```
!git clone https://github.com/Atharva-Malode/ML-Bootcamp.git
```

```
Cloning into 'ML-Bootcamp'...
remote: Enumerating objects: 551, done.
remote: Counting objects: 100% (226/226), done.
remote: Compressing objects: 100% (167/167), done.
remote: Total 551 (delta 100), reused 129 (delta 51), pack-reused 325
Receiving objects: 100% (551/551), 13.46 MiB | 12.86 MiB/s, done.
Resolving deltas: 100% (186/186), done.
```

```
!pip install pandas numpy seaborn matplotlib pycaret
```

```
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (1.5.3)
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (1.22.4)
Requirement already satisfied: seaborn in /usr/local/lib/python3.10/dist-packages (0.12.2)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.10/dist-packages (3.7.1)
Collecting pycaret
  Downloading pycaret-3.0.4-py3-none-any.whl (484 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 484.4/484.4 kB 17.9 MB/s eta 0:00:00
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10/dist-packages (from pandas)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas) (2022.7.1
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (0.11
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (2
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (8.4
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (
Requirement already satisfied: ipython>=5.5.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (7.34.
Requirement already satisfied: ipywidgets>=7.6.5 in /usr/local/lib/python3.10/dist-packages (from pycaret) (7.
Requirement already satisfied: tqdm>=4.62.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (4.65.0)
Requirement already satisfied: jinja2>=1.2 in /usr/local/lib/python3.10/dist-packages (from pycaret) (3.1.2)
Requirement already satisfied: scipy<2.0.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (1.10.1)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (1.2.0)
Requirement already satisfied: scikit-learn<1.3.0,>=1.0 in /usr/local/lib/python3.10/dist-packages (from pycar
Collecting pyod>=1.0.8 (from pycaret)
  Downloading pyod-1.1.0.tar.gz (153 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 153.4/153.4 kB 16.3 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: imbalanced-learn>=0.8.1 in /usr/local/lib/python3.10/dist-packages (from pycare
Collecting category-encoders>=2.4.0 (from pycaret)
  Downloading category_encoders-2.6.1-py2.py3-none-any.whl (81 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 81.9/81.9 kB 8.1 MB/s eta 0:00:00
Requirement already satisfied: lightgbm>=3.0.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (3.3.
Requirement already satisfied: numba>=0.55.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (0.56.4
Requirement already satisfied: requests>=2.27.1 in /usr/local/lib/python3.10/dist-packages (from pycaret) (2.2
Requirement already satisfied: psutil>=5.9.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (5.9.5)
Requirement already satisfied: markupsafe>=2.0.1 in /usr/local/lib/python3.10/dist-packages (from pycaret) (2.
Collecting importlib-metadata>=4.12.0 (from pycaret)
  Downloading importlib_metadata-6.8.0-py3-none-any.whl (22 kB)
Requirement already satisfied: nbformat>=4.2.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (5.9.
Requirement already satisfied: cloudpickle in /usr/local/lib/python3.10/dist-packages (from pycaret) (2.2.1)
Collecting deprecation>=2.1.0 (from pycaret)
  Downloading deprecation-2.1.0-py2.py3-none-any.whl (11 kB)
Collecting xxhash (from pycaret)
  Downloading xxhash-3.2.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (212 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 212.5/212.5 kB 21.9 MB/s eta 0:00:00
Collecting scikit-plot>=0.3.7 (from pycaret)
  Downloading scikit_plot-0.3.7-py3-none-any.whl (33 kB)
Requirement already satisfied: yellowbrick>=1.4 in /usr/local/lib/python3.10/dist-packages (from pycaret) (1.5
Requirement already satisfied: plotly>=5.0.0 in /usr/local/lib/python3.10/dist-packages (from pycaret) (5.13.1
Collecting kaleido>=0.2.1 (from pycaret)
  Downloading kaleido-0.2.1-py2.py3-none-manylinux1_x86_64.whl (79.9 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 79.9/79.9 MB 10.9 MB/s eta 0:00:00
Collecting schemdraw==0.15 (from pycaret)
  Downloading schemdraw-0.15-py3-none-any.whl (106 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 106.8/106.8 kB 13.9 MB/s eta 0:00:00
Collecting plotly-resampler>=0.8.3.1 (from pycaret)
  Downloading plotly_resampler-0.8.3.2.tar.gz (46 kB)
```

```
#Ashiya-Thakur
#IMPORTING ALL NECESSARY LIBRARIES & PACKAGES
import pandas as pd
```

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pickle
from pycaret.classification import *


df = pd.read_csv("/content/ML-Bootcamp/Week-4/Day-1/Dataset/dataset_full.csv")
df
```

	qty_dot_url	qty_hyphen_url	qty_underline_url	qty_slash_url	qty_questionmark_url	qty_equal_u
0	3	0	0	1	0	
1	5	0	1	3	0	
2	2	0	0	1	0	
3	4	0	2	5	0	
4	2	0	0	0	0	
...	...	...	...	...	...	
88642	3	1	0	0	0	
88643	2	0	0	0	0	
88644	2	1	0	5	0	
88645	2	0	0	1	0	
88646	2	0	0	0	0	

88647 rows × 112 columns

```
# Summary statistics of the dataset

print(df.describe())
```

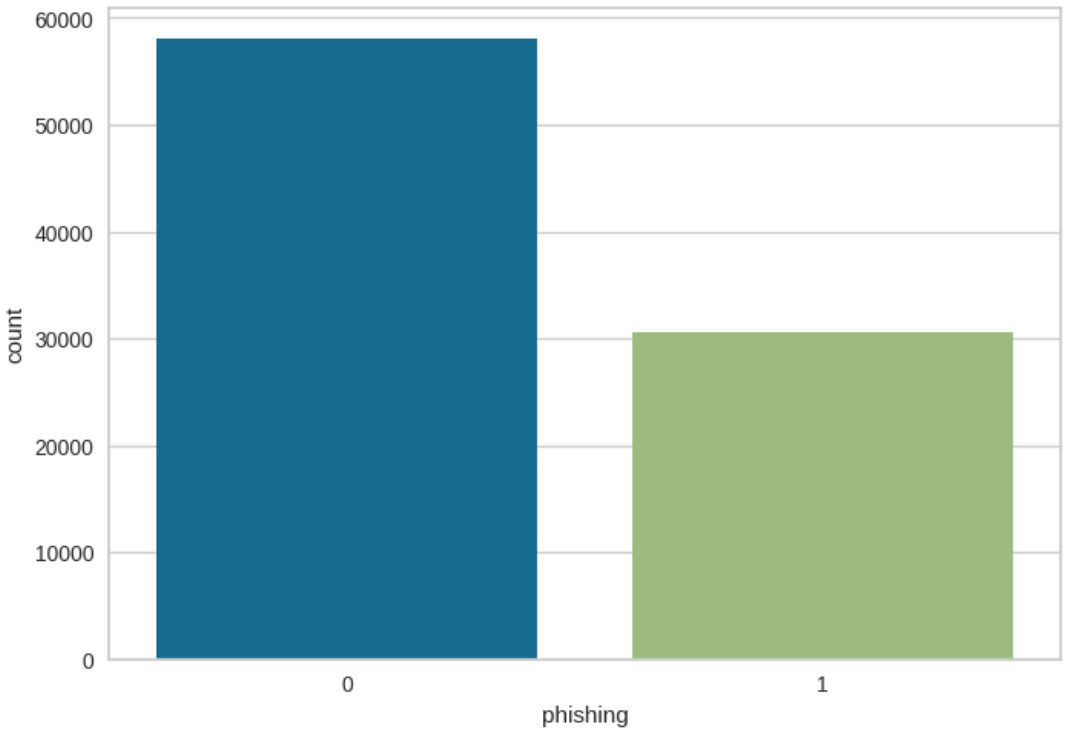
```
        qty_dot_url  qty_hyphen_url  qty_underline_url  qty_slash_url  \
count  88647.000000    88647.000000       88647.000000   88647.000000
mean       2.191343        0.328810           0.113879       1.281781
std        1.235636        1.119286           0.657767       1.893929
min        1.000000        0.000000           0.000000       0.000000
25%        2.000000        0.000000           0.000000       0.000000
50%        2.000000        0.000000           0.000000       0.000000
75%        2.000000        0.000000           0.000000       2.000000
max       24.000000       35.000000          21.000000      44.000000

       qty_questionmark_url  qty_equal_url    qty_at_url   qty_and_url  \
count          88647.000000   88647.000000  88647.000000  88647.000000
mean               0.009329       0.205861      0.022133      0.140885
std                0.112568       0.954272      0.279652      0.924864
min                0.000000       0.000000      0.000000      0.000000
25%                0.000000       0.000000      0.000000      0.000000
50%                0.000000       0.000000      0.000000      0.000000
75%                0.000000       0.000000      0.000000      0.000000
max                9.000000      23.000000     43.000000     26.000000

       qty_exclamation_url  qty_space_url  ...  qty_ip_resolved  \
count         88647.000000   88647.000000  ...     88647.000000
mean              0.002944       0.001015  ...         1.136564
std               0.087341       0.072653  ...         0.895146
min               0.000000       0.000000  ...        -1.000000
25%               0.000000       0.000000  ...         1.000000
50%               0.000000       0.000000  ...         1.000000
75%               0.000000       0.000000  ...         1.000000
max              10.000000       9.000000  ...        24.000000

       qty_nameservers  qty_mx_servers   ttl_hostname  tls_ssl_certificate  \
count     88647.000000    88647.000000   88647.000000         88647.000000
mean          2.772412        1.742428    6159.877514             0.506447
```

```
std           1.322999       1.706705   11465.583810             0.499961
min           0.000000       0.000000      -1.000000             0.000000
25%           2.000000       1.000000     292.000000             0.000000
50%           2.000000       1.000000    2029.000000             1.000000
75%           4.000000       2.000000   10798.000000             1.000000
max          20.000000      20.000000  604800.000000             1.000000

       qty_redirects  url_google_index  domain_google_index  url_shortened  \
count   88647.000000       88647.00000         88647.000000   88647.000000
mean        0.343903           0.00141             0.002019       0.005482
std         0.783892           0.05864             0.063250       0.073841
min        -1.000000          -1.00000            -1.000000       0.000000
25%         0.000000           0.00000             0.000000       0.000000
50%         0.000000           0.00000             0.000000       0.000000
75%         1.000000           0.00000             0.000000       0.000000
max        17.000000           1.00000             1.000000       1.000000

           phishing
count  88647.000000
mean       0.345720
std        0.475605
min        0.000000
25%        0.000000
50%        0.000000
75%        1.000000
```

```
df.isnull().sum()
```

```
qty_dot_url             0
qty_hyphen_url          0
qty_underline_url       0
qty_slash_url           0
qty_questionmark_url    0
                       ..
qty_redirects           0
url_google_index        0
domain_google_index     0
url_shortened           0
phishing                0
Length: 112, dtype: int64
```

```
#Visualize the given dataset

sns.countplot(x='phishing',data=df)        # Y(target value)
plt.show()
```



```
cols_to_drop = ['url_google_index',
                  'domain_google_index',
                  'server_client_domain',
                  'tld_present_params',
                  'time_response',
                  'domain_spf',
                  'qty_ip_resolved',
                  'qty_nameservers',
                  'qty_mx_servers',
                  'ttl_hostname',
                  'url_shortened']
df = df.drop(cols_to_drop,axis=1)


extra_one = ['qty_vowels_domain']
df = df.drop(extra_one,axis=1)


rows, columns = df.shape
print("Number of rows : ",rows)
print("Number of columns : ",columns)
```

```
Number of rows :  88647
Number of columns :  100
```

PERFORM FEATURE ENGINEERING ON OUR DATA SET AND WILL SHOW IT

```
original_features = list(df.columns)


original_features
```

```
['qty_dot_url',
 'qty_hyphen_url',
 'qty_underline_url',
 'qty_slash_url',
 'qty_questionmark_url',
 'qty_equal_url',
 'qty_at_url',
 'qty_and_url',
 'qty_exclamation_url',
 'qty_space_url',
 'qty_tilde_url',
 'qty_comma_url',
 'qty_plus_url',
 'qty_asterisk_url',
 'qty_hashtag_url',
 'qty_dollar_url',
 'qty_percent_url',
 'qty_tld_url',
 'length_url',
 'qty_dot_domain',
 'qty_hyphen_domain',
 'qty_underline_domain',
 'qty_slash_domain',
 'qty_questionmark_domain',
 'qty_equal_domain',
 'qty_at_domain',
 'qty_and_domain',
 'qty_exclamation_domain',
 'qty_space_domain',
 'qty_tilde_domain',
 'qty_comma_domain',
 'qty_plus_domain',
 'qty_asterisk_domain',
 'qty_hashtag_domain',
 'qty_dollar_domain',
 'qty_percent_domain',
 'domain_length',
 'domain_in_ip',
```

```
        'qty_dot_directory',
        'qty_hyphen_directory',
        'qty_underline_directory',
        'qty_slash_directory',
        'qty_questionmark_directory',
        'qty_equal_directory',
        'qty_at_directory',
        'qty_and_directory',
        'qty_exclamation_directory',
        'qty_space_directory',
        'qty_tilde_directory',
        'qty_comma_directory',
        'qty_plus_directory',
        'qty_asterisk_directory',
        'qty_hashtag_directory',
        'qty_dollar_directory',
        'qty_percent_directory',
        'directory_length',
        'qty_dot_file',
        'qty_hyphen_file',
```

```
dataset_array = np.array(df)
```

```
print(dataset_array)
```

```
    [[3 0 0 ... 0 0 1]
     [5 0 1 ... 1 0 1]
     [2 0 0 ... 1 0 0]
     ...
     [2 1 0 ... 1 0 1]
     [2 0 0 ... 1 0 1]
     [2 0 0 ... 0 0 0]]
```

```
features_indices = []
attributes = ['url','domain','file','params']
```

```
new_dataset = {}
```

```
for index, name in enumerate(original_features):
  if 'qty' in name and name.split('_')[-1] in attributes:
    features_indices.append([index,name.split('_')[-1]])
  else:
    new_dataset[name] = dataset_array[:,index]
```

```
for index, attribute in features_indices:
  if attribute == 'domain':
    if f'qty_char_{attribute}' not in new_dataset.keys():
      new_dataset[f'qty_char_{attribute}'] = np.zeros(rows)

    new_dataset[f'qty_char_{attribute}'] += dataset_array[:,index]
```

```
df1 = pd.DataFrame(new_dataset).astype(int)
```

```
df1[df1<-1] = -1
df1
```

	length_url	domain_length	domain_in_ip	qty_dot_directory	qty_hyphen_directory	qty_underline_
0	25	17	0	1	0	
1	223	16	0	3	0	
2	15	14	0	0	0	
3	81	19	0	2	0	
4	19	19	0	-1	-1	
...	...	...	...	...	...	
88642	23	23	0	-1	-1	
88643	34	34	0	-1	-1	

```
# Summary statistics of our dataset

print(df1.describe())
```

```
         length_url  domain_length  domain_in_ip  qty_dot_directory  \
count  88647.000000   88647.000000  88647.000000       88647.000000
mean      36.347615      18.560820      0.002267          -0.323666
std       46.191590       6.598694      0.047564           0.899499
min        4.000000       4.000000      0.000000          -1.000000
25%       17.000000      14.000000      0.000000          -1.000000
50%       22.000000      18.000000      0.000000          -1.000000
75%       38.000000      22.000000      0.000000           0.000000
max     4165.000000     231.000000      1.000000          19.000000

       qty_hyphen_directory  qty_underline_directory  qty_slash_directory  \
count          88647.000000             88647.000000         88647.000000
mean              -0.360813                -0.477997             0.713685
std                1.101398                 0.682409             2.216137
min               -1.000000                -1.000000            -1.000000
25%               -1.000000                -1.000000            -1.000000
50%               -1.000000                -1.000000            -1.000000
75%                0.000000                 0.000000             2.000000
max               23.000000                17.000000            22.000000

       qty_questionmark_directory  qty_equal_directory  qty_at_directory  ...  \
count                88647.000000         88647.000000      88647.000000  ...
mean                    -0.535935            -0.528343         -0.532550  ...
std                      0.498710             0.517986          0.551786  ...
min                     -1.000000            -1.000000         -1.000000  ...
25%                     -1.000000            -1.000000         -1.000000  ...
50%                     -1.000000            -1.000000         -1.000000  ...
75%                      0.000000             0.000000          0.000000  ...
max                      0.000000             5.000000         43.000000  ...

       params_length  email_in_url         asn_ip  time_domain_activation  \
count   88647.000000  88647.000000   88647.000000            88647.000000
mean        5.273185      0.018331   31131.152763             3389.676661
std        34.937007      0.134147   45261.502645             3044.165723
min        -1.000000      0.000000      -1.000000               -1.000000
25%        -1.000000      0.000000   13335.000000               -1.000000
50%        -1.000000      0.000000   20013.000000             3046.000000
75%        -1.000000      0.000000   34922.000000             6423.000000
max      4094.000000      1.000000  395754.000000            17775.000000

       time_domain_expiration  tls_ssl_certificate  qty_redirects  \
count            88647.000000         88647.000000   88647.000000
mean               352.043250             0.506447       0.343903
std                598.264801             0.499961       0.783892
min                 -1.000000             0.000000      -1.000000
25%                 -1.000000             0.000000       0.000000
50%                168.000000             1.000000       0.000000
75%                354.000000             1.000000       1.000000
max              22574.000000             1.000000      17.000000

           phishing  qty_chardomain  qty_char_domain
count  88647.000000    88647.000000     88647.000000
mean       0.345720        1.985967         5.957900
std        0.475605        0.836865         2.510594
min        0.000000        0.000000         0.000000
25%        0.000000        2.000000         6.000000
```

```
50%       0.000000       2.000000       6.000000
75%       1.000000       2.000000       6.000000
```

```
# Setting up the data for for modelling

setup(data = df1, target = "phishing")
```

	Description	Value
0	Session id	1413
1	Target	phishing
2	Target type	Binary
3	Original data shape	(88647, 32)
4	Transformed data shape	(88647, 32)
5	Transformed train set shape	(62052, 32)
6	Transformed test set shape	(26595, 32)
7	Numeric features	31
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Fold Generator	StratifiedKFold
13	Fold Number	10
14	CPU Jobs	-1
15	Use GPU	False
16	Log Experiment	False
17	Experiment Name	clf-default-name
18	USI	f19b

```
<pycaret.classification.oop.ClassificationExperiment at 0x7fc9dfb38d30>
```

```
# Comparing and select the best data model

best_model = compare_models()
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
xgboost	Extreme Gradient Boosting	0.9658	0.9936	0.9527	0.9485	0.9506	0.9245	0.92
rf	Random Forest Classifier	0.9637	0.9918	0.9521	0.9436	0.9478	0.9200	0.92
et	Extra Trees Classifier	0.9622	0.9897	0.9496	0.9417	0.9456	0.9167	0.91
lightgbm	Light Gradient Boosting	0.9611	0.9926	0.9469	0.9409	0.9439	0.9141	0.91

```
tuned_model = tune_model(best_model, n_iter = 1, optimize = 'F1')
```

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.8832	0.9768	0.9907	0.7510	0.8543	0.7599	0.7802
1	0.8817	0.9792	0.9897	0.7489	0.8527	0.7570	0.7776
2	0.8791	0.9798	0.9907	0.7444	0.8501	0.7522	0.7738
3	0.8824	0.9779	0.9902	0.7497	0.8534	0.7582	0.7787
4	0.8788	0.9793	0.9911	0.7436	0.8497	0.7516	0.7734
5	0.8786	0.9785	0.9916	0.7432	0.8496	0.7513	0.7733
6	0.8714	0.9786	0.9925	0.7314	0.8422	0.7378	0.7623
7	0.8783	0.9779	0.9911	0.7428	0.8492	0.7507	0.7727
8	0.8798	0.9777	0.9893	0.7459	0.8505	0.7532	0.7743
9	0.8804	0.9801	0.9939	0.7452	0.8518	0.7550	0.7770
Mean	0.8794	0.9786	0.9911	0.7446	0.8503	0.7527	0.7743
Std	0.0031	0.0010	0.0013	0.0052	0.0032	0.0058	0.0047

```
predictions = predict_model(tuned_model, data = df1)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Extreme Gradient	0.9731	0.9958	0.9621	0.9602	0.9611	0.9406	0.9406