109          -0.016300
0.023784

                     DisableRightClick  UsingPopupWindow
IframeRedirection  \
class                         0.012653          0.000086          -
0.003394
HTTPS                         0.015854         -0.013005          -
0.002773
AnchorURL                     0.022168          0.041150
0.013403
PrefixSuffix-                -0.024868         -0.014733          -
0.036904
WebsiteTraffic               -0.013594         -0.043190          -
0.022080
SubDomains                    0.018230         -0.025312
0.010637
RequestURL                   -0.020452         -0.004622
0.016934
LinksInScriptTags            -0.037469         -0.112282          -
0.070030
ServerFormHandler             0.008467         -0.004863
0.007067
GoogleIndex                  -0.008066         -0.010256          -
0.003519
AgeofDomain                   0.006764         -0.000948
0.018848
PageRank                      0.025341          0.017114
0.022407
UsingIP                       0.042881          0.096882
0.054694
StatsReport                   0.204409          0.285261
0.268418
DNSRecording                  0.038255          0.098658
0.047293
LongURL                      -0.013613         -0.049381          -
0.013838
Symbol@                       0.219503          0.290893
0.284410
StatusBarCust                 0.474054          0.733629
```

```
0.659478
NonStdPort                    0.481631          0.748517
0.687044
LinksPointingToPage          -0.119831         -0.121325          -
0.140824
InfoEmail                     0.398629          0.629462
0.577490
DisableRightClick             1.000000          0.415268
0.655863
UsingPopupWindow              0.415268          1.000000
0.629406
Favicon                       0.414382          0.939633
0.627607
IframeRedirection             0.655863          0.629406
1.000000
WebsiteForwarding            -0.023586         -0.026327          -
0.012668
Redirecting//                 0.025863          0.054463
0.010459
HTTPSDomainURL                0.009265          0.066957
0.017509
AbnormalURL                   0.023710          0.091188
0.017590
ShortURL                      0.038118          0.036616
0.016581
DomainRegLen                  0.023520          0.051410
0.004393

                     AgeofDomain  DNSRecording  WebsiteTraffic
PageRank  \
class                   0.121496      0.075718        0.346103
0.104645
HTTPS                   0.162809      0.050972        0.258768
0.074545
AnchorURL               0.075508      0.093288        0.326293
0.099261
PrefixSuffix-           0.074116     -0.016556        0.110598 -
0.006834
WebsiteTraffic          0.089949      0.048650        1.000000
0.030984
SubDomains              0.119254      0.125493       -0.005764
0.120730
RequestURL              0.090455      0.015933        0.161166
0.055734
LinksInScriptTags       0.078057     -0.038545        0.064548 -
0.006450
ServerFormHandler      -0.015840      0.034440        0.052706
0.001979
GoogleIndex            -0.028471      0.137509       -0.012584
0.032431
```

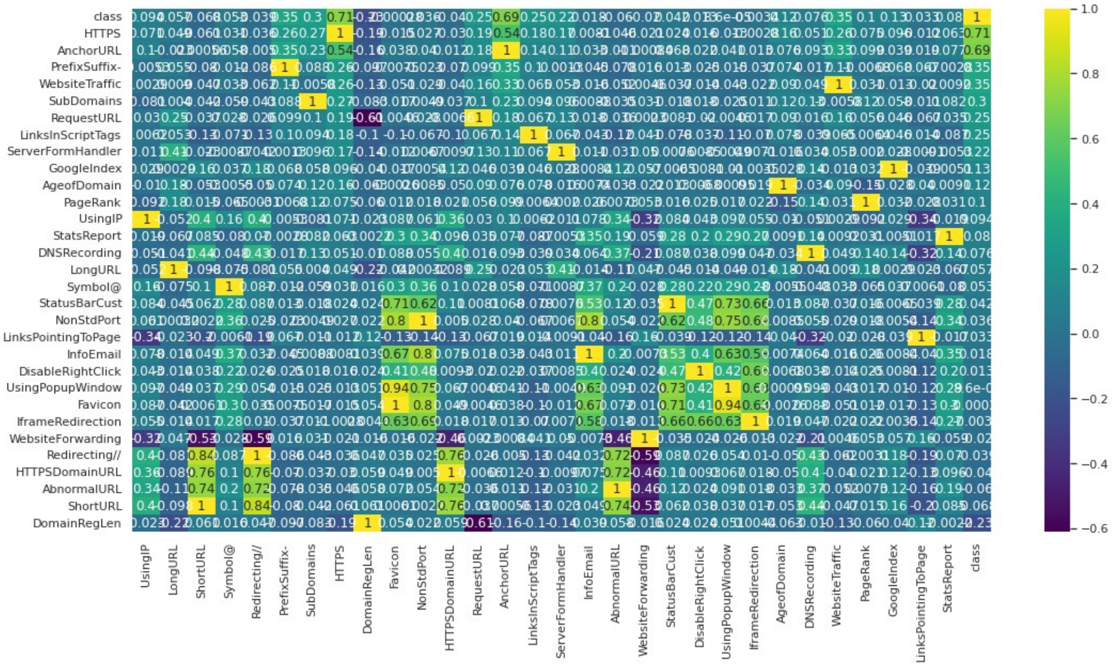
```
AgeofDomain             1.000000     -0.034082        0.089949 -
0.147194
PageRank               -0.147194      0.137860        0.030984
1.000000
UsingIP                -0.010446     -0.050733        0.002922 -
0.091774
StatsReport             0.009115      0.136860        0.009223
0.031049
DNSRecording           -0.034082      1.000000        0.048650
0.137860
LongURL                 0.179426     -0.040823        0.008993
0.183518
Symbol@                -0.005499     -0.047872        0.032918 -
0.064735
StatusBarCust           0.013306      0.087161       -0.036531
0.015634
NonStdPort              0.008459      0.054849       -0.028543
0.017954
LinksPointingToPage     0.040407     -0.318266       -0.019860 -
0.028216
InfoEmail               0.007357      0.064145       -0.015685
0.026208
DisableRightClick       0.006764      0.038255       -0.013594
0.025341
UsingPopupWindow       -0.000948      0.098658       -0.043190
0.017114
Favicon                -0.002628      0.088211       -0.050922
0.011699
IframeRedirection       0.018848      0.047293       -0.022080
0.022407
WebsiteForwarding      -0.022476     -0.211096        0.004631
0.052867
Redirecting//          -0.050107      0.431409       -0.062369 -
0.003132
HTTPSDomainURL         -0.049632      0.395387       -0.039708
0.021104
AbnormalURL            -0.032533      0.366833       -0.052416
0.007318
ShortURL               -0.052596      0.436064       -0.047074
0.014591
DomainRegLen           -0.062851     -0.010477       -0.134454 -
0.059898

                     GoogleIndex  LinksPointingToPage  StatsReport
class
class                   0.128950             0.032574     0.079857
1.000000
HTTPS                   0.096051            -0.011710     0.063411
0.714741
AnchorURL               0.038816             0.018651     0.077377
```

```
0.692935
PrefixSuffix-           0.067781             0.067423    -0.002763
0.348606
WebsiteTraffic         -0.012584            -0.019860     0.009223
0.346103
SubDomains              0.057673            -0.010526     0.081627
0.298323
RequestURL              0.046409            -0.067109     0.035412
0.253372
LinksInScriptTags       0.045557             0.013561    -0.087343
0.248229
ServerFormHandler       0.027588            -0.009068    -0.005289
0.221419
GoogleIndex             1.000000            -0.038777    -0.005103
0.128950
AgeofDomain            -0.028471             0.040407     0.009115
0.121496
PageRank                0.032431            -0.028216     0.031049
0.104645
UsingIP                 0.029153            -0.339065    -0.019103
0.094160
StatsReport            -0.005103            -0.016817     1.000000
0.079857
DNSRecording            0.137509            -0.318266     0.136860
0.075718
LongURL                 0.002902            -0.022987    -0.067153
0.057430
Symbol@                 0.037061            -0.006080    -0.080357
0.052948
StatusBarCust          -0.006510            -0.038551     0.277347
0.041838
NonStdPort             -0.005413            -0.139104     0.343987
0.036419
LinksPointingToPage    -0.038777             1.000000    -0.016817
0.032574
InfoEmail              -0.008378            -0.039956     0.352074
0.018249
DisableRightClick      -0.008066            -0.119831     0.204409
0.012653
UsingPopupWindow       -0.010256            -0.121325     0.285261
0.000086
Favicon                -0.016668            -0.127243     0.300917 -
0.000280
IframeRedirection      -0.003519            -0.140824     0.268418 -
0.003394
WebsiteForwarding       0.057230             0.161278    -0.059194 -
0.020113
Redirecting//           0.178415            -0.194165     0.070390 -
0.038608
HTTPSDomainURL          0.115450            -0.128724     0.096187 -
```

```
0.039854
AbnormalURL             0.124751            -0.161027     0.186399 -
0.060488
ShortURL                0.155844            -0.198410     0.085461 -
0.067966
DomainRegLen           -0.039766             0.122672    -0.002212 -
0.225789
```

```
sns.set(rc={'figure.figsize':(18,9)})
sns.heatmap(df_phishing.corr().sort_values(by='class',ascending=False)
,annot = True,cmap='viridis')
plt.show()
```



HTTPS , AnchorURL has significant relation with class and PrefixSuffix- , WebsiteTraffic has some relationship with class

```
# Importing algorithims and sklearn libraries

from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import
confusion_matrix,classification_report,f1_score,accuracy_score

logreg = LogisticRegression(solver = 'lbfgs',C = 100 , penalty = 'l2')
# C is hyperparameter tuning.

X = df_phishing.drop('class',axis=1)
y = df_phishing['class']
```

```
X_train , X_test , y_train , y_test =
train_test_split(X,y,test_size=0.3,random_state=101)
```

```
print(X_train.shape , X_test.shape , y_train.shape , y_test.shape)
```

```
(7738, 30) (3317, 30) (7738,) (3317,)
```

```
X_train.head(1)
```

```
      UsingIP  LongURL  ShortURL  Symbol@  Redirecting//
PrefixSuffix-  \
4635        1       -1         1        1              1             -
1

      SubDomains  HTTPS  DomainRegLen  Favicon  NonStdPort
HTTPSDomainURL  \
4635           1      1            -1        1           1
1

      RequestURL  AnchorURL  LinksInScriptTags  ServerFormHandler
InfoEmail  \
4635          -1          0                  0                 -1
1

      AbnormalURL  WebsiteForwarding  StatusBarCust  DisableRightClick
\
4635            1                  0              1                  1


      UsingPopupWindow  IframeRedirection  AgeofDomain
DNSRecording  \
4635                 1                  1            1             1


      WebsiteTraffic  PageRank  GoogleIndex  LinksPointingToPage
StatsReport
4635               1        -1            1                    0
1
```

```
 X_test.head(1)
```

```
      UsingIP  LongURL  ShortURL  Symbol@  Redirecting//
PrefixSuffix-  \
6923        1       -1         1        1              1
1

      SubDomains  HTTPS  DomainRegLen  Favicon  NonStdPort
HTTPSDomainURL  \
6923           1      1             1        1           1
1
```

```
      RequestURL  AnchorURL  LinksInScriptTags  ServerFormHandler
InfoEmail  \
6923         -1          1                 -1                 -1
1

      AbnormalURL  WebsiteForwarding  StatusBarCust  DisableRightClick
\
6923            1                  0              1                  1


      UsingPopupWindow  IframeRedirection  AgeofDomain
DNSRecording  \
6923                -1                  1           -1             1


      WebsiteTraffic  PageRank  GoogleIndex  LinksPointingToPage
StatsReport
6923               1         1            1                    0
1
```

```
-y_train.head(1)
```

```
4635   -1
Name: class, dtype: int64
```

```
y_test.head(1)
```

```
6923    1
Name: class, dtype: int64
```

```
model_logreg = logreg.fit(X_train,y_train)
model_logreg
```

```
LogisticRegression(C=100)
```

```
y_pred_logreg = model_logreg.predict(X_test)
y_pred_logreg
```

```
array([ 1, -1,  1, ...,  1, -1, -1])
```

```
confusion_matrix(y_test,y_pred_logreg)
```

```
array([[1327,  133],
       [ 123, 1734]])
```

```
print(classification_report(y_test,y_pred_logreg))
```

```
              precision    recall  f1-score   support

          -1       0.92      0.91      0.91      1460
           1       0.93      0.93      0.93      1857
```

```
    accuracy                           0.92      3317
   macro avg       0.92      0.92      0.92      3317
weighted avg       0.92      0.92      0.92      3317
```

```
print('Accuracy of the Model is ' ,
accuracy_score(y_test,y_pred_logreg)*100)
```

```
Accuracy of the Model is  92.28218269520652
```

```
print('F1 score is ' , f1_score(y_test,y_pred_logreg))
```

```
F1 score is  0.9312567132116004
```

Exercise - 2

```
# classify as features(Prefix_Suffix and URL_of_Anchor) and label with
index 5

X_new = df_phishing.iloc[:,[5,13]].values
y_new = df_phishing.iloc[:,30].values
```

```
X_new
```

```
array([[-1, -1],
       [-1,  0],
       [-1,  0],
       ...,
       [-1,  0],
       [-1, -1],
       [-1, -1]])
```

```
y_new
```

```
array([-1, -1, -1, ..., -1, -1, -1])
```

```
X_new_train , X_new_test , y_new_train , y_new_test =
train_test_split(X_new,y_new,test_size = 0.3 , random_state = 4)
```

```
print(X_new_train.shape , X_new_test.shape , y_new_train.shape ,
y_new_test.shape)
```

```
(7738, 2) (3317, 2) (7738,) (3317,)
```

```
# Perform the standard scaling.
sc = StandardScaler()
X_new_train = sc.fit_transform(X_new_train)
X_new_test = sc.fit_transform(X_new_test)
```

```
model_new = logreg.fit(X_new_train,y_new_train)
```

```
y_pred_new = logreg.predict(X_new_test)
y_pred_new
```

```
array([-1,  1,  1, ...,  1,  1,  1])
```

```
y_pred_prob = logreg.predict_proba(X_new_test)
```

```
y_pred_prob
```

```
array([[9.56088123e-01, 4.39118768e-02],
       [2.02467287e-02, 9.79753271e-01],
       [4.33706746e-07, 9.99999566e-01],
       ...,
       [4.33706746e-07, 9.99999566e-01],
       [2.02467287e-02, 9.79753271e-01],
       [4.01475454e-01, 5.98524546e-01]])
```

```
#confusion matrix for printing count of misclassified samples in the test data prediction

confusion_matrix(y_new_test,y_pred_new)
```

```
array([[ 966,  463],
       [   3, 1885]])
```

```
print(classification_report(y_new_test,y_pred_new))
```

```
              precision    recall  f1-score   support

          -1       1.00      0.68      0.81      1429
           1       0.80      1.00      0.89      1888

    accuracy                           0.86      3317
   macro avg       0.90      0.84      0.85      3317
weighted avg       0.89      0.86      0.85      3317
```

```
training_score = model_new.score(X_new_train,y_new_train)
```

```
training_score
```

```
0.8444042388214009
```

```
test_score = model_new.score(X_new_test,y_new_test)
test_score
```

```
0.859511606873681
```

Model is not predicting good for class 1 and biased only towards class -1

Plot the test samples along with the decision boundary when trained with index 5 and index 13 parameters.

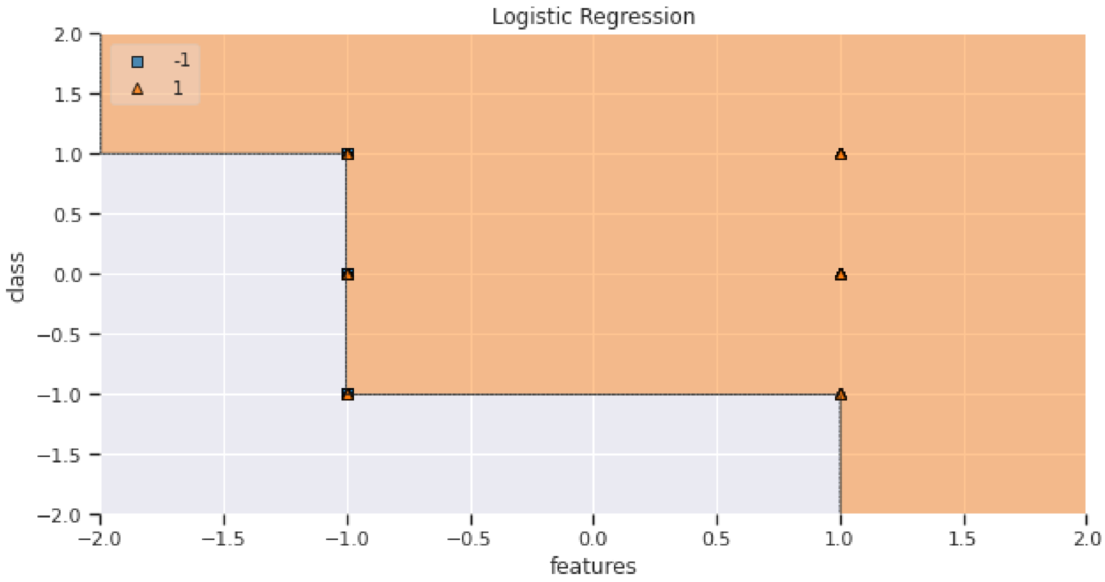
```
from mlxtend.plotting import plot_decision_regions
```

```
#X_new = X_new.values
```

```
#y_new = y_new.to_numpy()

x = df_phishing[['PrefixSuffix-','AnchorURL']].values
y = df_phishing['class'].values

plt.figure(figsize=(10,5))
plot_decision_regions(x,y,clf=logreg,legend=2)
plt.title('Logistic Regression')
plt.xlabel('features')
plt.ylabel('class')
sns.set_style('white')
plt.show()
```


Logistic Regression
■ -1
▲ 1
class
features