Machine:

- Run time -
- Intel 12 Cores (6 CPU)
- 16GB RAM
- 256GB SSD
- 1TB HDD

Contents:

1. Feature Selection
2. Validating selected features
3. Final Data set preperation

In this Notebook I tried various ways suggestted by other kaggle winners and Srikanth sir in case study videos to select best features based on their importances. I have referenced the code blocks which I have taken from other sources. I have tried various time complexity reduction techniques (using various parallelizing techniques) and equipped the best working ones according to my system specs.

In [1]:

```
import matplotlib
matplotlib.use(u'nbAgg')
import matplotlib.pyplot as plt
from sklearn.manifold import TSNE
import seaborn as sns
import pandas as pd
import numpy as np
from tqdm import tqdm
tqdm.pandas()
import pickle as pk
import hickle as hk
import joblib as jb
import random
import string
import array
import math
import sys
import gc

import warnings
warnings.filterwarnings("ignore")
```

In [2]:

```
import warnings
warnings.filterwarnings("ignore")
```

In [3]:

```
from sklearn import preprocessing as pre
# from xgboost import XGBClassifier
# from sklearn.model_selection import RandomizedSearchCV
from sklearn.tree import DecisionTreeClassifier
# from sklearn.calibration import CalibratedClassifierCV
# from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import log_loss
# from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
# from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier,ExtraTreesClassifier
```

In [4]:

```
# from dask.distributed import Client
# client = Client()
```

In [4]:

```
from dask import delayed
# import dask.array as da
import dask.bag as db
import dask.dataframe as ddf
```

In [6]:

```
# client
```

In [5]:

```
from sklearn.feature_selection import SelectKBest, SelectPercentile,SelectFromModel
from sklearn.feature_selection import chi2
from sklearn.pipeline import Pipeline
```

In [6]:

```
import os
os.chdir("D:/LargeDatasets/MicrosoftMalware/")
# os.mkdir("final_features")
```

In [7]:

```
def norm_util(feature):
    feature = feature.apply(float)
    max_value = feature.max()
    min_value = feature.min()
    feature = (feature - min_value) / (max_value - min_value)
    return
def normalize(df):
#     result1 = df.copy()
    jb.Parallel(n_jobs=-2,verbose=3)(jb.delayed(norm_util)(df[feature_name]) for feature_name in df.col
umns if (str(feature_name) != str('ID') and str(feature_name)!=str('Class')))
    return
```

In [59]:

```
def plot_feature_imp(clf,n_disp):
    plt.bar(np.arange(n_disp),sorted(clf.feature_importances_)[::-1][:n_disp])
    plt.title("Feature Importances")
#     plt.xticks(None)
    plt.xlabel("Feature Indices")
    plt.ylabel("Feature Importances")
    plt.show()



def validation_pipeline(X_df,y_df,X_tes=None,y_tes=None,n_disp=30,tr=None):

    print("Feature Selection")
    if (X_tes is None) and (y_tes is None):
        print("Splitting Input Data")
        X_df,X_tes,y_df,y_tes = train_test_split(X_df,y_df,test_size=0.18,random_state=13,stratify=y_df
)

    clf = ExtraTreesClassifier(bootstrap=True,random_state=13,verbose=1,n_jobs=-2)
    clf.fit(X_df,y_df)
    y_pred = clf.predict_proba(X_tes)
    initial = log_loss(y_tes,y_pred)

    clf_new = SelectFromModel(ExtraTreesClassifier(bootstrap=True,random_state=13,verbose=1,n_jobs=-2),
threshold=tr)
    X_new = clf_new.fit_transform(X_df,y_df)
    X_tes = clf_new.transform(X_tes)
```

```
    new_features =clf_new.get_support(indices=True)
#     print(X_new.shape)

    clf_new = ExtraTreesClassifier(bootstrap=True,random_state=13,verbose=1,n_jobs=-2)
    clf_new.fit(X_new,y_df)

#     pipe = Pipeline([('FeatureSelection',SelectFromModel(ExtraTreesClassifier(bootstrap=True,random_s
tate=13,n_jobs=-2),threshold=1e-3))\
#                      ,('Model',ExtraTreesClassifier(bootstrap=True,random_state=13,n_jobs=-2))],verbos
e=2)
#     pipe.fit(X_df,y_df)
    y_pred = clf_new.predict_proba(X_tes)
    final = log_loss(y_tes,y_pred)
    print("="*100)
    print("No.Of features initially:",clf.n_features_,"Initial Logloss:",initial)
    print("No.Of features features selected by their importances:",clf_new.n_features_,"Final Logloss:"
,final)
    print("="*100)
    if final<initial:
        print("Done!!")
        plot_feature_imp(clf,n_disp)
        return new_features
    else :
        print("Done!!")
        plot_feature_imp(clf,n_disp)
        return X_df.columns
```

In [15]:

```
def feature_reduction(X,y):
#     normalize(X)
    X_tr,X_tes,y_tr,y_tes = train_test_split(X,y,test_size=0.18,random_state=13,stratify=y)
    percentiles = np.arange(5,101,5)
    scaler = pre.StandardScaler()
    scaler.fit(X_tr)
    scaler.transform(X_tr)
    scaler.transform(X_tes)
    score_tr=[]
    score_tes=[]

    for p in tqdm(percentiles):
        pipe = Pipeline([('clf1',SelectPercentile(percentile=p)),\
                         ('clf2',ExtraTreesClassifier(n_estimators=500,bootstrap=True,random_state=13,v
erbose=1,n_jobs=-2))],verbose=1)
        pipe.fit(X_tr,y_tr)
        score_tr.append(log_loss(y_tr,pipe.predict_proba(X_tr)))
        score_tes.append(log_loss(y_tes,pipe.predict_proba(X_tes)))
    return score_tr,score_tes,percentiles
```

In [20]:

```
def feature_selection_pipeline2(X,y,ptile=100,tresh=None):
    n1=X.shape[1]
    X_tr,X_tes,y_tr,y_tes = train_test_split(X,y,test_size=0.18,random_state=13,stratify=y)
    print("Standardize data")
    scaler = pre.StandardScaler()
    X_tr = scaler.fit_transform(X_tr)
    X_tes = scaler.transform(X_tes)
    print("Done!!")
    print("Feature Reduction")
    clf = SelectPercentile(percentile=ptile)
    X_tr = clf.fit_transform(X_tr,y_tr)
    X_tes = clf.transform(X_tes)
    print("Done!!")
    print("="*100)
    print("Reduced No.of features to {} from {} by selecting {}percentile of feature scores using ANOVA
test".format(n1,X_tr.shape[1],ptile))
    print("="*100)
    return validation_pipeline(X_tr,y_tr,X_tes,y_tes,n_disp=80,tr=tresh)
```

Feature Selection of Byte Features

Feature reduction of byte file one grams

In [13]:

```
df_byte = pd.read_csv("one_gram_byte_features.csv")
```

In [ ]:

```
df_byte.columns
```

In [50]:

```
y_df = df_byte["Class"]
X_df = df_byte.drop(["ID","Class"],axis=1)
```

In [42]:

```
features = validation_pipeline(X_df,y_df)
```

```
========================================================================================================
Initaial Feature set Shape: (10868, 259)
========================================================================================================
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=-2)]: Done 100 out of 100 | elapsed:    0.4s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 100 out of 100 | elapsed:    0.0s finished
```

```
========================================================================================================
No.Of features initially: 259 Initial Logloss: 0.026727428442270905
No.Of features after feature reduction: 83 Final Logloss: 0.023362600704307644
========================================================================================================
```

In [48]:

```
new_df = df_byte[np.concatenate((features,["ID","Class"]))]
```

In [52]:

```
print(new_df.shape)
# new_df.to_csv("final_features/one_gram_byte.csv",index=False)
```

```
(10868, 85)
```

Feature selection of byte file image features

In [54]:

```
df_byte = pd.read_csv("byte_img_features.csv")
y_df=df_byte["Class"]
X_df=df_byte.drop(["Unnamed: 0","ID","Class","size"],axis=1)
X_df.head()
```

Out[54]:

	BYTE_0	BYTE_1	BYTE_2	BYTE_3	BYTE_4	BYTE_5	BYTE_6	BYTE_7	BYTE_8	BYTE_9	...	BYTE_990	BYTE_991	BYTE_992	BYTE
0	69	56	32	48	66	32	48	48	32	48	...	48	32	48	
1	67	55	32	48	49	32	50	52	32	48	...	48	32	48	

	BYTE_0	BYTE_1	BYTE_2	BYTE_3	BYTE_4	BYTE_5	BYTE_6	BYTE_7	BYTE_8	BYTE_9	...	BYTE_990	BYTE_991	BYTE_992	BYTE
2	67	60	32	67	66	32	67	66	32	67	...	33	32	54	
3	54	65	32	70	70	32	54	56	32	65	...	48	32	49	
4	65	52	32	65	67	32	52	65	32	48	...	52	32	54	

5 rows × 1000 columns

In [55]:

```
new_features=validation_pipeline(X_df,y_df)
```

```
Feature Selection
Splitting Input Data
building tree 1 of 100
building tree 2 of 100
building tree 3 of 100
building tree 4 of 100
building tree 5 of 100
building tree 6 of 100
building tree 7 of 100
building tree 8 of 100
building tree 9 of 100
building tree 10 of 100
building tree 11 of 100
```

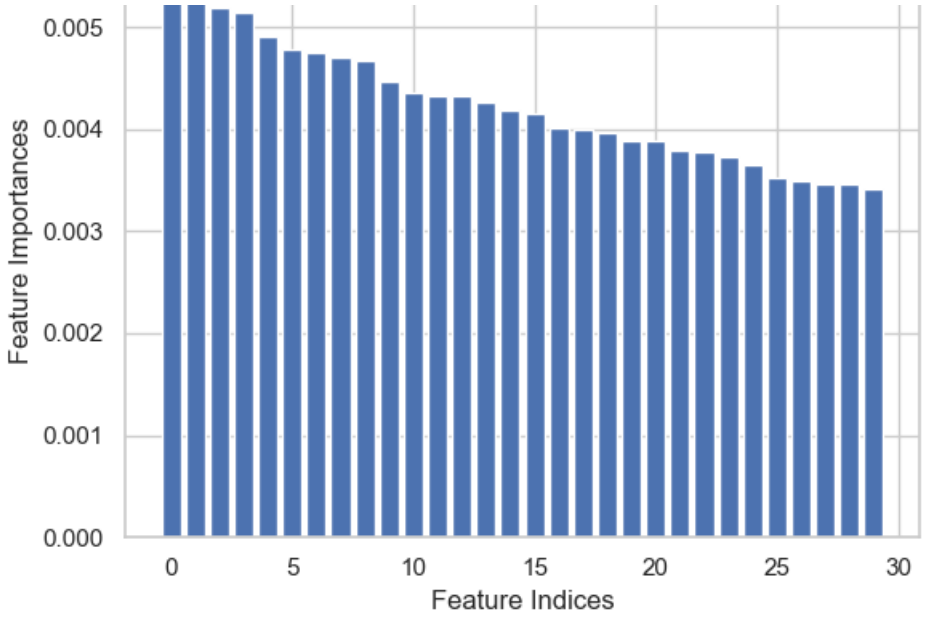
```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
```

```
building tree 12 of 100
building tree 13 of 100
building tree 14 of 100
building tree 15 of 100
building tree 16 of 100
building tree 17 of 100
building tree 18 of 100
building tree 19 of 100
building tree 20 of 100
building tree 21 of 100
building tree 22 of 100
building tree 23 of 100building tree 24 of 100

building tree 25 of 100
building tree 26 of 100
building tree 27 of 100
building tree 28 of 100building tree 29 of 100

building tree 30 of 100building tree 31 of 100
building tree 32 of 100building tree 33 of 100


building tree 34 of 100building tree 35 of 100
```

```
[Parallel(n_jobs=-2)]: Done  19 tasks      | elapsed:    0.1s
```

```
building tree 36 of 100

building tree 37 of 100building tree 38 of 100
building tree 39 of 100
building tree 40 of 100

building tree 41 of 100
building tree 42 of 100building tree 43 of 100building tree 44 of 100


building tree 45 of 100
building tree 46 of 100
building tree 47 of 100
building tree 48 of 100
building tree 49 of 100
building tree 50 of 100
building tree 51 of 100building tree 52 of 100building tree 53 of 100
```

```
building tree 54 of 100

building tree 55 of 100
building tree 56 of 100
building tree 57 of 100
building tree 58 of 100
building tree 59 of 100building tree 60 of 100

building tree 61 of 100
building tree 62 of 100
building tree 63 of 100building tree 64 of 100

building tree 65 of 100building tree 66 of 100

building tree 67 of 100
building tree 68 of 100building tree 69 of 100

building tree 70 of 100
building tree 71 of 100building tree 72 of 100
building tree 73 of 100

building tree 74 of 100
building tree 75 of 100
building tree 76 of 100building tree 77 of 100

building tree 78 of 100
building tree 79 of 100
building tree 80 of 100
building tree 81 of 100
building tree 82 of 100
building tree 83 of 100
building tree 84 of 100building tree 85 of 100

building tree 86 of 100building tree 87 of 100building tree 88 of 100


building tree 89 of 100
building tree 90 of 100
building tree 91 of 100building tree 92 of 100
building tree 93 of 100

building tree 94 of 100building tree 95 of 100

building tree 96 of 100
building tree 97 of 100
building tree 98 of 100
building tree 99 of 100
building tree 100 of 100

[Parallel(n_jobs=-2)]: Done 100 out of 100 | elapsed:    0.9s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  19 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 100 out of 100 | elapsed:    0.0s finished
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    0.2s
[Parallel(n_jobs=-2)]: Done 100 out of 100 | elapsed:    1.0s finished
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    0.1s
[Parallel(n_jobs=-2)]: Done 100 out of 100 | elapsed:    0.6s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 100 out of 100 | elapsed:    0.0s finished

=============================================================================================
No.Of features initially: 1000 Initial Logloss: 0.8515301362797831
No.Of features features selected by their importances: 519 Final Logloss: 0.8495608815652693
=============================================================================================
Done!!
```



In [56]:

```
new_df = X_df.iloc[:,new_features]
new_df.shape
```

Out[56]:

```
(10868, 519)
```

In [57]:

```
new_df["ID"]=ID
new_df["Class"]=y_df
new_df.head()
```

Out[57]:

	BYTE_0	BYTE_1	BYTE_3	BYTE_4	BYTE_6	BYTE_7	BYTE_9	BYTE_10	BYTE_12	BYTE_13	...	BYTE_980	BYTE_983	BYTE_984	B
0	69	56	48	66	48	48	48	48	48	48	...	66	49	49	
1	67	55	48	49	50	52	48	52	53	67	...	69	65	50	
2	67	66	67	66	67	66	67	66	48	48	...	66	66	55	
3	54	65	70	70	54	56	65	51	49	54	...	49	51	48	
4	65	52	65	67	52	65	48	48	65	67	...	69	65	52	

5 rows × 521 columns

In [58]:

```
new_df.to_csv("final_features/byte_img_features.csv",index=False)
# new_df = pd.read_csv("final_features/byte_img_features.csv")
# new_df.shape
```

In [ ]:

```

```

Feature reduction of byte file two grams

In [15]:

```
cols = pd.read_csv("two_gram_byte_features.csv",nrows=1)
```

In [18]:

```
dtyp = {}
for a,b in zip(cols.columns[2:-2],cols.dtypes[2:-2]):
    dtyp[a] = str(b)
```

In [21]:

```
%%time
X_df = pd.read_csv("two_gram_byte_features.csv",dtype=dtyp,usecols=cols.columns[2:-2])
print(sys.getsizeof(X_df)/(1024**3))
```

```
5.306640766561031
Wall time: 31min 48s
```

In [13]:

```
y_df = pd.read_csv("two_gram_byte_features.csv",usecols=["Class"])
ID= pd.read_csv("two_gram_byte_features.csv",usecols=["ID"])
```

In [24]:

```
print(X_df.shape,y_df.shape)
```

```
(10868, 65536) (10868, 1)
```

In [26]:

```
X_df.head()
```

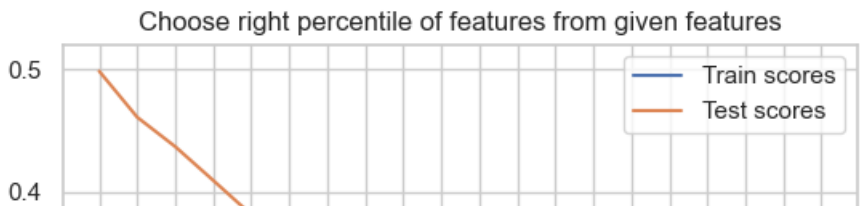
Out[26]:

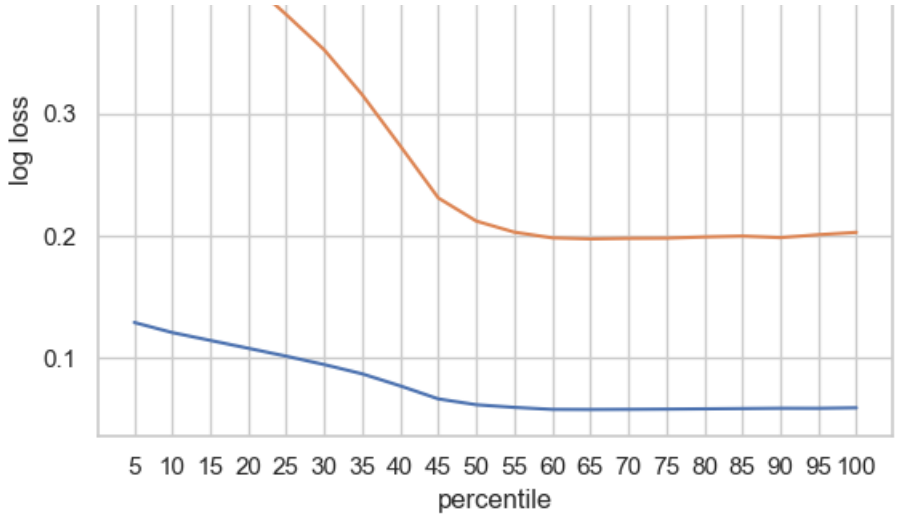
	0	1	2	3	4	5	6	7	8	9	...	65526	65527	65528	65529	65530	65531	65532	65533	65534
0	273053	1002	801	1170	943	840	1125	1003	860	987	...	10	10	9	7	5	7	11	9	6
1	19852	719	64	43	159	10	6	10	35	8	...	35	68	23	72	45	65	15	101	125
2	16032	592	157	144	509	590	551	146	523	154	...	118	73	82	81	108	118	66	97	84
3	9903	204	59	69	103	34	19	21	55	14	...	18	9	55	9	13	17	86	24	63
4	15288	58	20	110	8	11	3	5	8	2	...	52	159	108	24	2	1	0	4	3

5 rows × 65536 columns

In [43]:

```
# score_tr,score_tes,percentiles = feature_reduction(X_df,y_df)
sns.set(style="whitegrid")
plt.plot(percentiles,score_tr,label="Train scores")
plt.plot(percentiles,score_tes,label = "Test scores")
plt.title("Choose right percentile of features from given features")
plt.xticks(percentiles)
plt.xlabel("percentile")
plt.legend()
plt.ylabel("log loss")
plt.show()
```





In [79]:

```
final_features = feature_selection_pipeline2(X_df,y_df,ptile=60)
```

```
Standardize data
Done!!
Feature Reduction
Done!!
==========================================================================================
Reduced No.of features to 65536 from 39321 by selecting 60percentile of feature scores using ANOVA test
==========================================================================================
Feature Selection
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
```

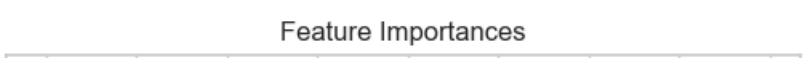
```
building tree 1 of 100building tree 2 of 100building tree 3 of 100
building tree 4 of 100
building tree 5 of 100
building tree 6 of 100building tree 7 of 100
building tree 8 of 100
building tree 9 of 100building tree 10 of 100
building tree 11 of 100
building tree 12 of 100
building tree 13 of 100
building tree 14 of 100
building tree 15 of 100building tree 16 of 100
building tree 17 of 100
building tree 18 of 100
building tree 19 of 100
building tree 20 of 100
building tree 21 of 100
building tree 22 of 100
building tree 23 of 100
building tree 24 of 100
building tree 25 of 100
building tree 26 of 100
building tree 27 of 100
building tree 28 of 100
building tree 29 of 100
building tree 30 of 100
building tree 31 of 100
building tree 32 of 100
building tree 33 of 100
```

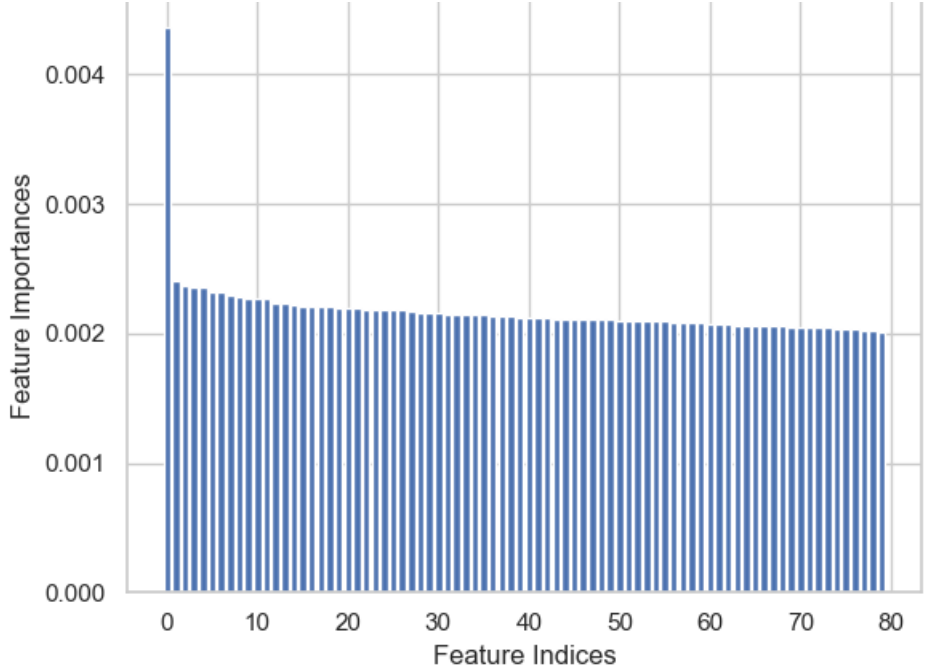
```
[Parallel(n_jobs=-2)]: Done  19 tasks      | elapsed:    1.3s
```

```
building tree 34 of 100
```

```
building tree 35 of 100
building tree 36 of 100
building tree 37 of 100
building tree 38 of 100
building tree 39 of 100
building tree 40 of 100
building tree 41 of 100
building tree 42 of 100
building tree 43 of 100
building tree 44 of 100
building tree 45 of 100
building tree 46 of 100
building tree 47 of 100
building tree 48 of 100
building tree 49 of 100
building tree 50 of 100
building tree 51 of 100
building tree 52 of 100
building tree 53 of 100
building tree 54 of 100
building tree 55 of 100
building tree 56 of 100
building tree 57 of 100
building tree 58 of 100
building tree 59 of 100
building tree 60 of 100
building tree 61 of 100
building tree 62 of 100
building tree 63 of 100
building tree 64 of 100
building tree 65 of 100
building tree 66 of 100
building tree 67 of 100
building tree 68 of 100
building tree 69 of 100
building tree 70 of 100
building tree 71 of 100
building tree 72 of 100
building tree 73 of 100
building tree 74 of 100
building tree 75 of 100
building tree 76 of 100
building tree 77 of 100
building tree 78 of 100
building tree 79 of 100
building tree 80 of 100
building tree 81 of 100
building tree 82 of 100
building tree 83 of 100
building tree 84 of 100
building tree 85 of 100
building tree 86 of 100
building tree 87 of 100
building tree 88 of 100
building tree 89 of 100
building tree 90 of 100
building tree 91 of 100
building tree 92 of 100
building tree 93 of 100
building tree 94 of 100
building tree 95 of 100
building tree 96 of 100
building tree 97 of 100
building tree 98 of 100
building tree 99 of 100
building tree 100 of 100
```

```
[Parallel(n_jobs=-2)]: Done 100 out of 100 | elapsed:    5.4s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  19 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 100 out of 100 | elapsed:    0.0s finished
```






```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.6s
[Parallel(n_jobs=-2)]: Done 100 out of 100 | elapsed:    5.2s finished
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    0.5s
[Parallel(n_jobs=-2)]: Done 100 out of 100 | elapsed:    1.7s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 100 out of 100 | elapsed:    0.0s finished
```

```
=============================================================================================================
No.Of features initially: 39321 Initial Logloss: 0.19840857899840522
No.Of features features selected by their importances: 5654 Final Logloss: 0.1398467454538826
=============================================================================================================
Done!!
```

In [89]:

```
# new_df = X_df.iloc[:,final_features]
new_df.shape
```

Out[89]:

```
(10868, 5656)
```

In [87]:

```
new_df["ID"]=ID
new_df["Class"]=y_df
```

In [88]:

```
new_df.head()
```

Out[88]:

	0	1	2	3	4	5	6	7	8	9	...	39313	39314	39315	39316	39317	39318	39319	39320	
0	273053	1002	801	1170	943	840	1125	1003	860	987	...	7	8	6	5	8	6	4	0	01az
1	19852	719	64	43	159	10	6	10	35	8	...	0	2	15	0	0	15	2	7	01l
2	16032	592	157	144	509	590	551	146	523	154	...	1	4	6	9	4	2	4	6	01js
3	9903	204	59	69	103	34	19	21	55	14	...	1	1	1	2	0	0	2	2	01kcP
4	15288	58	20	110	8	11	3	5	8	2	...	0	0	1	2	0	0	1	0	01S

0	1	2	3	4	5	6	7	8	9	...	39313	39314	39315	39316	39317	39318	39319	39320

5 rows × 5656 columns

In [91]:

```
# new_df.to_csv("final_features/byte_two_gram_features.csv",index=False)
new_df = pd.read_csv("final_features/byte_two_gram_features.csv")
new_df.shape
```

Out[91]:

```
(10868, 5656)
```

Feature reduction of byte file four gram hash Encoded

In [10]:

```
cols = pd.read_csv("four_gram_hash_encoded_byte_features.csv",nrows=1)
dtyp = {}
for a,b in zip(cols.columns[2:-2],cols.dtypes[2:-2]):
    dtyp[a] = str(b)
```

In [11]:

```
%%time
X_df = pd.read_csv("four_gram_hash_encoded_byte_features.csv",dtype=dtyp,usecols=cols.columns[2:-2])
y_df = pd.read_csv("four_gram_hash_encoded_byte_features.csv",usecols=["Class"])
ID= pd.read_csv("four_gram_hash_encoded_byte_features.csv",usecols=["ID"])
print(sys.getsizeof(X_df)/(1024**3))
```

```
5.306640766561031
Wall time: 22min 40s
```

In [12]:

```
print(X_df.shape,y_df.shape)
X_df.head()
```

```
(10868, 65536) (10868, 1)
```

Out[12]:

	0	1	2	3	4	5	6	7	8	9	...	65526	65527	65528	65529	65530	65531	65532	65533	6
0	16236	1831	1510	1959	1696	1562	2024	1721	1506	1741	...	17	4	23	2	5	5	8	2	
1	16188	594	166	159	393	167	97	75	108	18	...	30	81	123	109	43	21	131	105	
2	9633	391	253	258	281	326	286	256	298	257	...	146	148	145	142	173	191	159	157	
3	8051	194	69	49	98	37	33	32	64	13	...	8	5	42	9	12	16	92	27	
4	12582	75	36	38	112	59	8	12	25	5	...	53	162	108	23	0	0	1	0	

5 rows × 65536 columns

In [16]:

```
%time score_tr,score_tes,percentiles = feature_reduction(X_df,y_df)
```

```
  0%|                                                                       | 0/20
[00:00<?, ?it/s]
```

```
[Pipeline] .............. (step 1 of 2) Processing clf1, total=  29.3s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    0.7s
```

```
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    3.2s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:    7.0s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:    8.1s finished
```

```
[Pipeline] ..............  (step 2 of 2) Processing clf2, total=   9.1s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.3s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.4s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.2s finished
  5%|▌                                                                                | 1/20 [00:40<1
2:49, 40.50s/it]
```

```
[Pipeline] ..............  (step 1 of 2) Processing clf1, total=  29.7s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    0.6s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    3.7s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:    8.7s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   10.2s finished
```

```
[Pipeline] ..............  (step 2 of 2) Processing clf2, total=  10.7s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.3s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.4s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.2s finished
 10%|▉                                                                                | 2/20 [01:24<1
2:45, 42.55s/it]
```

```
[Pipeline] ..............  (step 1 of 2) Processing clf1, total=  28.4s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    0.8s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    4.4s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   10.2s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   11.8s finished
```

```
[Pipeline] ..............  (step 2 of 2) Processing clf2, total=  12.4s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.3s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.2s finished
 15%|▌                                                                                | 3/20 [02:08<1
2:14, 43.20s/it]
```

```
[Pipeline] ..............  (step 1 of 2) Processing clf1, total=  29.4s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
```

```
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    0.9s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    4.8s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   11.1s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   12.8s finished
```

```
[Pipeline] ..... (step 2 of 2) Processing clf2, total=  13.4s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 20%|                                                                                | 4/20 [02:54<1
1:49, 44.35s/it]
```

```
[Pipeline] ..... (step 1 of 2) Processing clf1, total=  28.6s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    0.9s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    5.0s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   11.8s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   13.7s finished
```

```
[Pipeline] ..... (step 2 of 2) Processing clf2, total=  14.4s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 25%|                                                                                | 5/20 [03:41<1
1:20, 45.35s/it]
```

```
[Pipeline] ..... (step 1 of 2) Processing clf1, total=  28.1s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    0.9s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    5.5s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   12.9s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   15.0s finished
```

```
[Pipeline] ..... (step 2 of 2) Processing clf2, total=  15.8s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 30%|                                                                                | 6/20 [04:29<1
0:44, 46.03s/it]
```

```
[Pipeline] ..... (step 1 of 2) Processing clf1, total=  29.6s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
```

```
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.0s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    6.0s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   14.5s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   16.7s finished
```

```
[Pipeline] ..............  (step 2 of 2) Processing clf2, total=  17.6s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 35%|                                                                                | 7/20 [05:19<1
0:18, 47.59s/it]
```

```
[Pipeline] ..............  (step 1 of 2) Processing clf1, total=  31.3s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.1s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    6.3s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   14.9s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   17.2s finished
```

```
[Pipeline] ..............  (step 2 of 2) Processing clf2, total=  18.2s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 40%|                                                                                | 8/20 [06:13<0
9:52, 49.42s/it]
```

```
[Pipeline] ..............  (step 1 of 2) Processing clf1, total=  29.8s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.2s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    6.7s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   16.0s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   18.4s finished
```

```
[Pipeline] ..............  (step 2 of 2) Processing clf2, total=  19.6s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 45%|                                                                                | 9/20 [07:07<0
9:20, 50.94s/it]
```

```
[Pipeline] ..............  (step 1 of 2) Processing clf1, total=  31.3s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.3s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    7.1s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   16.8s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   19.5s finished

[Pipeline] ..... (step 2 of 2) Processing clf2, total=  20.7s

[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 50%|                                                                                | 10/20 [08:03<0
8:46, 52.61s/it]

[Pipeline] ..... (step 1 of 2) Processing clf1, total=  29.1s

[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.3s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    7.3s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   20.3s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   23.8s finished

[Pipeline] ..... (step 2 of 2) Processing clf2, total=  24.9s

[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.3s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 55%|                                                                                | 11/20 [09:03<0
8:12, 54.70s/it]

[Pipeline] ..... (step 1 of 2) Processing clf1, total=  31.5s

[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.4s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    7.7s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   18.2s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   21.2s finished

[Pipeline] ..... (step 2 of 2) Processing clf2, total=  22.4s

[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 60%|                                                                                | 12/20 [10:02<0
7:27, 55.95s/it]

[Pipeline] ..... (step 1 of 2) Processing clf1, total=  31.9s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.4s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    8.0s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   18.8s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   21.8s finished
```

```
[Pipeline] ............. (step 2 of 2) Processing clf2, total=  23.2s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 65%|█████████████████████████████████████████████████████                              | 13/20 [11:02<0
6:40, 57.16s/it]
```

```
[Pipeline] ............. (step 1 of 2) Processing clf1, total=  34.6s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.5s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    8.4s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   19.7s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   22.9s finished
```

```
[Pipeline] ............. (step 2 of 2) Processing clf2, total=  24.4s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.3s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 70%|██████████████████████████████████████████████████████████                         | 14/20 [12:05<0
5:55, 59.21s/it]
```

```
[Pipeline] ............. (step 1 of 2) Processing clf1, total=  32.4s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.6s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    8.8s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   20.8s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   24.3s finished
```

```
[Pipeline] ............. (step 2 of 2) Processing clf2, total=  25.9s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 75%|██████████████████████████████████████████████████████████████                     | 15/20 [13:11<0
5:05, 61.12s/it]
```

```
[Pipeline] ............. (step 1 of 2) Processing clf1, total=  30.5s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.6s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    8.8s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   21.4s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   24.9s finished

[Pipeline] ..... (step 2 of 2) Processing clf2, total=  26.5s

[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.3s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.2s finished
 80%|                                                                                | 16/20 [14:15<0
4:07, 61.91s/it]

[Pipeline] ..... (step 1 of 2) Processing clf1, total=  34.9s

[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.7s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    9.3s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   21.9s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   25.4s finished

[Pipeline] ..... (step 2 of 2) Processing clf2, total=  26.9s

[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 85%|                                                                                | 17/20 [15:22<0
3:10, 63.56s/it]

[Pipeline] ..... (step 1 of 2) Processing clf1, total=  35.4s

[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.7s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    9.5s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   22.7s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   26.1s finished

[Pipeline] ..... (step 2 of 2) Processing clf2, total=  27.9s

[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
 90%|                                                                                | 18/20 [16:31<0
2:10, 65.21s/it]

[Pipeline] ..... (step 1 of 2) Processing clf1, total=  37.7s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.7s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:    9.7s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   23.2s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   27.0s finished
```

```
[Pipeline] ..............  (step 2 of 2) Processing clf2, total=  28.8s
```

```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.3s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.1s finished
 95%|                                                                                                  | 19/20 [17:43<0
1:07, 67.33s/it]
```

```
[Pipeline] ..............  (step 1 of 2) Processing clf1, total=  41.6s
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    2.2s
[Parallel(n_jobs=-2)]: Done 178 tasks      | elapsed:   12.1s
[Parallel(n_jobs=-2)]: Done 428 tasks      | elapsed:   28.1s
[Parallel(n_jobs=-2)]: Done 500 out of 500 | elapsed:   33.0s finished
```

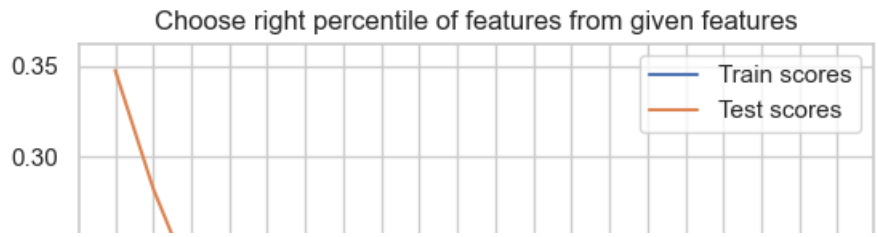
```
[Pipeline] ..............  (step 2 of 2) Processing clf2, total=  35.0s
```

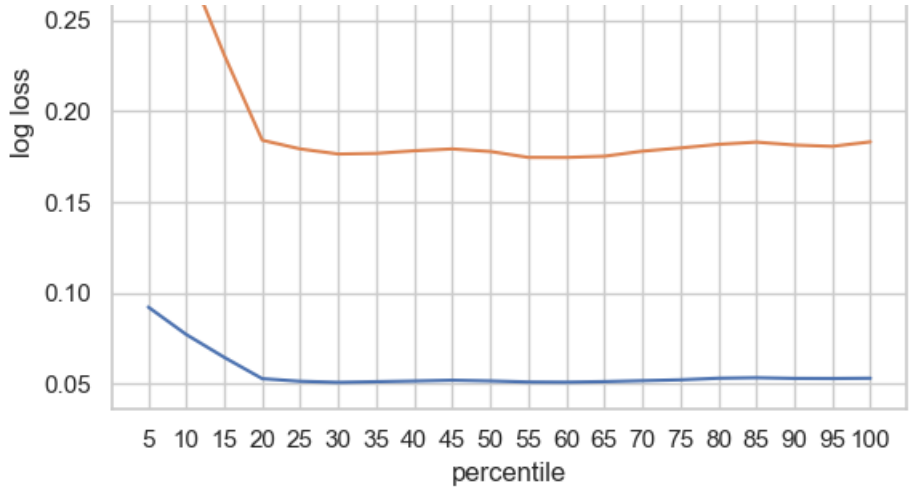
```
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.4s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.4s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 178 tasks      | elapsed:    0.1s
[Parallel(n_jobs=11)]: Done 428 tasks      | elapsed:    0.2s
[Parallel(n_jobs=11)]: Done 500 out of 500 | elapsed:    0.2s finished
100%|                                                                                                  | 20/20 [19:08<0
0:00, 57.44s/it]
```

```
Wall time: 20min 2s
```

In [22]:

```
sns.set(style="whitegrid")
plt.plot(percentiles,score_tr,label="Train scores")
plt.plot(percentiles,score_tes,label = "Test scores")
plt.title("Choose right percentile of features from given features")
plt.xticks(percentiles)
plt.xlabel("percentile")
plt.legend()
plt.ylabel("log loss")
plt.show()
```





In [23]:

```
%time final_features = feature_selection_pipeline2(X_df,y_df,ptile=30)
```

```
Standardize data
Done!!
Feature Reduction
Done!!
=================================================================================================
Reduced No.of features to 65536 from 19661 by selecting 30percentile of feature scores using ANOVA test
=================================================================================================
Feature Selection
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
```

```
building tree 1 of 100building tree 2 of 100building tree 3 of 100
building tree 4 of 100
building tree 5 of 100

building tree 6 of 100
building tree 7 of 100
building tree 8 of 100
building tree 9 of 100
building tree 10 of 100

building tree 11 of 100
building tree 12 of 100
building tree 13 of 100
building tree 14 of 100
building tree 15 of 100
building tree 16 of 100
building tree 17 of 100
building tree 18 of 100
building tree 19 of 100
building tree 20 of 100
building tree 21 of 100
building tree 22 of 100
building tree 23 of 100
building tree 24 of 100
building tree 25 of 100
building tree 26 of 100
building tree 27 of 100
building tree 28 of 100
building tree 29 of 100
building tree 30 of 100
building tree 31 of 100
building tree 32 of 100
building tree 33 of 100
```

```
[Parallel(n_jobs=-2)]: Done  19 tasks      | elapsed:    1.1s
```

```
building tree 34 of 100
building tree 35 of 100
```

```
building tree 36 of 100
building tree 37 of 100
building tree 38 of 100
building tree 39 of 100building tree 40 of 100

building tree 41 of 100
building tree 42 of 100
building tree 43 of 100
building tree 44 of 100
building tree 45 of 100
building tree 46 of 100
building tree 47 of 100
building tree 48 of 100
building tree 49 of 100
building tree 50 of 100
building tree 51 of 100
building tree 52 of 100
building tree 53 of 100
building tree 54 of 100
building tree 55 of 100
building tree 56 of 100
building tree 57 of 100building tree 58 of 100

building tree 59 of 100
building tree 60 of 100
building tree 61 of 100
building tree 62 of 100
building tree 63 of 100
building tree 64 of 100
building tree 65 of 100
building tree 66 of 100
building tree 67 of 100
building tree 68 of 100
building tree 69 of 100
building tree 70 of 100
building tree 71 of 100
building tree 72 of 100
building tree 73 of 100building tree 74 of 100

building tree 75 of 100
building tree 76 of 100
building tree 77 of 100
building tree 78 of 100
building tree 79 of 100
building tree 80 of 100
building tree 81 of 100
building tree 82 of 100
building tree 83 of 100
building tree 84 of 100
building tree 85 of 100
building tree 86 of 100
building tree 87 of 100
building tree 88 of 100
building tree 89 of 100
building tree 90 of 100
building tree 91 of 100
building tree 92 of 100building tree 93 of 100

building tree 94 of 100
building tree 95 of 100
building tree 96 of 100
building tree 97 of 100
building tree 98 of 100
building tree 99 of 100
building tree 100 of 100

[Parallel(n_jobs=-2)]: Done 100 out of 100 | elapsed:    4.7s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  19 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 100 out of 100 | elapsed:    0.0s finished
```

```
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    1.4s
[Parallel(n_jobs=-2)]: Done 100 out of 100 | elapsed:    4.6s finished
[Parallel(n_jobs=-2)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    0.5s
[Parallel(n_jobs=-2)]: Done 100 out of 100 | elapsed:    1.7s finished
[Parallel(n_jobs=11)]: Using backend ThreadingBackend with 11 concurrent workers.
[Parallel(n_jobs=11)]: Done  28 tasks      | elapsed:    0.0s
[Parallel(n_jobs=11)]: Done 100 out of 100 | elapsed:    0.0s finished
```

```
=============================================================================================================
No.Of features initially: 19661 Initial Logloss: 0.17792112555449863
No.Of features features selected by their importances: 2996 Final Logloss: 0.13391851279375921
=============================================================================================================
Done!!
Wall time: 2min 46s
```

In [24]:

```
new_df = X_df.iloc[:,final_features]
new_df.shape
```

Out[24]:

```
(10868, 2996)
```

In [25]:

```
new_df["ID"]=ID
new_df["Class"]=y_df
new_df.head()
```

Out[25]:

	0	1	2	3	4	5	6	7	8	9	...	19648	19651	19652	19655	19656	19658	19659	19660
0	16236	1831	1510	1959	1696	1562	2024	1721	1506	1741	...	2	2	2	3	7	2	5	3
1	16188	594	166	159	393	167	97	75	108	18	...	0	1	0	1	0	0	0	1
2	9633	391	253	258	281	326	286	256	298	257	...	18	0	0	0	0	0	503	0
3	8051	194	69	49	98	37	33	32	64	13	...	4	4	2	1	0	0	0	1
4	12582	75	36	38	112	59	8	12	25	5	...	0	0	0	0	1	0	0	1

5 rows × 2998 columns

In [27]:

```
# new_df.to_csv("final_features/byte_four_gram_hash_encoded_features.csv",index=False)
new_df = pd.read_csv("final_features/byte_four_gram_hash_encoded_features.csv")
new_df.shape
```

Out[27]:

```
(10868, 2998)
```

Feature selection of asm file image features

In [44]:

```
df_byte = pd.read_csv("asm_img_features.csv")
y_df=df_byte["Class"]
X_df=df_byte.drop(["Unnamed: 0","ID","Class","size"],axis=1)
X_df.head()
```

Out[44]:

	ASM_0	ASM_1	ASM_2	ASM_3	ASM_4	ASM_5	ASM_6	ASM_7	ASM_8	ASM_9	...	ASM_990	ASM_991	ASM_992	ASM_993	A
0	72	69	65	68	69	82	58	48	48	52	...	116	101	120	116	
1	46	116	101	120	116	58	48	48	52	48	...	10	46	116	101	
2	72	69	65	68	69	82	58	48	48	52	...	116	101	120	116	
3	72	69	65	68	69	82	58	49	48	48	...	71	77	69	78	
4	72	69	65	68	69	82	58	48	48	52	...	116	101	120	116	

5 rows × 1000 columns

In [ ]:

```
new_features=validation_pipeline(X_df,y_df)
```

In [46]:

```
new_df = X_df.iloc[:,new_features]
new_df.shape
```

Out[46]:

```
(10868, 253)
```

In [47]:

```
new_df["ID"]=ID
new_df["Class"]=y_df
new_df.head()
```

Out[47]:

	ASM_3	ASM_6	ASM_7	ASM_8	ASM_9	ASM_20	ASM_31	ASM_32	ASM_33	ASM_34	...	ASM_991	ASM_992	ASM_993	ASM_
0	68	58	48	48	52	9	58	48	48	52	...	101	120	116	
1	120	48	48	52	48	32	116	101	120	116	...	46	116	101	
2	68	58	48	48	52	9	58	48	48	52	...	101	120	116	
3	68	58	49	48	48	9	58	49	48	48	...	77	69	78	
4	68	58	48	48	52	9	58	48	48	52	...	101	120	116	

5 rows × 255 columns

In [52]:

```
# new_df.to_csv("final_features/asm_img_features.csv",index=False)
new_df = pd.read_csv("final_features/asm_img_features.csv")
new_df.shape
```

Out[52]:

```
(10868, 255)
```

In [ ]:

```

```