

# Ensemble Boosting Models

## Import

```
In [1]: import pandas as pd
from pycaret.classification import *
import time
from DB_scripts import rnd_bln_split_CSV as shf
```

## Settings

```
In [2]: # set constants
target_label = 'tuple'
learning_model = ['rf', 'et', 'lightgbm', 'xgboost']
num_features = ['min_packet_size', 'min_fpkt', 'min_bpkt']
file_name = "all_features_"
path = target_label + "_dataset/"
ensemble_method='Boosting'
```

```
In [3]: # function for making model-prediction over the data set and measure the run time
def timed_prediction(in_data,in_model):
    t = time.process_time()
    predicted = predict_model(in_model, data=in_data)
    elapsed_time = time.process_time() - t
    print("prediction took: " + str(elapsed_time))
    return predicted
```

```
In [4]: # function for checkign the correction of the model-prediction over the data
def check_correction(in_predicted):
    count=0
    index = in_predicted.index
    number_of_rows = len(index)
    for i in range(0,number_of_rows):
        if str(int(in_predicted.iloc[i][target_label])) != str(int(in_predicted.iloc[i]['Label'])):
            #print("prediction not matched in Line " + str(i) + " as " + str(in_predicted.iloc[i]['app']) + "!=" + str(in_predicted.iloc[i]['Label']))
            count=count+1
    print("number of error: " + str(count) + " from " + str(number_of_rows) +
" test samples \n which is " + str(count/number_of_rows) + " precent of erro
r.")
```

```
In [5]: # compare answers and labeled test
def compare_prediction_with_answers(in_predicted, in_answers):
    count=0
    index = in_predicted.index
    number_of_rows = len(index)
    for i in range(0,number_of_rows):
        if str(in_answers.iloc[i]) != str(int(in_predicted.iloc[i]['Label'])):
            count=count+1
            # print the unmatched answers
            #print("answer os and test Label are not matched in Line " + str(i) +
            " as " + str(answers.iloc[i]['os']) + "!=" + str(predict_test.iloc[i]['Label']))
            print("number of error: " + str(count) + " from " + str(number_of_rows) +
            " test samples \n which is " + str(count/number_of_rows) + " precent of erro
r.")
```

```
In [6]: # activating balanced random data shuffling
shf.split_CSV_randomly_balanced(target_label,file_name)
```

## Read Data

```
In [38]: data = pd.read_csv(path+file_name+target_label+'_train.csv',
                         sep='\t',
                         skiprows=[1])
```

In [39]: `data.info()`

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14442 entries, 0 to 14441
Data columns (total 70 columns):
 #   Column           Non-Null Count Dtype  
--- 
 0   fSSL_session_id_len    14442 non-null float64
 1   fSSL_num_extensions   14442 non-null float64
 2   fSSL_num_compression_methods 14442 non-null float64
 3   SYN_tcp_scale        14442 non-null float64
 4   SYN_MSS              14442 non-null float64
 5   SYN_tcp_winsize      14442 non-null float64
 6   fcipher_suites_1     14442 non-null float64
 7   fcipher_suites_2     14442 non-null float64
 8   fcipher_suites_3     14442 non-null float64
 9   fSSLv_1               14442 non-null float64
 10  fSSLv_2               14442 non-null float64
 11  fSSLv_3               14442 non-null float64
 12  fSSLv_4               14442 non-null float64
 13  size_histogram_1     14442 non-null float64
 14  size_histogram_2     14442 non-null float64
 15  size_histogram_3     14442 non-null float64
 16  size_histogram_4     14442 non-null float64
 17  size_histogram_5     14442 non-null float64
 18  size_histogram_6     14442 non-null float64
 19  size_histogram_7     14442 non-null float64
 20  size_histogram_8     14442 non-null float64
 21  size_histogram_9     14442 non-null float64
 22  size_histogram_10    14442 non-null float64
 23  fpeak_features_1    14442 non-null float64
 24  fpeak_features_2    14442 non-null float64
 25  fpeak_features_3    14442 non-null float64
 26  fpeak_features_4    14442 non-null float64
 27  fpeak_features_5    14442 non-null float64
 28  fpeak_features_6    14442 non-null float64
 29  fpeak_features_7    14442 non-null float64
 30  fpeak_features_8    14442 non-null float64
 31  fpeak_features_9    14442 non-null float64
 32  bpeak_features_1    14442 non-null float64
 33  bpeak_features_2    14442 non-null float64
 34  bpeak_features_3    14442 non-null float64
 35  bpeak_features_4    14442 non-null float64
 36  bpeak_features_5    14442 non-null float64
 37  bpeak_features_6    14442 non-null float64
 38  bpeak_features_7    14442 non-null float64
 39  bpeak_features_8    14442 non-null float64
 40  bpeak_features_9    14442 non-null float64
 41  packet_count        14442 non-null float64
 42  min_packet_size     14442 non-null float64
 43  max_packet_size     14442 non-null float64
 44  mean_packet_size    14442 non-null float64
 45  sizevar             14442 non-null float64
 46  std_fiat            14442 non-null float64
 47  fpackets            14442 non-null float64
 48  bpackets            14442 non-null float64
 49  fbytes              14442 non-null float64
 50  bbytes              14442 non-null float64
 51  min_fiat            14442 non-null float64

```

```

52 min_biat           14442 non-null float64
53 max_fiat          14442 non-null float64
54 max_biat          14442 non-null float64
55 std_biat          14442 non-null float64
56 mean_fiat         14442 non-null float64
57 mean_biat         14442 non-null float64
58 min_fpkt          14442 non-null float64
59 min_bpkt          14442 non-null float64
60 max_fpkt          14442 non-null float64
61 max_bpkt          14442 non-null float64
62 std_fpkt          14442 non-null float64
63 std_bpkt          14442 non-null float64
64 mean_fpkt         14442 non-null float64
65 mean_bpkt         14442 non-null float64
66 mean_fttl_1       14442 non-null float64
67 mean_fttl_2       14442 non-null float64
68 num_keep_alive    14442 non-null float64
69 tuple              14442 non-null int64
dtypes: float64(69), int64(1)
memory usage: 7.7 MB

```

In [40]: `data[target_label].unique()`

Out[40]: `array([16102, 14602, 15602, 18102, 13102, 11602, 16201, 13201, 18201,
 16101, 13101, 18101, 16202, 18202, 13202, 16302, 18302, 13302,
 13403, 18403, 17403, 14601, 17101, 12101, 17201, 12201, 16403,
 16103, 13103, 18103], dtype=int64)`

## Read Unseen Test

In [41]: `unseen_data = pd.read_csv(path+file_name+target_label+'_test.csv',
 sep='\t',
 skiprows=[1])`

In [42]: `# check for target column values (we got all wanted values...)
unseen_data[target_label].unique()`

Out[42]: `array([16102, 14602, 15602, 18102, 13102, 11602, 16201, 13201, 18201,
 16101, 13101, 18101, 16202, 18202, 13202, 16302, 18302, 13302,
 13403, 18403, 17403, 14601, 17101, 12101, 17201, 12201, 16403,
 16103, 13103, 18103], dtype=int64)`

In [43]: `# saving the target column
answers = unseen_data[target_label]`

In [44]: `# dropping target column from test.
unseen_data = unseen_data.drop(columns=[target_label])`

```
In [45]: unseen_data.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6189 entries, 0 to 6188
Data columns (total 69 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   fSSL_session_id_len    6189 non-null   float64
 1   fSSL_num_extensions   6189 non-null   float64
 2   fSSL_num_compression_methods 6189 non-null   float64
 3   SYN_tcp_scale        6189 non-null   float64
 4   SYN_MSS              6189 non-null   float64
 5   SYN_tcp_winsize      6189 non-null   float64
 6   fcipher_suites_1    6189 non-null   float64
 7   fcipher_suites_2    6189 non-null   float64
 8   fcipher_suites_3    6189 non-null   float64
 9   fSSLv_1               6189 non-null   float64
 10  fSSLv_2               6189 non-null   float64
 11  fSSLv_3               6189 non-null   float64
 12  fSSLv_4               6189 non-null   float64
 13  size_histogram_1     6189 non-null   float64
 14  size_histogram_2     6189 non-null   float64
 15  size_histogram_3     6189 non-null   float64
 16  size_histogram_4     6189 non-null   float64
 17  size_histogram_5     6189 non-null   float64
 18  size_histogram_6     6189 non-null   float64
 19  size_histogram_7     6189 non-null   float64
 20  size_histogram_8     6189 non-null   float64
 21  size_histogram_9     6189 non-null   float64
 22  size_histogram_10    6189 non-null   float64
 23  fpeak_features_1    6189 non-null   float64
 24  fpeak_features_2    6189 non-null   float64
 25  fpeak_features_3    6189 non-null   float64
 26  fpeak_features_4    6189 non-null   float64
 27  fpeak_features_5    6189 non-null   float64
 28  fpeak_features_6    6189 non-null   float64
 29  fpeak_features_7    6189 non-null   float64
 30  fpeak_features_8    6189 non-null   float64
 31  fpeak_features_9    6189 non-null   float64
 32  bpeak_features_1    6189 non-null   float64
 33  bpeak_features_2    6189 non-null   float64
 34  bpeak_features_3    6189 non-null   float64
 35  bpeak_features_4    6189 non-null   float64
 36  bpeak_features_5    6189 non-null   float64
 37  bpeak_features_6    6189 non-null   float64
 38  bpeak_features_7    6189 non-null   float64
 39  bpeak_features_8    6189 non-null   float64
 40  bpeak_features_9    6189 non-null   float64
 41  packet_count        6189 non-null   float64
 42  min_packet_size     6189 non-null   float64
 43  max_packet_size     6189 non-null   float64
 44  mean_packet_size    6189 non-null   float64
 45  sizevar             6189 non-null   float64
 46  std_fiat            6189 non-null   float64
 47  fpackets            6189 non-null   float64
 48  bpackets            6189 non-null   float64
 49  fbytes              6189 non-null   float64
 50  bbytes              6189 non-null   float64
 51  min_fiat            6189 non-null   float64

```

```
ensemble_boosting  
52 min_biat          6189 non-null float64  
53 max_fiat          6189 non-null float64  
54 max_biat          6189 non-null float64  
55 std_biat          6189 non-null float64  
56 mean_fiat         6189 non-null float64  
57 mean_biat         6189 non-null float64  
58 min_fpkt          6189 non-null float64  
59 min_bpkt          6189 non-null float64  
60 max_fpkt          6189 non-null float64  
61 max_bpkt          6189 non-null float64  
62 std_fpkt          6189 non-null float64  
63 std_bpkt          6189 non-null float64  
64 mean_fpkt         6189 non-null float64  
65 mean_bpkt         6189 non-null float64  
66 mean_fttl_1       6189 non-null float64  
67 mean_fttl_2       6189 non-null float64  
68 num_keep_alive    6189 non-null float64  
dtypes: float64(69)  
memory usage: 3.3 MB
```

## Setup Classifier

which splits the data into train and test for model building

```
In [46]: setup(data=data,
             target=target_label,
             numeric_features=num_features,
             silent=True)
# train_size=0.99
```

Setup Successfully Completed!

	Description	Value
0	session_id	8849
1	Target Type	Multiclass
2	Label Encoded	None
3	Original Data	(14442, 70)
4	Missing Values	False
5	Numeric Features	61
6	Categorical Features	8
7	Ordinal Features	False
8	High Cardinality Features	False
9	High Cardinality Method	None
10	Sampled Data	(14442, 70)
11	Transformed Train Set	(10109, 77)
12	Transformed Test Set	(4333, 77)
13	Numeric Imputer	mean
14	Categorical Imputer	constant
15	Normalize	False
16	Normalize Method	None
17	Transformation	False
18	Transformation Method	None
19	PCA	False
20	PCA Method	None
21	PCA Components	None
22	Ignore Low Variance	False
23	Combine Rare Levels	False
24	Rare Level Threshold	None
25	Numeric Binning	False
26	Remove Outliers	False
27	Outliers Threshold	None
28	Remove Multicollinearity	False
29	Multicollinearity Threshold	None
30	Clustering	False
31	Clustering Iteration	None
32	Polynomial Features	False
33	Polynomial Degree	None
34	Trigonometry Features	False

	Description	Value
35	Polynomial Threshold	None
36	Group Features	False
37	Feature Selection	False
38	Features Selection Threshold	None
39	Feature Interaction	False
40	Feature Ratio	False
41	Interaction Threshold	None
42	Fix Imbalance	False
43	Fix Imbalance Method	SMOTE

```
Out[46]: (   fSSL_num_extensions  fSSL_num_compression_methods  SYN_tcp_scale \
0           11.0                  1.0                  8.0
1           11.0                  1.0                  8.0
2           13.0                  1.0                  8.0
3           13.0                  1.0                  8.0
4           13.0                  1.0                  8.0
...
14437      ...                  ...                  ...
14438      12.0                  1.0                  5.0
14439      12.0                  1.0                  5.0
14440      12.0                  1.0                  5.0
14441      12.0                  1.0                  5.0

      SYN_MSS  SYN_tcp_winsize  fSSLv_1  fSSLv_3  size_histogram_1 \
0     1460.0       8192.0     0.0     0.0      121.0
1     1460.0       8192.0     0.0     0.0      60.0
2     1460.0       8192.0     0.0     0.0      37.0
3     1460.0       8192.0     0.0     0.0      77.0
4     1460.0       8192.0     0.0     0.0      60.0
...
14437    ...        ...      ...      ...      ...
14438    1460.0     65535.0     0.0     0.0      43.0
14439    1460.0     65535.0     0.0     0.0      8.0
14440    1460.0     65535.0     0.0     0.0      20.0
14441    1460.0     65535.0     0.0     0.0      20.0
14441    1460.0     65535.0     0.0     0.0      8.0

      size_histogram_2  size_histogram_3  ...  fcipher_suites_3_0.0 \
0           5.0          4.0  ...          0.0
1           4.0          3.0  ...          0.0
2           3.0          2.0  ...          1.0
3           8.0          2.0  ...          1.0
4           8.0          4.0  ...          1.0
...
14437    ...        ...      ...      ...
14438    2.0          5.0  ...          1.0
14439    4.0          1.0  ...          1.0
14440    4.0          1.0  ...          1.0
14441    4.0          1.0  ...          1.0

      fcipher_suites_3_1.0  fSSLv_2_0.0  fSSLv_2_1.0  fSSLv_4_0.0 \
0           1.0          0.0          1.0          1.0
1           1.0          0.0          1.0          1.0
2           0.0          0.0          1.0          1.0
3           0.0          0.0          1.0          1.0
4           0.0          0.0          1.0          1.0
...
14437    ...        ...      ...      ...
14438    0.0          0.0          1.0          1.0
14439    0.0          0.0          1.0          1.0
14440    0.0          0.0          1.0          1.0
14441    0.0          0.0          1.0          1.0

      fSSLv_4_1.0  mean_fttl_1_0.0  mean_fttl_1_1.0  mean_fttl_2_0.0 \
0           0.0          1.0          0.0          0.0
1           0.0          1.0          0.0          0.0
2           0.0          1.0          0.0          0.0
3           0.0          1.0          0.0          0.0
```

4	0.0	1.0	0.0	0.0
...	...	...	...	...
14437	0.0	0.0	1.0	1.0
14438	0.0	0.0	1.0	1.0
14439	0.0	0.0	1.0	1.0
14440	0.0	0.0	1.0	1.0
14441	0.0	0.0	1.0	1.0

mean_fttl_2_1.0	
0	1.0
1	1.0
2	1.0
3	1.0
4	1.0
...	...
14437	0.0
14438	0.0
14439	0.0
14440	0.0
14441	0.0

[14442 rows x 77 columns],

0	16102
1	16102
2	16102
3	16102
4	16102
...	...
14437	13103
14438	18103
14439	18103
14440	18103
14441	18103

Name: tuple, Length: 14442, dtype: int64,

	fSSL_num_extensions	fSSL_num_compression_methods	SYN_tcp_scale	\
2337	11.0		1.0	8.0
9280	7.0		1.0	8.0
14162	10.0		1.0	7.0
4051	12.0		1.0	7.0
1622	7.0		1.0	8.0
...	...		...	...
3292	10.0		1.0	7.0
518	13.0		1.0	8.0
13240	5.0		1.0	5.0
3641	12.0		1.0	7.0
6932	10.0		1.0	8.0

	SYN_MSS	SYN_tcp_winsize	fSSLv_1	fSSLv_3	size_histogram_1	\
2337	1460.0	8192.0	0.0	0.0		36.0
9280	1460.0	65535.0	0.0	0.0		84.0
14162	1460.0	29200.0	0.0	0.0		20.0
4051	1460.0	29200.0	0.0	0.0		44.0
1622	1460.0	8192.0	0.0	0.0		10.0
...	...	...	...	...		...
3292	1460.0	29200.0	0.0	0.0		43.0
518	1460.0	8192.0	0.0	0.0		62.0
13240	1460.0	65535.0	0.0	0.0		657.0

## ensemble\_boosting

3641	1460.0	29200.0	0.0	0.0	34.0
6932	1460.0	8192.0	0.0	0.0	289.0
2337	size_histogram_2	size_histogram_3	...	fcipher_suites_3_0.0	\
9280	2.0	5.0	...		0.0
14162	5.0	10.0	...		0.0
4051	1.0	0.0	...		1.0
1622	7.0	1.0	...		1.0
...	1.0	1.0	...		0.0
3292	...	...	...		...
518	5.0	2.0	...		1.0
13240	4.0	1.0	...		1.0
3641	30.0	11.0	...		1.0
6932	4.0	1.0	...		1.0
	3.0	12.0	...		1.0
2337	fcipher_suites_3_1.0	fSSLv_2_0.0	fSSLv_2_1.0	fSSLv_4_0.0	\
9280	1.0	0.0	1.0		1.0
14162	1.0	1.0	0.0		0.0
4051	0.0	0.0	1.0		1.0
1622	0.0	0.0	1.0		1.0
...	1.0	1.0	0.0		0.0
3292	...	...	...		...
518	0.0	0.0	1.0		1.0
13240	0.0	0.0	1.0		1.0
3641	0.0	0.0	1.0		1.0
6932	0.0	0.0	1.0		1.0
2337	fSSLv_4_1.0	mean_fttl_1_0.0	mean_fttl_1_1.0	mean_fttl_2_0.0	\
9280	0.0	1.0	0.0		0.0
14162	1.0	1.0	0.0		0.0
4051	0.0	0.0	1.0		1.0
1622	0.0	0.0	1.0		1.0
...	1.0	1.0	0.0		0.0
3292	...	...	...		...
518	0.0	0.0	1.0		1.0
13240	0.0	0.0	1.0		1.0
3641	0.0	0.0	1.0		1.0
6932	0.0	1.0	0.0		0.0
2337	mean_fttl_2_1.0				
9280	1.0				
14162	1.0				
4051	0.0				
1622	0.0				
...	0.0				
3292	...				
518	0.0				
13240	0.0				
3641	0.0				
6932	0.0				
2337	[10109 rows x 77 columns],				
9280	fSSL_num_extensions	fSSL_num_compression_methods	SYN_tcp_scale	\	

10305	7.0		1.0	8.0
9469	7.0		1.0	8.0
14057	10.0		1.0	7.0
7338	13.0		1.0	8.0
2212	7.0		1.0	8.0
...	...		...	...
2503	12.0		1.0	8.0
6057	13.0		1.0	7.0
3223	10.0		1.0	7.0
13719	5.0		1.0	5.0
6611	10.0		1.0	8.0

	SYN_MSS	SYN_tcp_winsize	fSSLv_1	fSSLv_3	size_histogram_1	\
10305	1460.0	65535.0	0.0	0.0	159.0	
9469	1460.0	65535.0	0.0	0.0	139.0	
14057	1460.0	29200.0	0.0	0.0	570.0	
7338	1460.0	8192.0	0.0	0.0	53.0	
2212	1460.0	8192.0	0.0	0.0	11.0	
...	...	...	...	...	...	
2503	1460.0	8192.0	0.0	0.0	43.0	
6057	1460.0	29200.0	0.0	0.0	65.0	
3223	1460.0	29200.0	0.0	0.0	14.0	
13719	1460.0	65535.0	0.0	0.0	164.0	
6611	1460.0	8192.0	0.0	0.0	42.0	

	size_histogram_2	size_histogram_3	...	fcipher_suites_3_0.0	\
10305	4.0	3.0	...	0.0	
9469	3.0	6.0	...	0.0	
14057	1.0	0.0	...	1.0	
7338	6.0	0.0	...	1.0	
2212	3.0	0.0	...	1.0	
...	...	...	...	...	
2503	4.0	5.0	...	1.0	
6057	9.0	4.0	...	1.0	
3223	1.0	1.0	...	1.0	
13719	16.0	7.0	...	1.0	
6611	1.0	1.0	...	1.0	

	fcipher_suites_3_1.0	fSSLv_2_0.0	fSSLv_2_1.0	fSSLv_4_0.0	\
10305	1.0	1.0	0.0	0.0	
9469	1.0	1.0	0.0	0.0	
14057	0.0	0.0	1.0	1.0	
7338	0.0	0.0	1.0	1.0	
2212	0.0	1.0	0.0	0.0	
...	...	...	...	...	
2503	0.0	0.0	1.0	1.0	
6057	0.0	0.0	1.0	1.0	
3223	0.0	0.0	1.0	1.0	
13719	0.0	0.0	1.0	1.0	
6611	0.0	0.0	1.0	1.0	

	fSSLv_4_1.0	mean_fttl_1_0.0	mean_fttl_1_1.0	mean_fttl_2_0.0	\
10305	1.0	1.0	0.0	0.0	
9469	1.0	1.0	0.0	0.0	
14057	0.0	0.0	1.0	1.0	
7338	0.0	1.0	0.0	0.0	
2212	1.0	1.0	0.0	0.0	

```

...
2503      0.0      1.0      0.0      0.0
6057      0.0      0.0      1.0      1.0
3223      0.0      0.0      1.0      1.0
13719     0.0      0.0      1.0      1.0
6611      0.0      1.0      0.0      0.0

    mean_fttl_2_1.0
10305      1.0
9469       1.0
14057     0.0
7338       1.0
2212      1.0
...
2503      1.0
6057      0.0
3223      0.0
13719     0.0
6611      1.0

[4333 rows x 77 columns],
2337      13102
9280      16302
14162     12201
4051      13201
1622      14602
...
3292      13201
518       16102
13240     18403
3641      13201
6932      16202
Name: tuple, Length: 10109, dtype: int64,
10305     16302
9469      16302
14057     17201
7338      13202
2212      15602
...
2503      13102
6057      18101
3223      16201
13719     17403
6611      16202
Name: tuple, Length: 4333, dtype: int64,
8849,
Pipeline(memory=None,
          steps=[('dtypes',
                  DataTypes_Auto_infer(categorical_features=[],
                                       display_types=False, features_todrop=
          []),
          ml_usecase='regression',
          numerical_features=['min_packet_size',
                               'min_fpkt',
                               'min_bpkt'],
          target='tuple', time_features=[])),
          ('imputer',

```

```

ensemble_boosting
    Simple_Imputer(categorical_strategy='not_available',
                    numeric_strategy='mean',
                    target_variable=None))...
    ('group', Empty()), ('nonliner', Empty()), ('scaling', Empty)
()),
    ('P_transform', Empty()), ('pt_target', Empty()),
    ('binn', Empty()), ('rem_outliers', Empty()),
    ('cluster_all', Empty()), ('dummy', Dummify(target='tuple')),
    ('fix_perfect', Empty()), ('clean_names', Clean_Colum_Names
()),
    ('feature_select', Empty()), ('fix_multi', Empty()),
    ('dfs', Empty()), ('pca', Empty())],
verbose=False),
[('Classification Setup Config',
    Description          Value
0                  session_id      8849
1                  Target Type   Multiclass
2                  Label Encoded    None
3                  Original Data (14442, 70)
4                  Missing Values  False
5                  Numeric Features 61
6                  Categorical Features 8
7                  Ordinal Features  False
8 High Cardinality Features  False
9 High Cardinality Method  None
10 Sampled Data (14442, 70)
11 Transformed Train Set (10109, 77)
12 Transformed Test Set (4333, 77)
13 Numeric Imputer mean
14 Categorical Imputer constant
15 Normalize  False
16 Normalize Method  None
17 Transformation  False
18 Transformation Method  None
19 PCA  False
20 PCA Method  None
21 PCA Components  None
22 Ignore Low Variance  False
23 Combine Rare Levels  False
24 Rare Level Threshold  None
25 Numeric Binning  False
26 Remove Outliers  False
27 Outliers Threshold  None
28 Remove Multicollinearity  False
29 Multicollinearity Threshold  None
30 Clustering  False
31 Clustering Iteration  None
32 Polynomial Features  False
33 Polynomial Degree  None
34 Trigonometry Features  False
35 Polynomial Threshold  None
36 Group Features  False
37 Feature Selection  False
38 Features Selection Threshold  None
39 Feature Interaction  False
40 Feature Ratio  False

```

```

41      Interaction Threshold      None
42          Fix Imbalance      False
43          Fix Imbalance Method      SMOTE),
('X_training Set',
     fSSL_num_extensions  fSSL_num_compression_methods  SYN_tcp_scale  \
2337           11.0                  1.0            8.0
9280            7.0                  1.0            8.0
14162          10.0                  1.0            7.0
4051            12.0                  1.0            7.0
1622            7.0                  1.0            8.0
...
3292           10.0                  1.0            7.0
518             13.0                  1.0            8.0
13240           5.0                  1.0            5.0
3641           12.0                  1.0            7.0
6932           10.0                  1.0            8.0

     SYN_MSS  SYN_tcp_winsize  fSSLv_1  fSSLv_3  size_histogram_1  \
2337   1460.0        8192.0    0.0    0.0            36.0
9280   1460.0       65535.0    0.0    0.0            84.0
14162   1460.0       29200.0    0.0    0.0            20.0
4051   1460.0       29200.0    0.0    0.0            44.0
1622   1460.0        8192.0    0.0    0.0            10.0
...
3292   1460.0       29200.0    0.0    0.0            43.0
518    1460.0        8192.0    0.0    0.0            62.0
13240   1460.0       65535.0    0.0    0.0            657.0
3641   1460.0       29200.0    0.0    0.0            34.0
6932   1460.0        8192.0    0.0    0.0            289.0

     size_histogram_2  size_histogram_3  ...  fcipher_suites_3_0.0  \
2337         2.0          5.0  ...            0.0
9280         5.0          10.0  ...            0.0
14162        1.0          0.0  ...            1.0
4051         7.0          1.0  ...            1.0
1622         1.0          1.0  ...            0.0
...
3292         5.0          2.0  ...            1.0
518          4.0          1.0  ...            1.0
13240        30.0         11.0  ...            1.0
3641         4.0          1.0  ...            1.0
6932         3.0          12.0  ...            1.0

     fcipher_suites_3_1.0  fSSLv_2_0.0  fSSLv_2_1.0  fSSLv_4_0.0  \
2337           1.0          0.0          1.0          1.0
9280           1.0          1.0          0.0          0.0
14162          0.0          0.0          1.0          1.0
4051          0.0          0.0          1.0          1.0
1622           1.0          1.0          0.0          0.0
...
3292          0.0          0.0          1.0          1.0
518            0.0          0.0          1.0          1.0
13240          0.0          0.0          1.0          1.0
3641          0.0          0.0          1.0          1.0
6932          0.0          0.0          1.0          1.0

     fSSLv_4_1.0  mean_fttl_1_0.0  mean_fttl_1_1.0  mean_fttl_2_0.0  \

```

```

ensemble_boosting
2337      0.0      1.0      0.0      0.0
9280      1.0      1.0      0.0      0.0
14162     0.0      0.0      1.0      1.0
4051      0.0      0.0      1.0      1.0
1622      1.0      1.0      0.0      0.0
...       ...
3292     0.0      0.0      1.0      1.0
518      0.0      1.0      0.0      0.0
13240     0.0      0.0      1.0      1.0
3641     0.0      0.0      1.0      1.0
6932     0.0      1.0      0.0      0.0

mean_fttl_2_1.0
2337      1.0
9280      1.0
14162     0.0
4051      0.0
1622      1.0
...
3292     0.0
518      1.0
13240     0.0
3641     0.0
6932     1.0

[10109 rows x 77 columns]),
('y_training Set',
2337      13102
9280      16302
14162     12201
4051      13201
1622      14602
...
3292     13201
518      16102
13240     18403
3641      13201
6932      16202
Name: tuple, Length: 10109, dtype: int64),
('X_test Set',
   fSSL_num_extensions  fSSL_num_compression_methods  SYN_tcp_scale \
10305          7.0                  1.0            8.0
9469          7.0                  1.0            8.0
14057         10.0                  1.0            7.0
7338          13.0                  1.0            8.0
2212          7.0                  1.0            8.0
...
2503          ...                  ...
2503          12.0                  1.0            8.0
6057          13.0                  1.0            7.0
3223          10.0                  1.0            7.0
13719         5.0                  1.0            5.0
6611          10.0                 1.0            8.0

   SYN_MSS  SYN_tcp_winsize  fSSLv_1  fSSLv_3  size_histogram_1 \
10305  1460.0      65535.0    0.0    0.0        159.0
9469   1460.0      65535.0    0.0    0.0        139.0
14057  1460.0     29200.0    0.0    0.0        570.0

```

## ensemble\_boosting

7338	1460.0	8192.0	0.0	0.0	53.0
2212	1460.0	8192.0	0.0	0.0	11.0
...	...	...	...	...	...
2503	1460.0	8192.0	0.0	0.0	43.0
6057	1460.0	29200.0	0.0	0.0	65.0
3223	1460.0	29200.0	0.0	0.0	14.0
13719	1460.0	65535.0	0.0	0.0	164.0
6611	1460.0	8192.0	0.0	0.0	42.0
size_histogram_2 size_histogram_3 ... fcipher_suites_3_0.0 \					
10305		4.0	3.0	...	0.0
9469		3.0	6.0	...	0.0
14057		1.0	0.0	...	1.0
7338		6.0	0.0	...	1.0
2212		3.0	0.0	...	1.0
...		...	...	...	...
2503		4.0	5.0	...	1.0
6057		9.0	4.0	...	1.0
3223		1.0	1.0	...	1.0
13719		16.0	7.0	...	1.0
6611		1.0	1.0	...	1.0
fcipher_suites_3_1.0 fSSLv_2_0.0 fSSLv_2_1.0 fSSLv_4_0.0 \					
10305		1.0	1.0	0.0	0.0
9469		1.0	1.0	0.0	0.0
14057		0.0	0.0	1.0	1.0
7338		0.0	0.0	1.0	1.0
2212		0.0	1.0	0.0	0.0
...		...	...	...	...
2503		0.0	0.0	1.0	1.0
6057		0.0	0.0	1.0	1.0
3223		0.0	0.0	1.0	1.0
13719		0.0	0.0	1.0	1.0
6611		0.0	0.0	1.0	1.0
fSSLv_4_1.0 mean_fttl_1_0.0 mean_fttl_1_1.0 mean_fttl_2_0.0 \					
10305		1.0	1.0	0.0	0.0
9469		1.0	1.0	0.0	0.0
14057		0.0	0.0	1.0	1.0
7338		0.0	1.0	0.0	0.0
2212		1.0	1.0	0.0	0.0
...		...	...	...	...
2503		0.0	1.0	0.0	0.0
6057		0.0	0.0	1.0	1.0
3223		0.0	0.0	1.0	1.0
13719		0.0	0.0	1.0	1.0
6611		0.0	1.0	0.0	0.0
mean_fttl_2_1.0					
10305		1.0			
9469		1.0			
14057		0.0			
7338		1.0			
2212		1.0			
...		...			
2503		1.0			
6057		0.0			

```

3223          0.0
13719         0.0
6611          1.0

[4333 rows x 77 columns]),
('y_test Set',
10305    16302
9469     16302
14057    17201
7338     13202
2212     15602
...
2503     13102
6057     18101
3223     16201
13719    17403
6611     16202
Name: tuple, Length: 4333, dtype: int64),
('Transformation Pipeline',
Pipeline(memory=None,
      steps=[('dtypes',
               DataTypes_Auto_infer(categorical_features=[],
                                     display_types=False, features_todrop
= [],
                                     ml_usecase='regression',
                                     numerical_features=['min_packet_siz
e',
                                     'min_fpkt',
                                     'min_bpkt'],
                                     target='tuple', time_features[])),
              ('imputer',
               Simple_Imputer(categorical_strategy='not_available',
                               numeric_strategy='mean',
                               target_variable=None))...
              ('group', Empty()), ('nonliner', Empty()), ('scaling', Emp
ty()),
              ('P_transform', Empty()), ('pt_target', Empty()),
              ('binn', Empty()), ('rem_outliers', Empty()),
              ('cluster_all', Empty()), ('dummy', Dummify(target='tupl
e')),
              ('fix_perfect', Empty()), ('clean_names', Clean_Colum_Nam
e),
              ('feature_select', Empty()), ('fix_multi', Empty()),
              ('dfs', Empty()), ('pca', Empty())],
      verbose=False),
False,
-1,
True,
[], []
[], []
[], []
['no_logging'],
False,
False,
'b0d7',
False,

```

```
None,
<Logger logs (DEBUG)>)
```

## Compare Models

In [47]: `compare_models(whitelist=learning_model)`

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
0 Extra Trees Classifier	0.9770	0.0000	0.8978	0.9771	0.9762	0.9730	0.9731	7.0587
1 Extreme Gradient Boosting	0.9768	0.0000	0.8759	0.9759	0.9756	0.9728	0.9728	13.0821
2 Random Forest Classifier	0.9681	0.0000	0.8623	0.9677	0.9667	0.9627	0.9628	4.4687
3 Light Gradient Boosting Machine	0.9653	0.0000	0.8207	0.9726	0.9678	0.9594	0.9595	6.5174

Out[47]: `OneVsRestClassifier(estimator=ExtraTreesClassifier(bootstrap=False,  
                         ccp_alpha=0.0,  
                         class_weight=None,  
                         criterion='gini',  
                         max_depth=None,  
                         max_features='auto',  
                         max_leaf_nodes=None,  
                         max_samples=None,  
                         min_impurity_decrease=0.0,  
                         min_impurity_split=None,  
                         min_samples_leaf=1,  
                         min_samples_split=2,  
                         min_weight_fraction_leaf=  
                         0.0,  
                         n_estimators=100, n_jobs=-  
                         1,  
                         oob_score=False,  
                         random_state=8849, verbose  
                         =0,  
                         warm_start=False),  
                         n_jobs=-1)`

## Creating Learning Model - Boosting RF

```
In [16]: model = create_model(learning_model[0], ensemble=True, method=ensemble_method)
```

	<b>Accuracy</b>	<b>AUC</b>	<b>Recall</b>	<b>Prec.</b>	<b>F1</b>	<b>Kappa</b>	<b>MCC</b>
<b>0</b>	0.9802	0.0000	0.8915	0.9818	0.9788	0.9769	0.9769
<b>1</b>	0.9733	0.0000	0.8873	0.9736	0.9722	0.9687	0.9688
<b>2</b>	0.9802	0.0000	0.9504	0.9805	0.9792	0.9768	0.9769
<b>3</b>	0.9723	0.0000	0.8748	0.9714	0.9708	0.9676	0.9677
<b>4</b>	0.9782	0.0000	0.8798	0.9775	0.9771	0.9745	0.9746
<b>5</b>	0.9703	0.0000	0.8495	0.9710	0.9675	0.9652	0.9653
<b>6</b>	0.9792	0.0000	0.8546	0.9776	0.9778	0.9757	0.9757
<b>7</b>	0.9703	0.0000	0.8589	0.9679	0.9682	0.9653	0.9653
<b>8</b>	0.9703	0.0000	0.8659	0.9694	0.9685	0.9652	0.9653
<b>9</b>	0.9713	0.0000	0.9027	0.9721	0.9704	0.9664	0.9664
<b>Mean</b>	0.9746	0.0000	0.8815	0.9743	0.9731	0.9702	0.9703
<b>SD</b>	0.0041	0.0000	0.0281	0.0045	0.0044	0.0048	0.0048

## Prediction

```
In [17]: predicted = timed_prediction(data,model)
```

prediction took: 5.296875

```
In [18]: check_correction(predicted)
```

number of error: 134 from 14442 test samples  
which is 0.009278493283478743 precent of error.

## Make Unseen Test

```
In [19]: predicted = predict_model(model, data=unseen_data)
```

```
In [20]: predicted['Label'].unique()
```

```
Out[20]: array([16102, 14602, 15602, 18102, 13102, 13202, 11602, 16201, 13201,
   18101, 13403, 13101, 18201, 16101, 17201, 12201, 17101, 14601,
   16202, 18202, 13302, 16302, 18302, 18403, 16403, 17403, 12101,
   16103, 13103, 18103])
```

check prediction correction

In [21]: `answers.unique()`

Out[21]: `array([16102, 14602, 15602, 18102, 13102, 11602, 16201, 13201, 18201, 16101, 13101, 18101, 16202, 18202, 13202, 16302, 18302, 13302, 13403, 18403, 17403, 14601, 17101, 12101, 17201, 12201, 16403, 16103, 13103, 18103], dtype=int64)`

In [22]: `compare_prediction_with_answers(predicted, answers)`

number of error: 180 from 6189 test samples  
which is 0.029083858458555503 precent of error.

## Creating Learning Model - Boosting ET

In [23]: `model = create_model(learning_model[1], ensemble=True, method=ensemble_method)`

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
<b>0</b>	0.9901	0.0000	0.9091	0.9896	0.9891	0.9884	0.9884
<b>1</b>	0.9773	0.0000	0.9073	0.9772	0.9761	0.9734	0.9734
<b>2</b>	0.9773	0.0000	0.9117	0.9778	0.9767	0.9734	0.9734
<b>3</b>	0.9723	0.0000	0.8724	0.9732	0.9718	0.9676	0.9676
<b>4</b>	0.9852	0.0000	0.9660	0.9865	0.9850	0.9827	0.9827
<b>5</b>	0.9782	0.0000	0.8737	0.9774	0.9763	0.9745	0.9746
<b>6</b>	0.9812	0.0000	0.8495	0.9777	0.9789	0.9780	0.9780
<b>7</b>	0.9733	0.0000	0.8671	0.9708	0.9716	0.9687	0.9688
<b>8</b>	0.9693	0.0000	0.8817	0.9690	0.9683	0.9641	0.9641
<b>9</b>	0.9772	0.0000	0.8934	0.9756	0.9760	0.9733	0.9734
<b>Mean</b>	0.9781	0.0000	0.8932	0.9775	0.9770	0.9744	0.9744
<b>SD</b>	0.0058	0.0000	0.0310	0.0061	0.0059	0.0068	0.0068

## Prediction

In [24]: `predicted = timed_prediction(data, model)`

prediction took: 6.859375

In [25]: `check_correction(predicted)`

number of error: 121 from 14442 test samples  
which is 0.008378340950006924 precent of error.

## Make Unseen Test

```
In [26]: predicted = predict_model(model, data=unseen_data)
```

```
In [27]: predicted['Label'].unique()
```

```
Out[27]: array([16102, 18102, 13102, 14602, 16302, 15602, 13202, 11602, 16201,
 13201, 13101, 18101, 18201, 16101, 17201, 12201, 14601, 12101,
 16202, 18202, 13302, 18302, 13403, 18403, 16403, 17403, 17101,
 16103, 13103, 18103])
```

check prediction correction

```
In [28]: answers.unique()
```

```
Out[28]: array([16102, 14602, 15602, 18102, 13102, 11602, 16201, 13201, 18201,
 16101, 13101, 18101, 16202, 18202, 13202, 16302, 18302, 13302,
 13403, 18403, 17403, 14601, 17101, 12101, 17201, 12201, 16403,
 16103, 13103, 18103], dtype=int64)
```

```
In [29]: compare_prediction_with_answers(predicted,answers)
```

```
number of error: 170 from 6189 test samples
which is 0.027468088544191308 precent of error.
```

## Creating Learning Model - Boosting lightgbm

```
In [30]: model = create_model(learning_model[2], ensemble=True, method=ensemble_method)
```

	<b>Accuracy</b>	<b>AUC</b>	<b>Recall</b>	<b>Prec.</b>	<b>F1</b>	<b>Kappa</b>	<b>MCC</b>
<b>0</b>	0.9881	0.0000	0.9466	0.9885	0.9878	0.9861	0.9861
<b>1</b>	0.9802	0.0000	0.8992	0.9796	0.9791	0.9769	0.9769
<b>2</b>	0.9773	0.0000	0.9182	0.9770	0.9761	0.9734	0.9734
<b>3</b>	0.9753	0.0000	0.9090	0.9767	0.9745	0.9711	0.9711
<b>4</b>	0.9782	0.0000	0.8426	0.9764	0.9763	0.9745	0.9746
<b>5</b>	0.9782	0.0000	0.8300	0.9758	0.9753	0.9745	0.9746
<b>6</b>	0.9842	0.0000	0.8680	0.9808	0.9821	0.9815	0.9815
<b>7</b>	0.9723	0.0000	0.8644	0.9701	0.9708	0.9676	0.9676
<b>8</b>	0.9733	0.0000	0.8645	0.9719	0.9714	0.9687	0.9688
<b>9</b>	0.9782	0.0000	0.8944	0.9769	0.9768	0.9745	0.9745
<b>Mean</b>	0.9785	0.0000	0.8837	0.9774	0.9770	0.9749	0.9749
<b>SD</b>	0.0045	0.0000	0.0342	0.0048	0.0048	0.0053	0.0053

## Prediction

```
In [31]: predicted = timed_prediction(data,model)
```

prediction took: 32.90625

```
In [32]: check_correction(predicted)
```

number of error: 112 from 14442 test samples  
which is 0.007755158565295665 precent of error.

## Make Unseen Test

```
In [33]: predicted = predict_model(model, data=unseen_data)
```

```
In [34]: predicted['Label'].unique()
```

```
Out[34]: array([16102, 13102, 13202, 14602, 18102, 15602, 11602, 16201, 13201,
       13101, 16101, 18201, 17201, 12201, 18101, 17101, 14601, 12101,
       16202, 18202, 13302, 16302, 18302, 13403, 18403, 17403, 16103,
       16403, 13103, 18103])
```

check prediction correction

In [35]: `answers.unique()`

Out[35]: `array([16102, 14602, 15602, 18102, 13102, 11602, 16201, 13201, 18201, 16101, 13101, 18101, 16202, 18202, 13202, 16302, 18302, 13302, 13403, 18403, 17403, 14601, 17101, 12101, 17201, 12201, 16403, 16103, 13103, 18103], dtype=int64)`

In [36]: `compare_prediction_with_answers(predicted, answers)`

number of error: 159 from 6189 test samples  
which is 0.025690741638390694 precent of error.

## Creating Learning Model - Boosting xgboost

In [48]: `model = create_model(learning_model[3], ensemble=True, method=ensemble_method)`

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
<b>0</b>	0.3234	0.0000	0.0345	0.1046	0.1581	0.0000	0.0000
<b>1</b>	0.3234	0.0000	0.0345	0.1046	0.1581	0.0000	0.0000
<b>2</b>	0.3234	0.0000	0.0345	0.1046	0.1581	0.0000	0.0000
<b>3</b>	0.3234	0.0000	0.0345	0.1046	0.1581	0.0000	0.0000
<b>4</b>	0.3234	0.0000	0.0345	0.1046	0.1581	0.0000	0.0000
<b>5</b>	0.3234	0.0000	0.0345	0.1046	0.1581	0.0000	0.0000
<b>6</b>	0.3234	0.0000	0.0345	0.1046	0.1581	0.0000	0.0000
<b>7</b>	0.3234	0.0000	0.0333	0.1046	0.1581	0.0000	0.0000
<b>8</b>	0.3234	0.0000	0.0333	0.1046	0.1581	0.0000	0.0000
<b>9</b>	0.3238	0.0000	0.0345	0.1048	0.1584	0.0000	0.0000
<b>Mean</b>	0.3235	0.0000	0.0343	0.1046	0.1581	0.0000	0.0000
<b>SD</b>	0.0001	0.0000	0.0005	0.0001	0.0001	0.0000	0.0000

## Prediction

In [49]: `predicted = timed_prediction(data, model)`

prediction took: 228.5625

In [50]: `check_correction(predicted)`

number of error: 9770 from 14442 test samples  
which is 0.6764990998476665 precent of error.

# Make Unseen Test

```
In [51]: predicted = predict_model(model, data=unseen_data)
```

```
In [52]: predicted['Label'].unique()
```

```
Out[52]: array([16302])
```

check prediction correction

```
In [53]: answers.unique()
```

```
Out[53]: array([16102, 14602, 15602, 18102, 13102, 11602, 16201, 13201, 18201,
       16101, 13101, 18101, 16202, 18202, 13202, 16302, 18302, 13302,
       13403, 18403, 17403, 14601, 17101, 12101, 17201, 12201, 16403,
       16103, 13103, 18103], dtype=int64)
```

```
In [54]: compare_prediction_with_answers(predicted,answers)
```

```
number of error: 4187 from 6189 test samples
which is 0.6765228631442882 precent of error.
```