Microsoft Malware detection

1.Business/Real-world Problem

1.1. What is Malware?

The term malware is a contraction of malicious software. Put simply, malware is any piece of software that was written with the intent of doing harm to data, devices or to people.
Source: https://www.avg.com/en/signal/what-is-malware

1.2. Problem Statement

In the past few years, the malware industry has grown very rapidly that, the syndicates invest heavily in technologies to evade traditional protection, forcing the anti-malware groups/communities to build more robust softwares to detect and terminate these attacks. The major part of protecting a computer system from a malware attack is to **identify whether a given piece of file/software is a malware.**

1.3 Source/Useful Links

Microsoft has been very active in building anti-malware products over the years and it runs it's anti-malware utilities over 150 million computers around the world. This generates tens of millions of daily data points to be analyzed as potential malware. In order to be effective in analyzing and classifying such large amounts of data, we need to be able to group them into groups and identify their respective families.

This dataset provided by Microsoft contains about 9 classes of malware. ,

Source: https://www.kaggle.com/c/malware-classification

1.4. Real-world/Business objectives and constraints.

1. **Minimize multi-class error.**
2. **Multi-class probability estimates.**
3. **Malware detection should not take hours and block the user's computer. It should fininsh in a few seconds or a minute.**

2. Machine Learning Problem

2.1. Data

2.1.1. Data Overview

- **Source : https://www.kaggle.com/c/malware-classification/data**
- **For every malware, we have two files**

  1. **.asm file (read more: https://www.reviversoft.com/file-extensions/asm)**
  2. **.bytes file (the raw data contains the hexadecimal representation of the file's binary content, without the PE header)**

- **Total train dataset consist of 200GB data out of which 50Gb of data is .bytes files and 150GB of data is .asm files:**
- **Lots of Data for a single-box/computer.**
- **There are total 10,868 .bytes files and 10,868 asm files total 21,736 files**
- **There are 9 types of malwares (9 classes) in our give data**
- **Types of Malware:**

  1. **Ramnit**
  2. **Lollipop**
  3. **Kelihos_ver3**
  4. **Vundo**
  5. **Simda**
  6. **Tracur**
  7. **Kelihos_ver1**
  8. **Obfuscator.ACY**
  9. **Gatak**

2.1.2. Example Data Point

.asm file

```
.text:00401000                                  assume es:nothing,
ss:nothing, ds:_data, fs:nothing, gs:nothing
.text:00401000 56                               push    esi
.text:00401001 8D 44 24 08                          lea     eax,
[esp+8]
.text:00401005 50                               push    eax
.text:00401006 8B F1                                mov     esi, ecx
.text:00401008 E8 1C 1B 00 00                           call    ??
0exception@std@@QAE@ABQBD@Z ; std::exception::exception(char const *
const &)
.text:0040100D C7 06 08 BB 42 00                            mov     dword
ptr [esi], offset off_42BB08
.text:00401013 8B C6                                mov     eax, esi
.text:00401015 5E                               pop     esi
.text:00401016 C2 04 00                             retn    4
.text:00401016                          ; ---------------------------
---------------------------------------------
.text:00401019 CC CC CC CC CC CC CC                         align 10h
.text:00401020 C7 01 08 BB 42 00                            mov     dword
```

```
ptr [ecx], offset off_42BB08
.text:00401026 E9 26 1C 00 00                          jmp
sub_402C51
.text:00401026                          ; ---------------------------
-----------------------------------------------
.text:0040102B CC CC CC CC CC                          align 10h
.text:00401030 56                                  push    esi
.text:00401031 8B F1                                   mov     esi, ecx
.text:00401033 C7 06 08 BB 42 00                       mov     dword
ptr [esi], offset off_42BB08
.text:00401039 E8 13 1C 00 00                          call
sub_402C51
.text:0040103E F6 44 24 08 01                          test    byte
ptr [esp+8], 1
.text:00401043 74 09                                   jz      short
loc_40104E
.text:00401045 56                                  push    esi
.text:00401046 E8 6C 1E 00 00                          call    ??
3@YAXPAX@Z    ; operator delete(void *)
.text:0040104B 83 C4 04                                add     esp, 4
.text:0040104E
.text:0040104E                          loc_40104E:                ;
CODE XREF: .text:00401043j
.text:0040104E 8B C6                                   mov     eax, esi
.text:00401050 5E                                  pop     esi
.text:00401051 C2 04 00                                retn    4
.text:00401051                          ; ---------------------------
-----------------------------------------------
```

.bytes file

```
00401000 00 00 80 40 40 28 00 1C 02 42 00 C4 00 20 04 20
00401010 00 00 20 09 2A 02 00 00 00 00 8E 10 41 0A 21 01
00401020 40 00 02 01 00 90 21 00 32 40 00 1C 01 40 C8 18
00401030 40 82 02 63 20 00 00 09 10 01 02 21 00 82 00 04
00401040 82 20 08 83 00 08 00 00 00 00 02 00 60 80 10 80
00401050 18 00 00 20 A9 00 00 00 00 04 04 78 01 02 70 90
00401060 00 02 00 08 20 12 00 00 00 40 10 00 80 00 40 19
00401070 00 00 00 00 11 20 80 04 80 10 00 20 00 00 25 00
00401080 00 00 01 00 00 04 00 10 02 C1 80 80 00 20 20 00
00401090 08 A0 01 01 44 28 00 00 08 10 20 00 02 08 00 00
004010A0 00 40 00 00 00 34 40 40 00 04 00 08 80 08 00 08
004010B0 10 00 40 00 68 02 40 04 E1 00 28 14 00 08 20 0A
004010C0 06 01 02 00 40 00 00 00 00 00 00 20 00 02 00 04
004010D0 80 18 90 00 00 10 A0 00 45 09 00 10 04 40 44 82
004010E0 90 00 26 10 00 00 04 00 82 00 00 00 20 40 00 00
004010F0 B4 00 00 40 00 02 20 25 08 00 00 00 00 00 00 00
00401100 08 00 00 50 00 08 40 50 00 02 06 22 08 85 30 00
00401110 00 80 00 80 60 00 09 00 04 20 00 00 00 00 00 00
00401120 00 82 40 02 00 11 46 01 4A 01 8C 01 E6 00 86 10
00401130 4C 01 22 00 64 00 AE 01 EA 01 2A 11 E8 10 26 11
00401140 4E 11 8E 11 C2 00 6C 00 0C 11 60 01 CA 00 62 10
00401150 6C 01 A0 11 CE 10 2C 11 4E 10 8C 00 CE 01 AE 01
```

```
00401160 6C 10 6C 11 A2 01 AE 00 46 11 EE 10 22 00 A8 00
00401170 EC 01 08 11 A2 01 AE 10 6C 00 6E 00 AC 11 8C 00
00401180 EC 01 2A 10 2A 01 AE 00 40 00 C8 10 48 01 4E 11
00401190 0E 00 EC 11 24 10 4A 10 04 01 C8 11 E6 01 C2 00
```

2.2. Mapping the real-world problem to an ML problem

2.2.1. Type of Machine Learning Problem

There are nine different classes of malware that we need to classify a given a data point => Multi class classification problem

2.2.2. Performance Metric

Source: https://www.kaggle.com/c/malware-classification#evaluation

Metric(s):

- **Multi class log-loss**
- **Confusion matrix**

2.2.3. Machine Learing Objectives and Constraints

Objective: Predict the probability of each data-point belonging to each of the nine classes.

Constraints:

- **Class probabilities are needed.**
- **Penalize the errors in class probabilites => Metric is Log-loss.**
- **Some Latency constraints.**

2.3. Train and Test Dataset

Split the dataset randomly into three parts train, cross validation and test with 64%,16%, 20% of data respectively

2.4. Useful blogs, videos and reference papers

http://blog.kaggle.com/2015/05/26/microsoft-malware-winners-interview-1st-place-no-to-overfitting/
https://arxiv.org/pdf/1511.04317.pdf
First place solution in Kaggle competition: https://www.youtube.com/watch?v=VLQTRlLGz5Y
https://github.com/dchad/malware-detection
http://vizsec.org/files/2011/Nataraj.pdf
https://www.dropbox.com/sh/gfqzv0ckgs4l1bf/AAB6EelnEjvvuQg2nu_pIB6ua?dl=0
" Cross validation is more trustworthy than domain knowledge."

3. Exploratory Data Analysis

In [1]:
```
#To fetch data from kaggle directly to jupyter on GCP


# !wget --header="Host: storage.googleapis.com" --header="User-Agent: Mozilla/5.0 (Wind
```

```
--2020-12-24 17:50:06--  https://storage.googleapis.com/kagglesdsdata/competitions/4117/
46665/train.7z?GoogleAccessId=web-data@kaggle-161607.iam.gserviceaccount.com&Expires=160
9091375&Signature=F8FsESyHgo3CcacEXOAma%2FbZ5zSaHlrsdjL1CBs8Y4HPbeyB3GgH4uVtc3aCJGPwKQUA
CyyhUr89Xb7cJllhnsRKTx%2FuVfYV2vxQ5dipQhIJWrlTzj9NQVkUKsy%2BmhPNLWCQ%2F06f0uYOY0bZKHUnLi
AOCM0OetnlukVcXXDyg9gPBgEulDivg4QshNQY%2BnsrL0VASR40YcLUybbp%2Fu4jU9HtdBqhpWpUj8a99QWZei
2TriEYJeu%2BJvoKlXnW7%2FVT8vE%2BuYUS2Wz9TKRj7ZTiVvmvAec%2FO4Hg54y3vbzbjQEf9xXZcRS4WoyZYl
7H3qffL%2BMChCTDsrHTeEu6xqokWg%3D%3D&response-content-disposition=attachment%3B+filenam
e%3Dtrain.7z
Resolving storage.googleapis.com (storage.googleapis.com)... 74.125.31.128, 74.125.139.1
28, 173.194.210.128, ...
Connecting to storage.googleapis.com (storage.googleapis.com)|74.125.31.128|:443... conn
ected.
HTTP request sent, awaiting response... 200 OK
Length: 18810691091 (18G) [application/x-7z-compressed]
Saving to: ‘CurlWget388’

CurlWget388         100%[===================>]  17.52G   120MB/s    in 2m 32s

2020-12-24 17:52:38 (118 MB/s) - ‘CurlWget388’ saved [18810691091/18810691091]
```

In [1]:
```
#Code to unzip train.7z

# !pip install py7zr
# !python -m py7zr x train.7z
```

In [3]:
```
# pip install xgboost
```

```
Collecting xgboost
  Downloading xgboost-1.3.1-py3-none-manylinux2010_x86_64.whl (157.5 MB)
     |████████████████████████████████| 157.5 MB 45 kB/s s eta 0:00:01
Requirement already satisfied: scipy in /opt/conda/lib/python3.7/site-packages (from xgb
oost) (1.5.3)
Requirement already satisfied: numpy in /opt/conda/lib/python3.7/site-packages (from xgb
oost) (1.19.4)
Requirement already satisfied: numpy in /opt/conda/lib/python3.7/site-packages (from xgb
oost) (1.19.4)
Installing collected packages: xgboost
Successfully installed xgboost-1.3.1
Note: you may need to restart the kernel to use updated packages.
```

In [1]:
```
import warnings
warnings.filterwarnings("ignore")
import shutil
import os
import pandas as pd
import matplotlib
matplotlib.use(u'nbAgg')
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pickle
from sklearn.manifold import TSNE
```

```
from sklearn import preprocessing
import pandas as pd
from multiprocessing import Process# this is used for multithreading
import multiprocessing
import codecs# this is used for file operations
import random as r
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
from sklearn.feature_extraction.text import CountVectorizer
from scipy.sparse import csr_matrix
from tqdm import tqdm
import scipy.sparse
```

In [ ]:

```
#separating byte files and asm files

source = 'train'
destination_1 = 'byteFiles'
destination_2 = 'asmFiles'

# we will check if the folder 'byteFiles' exists if it not there we will create a folde
if not os.path.isdir(destination_1):
    os.makedirs(destination_1)
if not os.path.isdir(destination_2):
    os.makedirs(destination_2)

# if we have folder called 'train' (train folder contains both .asm files and .bytes fi
# for every file that we have in our 'asmFiles' directory we check if it is ending with
# 'byteFiles' folder

# so by the end of this snippet we will separate all the .byte files and .asm files
if os.path.isdir(source):
    data_files = os.listdir(source)
    for file in data_files:
        print(file)
        if (file.endswith("bytes")):
            shutil.move(source+'/'+file,destination_1)
        if (file.endswith("asm")):
            shutil.move(source+'/'+file,destination_2)
```

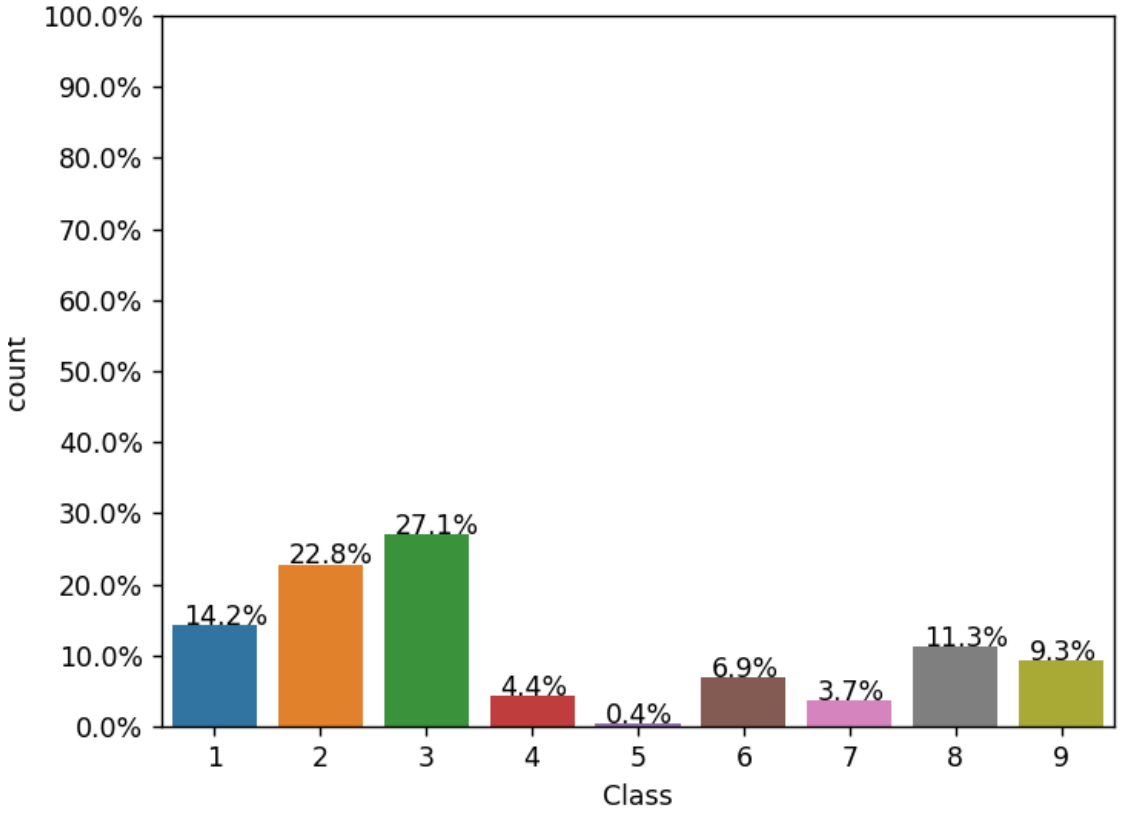
3.1. Distribution of malware classes in whole data set

In [2]:

```
Y=pd.read_csv("trainLabels.csv")
print(Y.head(10))
total = len(Y)*1.
ax=sns.countplot(x="Class", data=Y)
for p in ax.patches:
        ax.annotate('{:.1f}%'.format(100*p.get_height()/total), (p.get_x()+0.1, p.get_h

#put 11 ticks (therefore 10 steps), from 0 to the total number of rows in the dataframe
```

```
ax.yaxis.set_ticks(np.linspace(0, total, 11))

#adjust the ticklabel to the desired format, without changing the position of the ticks
ax.set_yticklabels(map('{:.1f}%'.format, 100*ax.yaxis.get_majorticklocs()/total))
plt.show()
```

	Id	Class
0	01kcPWA9K2BOxQeS5Rju	1
1	04EjIdbPV5e1XroFOpiN	1
2	05EeG39MTRrI6VY21DPd	1
3	05rJTUWYAKNegBk2wE8X	1
4	0AnoOZDNbPXIr2MRBSCJ	1
5	0AwWs42SUQ19mI7eDcTC	1
6	0cH8YeO15ZywEhPrJvmj	1
7	0DNVFKwYlcjO7bTfJ5p1	1
8	0DqUX5rkg3IbMY6BLGCE	1
9	0eaNKwluUmkYdIvZ923c	1



3.2. Feature extraction

3.2.1 File size of byte files as a feature

In [3]:

```
#file sizes of byte files

files=os.listdir('byteFiles')
filenames=Y['Id'].tolist()
class_y=Y['Class'].tolist()
class_bytes=[]
sizebytes=[]
```

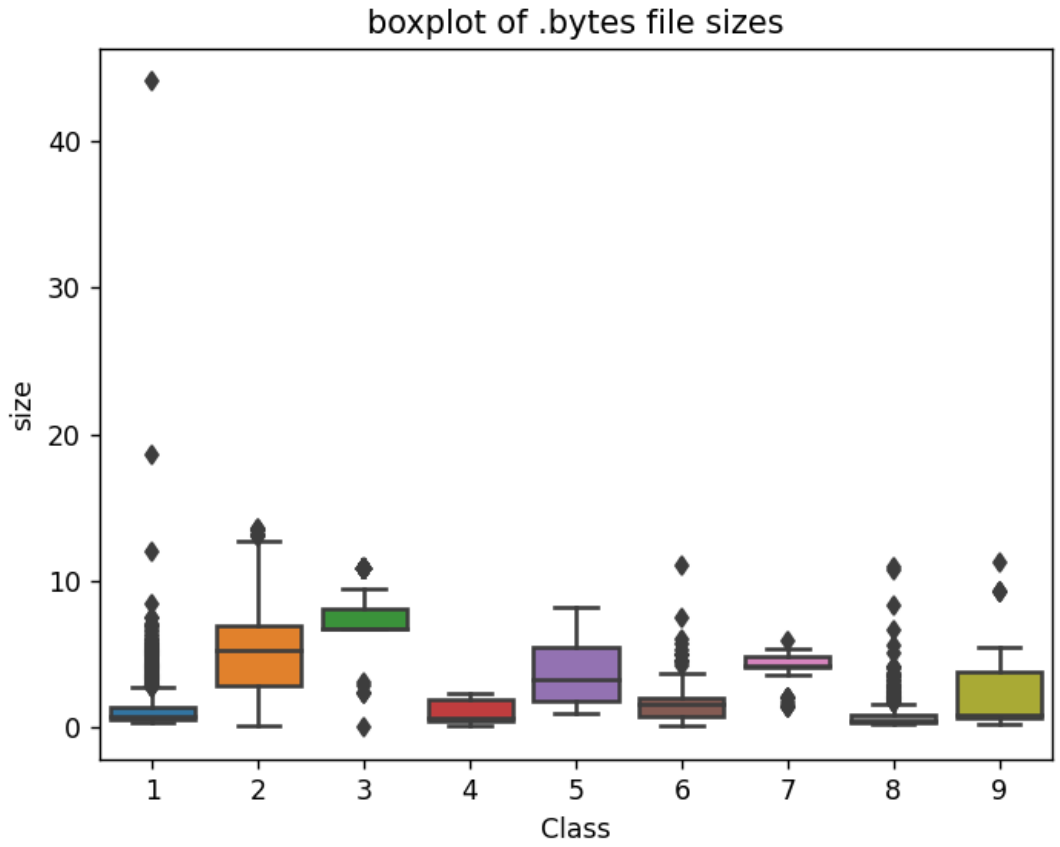
```
fnames=[]
for file in files:
    # print(os.stat('byteFiles/0A32eTdBKayjCWhZqDOQ.txt'))
    # os.stat_result(st_mode=33206, st_ino=1125899906874507, st_dev=3561571700, st_nlin
    # st_size=3680109, st_atime=1519638522, st_mtime=1519638522, st_ctime=1519638522)
    # read more about os.stat: here https://www.tutorialspoint.com/python/os_stat.htm
    statinfo=os.stat('byteFiles/'+file)
    # split the file name at '.' and take the first part of it i.e the file name
    file=file.split('.')[0]
    if any(file == filename for filename in filenames):
        i=filenames.index(file)
        class_bytes.append(class_y[i])
        # converting into Mb's
        sizebytes.append(statinfo.st_size/(1024.0*1024.0))
        fnames.append(file)
data_size_byte=pd.DataFrame({'ID':fnames,'size':sizebytes,'Class':class_bytes})
print (data_size_byte.head())
```

```
                     ID      size  Class
0  e5wb09ugRASHGFl3Zj7y  7.148438      2
1  6XBvjdoIDrAhG5fmqMFZ  6.703125      3
2  hmjcZH7rQ4qkLlbDzePy  6.046875      2
3  cNIGDKm5dUbMZjvnQX8R  6.714844      3
4  fgD5k647jbdSEWBZCQnh  6.714844      3
```

3.2.2 box plots of file size (.byte files) feature

In [4]:

```
#boxplot of byte files
ax = sns.boxplot(x="Class", y="size", data=data_size_byte)
plt.title("boxplot of .bytes file sizes")
plt.show()
```



3.2.3 feature extraction from byte files

In [ ]:

```
#removal of addres from byte files
# contents of .byte files
# ----------------
#00401000 56 8D 44 24 08 50 8B F1 E8 1C 1B 00 00 C7 06 08
#------------------
#we remove the starting address 00401000

files = os.listdir('byteFiles')
filenames=[]
array=[]
for file in files:
    if(file.endswith("bytes")):
        file=file.split('.')[0]
        text_file = open('byteFiles/'+file+".txt", 'w+')
        with open('byteFiles/'+file+".bytes","r") as fp:
            lines=""
            for line in fp:
                a=line.rstrip().split(" ")[1:]
                b=' '.join(a)
                b=b+"\n"
                text_file.write(b)
            fp.close()
            os.remove('byteFiles/'+file+".bytes")
        text_file.close()



files = os.listdir('byteFiles')
```

```
filenames2=[]
feature_matrix = np.zeros((len(files),257),dtype=int)
k=0


#program to convert into bag of words of bytefiles
#this is custom-built bag of words this is unigram bag of words
byte_feature_file=open('result.csv','w+')
byte_feature_file.write("ID,0,1,2,3,4,5,6,7,8,9,0a,0b,0c,0d,0e,0f,10,11,12,13,14,15,16,
byte_feature_file.write("\n")
for file in files:
    filenames2.append(file)
    byte_feature_file.write(file+",")
    if(file.endswith("txt")):
        with open('byteFiles/'+file,"r") as byte_flie:
            for lines in byte_flie:
                line=lines.rstrip().split(" ")
                for hex_code in line:
                    if hex_code=='??':
                        feature_matrix[k][256]+=1
                    else:
                        feature_matrix[k][int(hex_code,16)]+=1
        byte_flie.close()

    for i, row in enumerate(feature_matrix[k]):
        if i!=len(feature_matrix[k])-1:
            byte_feature_file.write(str(row)+",")
        else:
            byte_feature_file.write(str(row))
    byte_feature_file.write("\n")

    k += 1

byte_feature_file.close()
```

In [ ]:

```
#Dont run, result saved in npz

#bigram implementation

#list of byte Files
file_lst = os.listdir('byteFiles')

unigram_string = "0,1,2,3,4,5,6,7,8,9,0a,0b,0c,0d,0e,0f,10,11,12,13,14,15,16,17,18,19,1
uni_lst = unigram_string.split(",")

bigram_lst = []
#creating bigrams

for i in range(len(uni_lst)):
    for j in range(len(uni_lst)):
        bigram_lst.append(str(uni_lst[i])+" "+str(uni_lst[j]))

feature_matrix_bi = np.zeros((len(file_lst),len(bigram_lst)),dtype=int) #np array of ze

bow_bigram = CountVectorizer(ngram_range=(2,2), max_features = 500,vocabulary = bigram_

for i, j in tqdm(enumerate(file_lst)):

    byteFile_ite = open('byteFiles/' + j)
```

```
    hex_codes = [byteFile_ite.read().replace('\n', ' ').lower()]  #converting to lower

    feature_matrix_bi[i] +=csr_matrix(bow_bigram.fit_transform(hex_codes)) #fitting big

    byteFile_ite.close()
```

In [115…

```
feature_matrix_bi = scipy.sparse.load_npz('feature_matrix_bi.npz')

from sklearn.preprocessing import normalize
feature_matrix_bi = normalize(feature_matrix_bi, axis = 0)

feature_matrix_bi_sparse = pd.DataFrame(feature_matrix_bi.toarray())

file_lst = os.listdir('byteFiles')

unigram_string = "0,1,2,3,4,5,6,7,8,9,0a,0b,0c,0d,0e,0f,10,11,12,13,14,15,16,17,18,19,1
uni_lst = unigram_string.split(",")

bigram_lst = []
#creating bigrams

for i in range(len(uni_lst)):
    for j in range(len(uni_lst)):
        bigram_lst.append(str(uni_lst[i])+" "+str(uni_lst[j]))

#Adding column names
feature_matrix_bi_sparse.columns = bigram_lst



print(feature_matrix_bi_sparse.head(10))
```

```
   0 0  0 1  0 2  0 3  0 4  0 5  0 6  0 7  0 8  0 9  ...  ?? f7  ?? f8  ?? f9  \
0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
1  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
2  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
3  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
4  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
5  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
6  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
7  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
8  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
9  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0

   ?? fa  ?? fb  ?? fc  ?? fd  ?? fe  ?? ff  ?? ??
0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
1    0.0    0.0    0.0    0.0    0.0    0.0    0.0
2    0.0    0.0    0.0    0.0    0.0    0.0    0.0
3    0.0    0.0    0.0    0.0    0.0    0.0    0.0
4    0.0    0.0    0.0    0.0    0.0    0.0    0.0
5    0.0    0.0    0.0    0.0    0.0    0.0    0.0
6    0.0    0.0    0.0    0.0    0.0    0.0    0.0
7    0.0    0.0    0.0    0.0    0.0    0.0    0.0
8    0.0    0.0    0.0    0.0    0.0    0.0    0.0
9    0.0    0.0    0.0    0.0    0.0    0.0    0.0

[10 rows x 66049 columns]
```

In [80]:

```
#as per video using random forest for feature selection
```

```
data_y = Y['Class']
clf = RandomForestClassifier(n_jobs = -1)
clf.fit(feature_matrix_bi_sparse,data_y)
```

Out[80]:
```
RandomForestClassifier(n_jobs=-1)
```

In [116…
```
imp_bigrams = clf.feature_importances_
print(imp_bigrams)
```

```
[0. 0. 0. ... 0. 0. 0.]
```

In [117…
```
idx_bigram = np.argsort(imp_bigrams)[::-1]
print(idx_bigram)
print(imp_bigrams[idx_bigram[:500]])
print(feature_matrix_bi_sparse.columns[idx_bigram[500:]])
```

```
[65790 65019 65276 ... 26107 26131     0]
[2.30214419e-04 1.63968451e-04 1.44715034e-04 1.43177744e-04
 1.28347691e-04 1.24066562e-04 1.22437890e-04 1.15475012e-04
 1.13457446e-04 1.12934515e-04 1.08212578e-04 1.05231552e-04
 1.04086252e-04 1.03516171e-04 1.02611588e-04 1.02490254e-04
 9.96515431e-05 9.91227966e-05 9.84305292e-05 9.82711588e-05
 9.80425856e-05 9.78431577e-05 9.73874593e-05 9.63194747e-05
 9.55216255e-05 9.53571194e-05 9.30742304e-05 9.28777894e-05
 9.27622675e-05 9.15898757e-05 9.15649228e-05 9.12604836e-05
 9.10913295e-05 9.09654721e-05 9.07549934e-05 9.07135300e-05
 9.04673045e-05 9.02883945e-05 9.01226842e-05 9.00270068e-05
 8.99378626e-05 8.94668541e-05 8.88022751e-05 8.86268243e-05
 8.84266822e-05 8.81414183e-05 8.81119914e-05 8.71657376e-05
 8.70180229e-05 8.65588980e-05 8.57418812e-05 8.57336305e-05
 8.56906190e-05 8.50428004e-05 8.48858299e-05 8.42595614e-05
 8.40558347e-05 8.38959034e-05 8.35518249e-05 8.33301507e-05
 8.32267251e-05 8.32098074e-05 8.27177606e-05 8.24601758e-05
 8.20229767e-05 8.18327588e-05 8.17927071e-05 8.16378982e-05
 8.15910206e-05 8.11802293e-05 8.08355059e-05 8.07456116e-05
 8.06127762e-05 8.04588063e-05 8.02304187e-05 8.01452672e-05
 8.00202574e-05 7.97728714e-05 7.97490367e-05 7.93384215e-05
 7.92452262e-05 7.91745459e-05 7.91679783e-05 7.91371379e-05
 7.90977034e-05 7.90605452e-05 7.87606837e-05 7.87430819e-05
 7.87172535e-05 7.85396613e-05 7.84956195e-05 7.84277429e-05
 7.82633689e-05 7.74431210e-05 7.72862035e-05 7.68043433e-05
 7.66533708e-05 7.63843533e-05 7.60282608e-05 7.58482690e-05
 7.56479586e-05 7.55374651e-05 7.54009363e-05 7.50263737e-05
 7.48992507e-05 7.45773233e-05 7.45258004e-05 7.43419804e-05
 7.42471765e-05 7.41172786e-05 7.41063541e-05 7.40591127e-05
 7.40110388e-05 7.38954646e-05 7.36598161e-05 7.36513067e-05
 7.35957599e-05 7.34526050e-05 7.33552990e-05 7.32568213e-05
 7.31971038e-05 7.30513872e-05 7.28773531e-05 7.28204750e-05
 7.27748343e-05 7.27094133e-05 7.26210944e-05 7.25970709e-05
 7.25430978e-05 7.24276669e-05 7.23987253e-05 7.23309526e-05
 7.20489599e-05 7.19766409e-05 7.13823954e-05 7.10962537e-05
 7.10875521e-05 7.10706961e-05 7.10670644e-05 7.09453992e-05
 7.09173605e-05 7.08533517e-05 7.08124203e-05 7.07611338e-05
 7.06207152e-05 7.05851768e-05 7.04962621e-05 7.04688744e-05
 7.03225441e-05 7.03180946e-05 7.02305360e-05 7.01599886e-05
 7.01208883e-05 7.00186925e-05 7.00109598e-05 6.99476964e-05
 6.99409880e-05 6.99182100e-05 6.98506169e-05 6.97246020e-05
 6.97039956e-05 6.96349537e-05 6.94800344e-05 6.93582160e-05
 6.93489406e-05 6.92527088e-05 6.92115734e-05 6.91577240e-05
 6.91149836e-05 6.90676939e-05 6.88491279e-05 6.87218009e-05
 6.87180185e-05 6.85563758e-05 6.84464671e-05 6.84424593e-05
```

```
6.84099467e-05 6.84079753e-05 6.82332179e-05 6.81745890e-05
6.80847597e-05 6.79994768e-05 6.78692054e-05 6.78682353e-05
6.78529163e-05 6.77785411e-05 6.76518806e-05 6.75595016e-05
6.74942832e-05 6.74868210e-05 6.74437766e-05 6.74399907e-05
6.74169272e-05 6.73871753e-05 6.73605291e-05 6.73573630e-05
6.73357055e-05 6.72361594e-05 6.71124591e-05 6.70109237e-05
6.69114223e-05 6.67280856e-05 6.65620442e-05 6.64801822e-05
6.64214412e-05 6.64096608e-05 6.63950653e-05 6.63362842e-05
6.63237930e-05 6.62784791e-05 6.62716587e-05 6.62382882e-05
6.61557371e-05 6.61208063e-05 6.60478349e-05 6.59843198e-05
6.59704400e-05 6.59144900e-05 6.58998012e-05 6.57323597e-05
6.56893201e-05 6.56060320e-05 6.55512347e-05 6.55247605e-05
6.55236703e-05 6.55196333e-05 6.54884997e-05 6.54268756e-05
6.53791603e-05 6.53790952e-05 6.52736821e-05 6.51217289e-05
6.49584982e-05 6.49449191e-05 6.49218375e-05 6.49062707e-05
6.46073293e-05 6.46025461e-05 6.44639004e-05 6.44279088e-05
6.44140137e-05 6.44004730e-05 6.43342362e-05 6.42041543e-05
6.41122181e-05 6.41078936e-05 6.40676159e-05 6.40520034e-05
6.40211130e-05 6.39730901e-05 6.39375617e-05 6.39120651e-05
6.38923103e-05 6.38030151e-05 6.37738125e-05 6.37640698e-05
6.37566376e-05 6.36847932e-05 6.36216873e-05 6.35797033e-05
6.35396370e-05 6.35114940e-05 6.34942547e-05 6.34889262e-05
6.34169081e-05 6.34120343e-05 6.33742900e-05 6.33238340e-05
6.33197859e-05 6.32602751e-05 6.32579320e-05 6.32416058e-05
6.32258269e-05 6.32244927e-05 6.31413401e-05 6.31149110e-05
6.30844407e-05 6.30778540e-05 6.30406053e-05 6.29559568e-05
6.28295120e-05 6.27514642e-05 6.25424679e-05 6.25387689e-05
6.25306442e-05 6.25299994e-05 6.25208206e-05 6.25029308e-05
6.24690628e-05 6.24660849e-05 6.24633934e-05 6.23320925e-05
6.23261175e-05 6.23057664e-05 6.22961383e-05 6.22816331e-05
6.22807510e-05 6.22794210e-05 6.22711772e-05 6.22325884e-05
6.21673184e-05 6.21566009e-05 6.21497522e-05 6.20876902e-05
6.19492707e-05 6.19351228e-05 6.19175440e-05 6.19145723e-05
6.18680051e-05 6.18616240e-05 6.18107575e-05 6.18004641e-05
6.17876548e-05 6.17544279e-05 6.17279519e-05 6.17276953e-05
6.16999773e-05 6.16943069e-05 6.14708047e-05 6.14549287e-05
6.14341146e-05 6.13661802e-05 6.13422104e-05 6.13055689e-05
6.12704609e-05 6.11870500e-05 6.11682545e-05 6.11584790e-05
6.11506043e-05 6.11038982e-05 6.10872188e-05 6.10650209e-05
6.10126900e-05 6.09814462e-05 6.09723151e-05 6.09676677e-05
6.08836859e-05 6.07663670e-05 6.07352660e-05 6.07346987e-05
6.05868473e-05 6.05690681e-05 6.05342714e-05 6.04941794e-05
6.03871041e-05 6.03746373e-05 6.03740154e-05 6.03288561e-05
6.02752181e-05 6.02751186e-05 6.01724641e-05 6.01698942e-05
6.01623000e-05 6.00978480e-05 6.00863933e-05 6.00859899e-05
6.00430110e-05 6.00288484e-05 6.00215050e-05 5.99984254e-05
5.99915046e-05 5.99092170e-05 5.99002077e-05 5.98995670e-05
5.98264477e-05 5.98211320e-05 5.97444769e-05 5.97362607e-05
5.97069873e-05 5.96464616e-05 5.95845038e-05 5.95828460e-05
5.95541862e-05 5.95506452e-05 5.95216339e-05 5.94727587e-05
5.94121869e-05 5.93516668e-05 5.93412519e-05 5.93195491e-05
5.92849727e-05 5.92657657e-05 5.92491099e-05 5.92146160e-05
5.91841138e-05 5.91372806e-05 5.90971579e-05 5.90605530e-05
5.90034401e-05 5.89572194e-05 5.89459037e-05 5.89035959e-05
5.88452308e-05 5.88304962e-05 5.87679207e-05 5.87541671e-05
5.87294906e-05 5.87210944e-05 5.86837966e-05 5.85776498e-05
5.85082327e-05 5.84909698e-05 5.84774322e-05 5.84561887e-05
5.84416546e-05 5.84210016e-05 5.84137881e-05 5.83876787e-05
5.83443304e-05 5.83382400e-05 5.82846392e-05 5.81992365e-05
5.81926421e-05 5.81891606e-05 5.81715934e-05 5.81714618e-05
5.81700741e-05 5.81502441e-05 5.81282429e-05 5.81247266e-05
5.80936725e-05 5.80704023e-05 5.80284525e-05 5.80069231e-05
5.79379547e-05 5.79362416e-05 5.79171219e-05 5.79113134e-05
5.78529878e-05 5.78326299e-05 5.77734377e-05 5.77015321e-05
5.76976553e-05 5.76848037e-05 5.76771589e-05 5.76577626e-05
```

```
 5.76537368e-05 5.76139450e-05 5.76112082e-05 5.75836192e-05
 5.75139964e-05 5.74535692e-05 5.73965696e-05 5.73925457e-05
 5.73307057e-05 5.72712078e-05 5.72572939e-05 5.72518090e-05
 5.72502129e-05 5.72482350e-05 5.72374808e-05 5.72247575e-05
 5.72169600e-05 5.72149016e-05 5.70612010e-05 5.70443010e-05
 5.70442760e-05 5.70292809e-05 5.70258986e-05 5.70249403e-05
 5.70247318e-05 5.69811383e-05 5.69713320e-05 5.69674698e-05
 5.69575342e-05 5.69491080e-05 5.69397982e-05 5.68616317e-05
 5.68459412e-05 5.68309385e-05 5.68260129e-05 5.68026392e-05
 5.67992540e-05 5.67863427e-05 5.67801716e-05 5.67686679e-05
 5.67456591e-05 5.66940323e-05 5.66935262e-05 5.66922278e-05
 5.66672941e-05 5.66147821e-05 5.66137949e-05 5.66092118e-05
 5.65642421e-05 5.65350388e-05 5.65138713e-05 5.65127682e-05
 5.64969264e-05 5.64934216e-05 5.64842029e-05 5.64756874e-05
 5.64168417e-05 5.63892576e-05 5.63826517e-05 5.63437764e-05
 5.63218490e-05 5.63154285e-05 5.62730609e-05 5.62437403e-05]
Index(['e8 94', '8d be', 'aa aa', 'b8 89', 'ef ef', '4c 30', '13 ff', 'e8 8d',
       'a1 1c', 'e4 ff',
       ...
       '65 4', '65 5', '65 6', '65 7', '65 8', '65 9', '65 14', '65 96',
       '65 ae', '0 0'],
      dtype='object', length=65549)
```

In [118…

```
feature_matrix_bi_sparse.drop(feature_matrix_bi_sparse.columns[idx_bigram[500:]],axis =

f_name = []
for i in os.listdir('byteFiles'):
    f_name.append(i.replace('.txt',''))

#Adding file names
feature_matrix_bi_sparse["ID"] = f_name

feature_matrix_bi_sparse.shape
feature_matrix_bi_sparse.head(10)
```

Out[118…

	0c 10	10 20	10 59	10 c3	10 ff	14 30	14 ff	15 10	15 18	15
0	0.008595	0.003151	0.003519	0.005772	0.009100	0.000113	0.005553	0.004281	0.005773	0.000
1	0.000508	0.000540	0.000732	0.000366	0.000882	0.000367	0.001058	0.000507	0.000782	0.001
2	0.000000	0.000030	0.000000	0.000000	0.003915	0.000028	0.003808	0.000159	0.000060	0.000
3	0.000271	0.000210	0.000523	0.000610	0.000670	0.000451	0.000899	0.000507	0.000842	0.001
4	0.000440	0.000300	0.000523	0.000650	0.000705	0.000339	0.001375	0.000603	0.001143	0.002
5	0.000034	0.000030	0.000000	0.000000	0.000000	0.000028	0.000106	0.002156	0.002165	0.005
6	0.000135	0.000060	0.000070	0.000081	0.000247	0.000028	0.000053	0.000095	0.000241	0.000
7	0.003519	0.000030	0.007908	0.005935	0.005220	0.000000	0.000000	0.000095	0.000060	0.000
8	0.000034	0.000090	0.000035	0.000000	0.001799	0.000028	0.001851	0.000222	0.000481	0.000
9	0.000000	0.048856	0.000000	0.000000	0.000317	0.000000	0.000106	0.000000	0.000000	0.000

10 rows × 501 columns

In [119…

```
byte_features=pd.read_csv("result.csv")
```

```
byte_features['ID']  = byte_features['ID'].str.split('.').str[0]
byte_features.head(2)

data_size_byte.head(2)

byte_features_with_size = byte_features.merge(data_size_byte, on='ID')

byte_features_with_size.to_csv("result_with_size.csv")
byte_features_with_size.head(2)
```

Out[119…

	ID	0	1	2	3	4	5	6	7	8	...	f9	f
0	01azqd4InC7m9JpocGv5	601905	3905	2816	3832	3345	3242	3650	3201	2965	...	3101	321
1	01IsoiSMh5gxyDYTl4CB	39755	8337	7249	7186	8663	6844	8420	7589	9291	...	439	28

2 rows × 260 columns

In [120…

```
# https://stackoverflow.com/a/29651514
def normalize(df):
    result1 = df.copy()
    for feature_name in tqdm(df.columns):
        if (str(feature_name) != str('ID') and str(feature_name)!=str('Class')):
            max_value = df[feature_name].max()
            min_value = df[feature_name].min()
            result1[feature_name] = (df[feature_name] - min_value) / (max_value - min_v
    return result1
result = normalize(byte_features_with_size)
```

```
100%|██████████| 260/260 [00:00<00:00, 571.49it/s]
```

In [121…

```
result = result.merge(feature_matrix_bi_sparse, on='ID')
```

In [122…

```
result.head(2)
```

Out[122…

	ID	0	1	2	3	4	5	6
0	01azqd4InC7m9JpocGv5	0.262806	0.005498	0.001567	0.002067	0.002048	0.001835	0.002058	0.
1	01IsoiSMh5gxyDYTl4CB	0.017358	0.011737	0.004033	0.003876	0.005303	0.003873	0.004747	0.

2 rows × 760 columns

In [123…

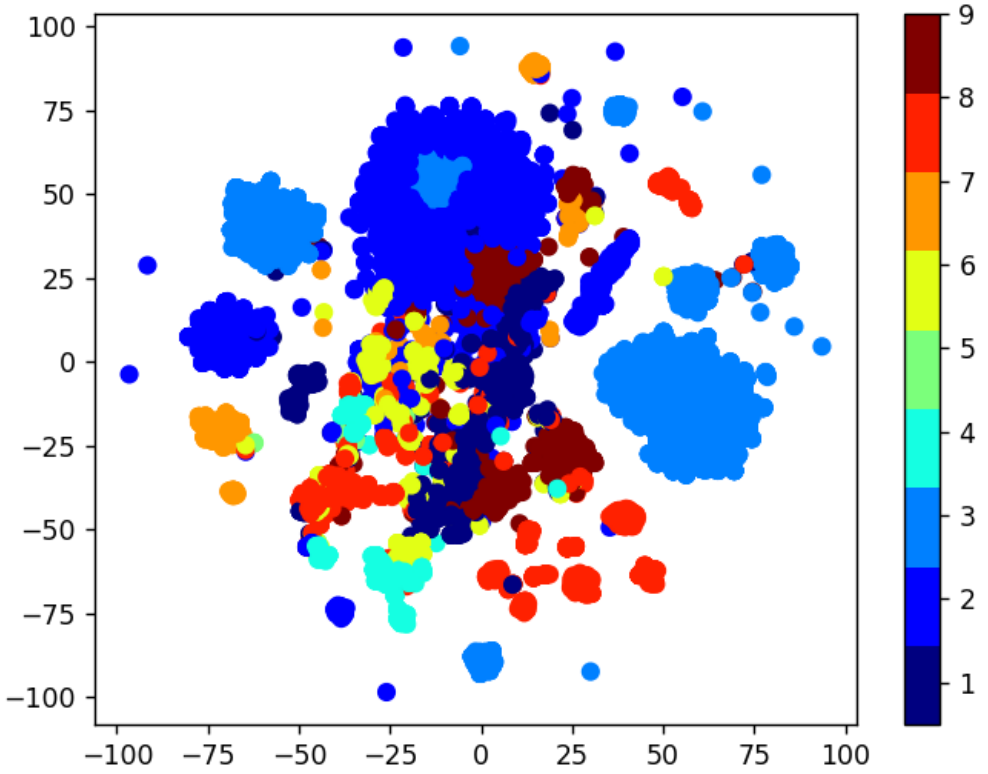
```
data_y = result['Class']
result.head()
print(data_y)
```

```
0        9
1        2
2        9
3        1
4        8
```

```
        ..
10863    4
10864    4
10865    4
10866    4
10867    4
Name: Class, Length: 10868, dtype: int64
```

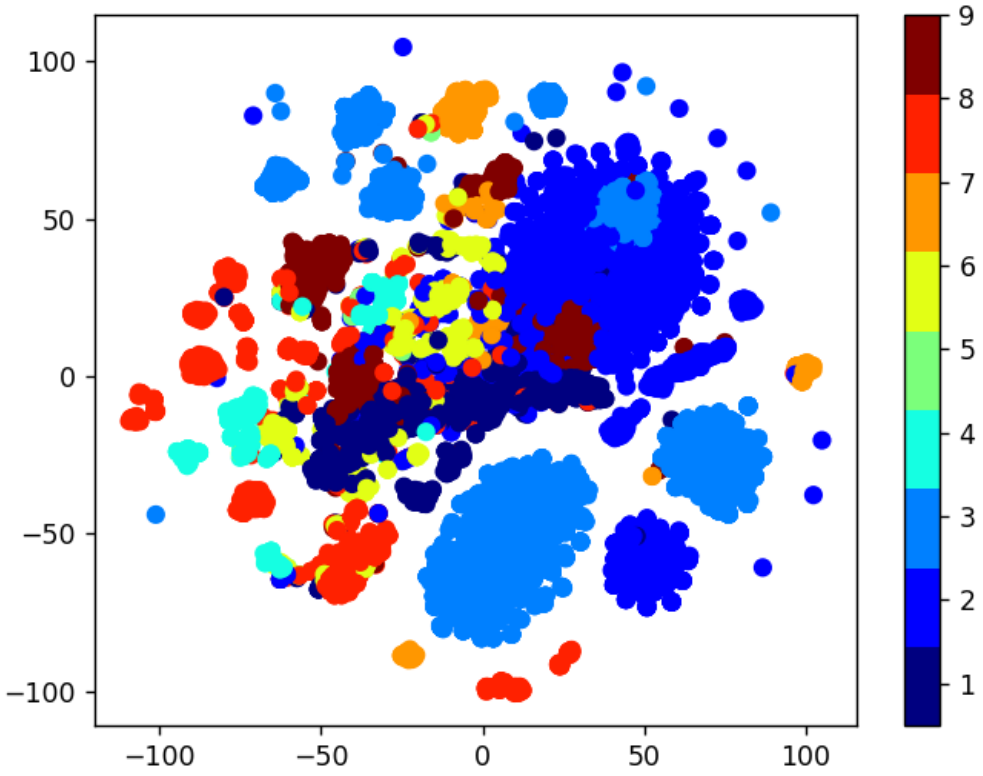
3.2.4 Multivariate Analysis

In [ ]:
```
#multivariate analysis on byte files
#this is with perplexity 50
xtsne=TSNE(perplexity=50)
results=xtsne.fit_transform(result.drop(['ID','Class'], axis=1))
vis_x = results[:, 0]
vis_y = results[:, 1]
plt.scatter(vis_x, vis_y, c=data_y, cmap=plt.cm.get_cmap("jet", 9))
plt.colorbar(ticks=range(10))
plt.clim(0.5, 9)
plt.show()
```



In [ ]:
```
#this is with perplexity 30
xtsne=TSNE(perplexity=30)
results=xtsne.fit_transform(result.drop(['ID','Class'], axis=1))
vis_x = results[:, 0]
vis_y = results[:, 1]
plt.scatter(vis_x, vis_y, c=data_y, cmap=plt.cm.get_cmap("jet", 9))
plt.colorbar(ticks=range(10))
```

```
plt.clim(0.5, 9)
plt.show()
```



Train Test split

In [124…

```
data_y = result['Class']
# split the data into test and train by maintaining same distribution of output varaibl
X_train, X_test, y_train, y_test = train_test_split(result.drop(['ID','Class'], axis=1)
# split the train data into train and cross validation by maintaining same distribution
X_train, X_cv, y_train, y_cv = train_test_split(X_train, y_train,stratify=y_train,test_
```

In [125…

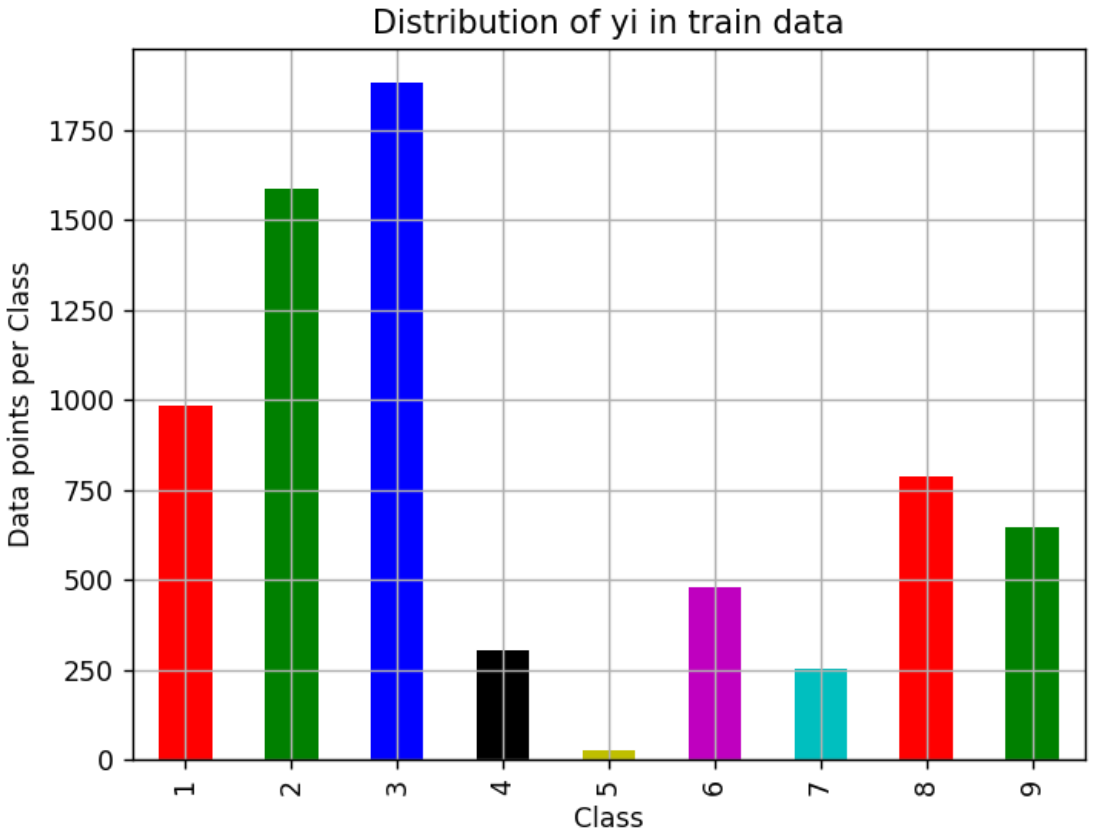
```
print('Number of data points in train data:', X_train.shape)
print('Number of data points in test data:', X_test.shape)
print('Number of data points in cross validation data:', X_cv.shape)
```

```
Number of data points in train data: (6955, 758)
Number of data points in test data: (2174, 758)
Number of data points in cross validation data: (1739, 758)
```

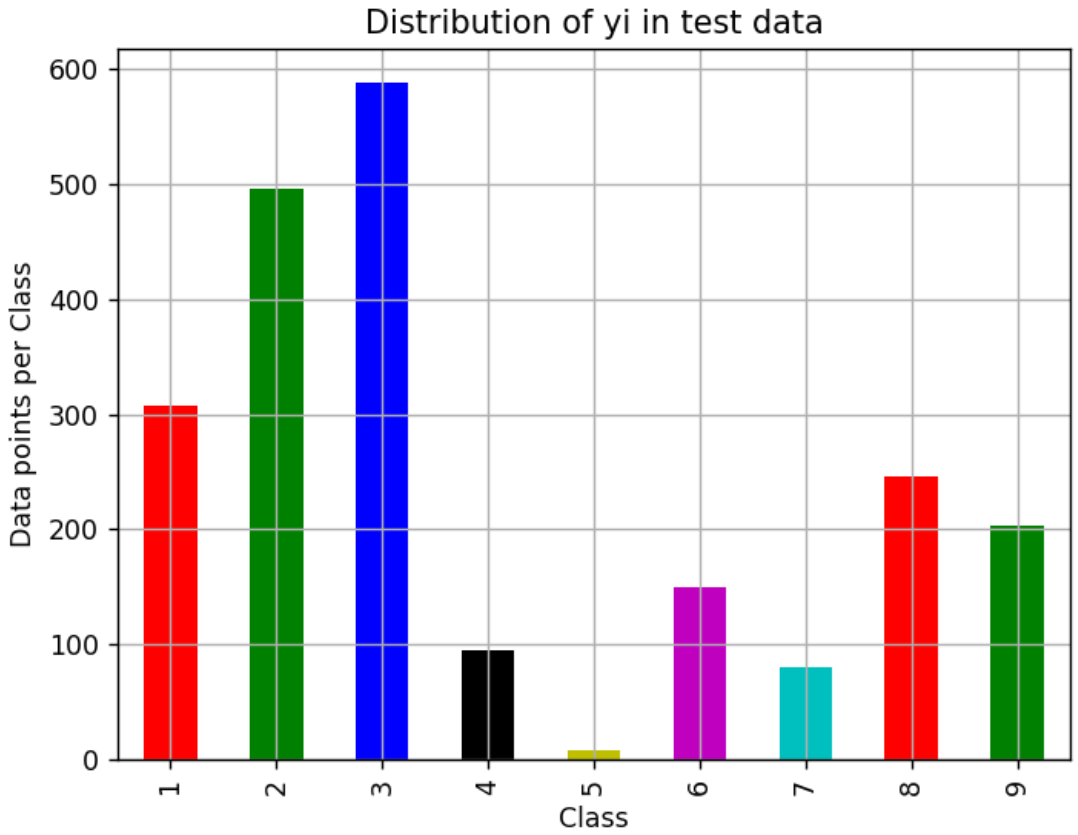
In [ ]:

```
# it returns a dict, keys as class labels and values as the number of data points in th
train_class_distribution = y_train.value_counts().sortlevel()
test_class_distribution = y_test.value_counts().sortlevel()
cv_class_distribution = y_cv.value_counts().sortlevel()

my_colors = 'rgbkymc'
train_class_distribution.plot(kind='bar', color=my_colors)
```

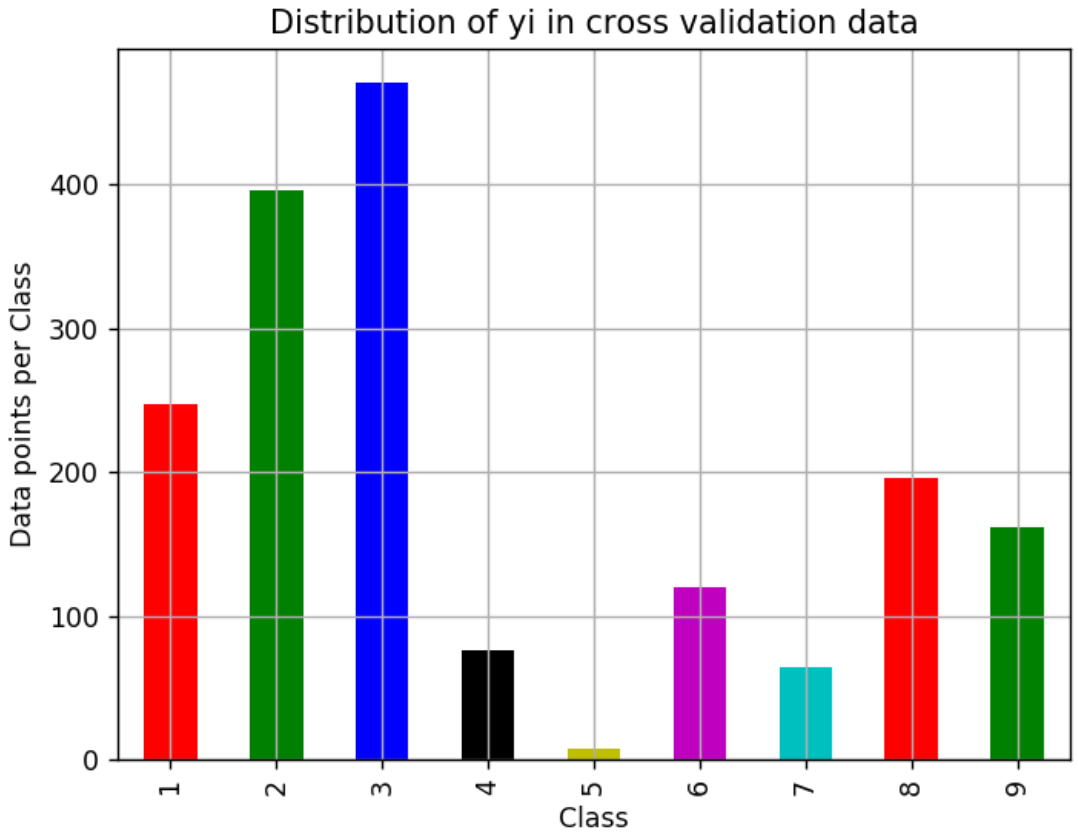
```
plt.xlabel('Class')
plt.ylabel('Data points per Class')
plt.title('Distribution of yi in train data')
plt.grid()
plt.show()

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/numpy.argsort.html
# -(train_class_distribution.values): the minus sign will give us in decreasing order
sorted_yi = np.argsort(-train_class_distribution.values)
for i in sorted_yi:
    print('Number of data points in class', i+1, ':',train_class_distribution.values[i]


print('-'*80)
my_colors = 'rgbkymc'
test_class_distribution.plot(kind='bar', color=my_colors)
plt.xlabel('Class')
plt.ylabel('Data points per Class')
plt.title('Distribution of yi in test data')
plt.grid()
plt.show()

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/numpy.argsort.html
# -(train_class_distribution.values): the minus sign will give us in decreasing order
sorted_yi = np.argsort(-test_class_distribution.values)
for i in sorted_yi:
    print('Number of data points in class', i+1, ':',test_class_distribution.values[i],

print('-'*80)
my_colors = 'rgbkymc'
cv_class_distribution.plot(kind='bar', color=my_colors)
plt.xlabel('Class')
plt.ylabel('Data points per Class')
plt.title('Distribution of yi in cross validation data')
plt.grid()
plt.show()

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/numpy.argsort.html
# -(train_class_distribution.values): the minus sign will give us in decreasing order
sorted_yi = np.argsort(-train_class_distribution.values)
for i in sorted_yi:
    print('Number of data points in class', i+1, ':',cv_class_distribution.values[i], '
```




```
Number of data points in class 3 : 1883 ( 27.074 %)
Number of data points in class 2 : 1586 ( 22.804 %)
Number of data points in class 1 : 986 ( 14.177 %)
Number of data points in class 8 : 786 ( 11.301 %)
Number of data points in class 9 : 648 ( 9.317 %)
Number of data points in class 6 : 481 ( 6.916 %)
Number of data points in class 4 : 304 ( 4.371 %)
Number of data points in class 7 : 254 ( 3.652 %)
Number of data points in class 5 : 27 ( 0.388 %)
--------------------------------------------------------------------------------
```




```
Number of data points in class 3 : 588 ( 27.047 %)
Number of data points in class 2 : 496 ( 22.815 %)
Number of data points in class 1 : 308 ( 14.167 %)
Number of data points in class 8 : 246 ( 11.316 %)
Number of data points in class 9 : 203 ( 9.338 %)
Number of data points in class 6 : 150 ( 6.9 %)
Number of data points in class 4 : 95 ( 4.37 %)
Number of data points in class 7 : 80 ( 3.68 %)
Number of data points in class 5 : 8 ( 0.368 %)
--------------------------------------------------------------------------------
```




```
Number of data points in class 3 : 471 ( 27.085 %)
Number of data points in class 2 : 396 ( 22.772 %)
Number of data points in class 1 : 247 ( 14.204 %)
Number of data points in class 8 : 196 ( 11.271 %)
Number of data points in class 9 : 162 ( 9.316 %)
Number of data points in class 6 : 120 ( 6.901 %)
Number of data points in class 4 : 76 ( 4.37 %)
Number of data points in class 7 : 64 ( 3.68 %)
Number of data points in class 5 : 7 ( 0.403 %)
```

In [ ]:

```
def plot_confusion_matrix(test_y, predict_y):
    C = confusion_matrix(test_y, predict_y)
    print("Number of misclassified points ",(len(test_y)-np.trace(C))/len(test_y)*100)
    # C = 9,9 matrix, each cell (i,j) represents number of points of class i are predic

    A =(((C.T)/(C.sum(axis=1))).T)
    #divid each element of the confusion matrix with the sum of elements in that column

    # C = [[1, 2],
    #     [3, 4]]
    # C.T = [[1, 3],
    #        [2, 4]]
    # C.sum(axis = 1)  axis=0 corresonds to columns and axis=1 corresponds to rows in t
    # C.sum(axix =1) = [[3, 7]]
    # ((C.T)/(C.sum(axis=1))) = [[1/3, 3/7]
    #                            [2/3, 4/7]]

    # ((C.T)/(C.sum(axis=1))).T = [[1/3, 2/3]
    #                              [3/7, 4/7]]
    # sum of row elements = 1

    B =(C/C.sum(axis=0))
```

```
    #divid each element of the confusion matrix with the sum of elements in that row
    # C = [[1, 2],
    #     [3, 4]]
    # C.sum(axis = 0)  axis=0 corresonds to columns and axis=1 corresponds to rows in t
    # C.sum(axix =0) = [[4, 6]]
    # (C/C.sum(axis=0)) = [[1/4, 2/6],
    #                      [3/4, 4/6]]

    labels = [1,2,3,4,5,6,7,8,9]
    cmap=sns.light_palette("green")
    # representing A in heatmap format
    print("-"*50, "Confusion matrix", "-"*50)
    plt.figure(figsize=(10,5))
    sns.heatmap(C, annot=True, cmap=cmap, fmt=".3f", xticklabels=labels, yticklabels=la
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.show()

    print("-"*50, "Precision matrix", "-"*50)
    plt.figure(figsize=(10,5))
    sns.heatmap(B, annot=True, cmap=cmap, fmt=".3f", xticklabels=labels, yticklabels=la
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.show()
    print("Sum of columns in precision matrix",B.sum(axis=0))

    # representing B in heatmap format
    print("-"*50, "Recall matrix"    , "-"*50)
    plt.figure(figsize=(10,5))
    sns.heatmap(A, annot=True, cmap=cmap, fmt=".3f", xticklabels=labels, yticklabels=la
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.show()
    print("Sum of rows in precision matrix",A.sum(axis=1))
```

4. Machine Learning Models

4.1. Machine Leaning Models on bytes files

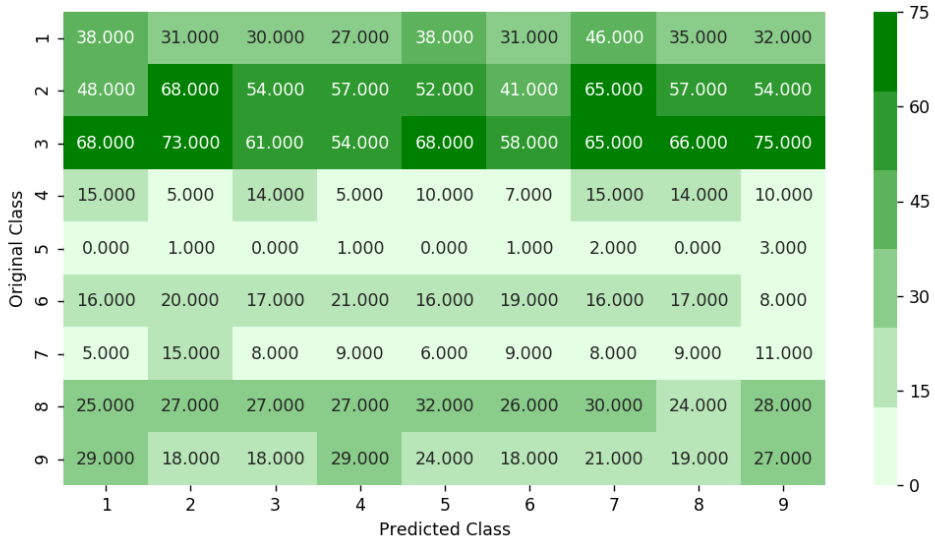
4.1.1. Random Model

In [ ]:

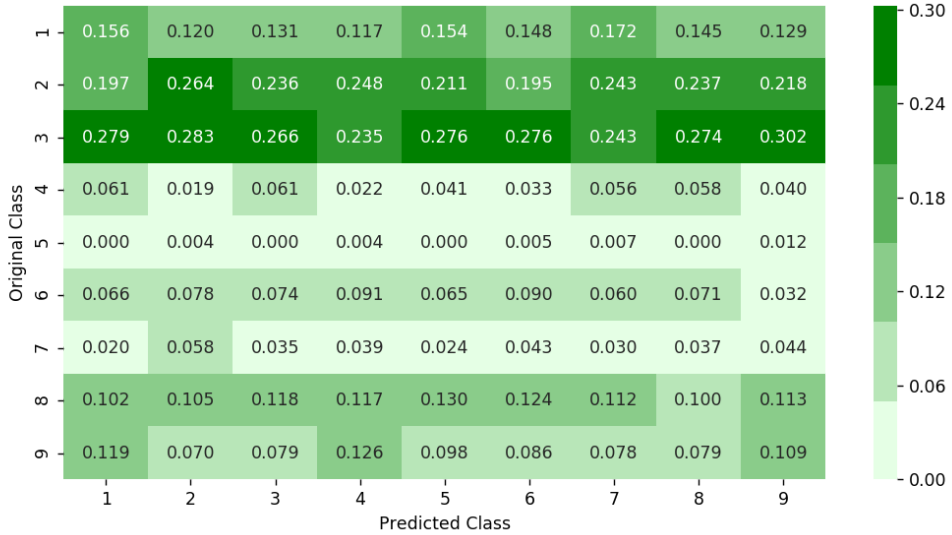
```
# we need to generate 9 numbers and the sum of numbers should be 1
# one solution is to genarate 9 numbers and divide each of the numbers by their sum
# ref: https://stackoverflow.com/a/18662466/4084039

test_data_len = X_test.shape[0]
cv_data_len = X_cv.shape[0]

# we create a output array that has exactly same size as the CV data
cv_predicted_y = np.zeros((cv_data_len,9))
for i in range(cv_data_len):
    rand_probs = np.random.rand(1,9)
    cv_predicted_y[i] = ((rand_probs/sum(sum(rand_probs)))[0])
print("Log loss on Cross Validation Data using Random Model",log_loss(y_cv,cv_predicted
```

```
# Test-Set error.
#we create a output array that has exactly same as the test data
test_predicted_y = np.zeros((test_data_len,9))
for i in range(test_data_len):
    rand_probs = np.random.rand(1,9)
    test_predicted_y[i] = ((rand_probs/sum(sum(rand_probs)))[0])
print("Log loss on Test Data using Random Model",log_loss(y_test,test_predicted_y, eps=

predicted_y =np.argmax(test_predicted_y, axis=1)
plot_confusion_matrix(y_test, predicted_y+1)
```

```
Log loss on Cross Validation Data using Random Model 2.45615644965
Log loss on Test Data using Random Model 2.48503905509
Number of misclassified points  88.5004599816
-------------------------------------------------- Confusion matrix --------------------
------------------------------
```

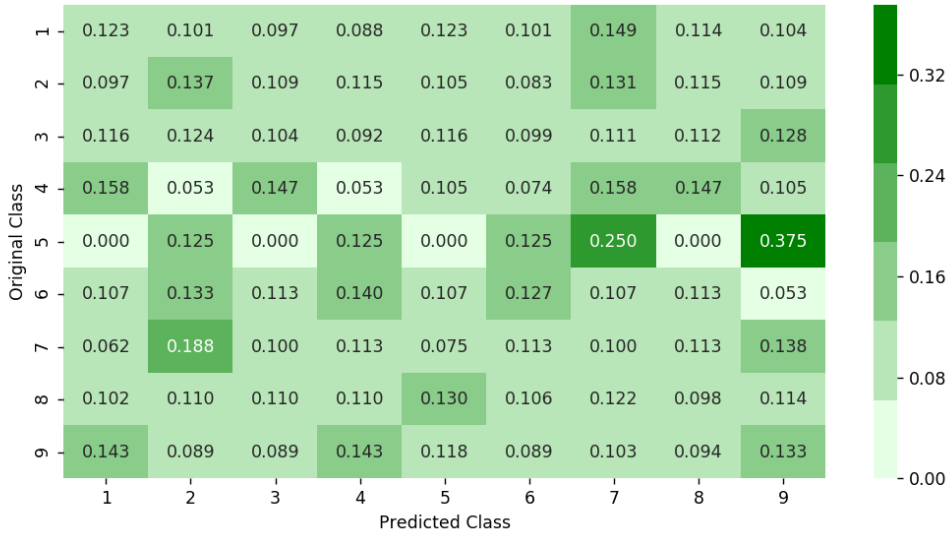



```
-------------------------------------------------- Precision matrix --------------------
------------------------------
```




```
Sum of columns in precision matrix [ 1.  1.  1.  1.  1.  1.  1.  1.  1.]
-------------------------------------------------- Recall matrix ----------------------
```

```
-------------------------
```




```
Sum of rows in precision matrix [ 1.  1.  1.  1.  1.  1.  1.  1.  1.]
```

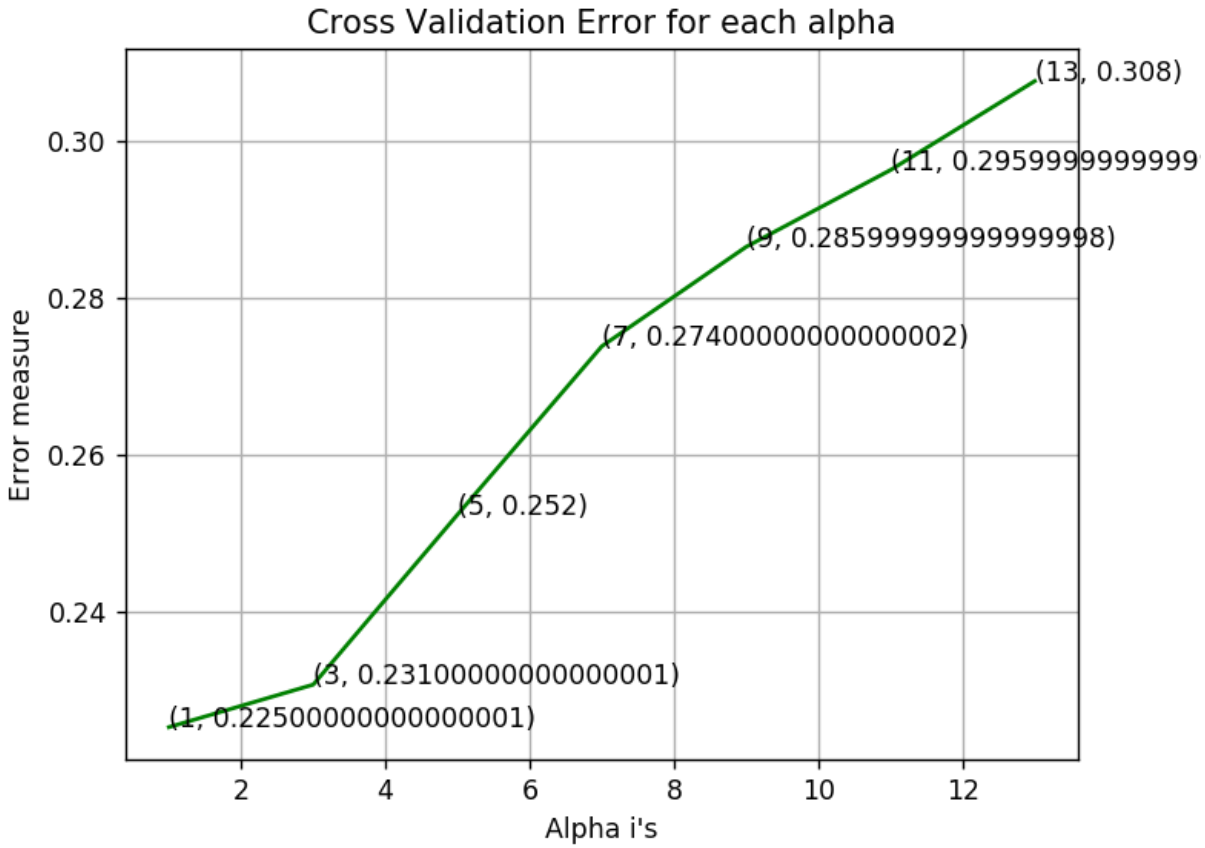
4.1.2. K Nearest Neighbour Classification

In [ ]:

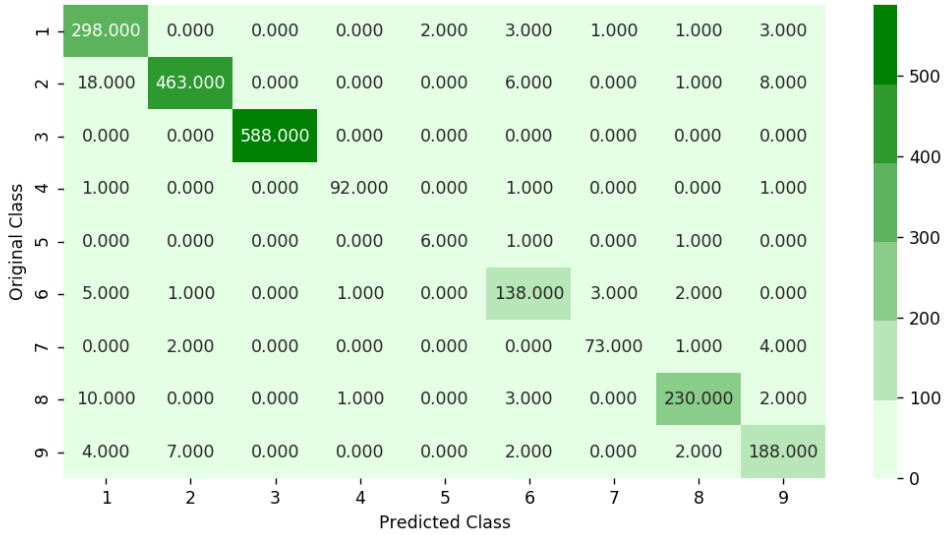
```
# find more about KNeighborsClassifier() here http://scikit-learn.org/stable/modules/ge
# ------------------------
# default parameter
# KNeighborsClassifier(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30
# metric='minkowski', metric_params=None, n_jobs=1, **kwargs)

# methods of
# fit(X, y) : Fit the model using X as training data and y as target values
# predict(X):Predict the class labels for the provided data
# predict_proba(X):Return probability estimates for the test data X.
#-----------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/k
#-----------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight])    Fit the calibrated model
# get_params([deep])    Get parameters for this estimator.
# predict(X)    Predict the target of new samples.
# predict_proba(X)      Posterior probabilities of classification
#-----------------------------------
# video link:
#-----------------------------------

alpha = [x for x in range(1, 15, 2)]
cv_log_error_array=[]
for i in alpha:
    k_cfl=KNeighborsClassifier(n_neighbors=i)
    k_cfl.fit(X_train,y_train)
```

```
    sig_clf = CalibratedClassifierCV(k_cfl, method="sigmoid")
    sig_clf.fit(X_train, y_train)
    predict_y = sig_clf.predict_proba(X_cv)
    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=k_cfl.classes_, eps=1e-1

for i in range(len(cv_log_error_array)):
    print ('log_loss for k = ',alpha[i],'is',cv_log_error_array[i])

best_alpha = np.argmin(cv_log_error_array)

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()

k_cfl=KNeighborsClassifier(n_neighbors=alpha[best_alpha])
k_cfl.fit(X_train,y_train)
sig_clf = CalibratedClassifierCV(k_cfl, method="sigmoid")
sig_clf.fit(X_train, y_train)

predict_y = sig_clf.predict_proba(X_train)
print ('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_l
predict_y = sig_clf.predict_proba(X_cv)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss
predict_y = sig_clf.predict_proba(X_test)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_los
plot_confusion_matrix(y_test, sig_clf.predict(X_test))
```

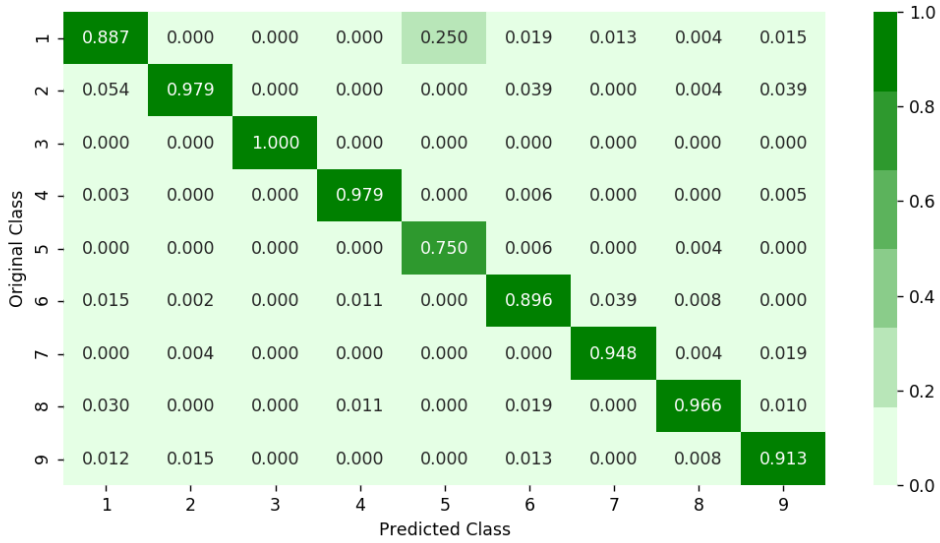
```
log_loss for k =  1 is 0.225386237304
log_loss for k =  3 is 0.230795229168
log_loss for k =  5 is 0.252421408646
log_loss for k =  7 is 0.273827486888
log_loss for k =  9 is 0.286469181555
log_loss for k =  11 is 0.29623391147
log_loss for k =  13 is 0.307551203154
```




```
For values of best alpha =  1 The train log loss is: 0.0782947669247
For values of best alpha =  1 The cross validation log loss is: 0.225386237304
For values of best alpha =  1 The test log loss is: 0.241508604195
Number of misclassified points  4.50781968721
---------------------------------------------------- Confusion matrix --------------------
------------------------------
```

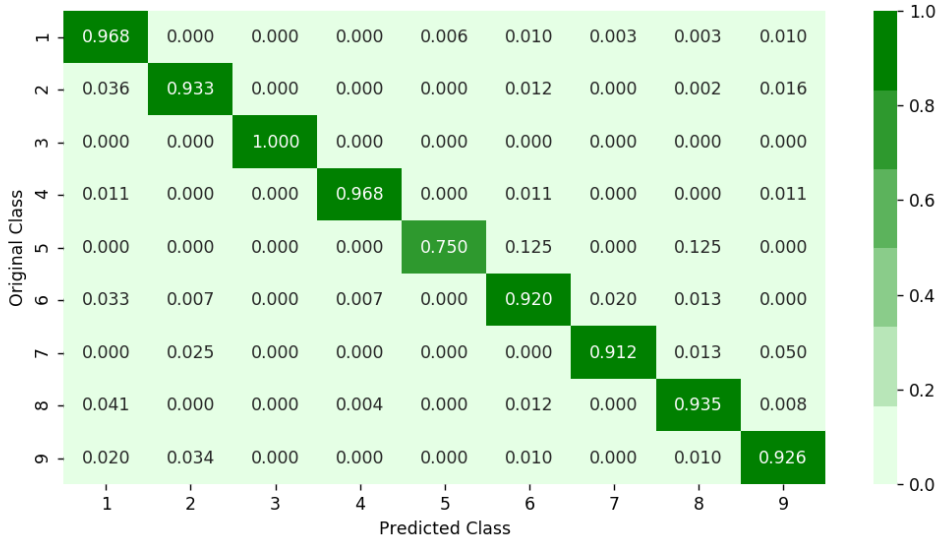



```
---------------------------------------------------- Precision matrix --------------------
------------------------------
```



Sum of columns in precision matrix [1. 1. 1. 1. 1. 1. 1. 1. 1.]

---------- Recall matrix ----------



Sum of rows in precision matrix [1. 1. 1. 1. 1. 1. 1. 1. 1.]

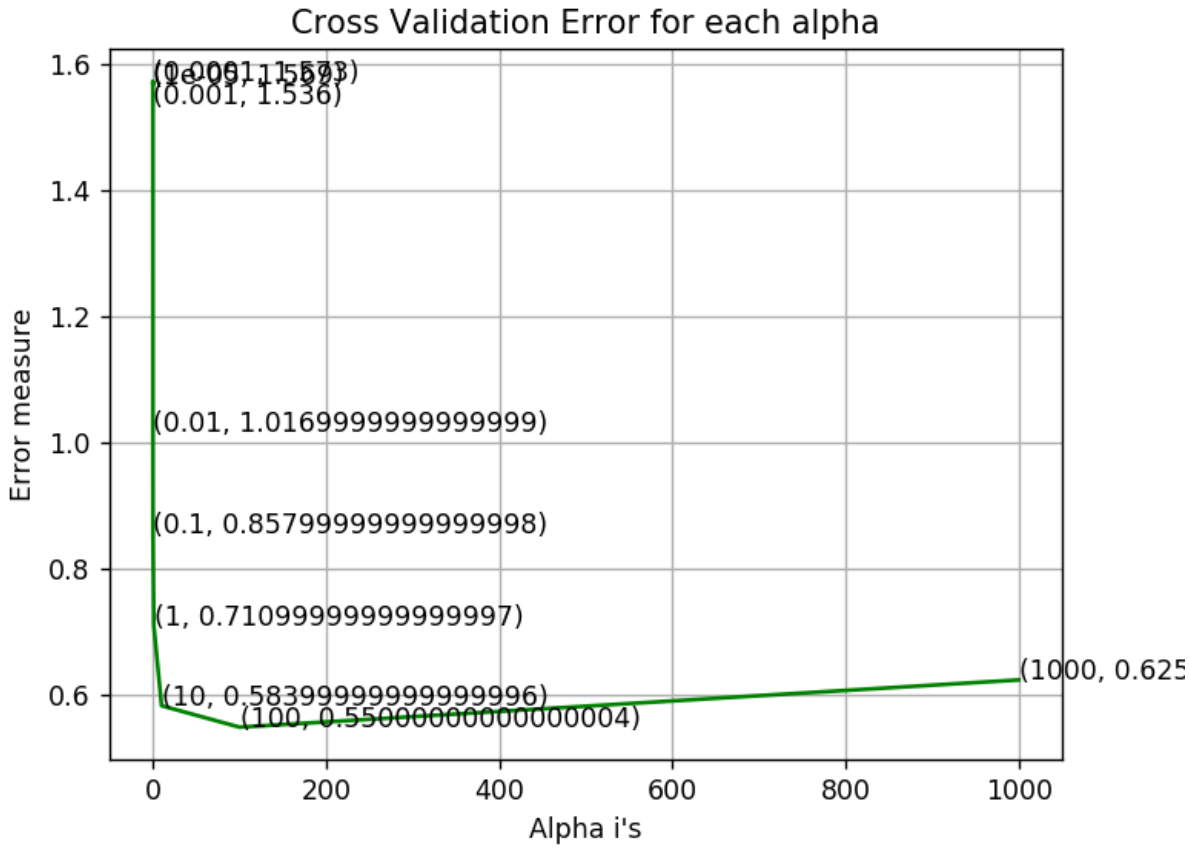
4.1.3. Logistic Regression

In [ ]:

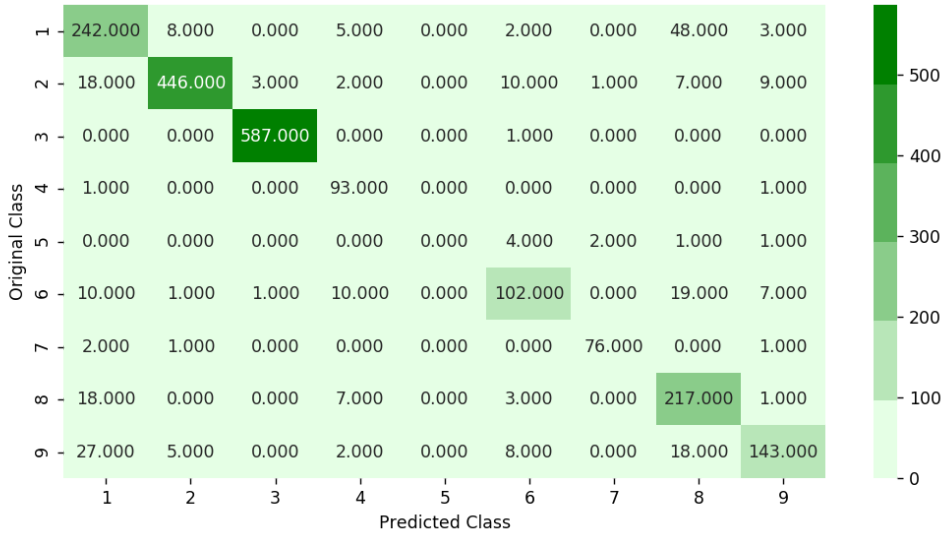
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/s
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='opt
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, …])     Fit linear model with Stochastic Gradie
# predict(X)    Predict class labels for samples in X.

#------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/g
```

```
#----------------------------

alpha = [10 ** x for x in range(-5, 4)]
cv_log_error_array=[]
for i in alpha:
    logisticR=LogisticRegression(penalty='l2',C=i,class_weight='balanced')
    logisticR.fit(X_train,y_train)
    sig_clf = CalibratedClassifierCV(logisticR, method="sigmoid")
    sig_clf.fit(X_train, y_train)
    predict_y = sig_clf.predict_proba(X_cv)
    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=logisticR.classes_, eps=

for i in range(len(cv_log_error_array)):
    print ('log_loss for c = ',alpha[i],'is',cv_log_error_array[i])

best_alpha = np.argmin(cv_log_error_array)

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()

logisticR=LogisticRegression(penalty='l2',C=alpha[best_alpha],class_weight='balanced')
logisticR.fit(X_train,y_train)
sig_clf = CalibratedClassifierCV(logisticR, method="sigmoid")
sig_clf.fit(X_train, y_train)
pred_y=sig_clf.predict(X_test)

predict_y = sig_clf.predict_proba(X_train)
print ('log loss for train data',log_loss(y_train, predict_y, labels=logisticR.classes_
predict_y = sig_clf.predict_proba(X_cv)
print ('log loss for cv data',log_loss(y_cv, predict_y, labels=logisticR.classes_, eps=
predict_y = sig_clf.predict_proba(X_test)
print ('log loss for test data',log_loss(y_test, predict_y, labels=logisticR.classes_,
plot_confusion_matrix(y_test, sig_clf.predict(X_test))
```

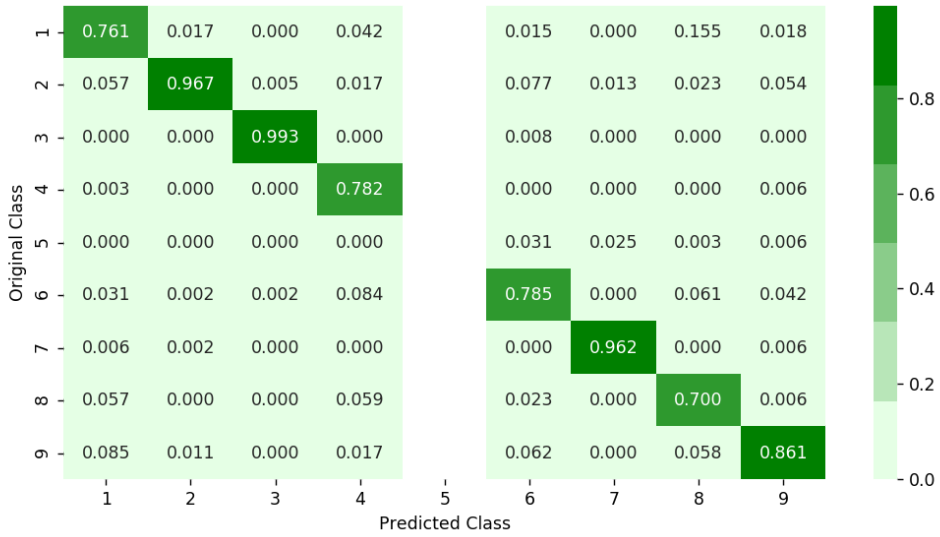
```
log_loss for c =  1e-05 is 1.56916911178
log_loss for c =  0.0001 is 1.57336384417
log_loss for c =  0.001 is 1.53598598273
log_loss for c =  0.01 is 1.01720972418
log_loss for c =  0.1 is 0.857766083873
log_loss for c =  1 is 0.711154393309
log_loss for c =  10 is 0.583929522635
log_loss for c =  100 is 0.549929846589
log_loss for c =  1000 is 0.624746769121
```




```
log loss for train data 0.498923428696
log loss for cv data 0.549929846589
log loss for test data 0.528347316704
Number of misclassified points  12.3275068997
---------------------------------------------------- Confusion matrix --------------------
------------------------------
```

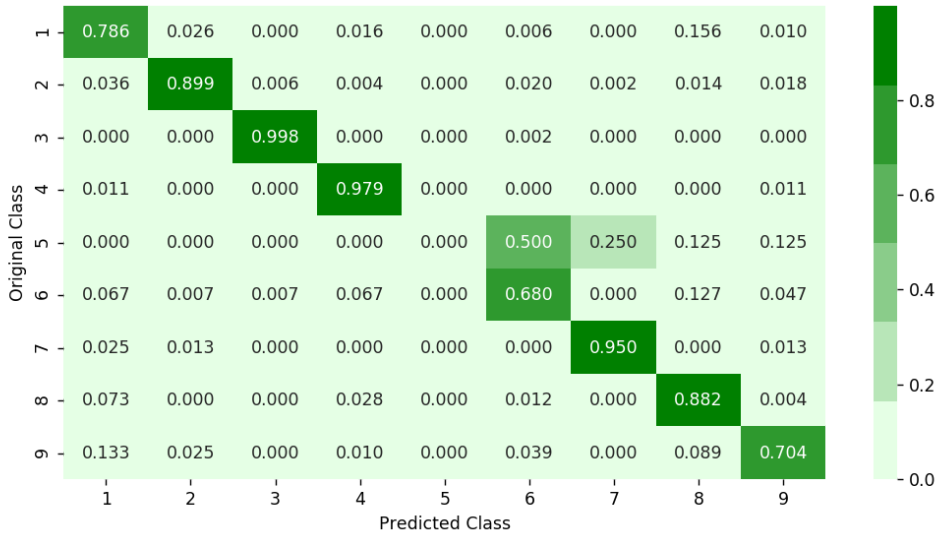



```
---------------------------------------------------- Precision matrix --------------------
------------------------------
```



Sum of columns in precision matrix [1. 1. 1. 1. nan 1. 1. 1. 1.]

---------- Recall matrix ----------



Sum of rows in precision matrix [1. 1. 1. 1. 1. 1. 1. 1. 1.]

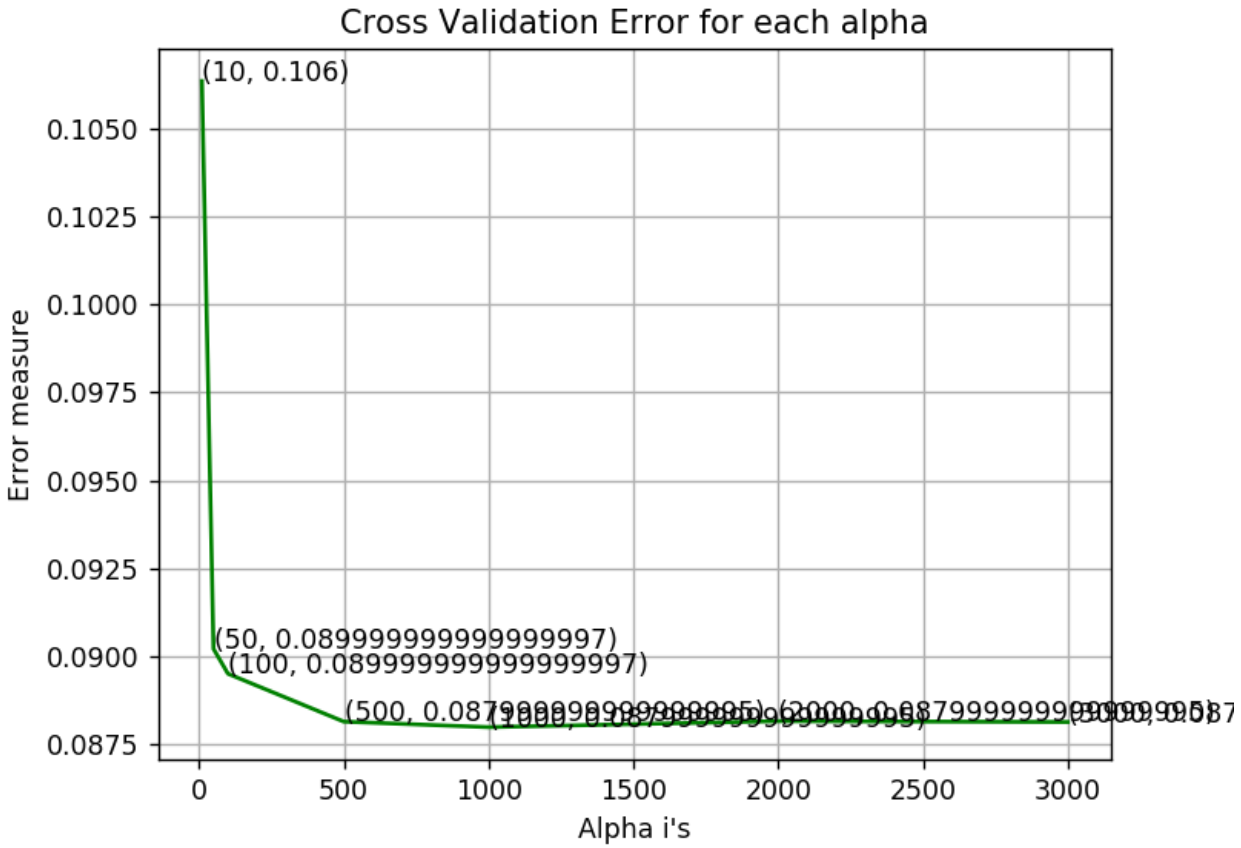
4.1.4. Random Forest Classifier

In [ ]:

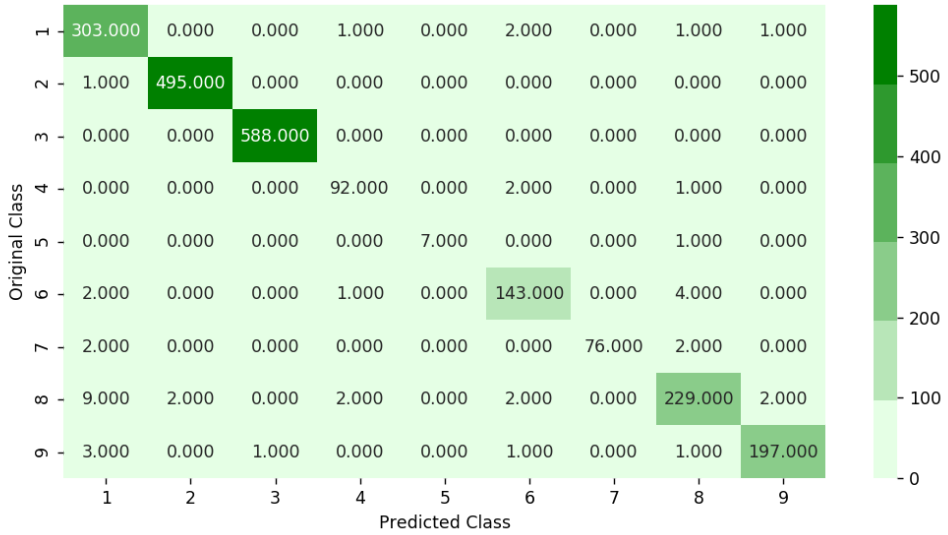
```
# ------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None
# class_weight=None)

# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight])    Fit the SVM model according to the given training data.
# predict(X)    Perform classification on samples in X.
# predict_proba (X)     Perform classification on samples in X.

# some of attributes of  RandomForestClassifier()
```

```
# feature_importances_ : array of shape = [n_features]
# The feature importances (the higher, the more important the feature).

# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/r
# --------------------------------

alpha=[10,50,100,500,1000,2000,3000]
cv_log_error_array=[]
train_log_error_array=[]
from sklearn.ensemble import RandomForestClassifier
for i in alpha:
    r_cfl=RandomForestClassifier(n_estimators=i,random_state=42,n_jobs=-1)
    r_cfl.fit(X_train,y_train)
    sig_clf = CalibratedClassifierCV(r_cfl, method="sigmoid")
    sig_clf.fit(X_train, y_train)
    predict_y = sig_clf.predict_proba(X_cv)
    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=r_cfl.classes_, eps=1e-1

for i in range(len(cv_log_error_array)):
    print ('log_loss for c = ',alpha[i],'is',cv_log_error_array[i])


best_alpha = np.argmin(cv_log_error_array)

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


r_cfl=RandomForestClassifier(n_estimators=alpha[best_alpha],random_state=42,n_jobs=-1)
r_cfl.fit(X_train,y_train)
sig_clf = CalibratedClassifierCV(r_cfl, method="sigmoid")
sig_clf.fit(X_train, y_train)

predict_y = sig_clf.predict_proba(X_train)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_lo
predict_y = sig_clf.predict_proba(X_cv)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss
predict_y = sig_clf.predict_proba(X_test)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_los
plot_confusion_matrix(y_test, sig_clf.predict(X_test))
```

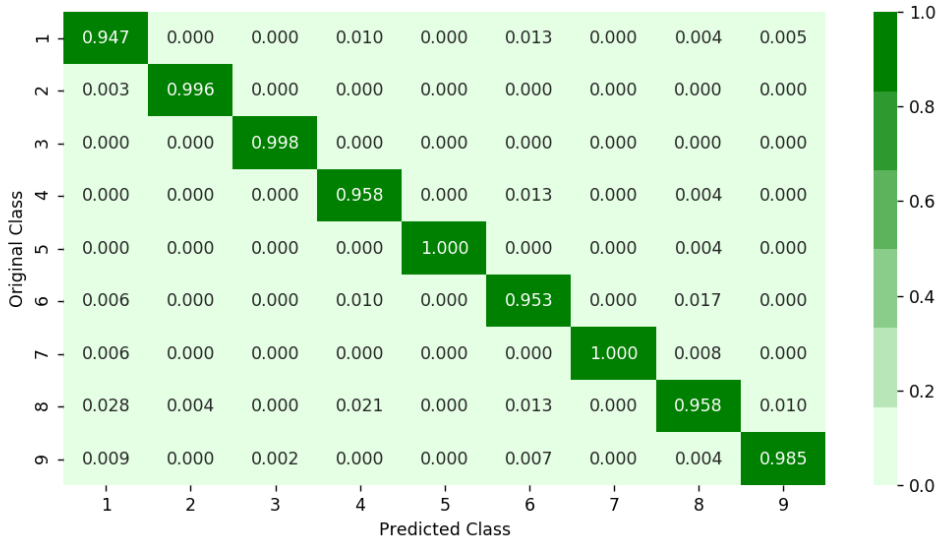
```
log_loss for c =  10 is 0.106357709164
log_loss for c =  50 is 0.0902124124145
log_loss for c =  100 is 0.0895043339776
log_loss for c =  500 is 0.0881420869288
log_loss for c =  1000 is 0.0879849524621
log_loss for c =  2000 is 0.0881566647295
log_loss for c =  3000 is 0.0881318948443
```




```
For values of best alpha =  1000 The train log loss is: 0.0266476291801
For values of best alpha =  1000 The cross validation log loss is: 0.0879849524621
For values of best alpha =  1000 The test log loss is: 0.0858346961407
Number of misclassified points  2.02391904324
---------------------------------------------------- Confusion matrix --------------------
------------------------------
```

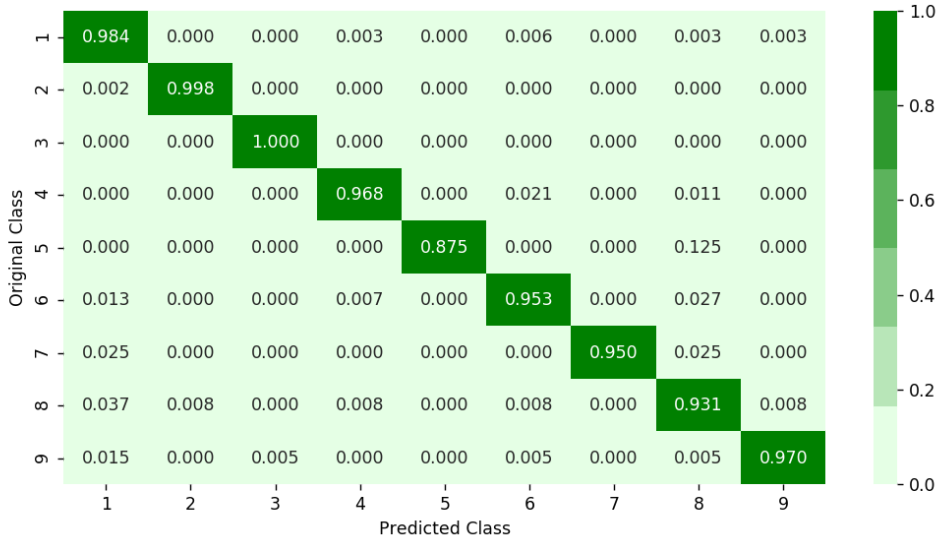



```
---------------------------------------------------- Precision matrix --------------------
------------------------------
```



Sum of columns in precision matrix [1. 1. 1. 1. 1. 1. 1. 1. 1.]

---------- Recall matrix ----------



Sum of rows in precision matrix [1. 1. 1. 1. 1. 1. 1. 1. 1.]

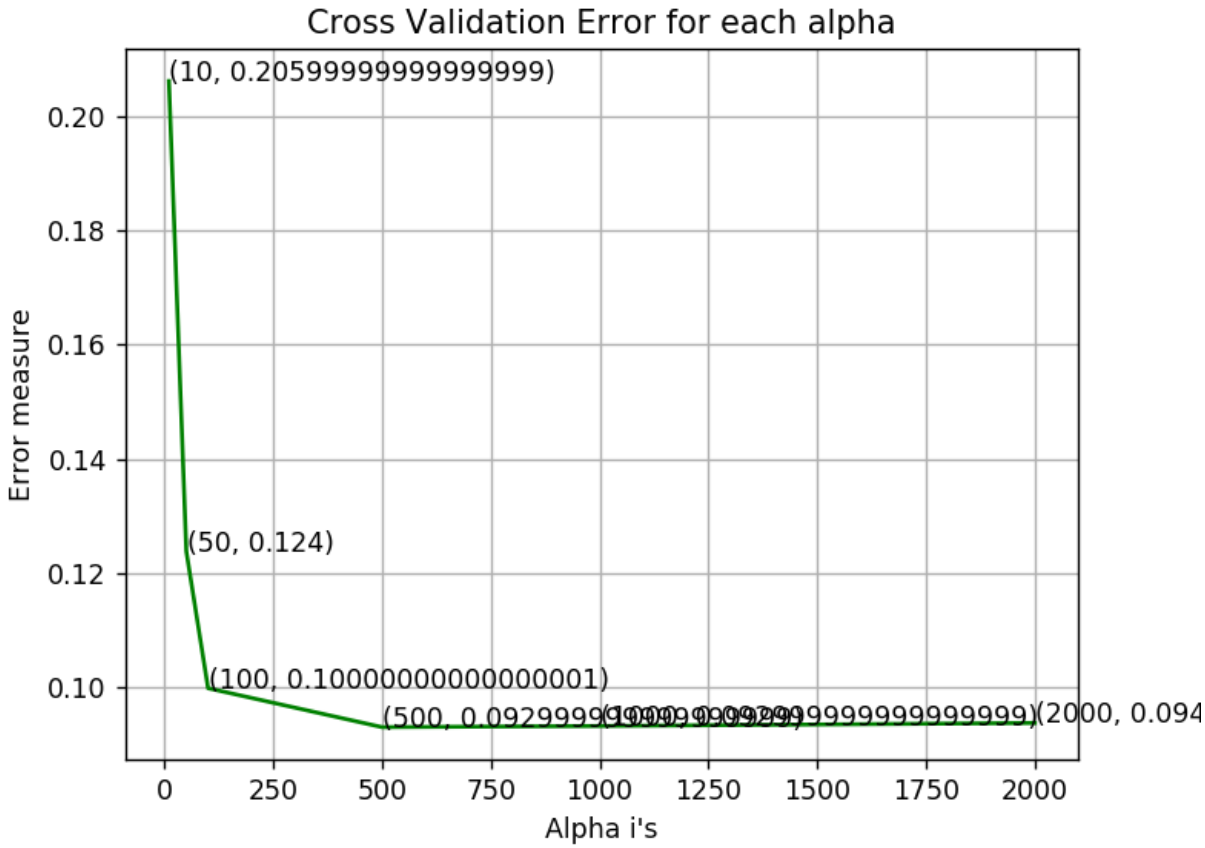
4.1.5. XgBoost Classification

In [ ]:

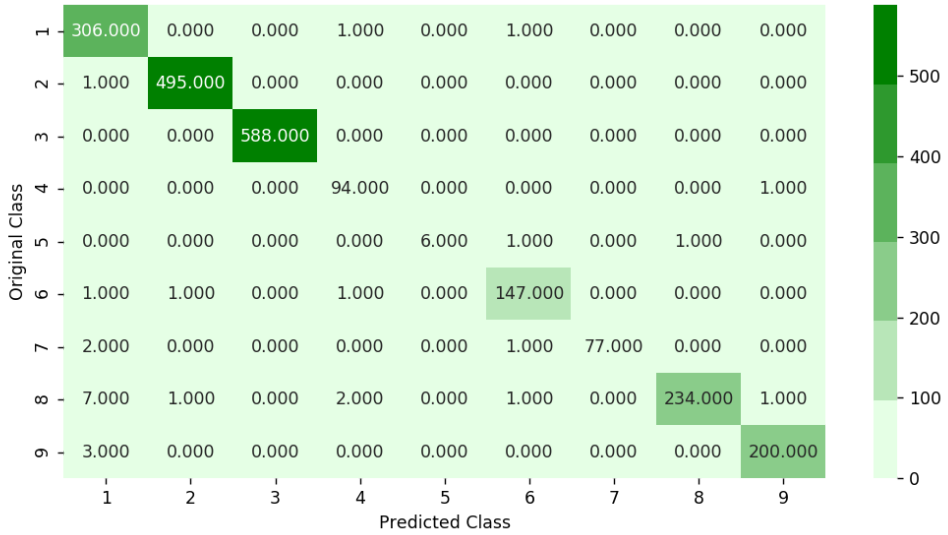
```
# Training a hyper-parameter tuned Xg-Boost regressor on our train data

# find more about XGBClassifier function here http://xgboost.readthedocs.io/en/latest/p
# ------------------------
# default paramters
# class xgboost.XGBClassifier(max_depth=3, learning_rate=0.1, n_estimators=100, silent=
# objective='binary:logistic', booster='gbtree', n_jobs=1, nthread=None, gamma=0, min_c
# max_delta_step=0, subsample=1, colsample_bytree=1, colsample_bylevel=1, reg_alpha=0,
# scale_pos_weight=1, base_score=0.5, random_state=0, seed=None, missing=None, **kwargs

# some of methods of RandomForestRegressor()
# fit(X, y, sample_weight=None, eval_set=None, eval_metric=None, early_stopping_rounds=
# get_params([deep])    Get parameters for this estimator.
```

```
# predict(data, output_margin=False, ntree_limit=0) : Predict with data. NOTE: This fun
# get_score(importance_type='weight') -> get the feature importance
# -----------------------
# video link1: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/
# video link2: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/
# -----------------------

alpha=[10,50,100,500,1000,2000]
cv_log_error_array=[]
for i in alpha:
    x_cfl=XGBClassifier(n_estimators=i,nthread=-1)
    x_cfl.fit(X_train,y_train)
    sig_clf = CalibratedClassifierCV(x_cfl, method="sigmoid")
    sig_clf.fit(X_train, y_train)
    predict_y = sig_clf.predict_proba(X_cv)
    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=x_cfl.classes_, eps=1e-1

for i in range(len(cv_log_error_array)):
    print ('log_loss for c = ',alpha[i],'is',cv_log_error_array[i])


best_alpha = np.argmin(cv_log_error_array)

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()

x_cfl=XGBClassifier(n_estimators=alpha[best_alpha],nthread=-1)
x_cfl.fit(X_train,y_train)
sig_clf = CalibratedClassifierCV(x_cfl, method="sigmoid")
sig_clf.fit(X_train, y_train)

predict_y = sig_clf.predict_proba(X_train)
print ('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_l
predict_y = sig_clf.predict_proba(X_cv)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss
predict_y = sig_clf.predict_proba(X_test)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_los
plot_confusion_matrix(y_test, sig_clf.predict(X_test))
```

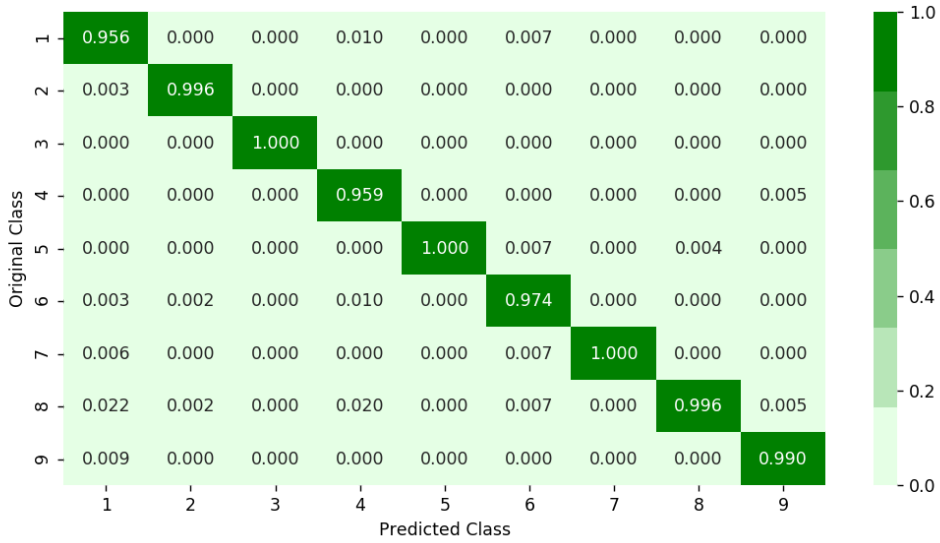
```
log_loss for c =  10 is 0.20615980494
log_loss for c =  50 is 0.123888382365
log_loss for c =  100 is 0.099919437112
log_loss for c =  500 is 0.0931035681289
log_loss for c =  1000 is 0.0933084876012
log_loss for c =  2000 is 0.0938395690309
```




```
For values of best alpha =  500 The train log loss is: 0.0225231805824
For values of best alpha =  500 The cross validation log loss is: 0.0931035681289
For values of best alpha =  500 The test log loss is: 0.0792067651731
Number of misclassified points  1.24195032199
---------------------------------------------------- Confusion matrix --------------------
------------------------------
```

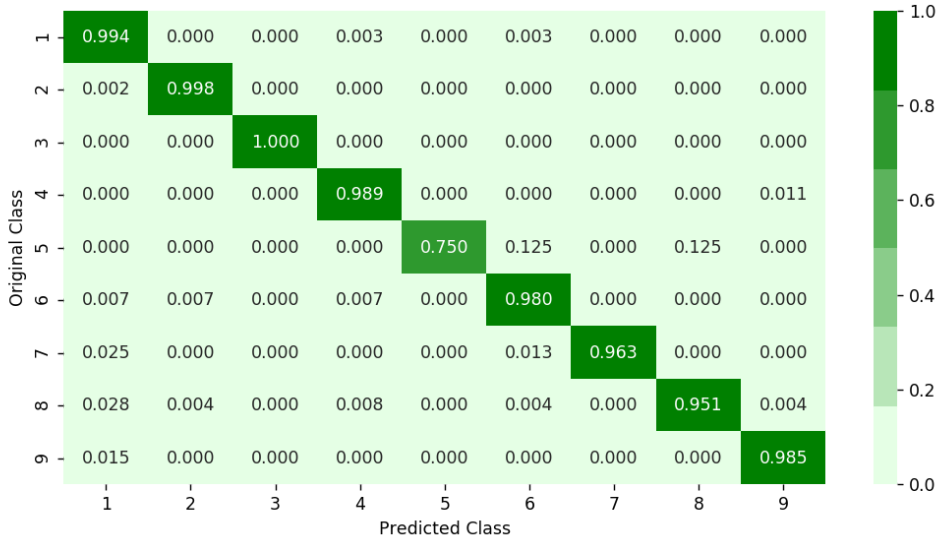



```
---------------------------------------------------- Precision matrix --------------------
------------------------------
```



Sum of columns in precision matrix [1. 1. 1. 1. 1. 1. 1. 1. 1.]

---------- Recall matrix ----------



Sum of rows in precision matrix [1. 1. 1. 1. 1. 1. 1. 1. 1.]

4.1.5. XgBoost Classification with best hyper parameters using RandomSearch

In [ ]:

```
# https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-
x_cfl=XGBClassifier()

prams={
    'learning_rate':[0.01,0.03,0.05,0.1,0.15,0.2],
     'n_estimators':[100,200,500,1000,2000],
     'max_depth':[3,5,10],
    'colsample_bytree':[0.1,0.3,0.5,1],
    'subsample':[0.1,0.3,0.5,1]
}
random_cfl1=RandomizedSearchCV(x_cfl,param_distributions=prams,verbose=10,n_jobs=-1,)
random_cfl1.fit(X_train,y_train)
```

In [ ]:
```
print (random_cfl1.best_params_)
```

In [126…
```
#assignment improving log loss by adding bigrams (top 500)
x_cfl=XGBClassifier(n_estimators=2000, learning_rate=0.05, colsample_bytree=1, max_dept
x_cfl.fit(X_train,y_train)
print ("XGB Done")
c_cfl=CalibratedClassifierCV(x_cfl,method='sigmoid')
c_cfl.fit(X_train,y_train)
print ("Calibration Done")

predict_y = c_cfl.predict_proba(X_train)
print ('train loss',log_loss(y_train, predict_y))
predict_y = c_cfl.predict_proba(X_cv)
print ('cv loss',log_loss(y_cv, predict_y))
predict_y = c_cfl.predict_proba(X_test)
print ('test loss',log_loss(y_test, predict_y))
```

```
[15:32:28] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
XGB Done
[15:35:48] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[15:38:28] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[15:41:10] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[15:43:48] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[15:46:31] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
Calibration Done
train loss 0.01780227884016285
cv loss 0.040421628779677095
test loss 0.03712580137786093
```

4.2 Modeling with .asm files

There are 10868 files of asm
All the files make up about 150 GB
The asm files contains :
1. Address
2. Segments
3. Opcodes
4. Registers
5. function calls
6. APIs

With the help of parallel processing we extracted all the features.In parallel we can use all the cores that are present in our computer.

Here we extracted 52 features from all the asm files which are important.

We read the top solutions and handpicked the features from those papers/videos/blogs.
Refer:https://www.kaggle.com/c/malware-classification/discussion

4.2.1 Feature extraction from asm files

- To extract the unigram features from the .asm files we need to process ~150GB of data
- **Note: Below two cells will take lot of time (over 48 hours to complete)**
- We will provide you the output file of these two cells, which you can directly use it

In [ ]:
```
#intially create five folders
#first
#second
#thrid
#fourth
#fifth
#this code tells us about random split of files into five folders
folder_1 ='first'
folder_2 ='second'
folder_3 ='third'
folder_4 ='fourth'
folder_5 ='fifth'
folder_6 = 'output'
for i in [folder_1,folder_2,folder_3,folder_4,folder_5,folder_6]:
    if not os.path.isdir(i):
        os.makedirs(i)

source='train/'
files = os.listdir('train')
#ID=df['Id'].tolist()
data=range(0,10868)
r.shuffle(data)
count=0
for i in range(0,10868):
    if i % 5==0:
        shutil.move(source+files[data[i]],'first')
    elif i%5==1:
        shutil.move(source+files[data[i]],'second')
    elif i%5 ==2:
        shutil.move(source+files[data[i]],'thrid')
    elif i%5 ==3:
        shutil.move(source+files[data[i]],'fourth')
    elif i%5==4:
        shutil.move(source+files[data[i]],'fifth')
```

In [ ]:
```
#http://flint.cs.yale.edu/cs421/papers/x86-asm/asm.html

def firstprocess():
    #The prefixes tells about the segments that are present in the asm files
    #There are 450 segments(approx) present in all asm files.
    #this prefixes are best segments that gives us best values.
    #https://en.wikipedia.org/wiki/Data_segment
```

```
prefixes = ['HEADER:','.text:','.Pav:','.idata:','.data:','.bss:','.rdata:','.edata
#this are opcodes that are used to get best results
#https://en.wikipedia.org/wiki/X86_instruction_listings

opcodes = ['jmp', 'mov', 'retf', 'push', 'pop', 'xor', 'retn', 'nop', 'sub', 'inc',
#best keywords that are taken from different blogs
keywords = ['.dll','std::',':dword']
#Below taken registers are general purpose registers and special registers
#All the registers which are taken are best
registers=['edx','esi','eax','ebx','ecx','edi','ebp','esp','eip']
file1=open("output\asmsmallfile.txt","w+")
files = os.listdir('first')
for f in files:
    #filling the values with zeros into the arrays
    prefixescount=np.zeros(len(prefixes),dtype=int)
    opcodescount=np.zeros(len(opcodes),dtype=int)
    keywordcount=np.zeros(len(keywords),dtype=int)
    registerscount=np.zeros(len(registers),dtype=int)
    features=[]
    f2=f.split('.')[0]
    file1.write(f2+",")
    opcodefile.write(f2+" ")
    # https://docs.python.org/3/library/codecs.html#codecs.ignore_errors
    # https://docs.python.org/3/library/codecs.html#codecs.Codec.encode
    with codecs.open('first/'+f,encoding='cp1252',errors ='replace') as fli:
        for lines in fli:
            # https://www.tutorialspoint.com/python3/string_rstrip.htm
            line=lines.rstrip().split()
            l=line[0]
            #counting the prefixs in each and every line
            for i in range(len(prefixes)):
                if prefixes[i] in line[0]:
                    prefixescount[i]+=1
            line=line[1:]
            #counting the opcodes in each and every line
            for i in range(len(opcodes)):
                if any(opcodes[i]==li for li in line):
                    features.append(opcodes[i])
                    opcodescount[i]+=1
            #counting registers in the line
            for i in range(len(registers)):
                for li in line:
                    # we will use registers only in 'text' and 'CODE' segments
                    if registers[i] in li and ('text' in l or 'CODE' in l):
                        registerscount[i]+=1
            #counting keywords in the line
            for i in range(len(keywords)):
                for li in line:
                    if keywords[i] in li:
                        keywordcount[i]+=1
    #pushing the values into the file after reading whole file
    for prefix in prefixescount:
        file1.write(str(prefix)+",")
    for opcode in opcodescount:
        file1.write(str(opcode)+",")
    for register in registerscount:
        file1.write(str(register)+",")
    for key in keywordcount:
        file1.write(str(key)+",")
    file1.write("\n")
file1.close()
```

```
#same as above
def secondprocess():
    prefixes = ['HEADER:','.text:','.Pav:','.idata:','.data:','.bss:','.rdata:','.edata
    opcodes = ['jmp', 'mov', 'retf', 'push', 'pop', 'xor', 'retn', 'nop', 'sub', 'inc',
    keywords = ['.dll','std::',':dword']
    registers=['edx','esi','eax','ebx','ecx','edi','ebp','esp','eip']
    file1=open("output\mediumasmfile.txt","w+")
    files = os.listdir('second')
    for f in files:
        prefixescount=np.zeros(len(prefixes),dtype=int)
        opcodescount=np.zeros(len(opcodes),dtype=int)
        keywordcount=np.zeros(len(keywords),dtype=int)
        registerscount=np.zeros(len(registers),dtype=int)
        features=[]
        f2=f.split('.')[0]
        file1.write(f2+",")
        opcodefile.write(f2+" ")
        with codecs.open('second/'+f,encoding='cp1252',errors ='replace') as fli:
            for lines in fli:
                line=lines.rstrip().split()
                l=line[0]
                for i in range(len(prefixes)):
                    if prefixes[i] in line[0]:
                        prefixescount[i]+=1
                line=line[1:]
                for i in range(len(opcodes)):
                    if any(opcodes[i]==li for li in line):
                        features.append(opcodes[i])
                        opcodescount[i]+=1
                for i in range(len(registers)):
                    for li in line:
                        if registers[i] in li and ('text' in l or 'CODE' in l):
                            registerscount[i]+=1
                for i in range(len(keywords)):
                    for li in line:
                        if keywords[i] in li:
                            keywordcount[i]+=1
        for prefix in prefixescount:
            file1.write(str(prefix)+",")
        for opcode in opcodescount:
            file1.write(str(opcode)+",")
        for register in registerscount:
            file1.write(str(register)+",")
        for key in keywordcount:
            file1.write(str(key)+",")
        file1.write("\n")
    file1.close()

# same as smallprocess() functions
def thirdprocess():
    prefixes = ['HEADER:','.text:','.Pav:','.idata:','.data:','.bss:','.rdata:','.edata
    opcodes = ['jmp', 'mov', 'retf', 'push', 'pop', 'xor', 'retn', 'nop', 'sub', 'inc',
    keywords = ['.dll','std::',':dword']
    registers=['edx','esi','eax','ebx','ecx','edi','ebp','esp','eip']
    file1=open("output\largeasmfile.txt","w+")
    files = os.listdir('thrid')
    for f in files:
        prefixescount=np.zeros(len(prefixes),dtype=int)
        opcodescount=np.zeros(len(opcodes),dtype=int)
```

```
        keywordcount=np.zeros(len(keywords),dtype=int)
        registerscount=np.zeros(len(registers),dtype=int)
        features=[]
        f2=f.split('.')[0]
        file1.write(f2+",")
        opcodefile.write(f2+" ")
        with codecs.open('thrid/'+f,encoding='cp1252',errors ='replace') as fli:
            for lines in fli:
                line=lines.rstrip().split()
                l=line[0]
                for i in range(len(prefixes)):
                    if prefixes[i] in line[0]:
                        prefixescount[i]+=1
                line=line[1:]
                for i in range(len(opcodes)):
                    if any(opcodes[i]==li for li in line):
                        features.append(opcodes[i])
                        opcodescount[i]+=1
                for i in range(len(registers)):
                    for li in line:
                        if registers[i] in li and ('text' in l or 'CODE' in l):
                            registerscount[i]+=1
                for i in range(len(keywords)):
                    for li in line:
                        if keywords[i] in li:
                            keywordcount[i]+=1
        for prefix in prefixescount:
            file1.write(str(prefix)+",")
        for opcode in opcodescount:
            file1.write(str(opcode)+",")
        for register in registerscount:
            file1.write(str(register)+",")
        for key in keywordcount:
            file1.write(str(key)+",")
        file1.write("\n")
    file1.close()


def fourthprocess():
    prefixes = ['HEADER:','.text:','.Pav:','.idata:','.data:','.bss:','.rdata:','.edata
    opcodes = ['jmp', 'mov', 'retf', 'push', 'pop', 'xor', 'retn', 'nop', 'sub', 'inc',
    keywords = ['.dll','std::',':dword']
    registers=['edx','esi','eax','ebx','ecx','edi','ebp','esp','eip']
    file1=open("output\hugeasmfile.txt","w+")
    files = os.listdir('fourth/')
    for f in files:
        prefixescount=np.zeros(len(prefixes),dtype=int)
        opcodescount=np.zeros(len(opcodes),dtype=int)
        keywordcount=np.zeros(len(keywords),dtype=int)
        registerscount=np.zeros(len(registers),dtype=int)
        features=[]
        f2=f.split('.')[0]
        file1.write(f2+",")
        opcodefile.write(f2+" ")
        with codecs.open('fourth/'+f,encoding='cp1252',errors ='replace') as fli:
            for lines in fli:
                line=lines.rstrip().split()
                l=line[0]
                for i in range(len(prefixes)):
                    if prefixes[i] in line[0]:
                        prefixescount[i]+=1
```

```
                line=line[1:]
                for i in range(len(opcodes)):
                    if any(opcodes[i]==li for li in line):
                        features.append(opcodes[i])
                        opcodescount[i]+=1
                for i in range(len(registers)):
                    for li in line:
                        if registers[i] in li and ('text' in l or 'CODE' in l):
                            registerscount[i]+=1
                for i in range(len(keywords)):
                    for li in line:
                        if keywords[i] in li:
                            keywordcount[i]+=1
        for prefix in prefixescount:
            file1.write(str(prefix)+",")
        for opcode in opcodescount:
            file1.write(str(opcode)+",")
        for register in registerscount:
            file1.write(str(register)+",")
        for key in keywordcount:
            file1.write(str(key)+",")
        file1.write("\n")
    file1.close()


def fifthprocess():
    prefixes = ['HEADER:','.text:','.Pav:','.idata:','.data:','.bss:','.rdata:','.edata
    opcodes = ['jmp', 'mov', 'retf', 'push', 'pop', 'xor', 'retn', 'nop', 'sub', 'inc',
    keywords = ['.dll','std::',':dword']
    registers=['edx','esi','eax','ebx','ecx','edi','ebp','esp','eip']
    file1=open("output\trainasmfile.txt","w+")
    files = os.listdir('fifth/')
    for f in files:
        prefixescount=np.zeros(len(prefixes),dtype=int)
        opcodescount=np.zeros(len(opcodes),dtype=int)
        keywordcount=np.zeros(len(keywords),dtype=int)
        registerscount=np.zeros(len(registers),dtype=int)
        features=[]
        f2=f.split('.')[0]
        file1.write(f2+",")
        opcodefile.write(f2+" ")
        with codecs.open('fifth/'+f,encoding='cp1252',errors ='replace') as fli:
            for lines in fli:
                line=lines.rstrip().split()
                l=line[0]
                for i in range(len(prefixes)):
                    if prefixes[i] in line[0]:
                        prefixescount[i]+=1
                line=line[1:]
                for i in range(len(opcodes)):
                    if any(opcodes[i]==li for li in line):
                        features.append(opcodes[i])
                        opcodescount[i]+=1
                for i in range(len(registers)):
                    for li in line:
                        if registers[i] in li and ('text' in l or 'CODE' in l):
                            registerscount[i]+=1
                for i in range(len(keywords)):
                    for li in line:
                        if keywords[i] in li:
                            keywordcount[i]+=1
```

```
            for prefix in prefixescount:
                file1.write(str(prefix)+",")
            for opcode in opcodescount:
                file1.write(str(opcode)+",")
            for register in registerscount:
                file1.write(str(register)+",")
            for key in keywordcount:
                file1.write(str(key)+",")
            file1.write("\n")
        file1.close()


def main():
    #the below code is used for multiprogramming
    #the number of process depends upon the number of cores present System
    #process is used to call multiprogramming
    manager=multiprocessing.Manager()
    p1=Process(target=firstprocess)
    p2=Process(target=secondprocess)
    p3=Process(target=thirdprocess)
    p4=Process(target=fourthprocess)
    p5=Process(target=fifthprocess)
    #p1.start() is used to start the thread execution
    p1.start()
    p2.start()
    p3.start()
    p4.start()
    p5.start()
    #After completion all the threads are joined
    p1.join()
    p2.join()
    p3.join()
    p4.join()
    p5.join()

if __name__=="__main__":
    main()
```

In [166…

```
# asmoutputfile.csv(output genarated from the above two cells) will contain all the ext
# this file will be uploaded in the drive, you can directly use this
dfasm=pd.read_csv("asmoutputfile.csv")
Y.columns = ['ID', 'Class']
result_asm = pd.merge(dfasm, Y,on='ID', how='left')
result_asm.head()
```

Out[166…

	ID	HEADER:	.text:	.Pav:	.idata:	.data:	.bss:	.rdata:	.edata:	.rsrc:	...	ed
0	01kcPWA9K2BOxQeS5Rju	19	744	0	127	57	0	323	0	3	...	18
1	1E93CpP60RHFNiT5Qfvn	17	838	0	103	49	0	0	0	3	...	18
2	3ekVow2ajZHbTnBcsDfX	17	427	0	50	43	0	145	0	3	...	13
3	3X2nY7iQaPBIWDrAZqJe	17	227	0	43	19	0	0	0	3	...	6
4	46OZzdsSKDCFV8h7XWxf	17	402	0	59	170	0	0	0	3	...	12

5 rows × 53 columns

4.2.1.1 Files sizes of each .asm file

In [167…

```
#file sizes of byte files

files=os.listdir('asmFiles')
filenames=Y['ID'].tolist()
class_y=Y['Class'].tolist()
class_bytes=[]
sizebytes=[]
fnames=[]
for file in files:
    # print(os.stat('byteFiles/0A32eTdBKayjCWhZqDOQ.txt'))
    # os.stat_result(st_mode=33206, st_ino=1125899906874507, st_dev=3561571700, st_nlin
    # st_size=3680109, st_atime=1519638522, st_mtime=1519638522, st_ctime=1519638522)
    # read more about os.stat: here https://www.tutorialspoint.com/python/os_stat.htm
    statinfo=os.stat('asmFiles/'+file)
    # split the file name at '.' and take the first part of it i.e the file name
    file=file.split('.')[0]
    if any(file == filename for filename in filenames):
        i=filenames.index(file)
        class_bytes.append(class_y[i])
        # converting into Mb's
        sizebytes.append(statinfo.st_size/(1024.0*1024.0))
        fnames.append(file)
asm_size_byte=pd.DataFrame({'ID':fnames,'size':sizebytes,'Class':class_bytes})
print (asm_size_byte.head())
```

```
                     ID       size  Class
0  AhEcNQy9nYHdfzZreoX1   8.670227      8
1  3ufcA2y6mtDTZFXiIoL8   0.171723      3
2  deou4UEvA59Vi2DZySL8   1.106604      9
3  kKBJnDycpQmzPol5OU6Z  17.764704      2
4  av90QiV8gsYkEpcKReuq   0.179785      3
```

In [20]:

```
#!pip install imageio
```

```
Collecting imageio
  Downloading imageio-2.9.0-py3-none-any.whl (3.3 MB)
     |████████████████████████████████| 3.3 MB 4.3 MB/s eta 0:00:01
Requirement already satisfied: numpy in /opt/conda/lib/python3.7/site-packages (from ima
geio) (1.19.4)
Requirement already satisfied: pillow in /opt/conda/lib/python3.7/site-packages (from im
ageio) (8.0.1)
Installing collected packages: imageio
Successfully installed imageio-2.9.0
```

In [24]:

```
#Converting asm files to images
#from youtube video given for reference

import scipy.misc
import array
import imageio

for asmfile in tqdm(os.listdir("asmFiles")):

    filename = asmfile.split('.')[0]  #getting name of asm file
    file = codecs.open("asmFiles/" + asmfile, 'rb')
    filelen = os.path.getsize("asmFiles/" + asmfile)
    width = int(filelen ** 0.5)
```

```
    rem = int(filelen / width)
    arr = array.array('B')
    arr.frombytes(file.read())
    file.close()
    reshaped = np.reshape(arr[:width * width], (width, width))
    reshaped = np.uint8(reshaped)
    imageio.imwrite('asm_image/' + filename + '.png',reshaped)

# for file in os.listdir("asmFiles"):
#     file_open = open("asmFiles/" + file)
#     ln = os.path.getsize("asmFiles/" + file)
#     width = int(ln**0.5)
#     rem = ln%width
#     a = array.array('B')
#     a.fromfile(file_open,ln-rem)
#     file_open.close()
#     g= np.reshape(a,(len(a)/width,width))
#     g= np.uint8(g)
#     scipy.misc.imsave('asm_image/' + filename + '.png',reshaped)
```

```
100%|██████████| 10868/10868 [3:30:39<00:00,  1.16s/it]
```

In [154…

```
#Getting first 800 pixel values from each asm image
#https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_core/py_basic_

import cv2

pixelValues = np.zeros((10868, 800))

asmimageFiles = os.listdir("asm_image")

for i, f in tqdm(enumerate(asmimageFiles)):

    each_file = cv2.imread("asm_image/" + f)

    pixelValues[i, :] += each_file.flatten()[:800]  #taking value of first 800
```

```
10868it [29:35,  6.12it/s]
```

In [155…

```
pixelValues
```

Out[155…

```
array([[ 46.,  46.,  46., ...,  32.,  32.,  32.],
       [ 72.,  72.,  72., ..., 103., 104., 104.],
       [ 72.,  72.,  72., ..., 103., 104., 104.],
       ...,
       [ 72.,  72.,  72., ..., 103., 104., 104.],
       [ 72.,  72.,  72., ..., 103., 104., 104.],
       [ 72.,  72.,  72., ..., 103., 104., 104.]])
```

In [156…

```
# import scipy.sparse
# scipy.sparse.save_npz('pixelValues.npz', csr_matrix(pixelValues))
```

In [168…

```
pixelValues = scipy.sparse.load_npz('pixelValues.npz')
pixelValues = pixelValues.toarray()
```

In [169…

```
print(pixelValues)
```

```
[[ 46.  46.  46. ...  32.  32.  32.]
 [ 72.  72.  72. ... 103. 104. 104.]
 [ 72.  72.  72. ... 103. 104. 104.]
 ...
 [ 72.  72.  72. ... 103. 104. 104.]
 [ 72.  72.  72. ... 103. 104. 104.]
 [ 72.  72.  72. ... 103. 104. 104.]]
```

In [170…

```
pixel_no = []

for i in tqdm(range(800)):
    pixel_no.append('p' +' '+ str(i))

from sklearn.preprocessing import normalize
pixelValues = normalize(pixelValues)

pixeldf = pd.DataFrame(pixelValues)
pixeldf.columns = pixel_no

f_name = []
for i in os.listdir('asm_image'):
    f_name.append(i.replace('.png',''))

#Adding file names
pixeldf["ID"] = f_name

pixeldf.head(10)
```

```
100%|██████████| 800/800 [00:00<00:00, 711501.95it/s]
```

Out[170…

	p 0	p 1	p 2	p 3	p 4	p 5	p 6	p 7	p 8	
0	0.025038	0.025038	0.025038	0.063140	0.063140	0.063140	0.054975	0.054975	0.054975	0.065
1	0.039310	0.039310	0.039310	0.037672	0.037672	0.037672	0.035489	0.035489	0.035489	0.037
2	0.039310	0.039310	0.039310	0.037672	0.037672	0.037672	0.035489	0.035489	0.035489	0.037
3	0.039310	0.039310	0.039310	0.037672	0.037672	0.037672	0.035489	0.035489	0.035489	0.037
4	0.039310	0.039310	0.039310	0.037672	0.037672	0.037672	0.035489	0.035489	0.035489	0.037
5	0.039310	0.039310	0.039310	0.037672	0.037672	0.037672	0.035489	0.035489	0.035489	0.037
6	0.039310	0.039310	0.039310	0.037672	0.037672	0.037672	0.035489	0.035489	0.035489	0.037
7	0.039310	0.039310	0.039310	0.037672	0.037672	0.037672	0.035489	0.035489	0.035489	0.037
8	0.039310	0.039310	0.039310	0.037672	0.037672	0.037672	0.035489	0.035489	0.035489	0.037
9	0.025038	0.025038	0.025038	0.063140	0.063140	0.063140	0.054975	0.054975	0.054975	0.065

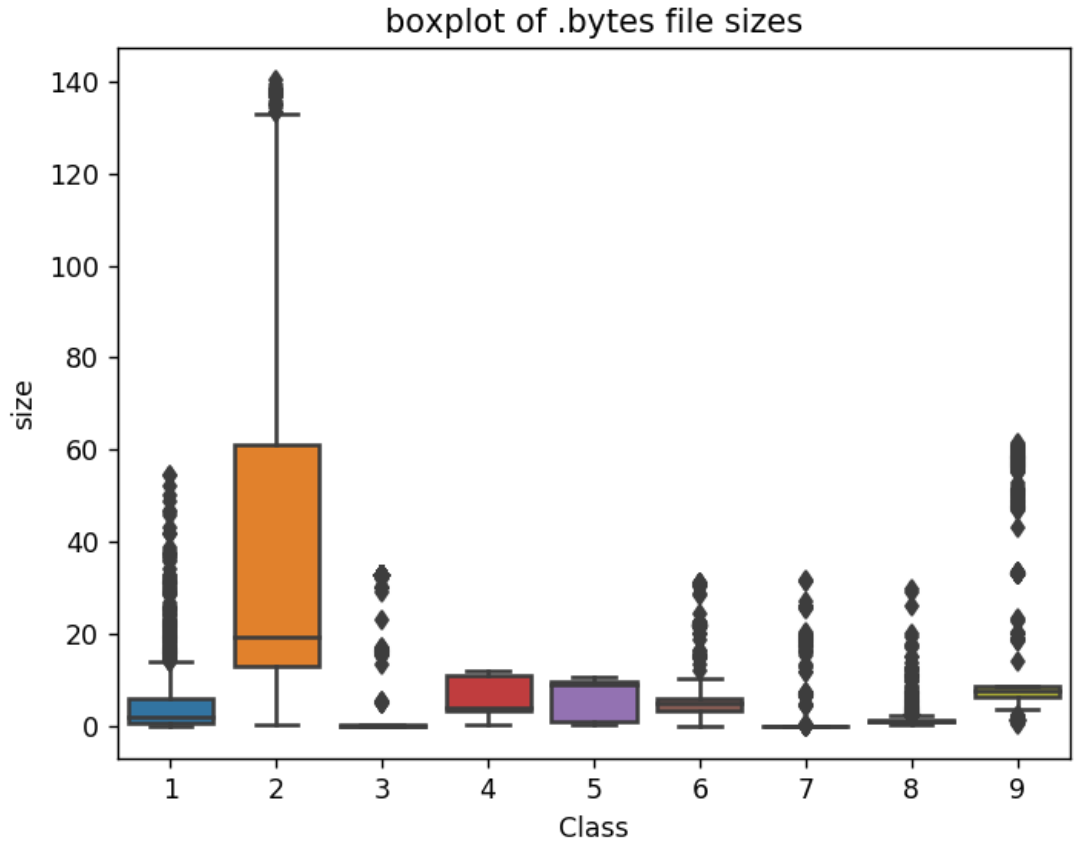
10 rows × 801 columns

4.2.1.2 Distribution of .asm file sizes

In [ ]:

```
#boxplot of asm files
ax = sns.boxplot(x="Class", y="size", data=asm_size_byte)
```

```
plt.title("boxplot of .bytes file sizes")
plt.show()
```



In [171…

```
# add the file size feature to previous extracted features
print(result_asm.shape)
print(asm_size_byte.shape)
result_asm = pd.merge(result_asm, asm_size_byte.drop(['Class'], axis=1),on='ID', how='l

result_asm.head(10)
```

```
(10868, 53)
(10868, 3)
```

Out[171…

	ID	HEADER:	.text:	.Pav:	.idata:	.data:	.bss:	.rdata:	.edata:	.rsrc:	...	es
0	01kcPWA9K2BOxQeS5Rju	19	744	0	127	57	0	323	0	3	...	6
1	1E93CpP60RHFNiT5Qfvn	17	838	0	103	49	0	0	0	3	...	2
2	3ekVow2ajZHbTnBcsDfX	17	427	0	50	43	0	145	0	3	...	4
3	3X2nY7iQaPBIWDrAZqJe	17	227	0	43	19	0	0	0	3	...	
4	46OZzdsSKDCFV8h7XWxf	17	402	0	59	170	0	0	0	3	...	
5	4QlfSpex83Nsh2oBcWXv	19	846	0	128	47	0	304	0	3	...	8
6	5mS2M1zjfQZ9HAVIFeLc	17	840	0	103	49	0	0	0	3	...	3
7	5NAQh3w9SULO24RtEM6b	17	459	0	84	26	0	0	0	3	...	2

	ID	HEADER:	.text:	.Pav:	.idata:	.data:	.bss:	.rdata:	.edata:	.rsrc:	...	es
8	5w6GuqFsfkxg7TWIDQBY	17	575	0	99	55	0	279	0	3	...	4(
9	63RfPeChL1tsJwy2lVHo	19	425	0	86	57	0	0	0	3	...	1!

10 rows × 54 columns

In [172…

```
# we normalize the data each column
# https://stackoverflow.com/a/29651514
def normalize(df):
    result1 = df.copy()
    for feature_name in tqdm(df.columns):
        if (str(feature_name) != str('ID') and str(feature_name)!=str('Class')):
            max_value = df[feature_name].max()
            min_value = df[feature_name].min()
            result1[feature_name] = (df[feature_name] - min_value) / (max_value - min_v
    return result1

result_asm = normalize(result_asm)
result_asm.head()
```

```
100%|██████████| 54/54 [00:00<00:00, 1337.64it/s]
```

Out[172…

	ID	HEADER:	.text:	.Pav:	.idata:	.data:	.bss:	.rdata:	.edata:	
0	01kcPWA9K2BOxQeS5Rju	0.107345	0.001092	0.0	0.000761	0.000023	0.0	0.000084	0.0	(
1	1E93CpP60RHFNiT5Qfvn	0.096045	0.001230	0.0	0.000617	0.000019	0.0	0.000000	0.0	(
2	3ekVow2ajZHbTnBcsDfX	0.096045	0.000627	0.0	0.000300	0.000017	0.0	0.000038	0.0	(
3	3X2nY7iQaPBIWDrAZqJe	0.096045	0.000333	0.0	0.000258	0.000008	0.0	0.000000	0.0	(
4	46OZzdsSKDCFV8h7XWxf	0.096045	0.000590	0.0	0.000353	0.000068	0.0	0.000000	0.0	(

5 rows × 54 columns

In [173…

```
result_asm = pd.merge(result_asm, pixeldf,on='ID', how='left')
result_asm.head(10)
```

Out[173…

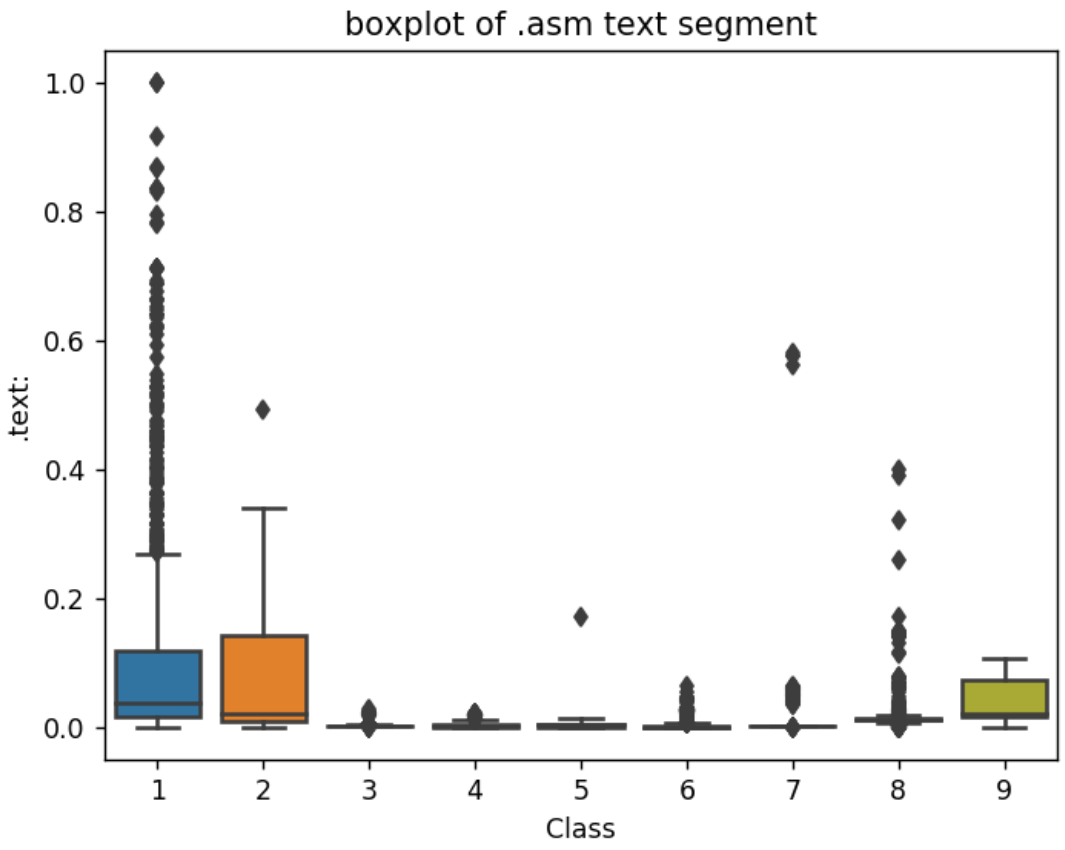
	ID	HEADER:	.text:	.Pav:	.idata:	.data:	.bss:	.rdata:	.edata:
0	01kcPWA9K2BOxQeS5Rju	0.107345	0.001092	0.0	0.000761	0.000023	0.0	0.000084	0.0
1	1E93CpP60RHFNiT5Qfvn	0.096045	0.001230	0.0	0.000617	0.000019	0.0	0.000000	0.0
2	3ekVow2ajZHbTnBcsDfX	0.096045	0.000627	0.0	0.000300	0.000017	0.0	0.000038	0.0
3	3X2nY7iQaPBIWDrAZqJe	0.096045	0.000333	0.0	0.000258	0.000008	0.0	0.000000	0.0
4	46OZzdsSKDCFV8h7XWxf	0.096045	0.000590	0.0	0.000353	0.000068	0.0	0.000000	0.0
5	4QlfSpex83Nsh2oBcWXv	0.107345	0.001242	0.0	0.000767	0.000019	0.0	0.000079	0.0
6	5mS2M1zjfQZ9HAVIFeLc	0.096045	0.001233	0.0	0.000617	0.000019	0.0	0.000000	0.0

	ID	HEADER:	.text:	.Pav:	.idata:	.data:	.bss:	.rdata:	.edata:
7	5NAQh3w9SULO24RtEM6b	0.096045	0.000674	0.0	0.000503	0.000010	0.0	0.000000	0.0
8	5w6GuqFsfkxg7TWIDQBY	0.096045	0.000844	0.0	0.000593	0.000022	0.0	0.000073	0.0
9	63RfPeChL1tsJwy2lVHo	0.107345	0.000624	0.0	0.000515	0.000023	0.0	0.000000	0.0

10 rows × 854 columns

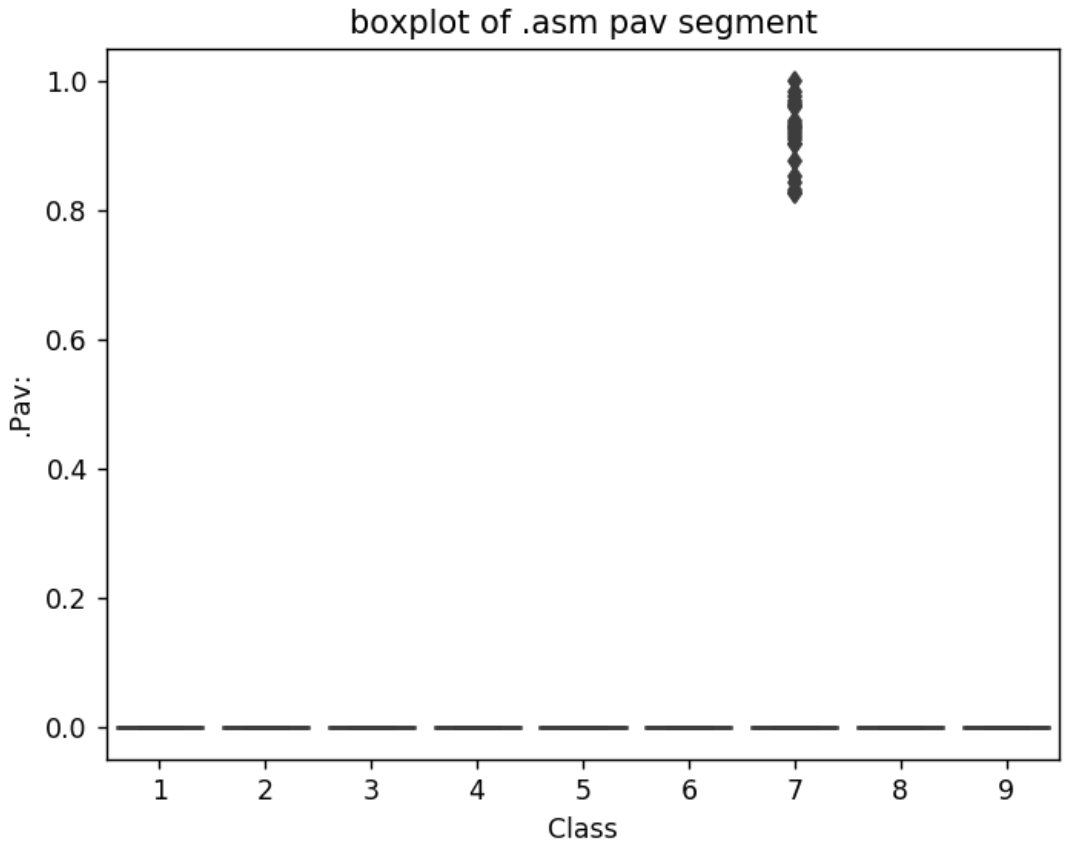
4.2.2 Univariate analysis on asm file features

In [ ]:
```
ax = sns.boxplot(x="Class", y=".text:", data=result_asm)
plt.title("boxplot of .asm text segment")
plt.show()
```



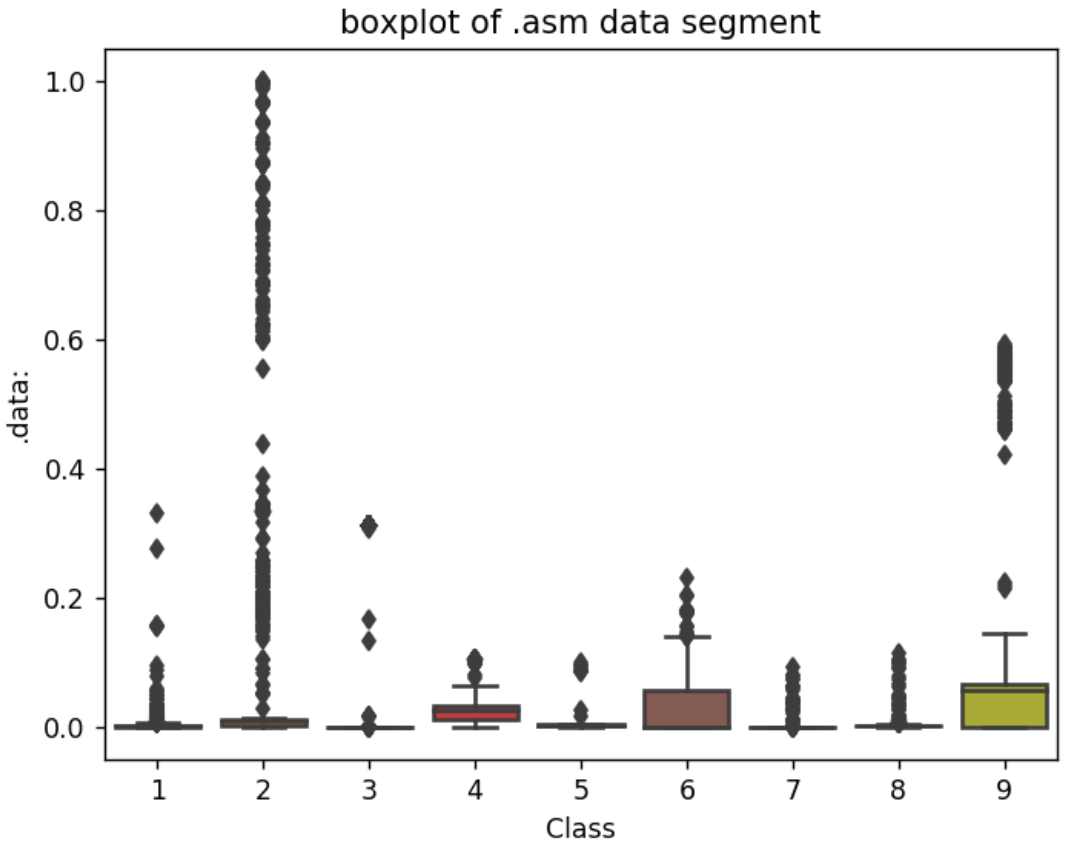
The plot is between Text and class
Class 1,2 and 9 can be easly separated

In [ ]:
```
ax = sns.boxplot(x="Class", y=".Pav:", data=result_asm)
plt.title("boxplot of .asm pav segment")
plt.show()
```



In [ ]:

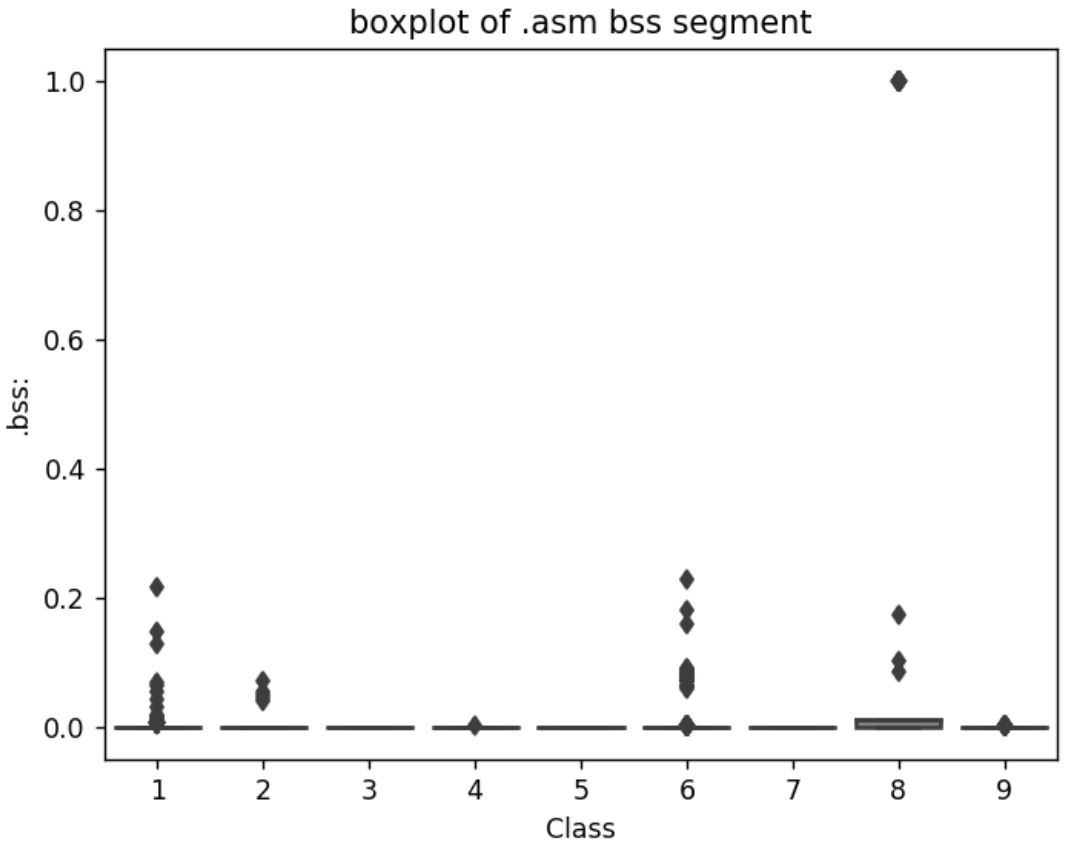
```
ax = sns.boxplot(x="Class", y=".data:", data=result_asm)
plt.title("boxplot of .asm data segment")
plt.show()
```



The plot is between data segment and class label
class 6 and class 9 can be easily separated from given points

In [ ]:

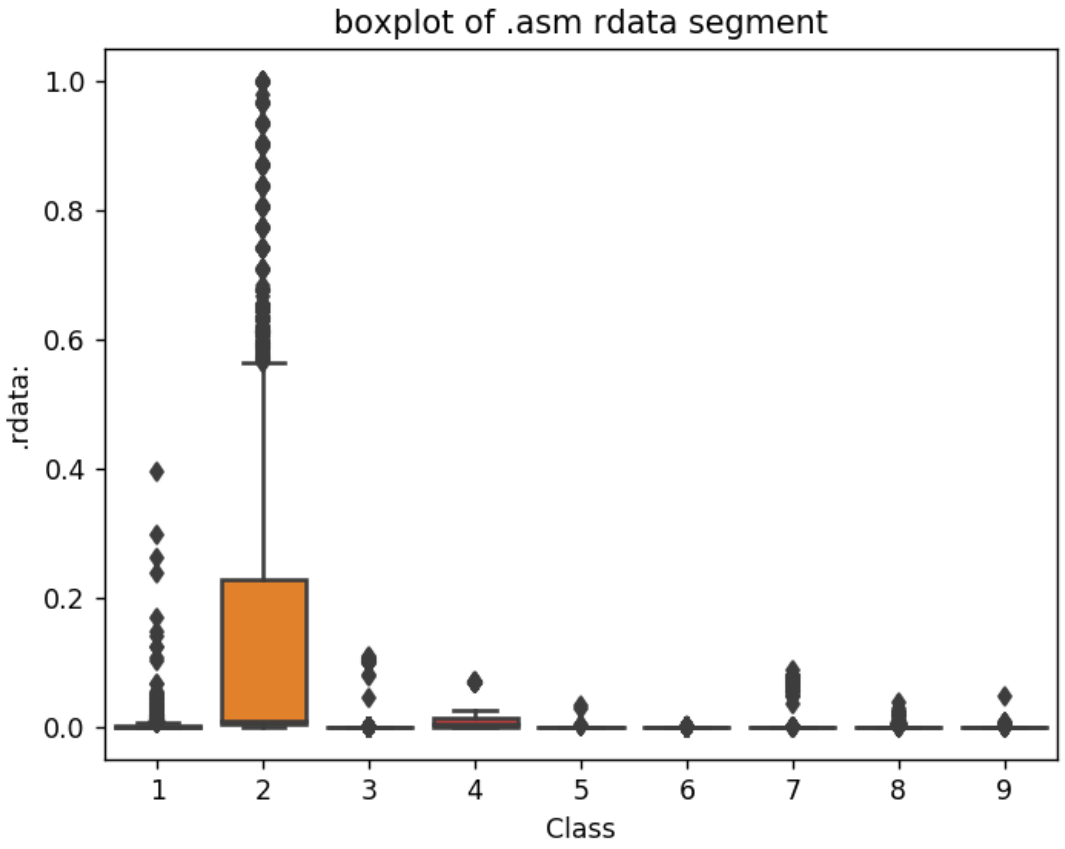
```
ax = sns.boxplot(x="Class", y=".bss:", data=result_asm)
plt.title("boxplot of .asm bss segment")
plt.show()
```



plot between bss segment and class label
very less number of files are having bss segment

In [ ]:

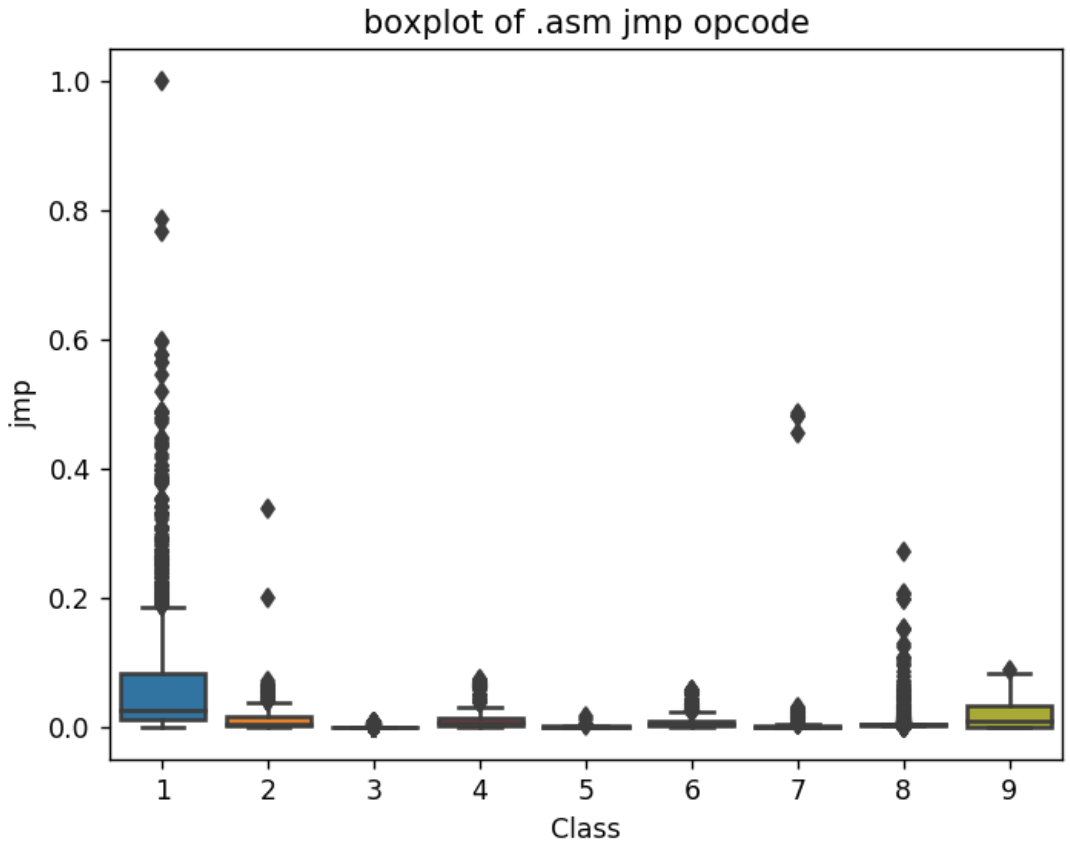
```
ax = sns.boxplot(x="Class", y=".rdata:", data=result_asm)
plt.title("boxplot of .asm rdata segment")
plt.show()
```



Plot between rdata segment and Class segment
Class 2 can be easily separated 75 pecentile files are having 1M rdata lines

In [ ]:

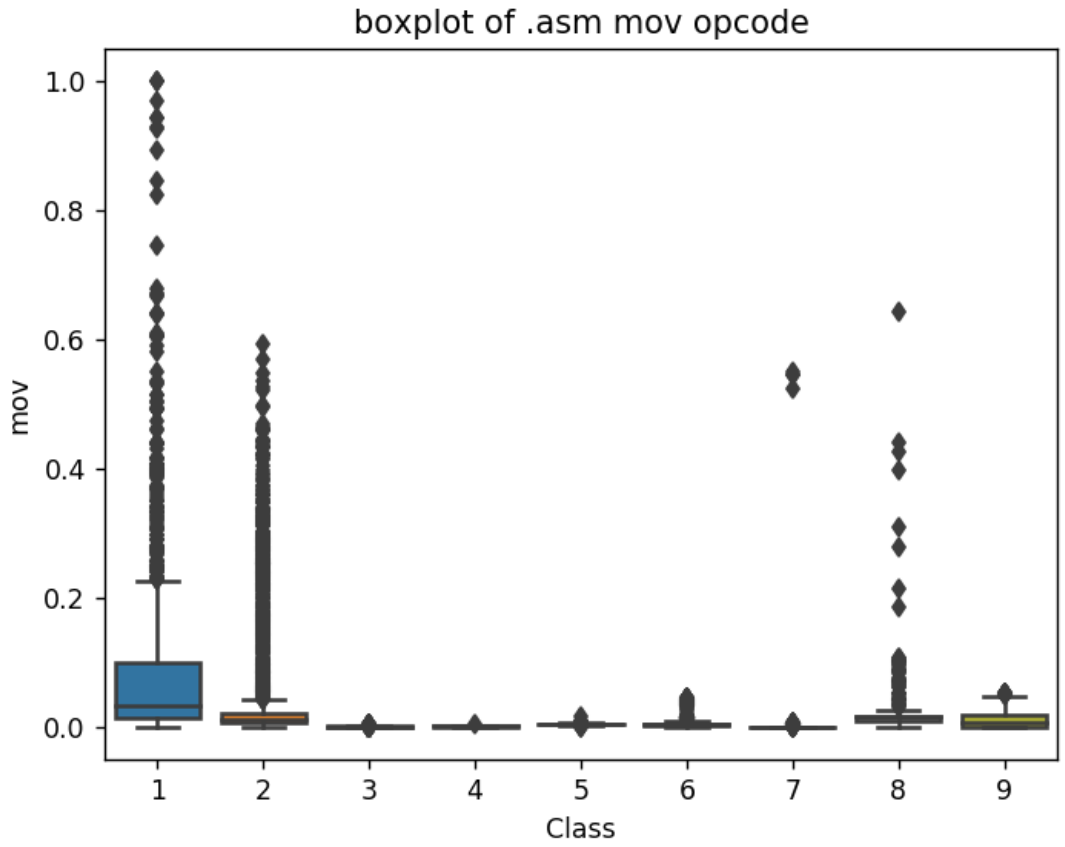
```
ax = sns.boxplot(x="Class", y="jmp", data=result_asm)
plt.title("boxplot of .asm jmp opcode")
plt.show()
```



plot between jmp and Class label
Class 1 is having frequency of 2000 approx in 75 perentile of files

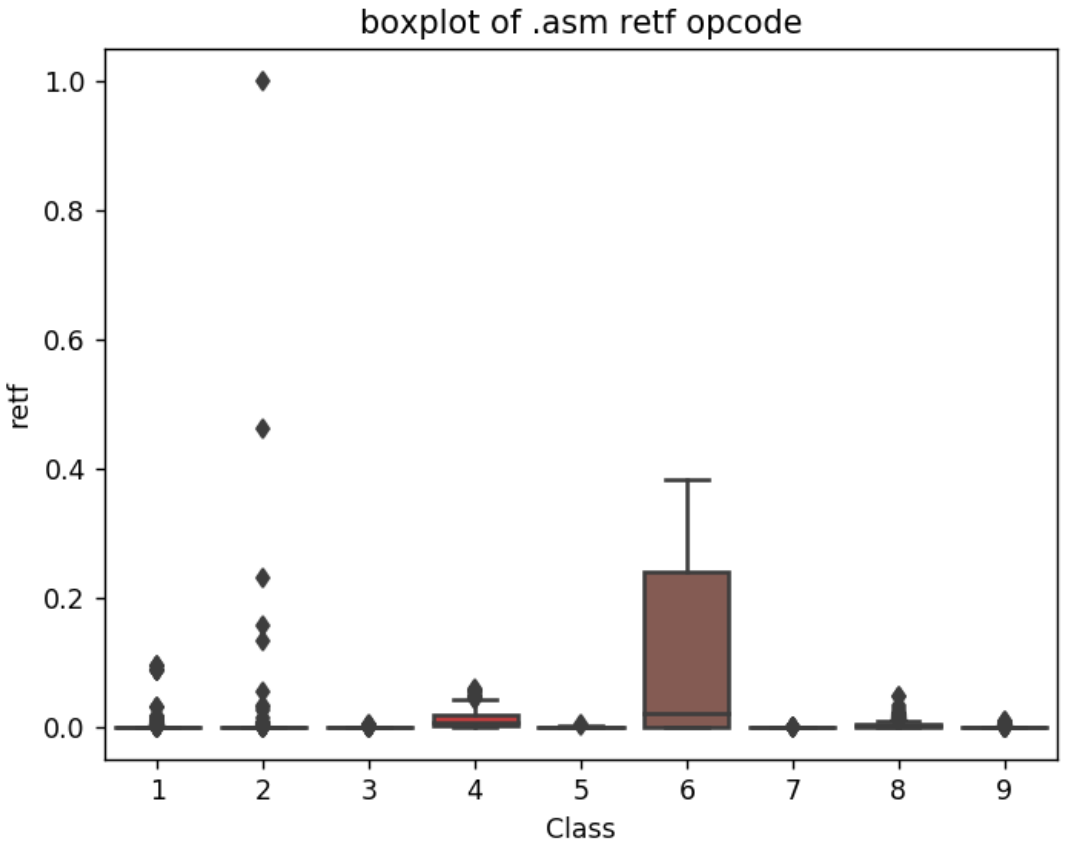
In [ ]:

```
ax = sns.boxplot(x="Class", y="mov", data=result_asm)
plt.title("boxplot of .asm mov opcode")
plt.show()
```



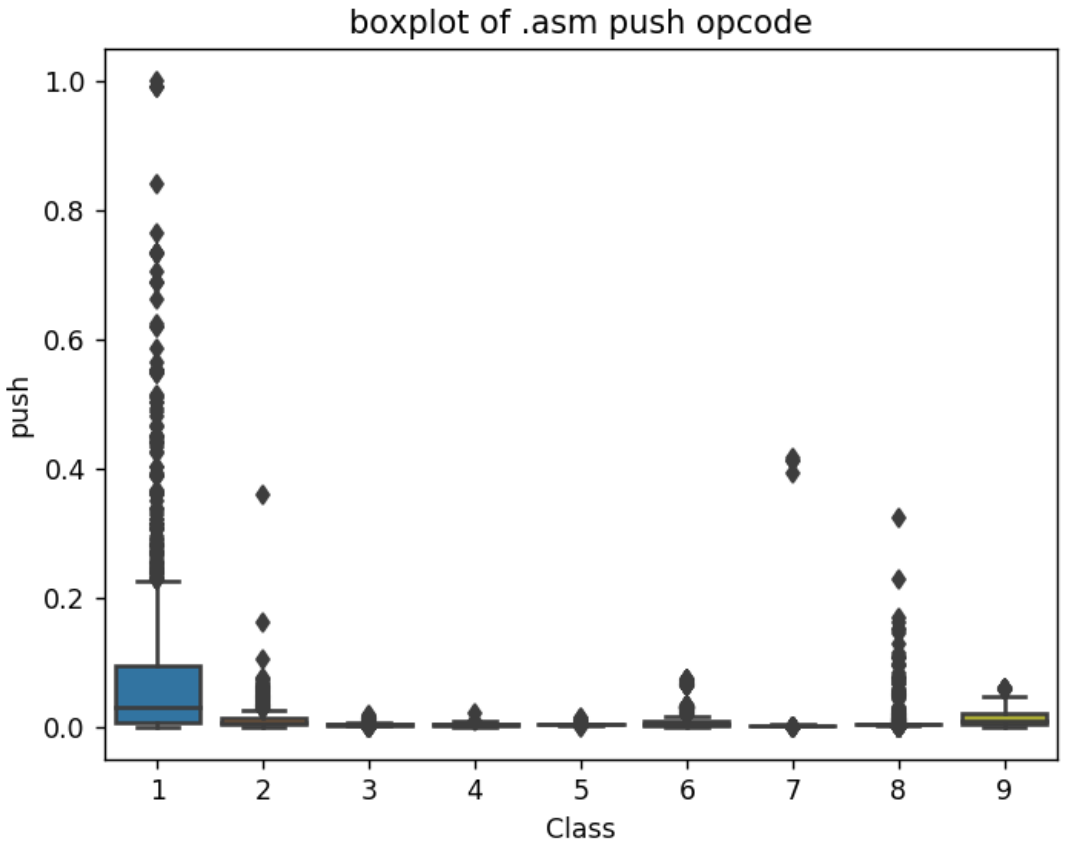
plot between Class label and mov opcode
Class 1 is having frequency of 2000 approx in 75 perentile of files

In [ ]:
```
ax = sns.boxplot(x="Class", y="retf", data=result_asm)
plt.title("boxplot of .asm retf opcode")
plt.show()
```



plot between Class label and retf
Class 6 can be easily separated with opcode retf
The frequency of retf is approx of 250.

In [ ]:
```
ax = sns.boxplot(x="Class", y="push", data=result_asm)
plt.title("boxplot of .asm push opcode")
plt.show()
```

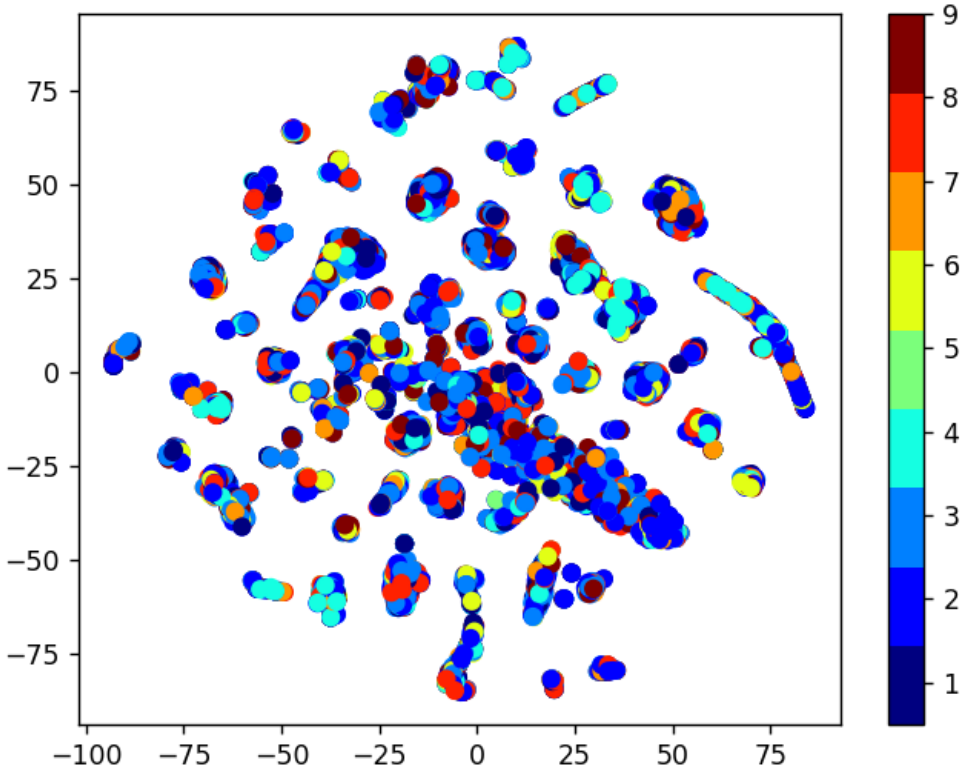


plot between push opcode and Class label
Class 1 is having 75 precentile files with push opcodes of frequency 1000

4.2.2 Multivariate Analysis on .asm file features

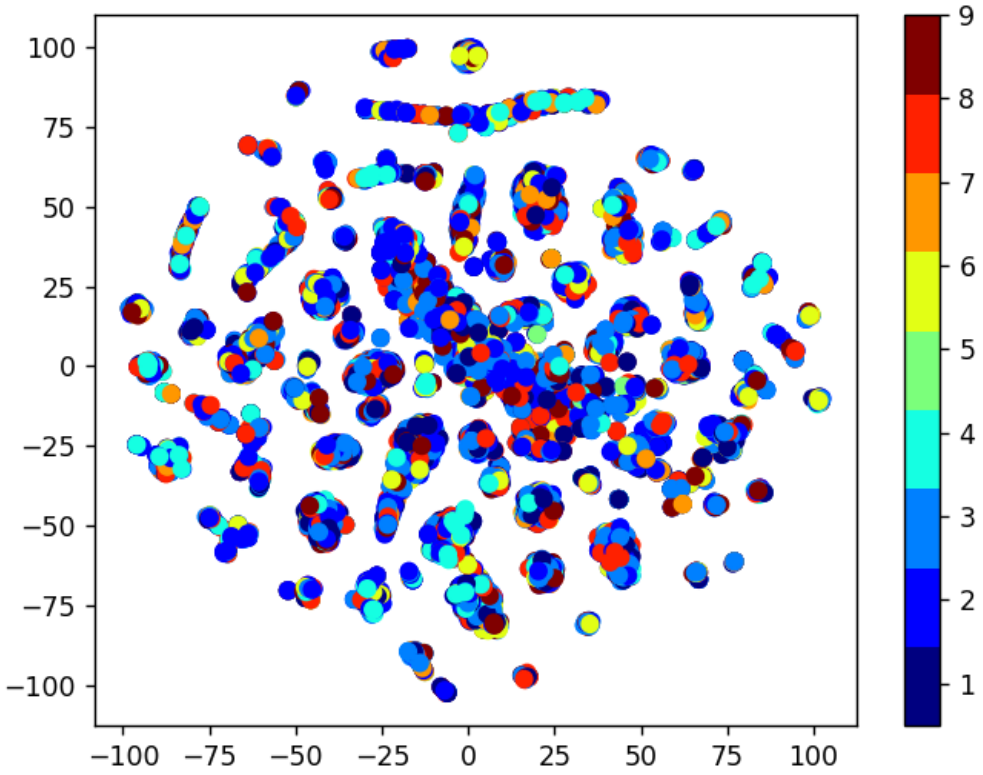
In [ ]:

```
# check out the course content for more explantion on tsne algorithm
# https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/t-distributed

#multivariate analysis on byte files
#this is with perplexity 50
xtsne=TSNE(perplexity=50)
results=xtsne.fit_transform(result_asm.drop(['ID','Class'], axis=1).fillna(0))
vis_x = results[:, 0]
vis_y = results[:, 1   ]
plt.scatter(vis_x, vis_y, c=data_y, cmap=plt.cm.get_cmap("jet", 9))
plt.colorbar(ticks=range(10))
plt.clim(0.5, 9)
plt.show()
```



In [ ]:

```
# by univariate analysis on the .asm file features we are getting very negligible infor
# 'rtn', '.BSS:' '.CODE' features, so heare we are trying multivariate analysis after r
# the plot looks very messy

xtsne=TSNE(perplexity=30)
results=xtsne.fit_transform(result_asm.drop(['ID','Class', 'rtn', '.BSS:', '.CODE','siz
vis_x = results[:, 0]
vis_y = results[:, 1]
plt.scatter(vis_x, vis_y, c=data_y, cmap=plt.cm.get_cmap("jet", 9))
plt.colorbar(ticks=range(10))
plt.clim(0.5, 9)
plt.show()
```



TSNE for asm data with perplexity 50

4.2.3 Conclusion on EDA

- **We have taken only 52 features from asm files (after reading through many blogs and research papers)**
- **The univariate analysis was done only on few important features.**
- **Take-aways**
  - **1. Class 3 can be easily separated because of the frequency of segments,opcodes and keywords being less**
  - **2. Each feature has its unique importance in separating the Class labels.**

4.3 Train and test split

In [174…
```
asm_y = result_asm['Class']
asm_x = result_asm.drop(['ID','Class','.BSS:','rtn','.CODE'], axis=1)
```

In [175…
```
X_train_asm, X_test_asm, y_train_asm, y_test_asm = train_test_split(asm_x,asm_y ,strati
X_train_asm, X_cv_asm, y_train_asm, y_cv_asm = train_test_split(X_train_asm, y_train_as
```

In [ ]:
```
print( X_cv_asm.isnull().all())
```

```
HEADER:    False
.text:     False
.Pav:      False
.idata:    False
.data:     False
.bss:      False
.rdata:    False
.edata:    False
.rsrc:     False
.tls:      False
.reloc:    False
jmp        False
mov        False
retf       False
push       False
pop        False
xor        False
retn       False
nop        False
sub        False
inc        False
dec        False
add        False
imul       False
xchg       False
or         False
shr        False
cmp        False
call       False
shl        False
ror        False
rol        False
jnb        False
jz         False
lea        False
movzx      False
.dll       False
std::      False
:dword     False
edx        False
esi        False
eax        False
ebx        False
ecx        False
edi        False
ebp        False
esp        False
eip        False
size       False
dtype: bool
```

4.4. Machine Learning models on features of .asm files

4.4.1 K-Nearest Neigbors

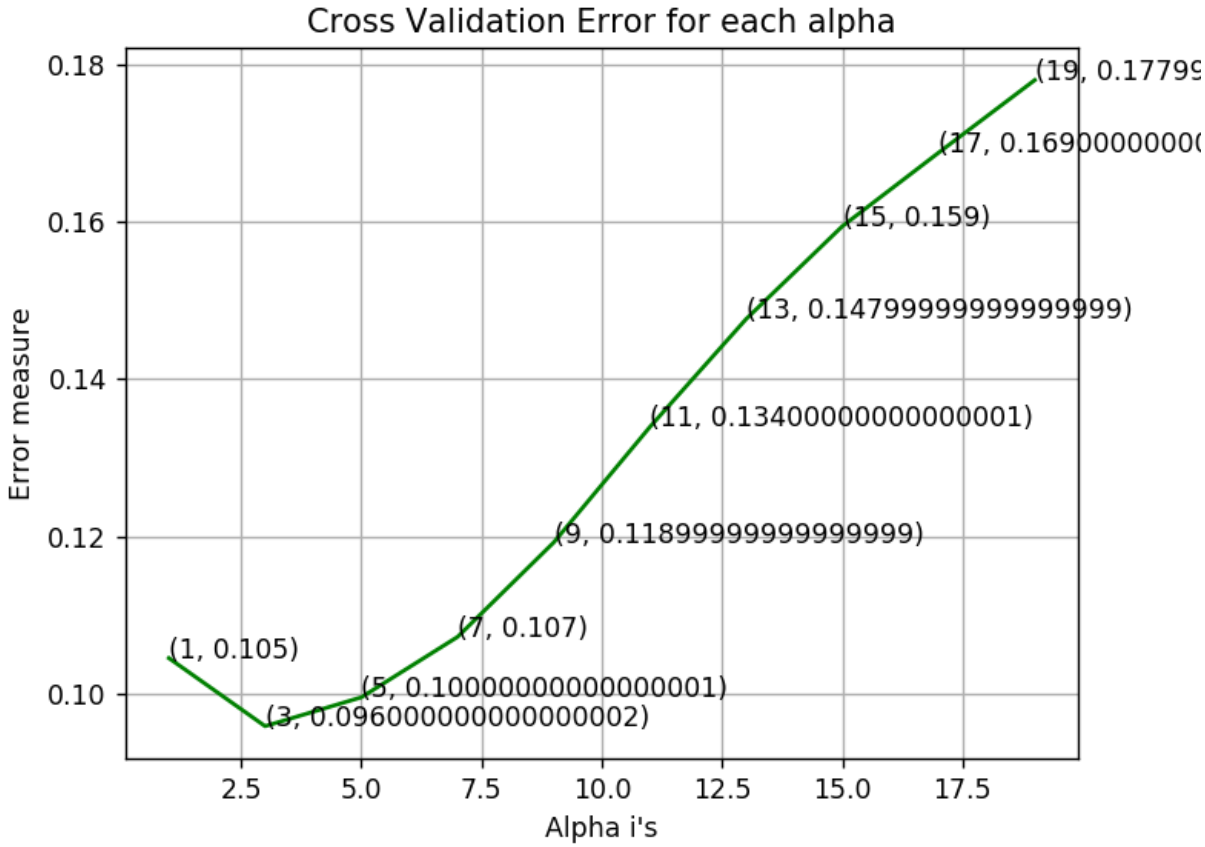
In [ ]:

```
# find more about KNeighborsClassifier() here http://scikit-learn.org/stable/modules/ge
# ------------------------
# default parameter
# KNeighborsClassifier(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30
# metric='minkowski', metric_params=None, n_jobs=1, **kwargs)
```

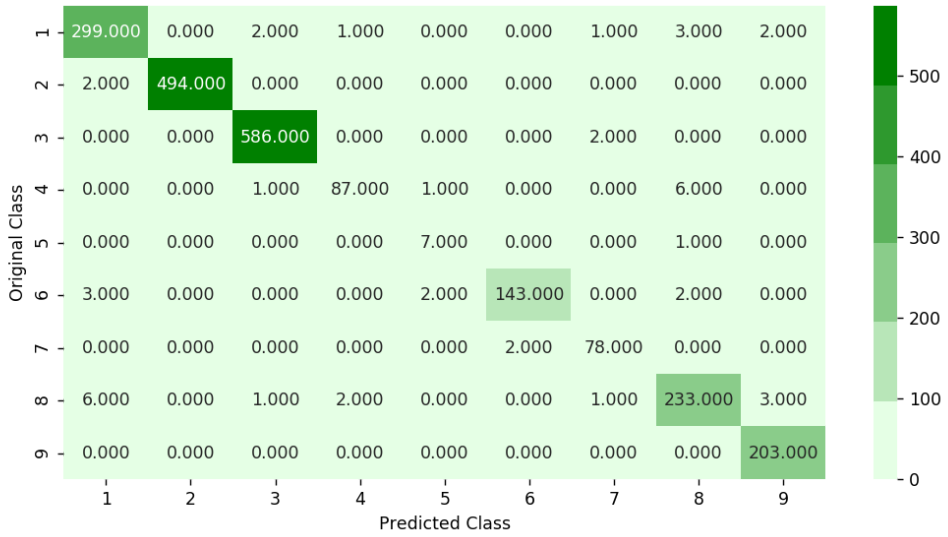
```
# methods of
# fit(X, y) : Fit the model using X as training data and y as target values
# predict(X):Predict the class labels for the provided data
# predict_proba(X):Return probability estimates for the test data X.
#--------------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/k
#--------------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight])    Fit the calibrated model
# get_params([deep])    Get parameters for this estimator.
# predict(X)    Predict the target of new samples.
# predict_proba(X)      Posterior probabilities of classification
#--------------------------------------
# video link:
#--------------------------------------

alpha = [x for x in range(1, 21,2)]
cv_log_error_array=[]
for i in alpha:
    k_cfl=KNeighborsClassifier(n_neighbors=i)
    k_cfl.fit(X_train_asm,y_train_asm)
    sig_clf = CalibratedClassifierCV(k_cfl, method="sigmoid")
    sig_clf.fit(X_train_asm, y_train_asm)
    predict_y = sig_clf.predict_proba(X_cv_asm)
    cv_log_error_array.append(log_loss(y_cv_asm, predict_y, labels=k_cfl.classes_, eps=

for i in range(len(cv_log_error_array)):
    print ('log_loss for k = ',alpha[i],'is',cv_log_error_array[i])

best_alpha = np.argmin(cv_log_error_array)

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()

k_cfl=KNeighborsClassifier(n_neighbors=alpha[best_alpha])
k_cfl.fit(X_train_asm,y_train_asm)
sig_clf = CalibratedClassifierCV(k_cfl, method="sigmoid")
sig_clf.fit(X_train_asm, y_train_asm)
pred_y=sig_clf.predict(X_test_asm)


predict_y = sig_clf.predict_proba(X_train_asm)
print ('log loss for train data',log_loss(y_train_asm, predict_y))
predict_y = sig_clf.predict_proba(X_cv_asm)
print ('log loss for cv data',log_loss(y_cv_asm, predict_y))
predict_y = sig_clf.predict_proba(X_test_asm)
```

```
print ('log loss for test data',log_loss(y_test_asm, predict_y))
plot_confusion_matrix(y_test_asm,sig_clf.predict(X_test_asm))
```

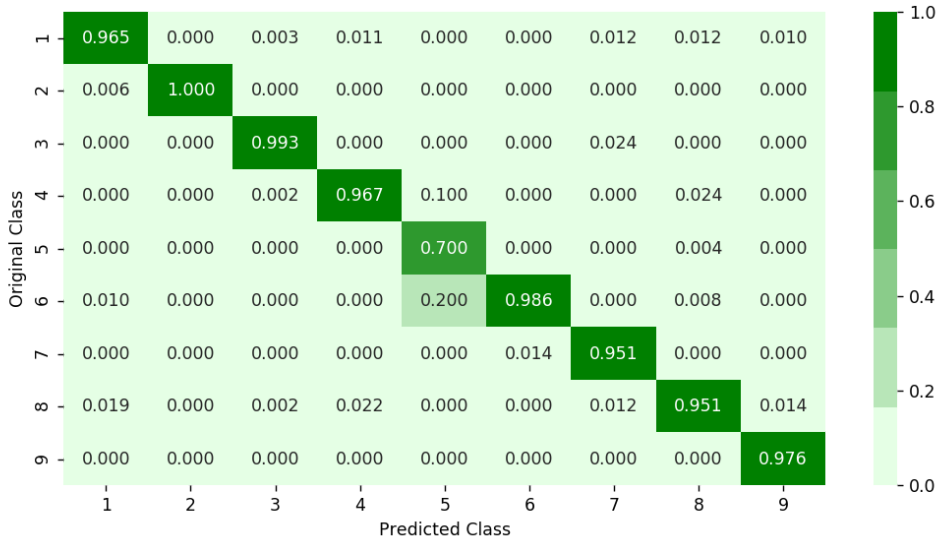
```
log_loss for k =  1 is 0.104531321344
log_loss for k =  3 is 0.0958800580948
log_loss for k =  5 is 0.0995466557335
log_loss for k =  7 is 0.107227274345
log_loss for k =  9 is 0.119239543547
log_loss for k =  11 is 0.133926642781
log_loss for k =  13 is 0.147643793967
log_loss for k =  15 is 0.159439699615
log_loss for k =  17 is 0.16878376444
log_loss for k =  19 is 0.178020728839
```




```
log loss for train data 0.0476773462198
log loss for cv data 0.0958800580948
log loss for test data 0.0894810720832
Number of misclassified points  2.02391904324
---------------------------------------------------- Confusion matrix --------------------
------------------------------
```

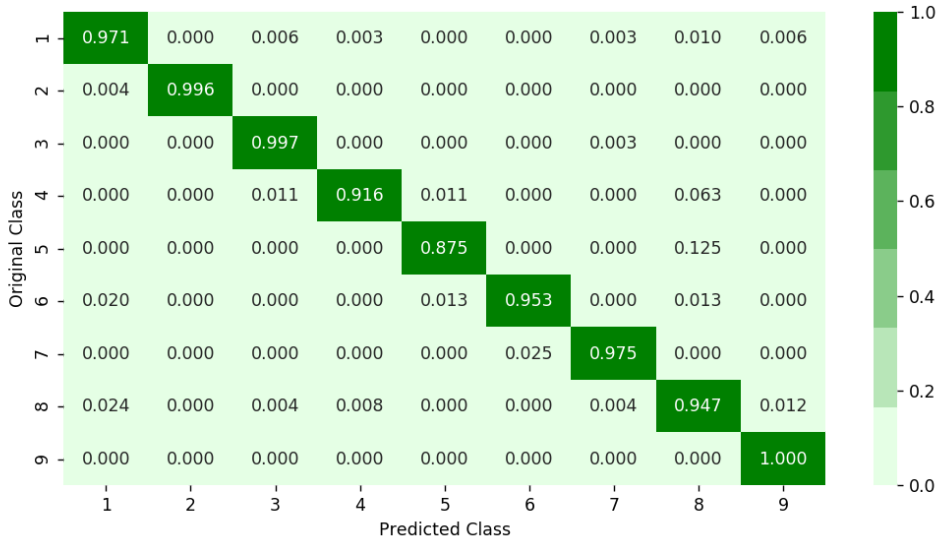


---------- Precision matrix ----------



Sum of columns in precision matrix [1. 1. 1. 1. 1. 1. 1. 1. 1.]

---------- Recall matrix ----------



Sum of rows in precision matrix [1. 1. 1. 1. 1. 1. 1. 1. 1.]

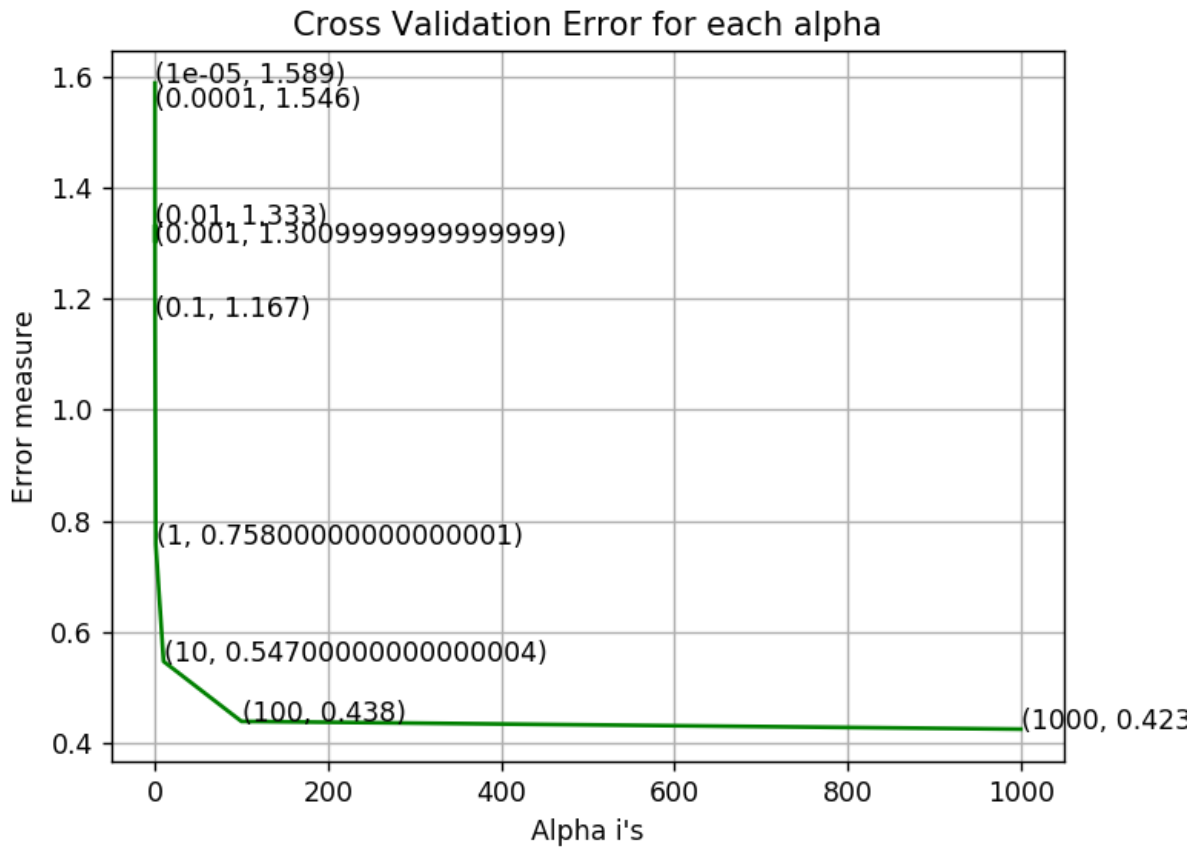
4.4.2 Logistic Regression

In [ ]:

```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/s
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='opt
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, ...])     Fit linear model with Stochastic Gradie
# predict(X)    Predict class labels for samples in X.

#-------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/g
#-------------------------------


alpha = [10 ** x for x in range(-5, 4)]
cv_log_error_array=[]
for i in alpha:
    logisticR=LogisticRegression(penalty='l2',C=i,class_weight='balanced')
    logisticR.fit(X_train_asm,y_train_asm)
    sig_clf = CalibratedClassifierCV(logisticR, method="sigmoid")
    sig_clf.fit(X_train_asm, y_train_asm)
    predict_y = sig_clf.predict_proba(X_cv_asm)
    cv_log_error_array.append(log_loss(y_cv_asm, predict_y, labels=logisticR.classes_,

for i in range(len(cv_log_error_array)):
    print ('log_loss for c = ',alpha[i],'is',cv_log_error_array[i])

best_alpha = np.argmin(cv_log_error_array)

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
```

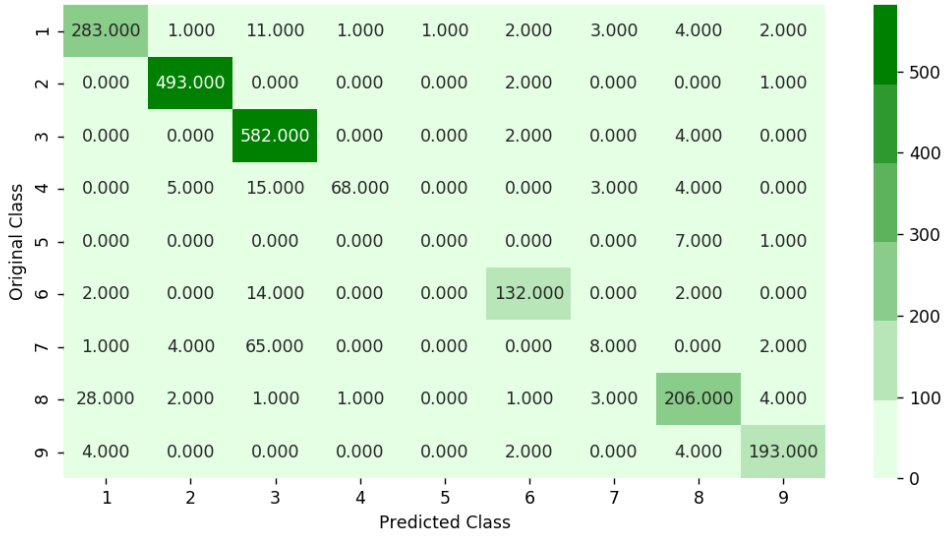
```
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()

logisticR=LogisticRegression(penalty='l2',C=alpha[best_alpha],class_weight='balanced')
logisticR.fit(X_train_asm,y_train_asm)
sig_clf = CalibratedClassifierCV(logisticR, method="sigmoid")
sig_clf.fit(X_train_asm, y_train_asm)

predict_y = sig_clf.predict_proba(X_train_asm)
print ('log loss for train data',(log_loss(y_train_asm, predict_y, labels=logisticR.cla
predict_y = sig_clf.predict_proba(X_cv_asm)
print ('log loss for cv data',(log_loss(y_cv_asm, predict_y, labels=logisticR.classes_,
predict_y = sig_clf.predict_proba(X_test_asm)
print ('log loss for test data',(log_loss(y_test_asm, predict_y, labels=logisticR.class
plot_confusion_matrix(y_test_asm,sig_clf.predict(X_test_asm))
```

```
log_loss for c =  1e-05 is 1.58867274165
log_loss for c =  0.0001 is 1.54560797884
log_loss for c =  0.001 is 1.30137786807
log_loss for c =  0.01 is 1.33317456931
log_loss for c =  0.1 is 1.16705751378
log_loss for c =  1 is 0.757667807779
log_loss for c =  10 is 0.546533939819
log_loss for c =  100 is 0.438414998062
log_loss for c =  1000 is 0.424423536526
```

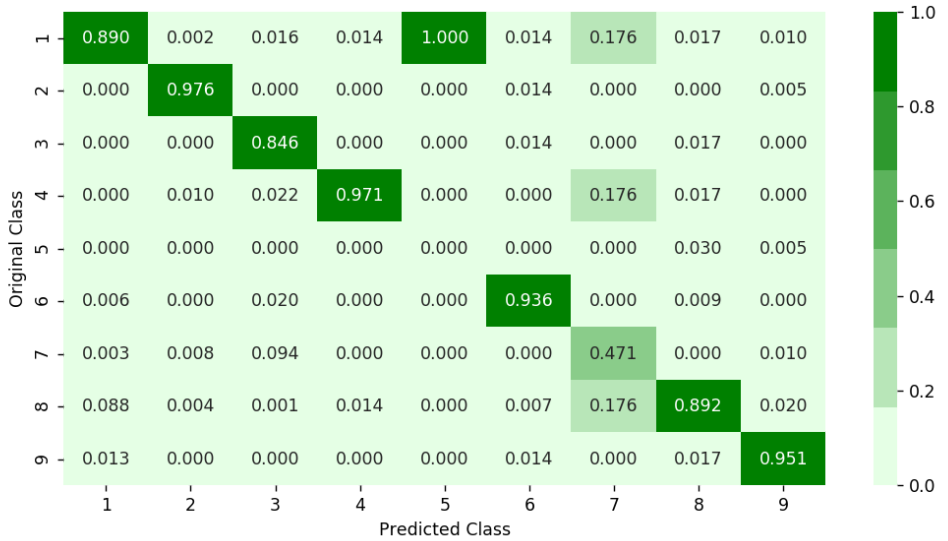



```
log loss for train data 0.396219394701
log loss for cv data 0.424423536526
log loss for test data 0.415685592517
Number of misclassified points  9.61361545538
```

---------- Confusion matrix ----------



---------- Precision matrix ----------



Sum of columns in precision matrix [1. 1. 1. 1. 1. 1. 1. 1. 1.]

---------- Recall matrix ----------

Sum of rows in precision matrix [1. 1. 1. 1. 1. 1. 1. 1. 1.]

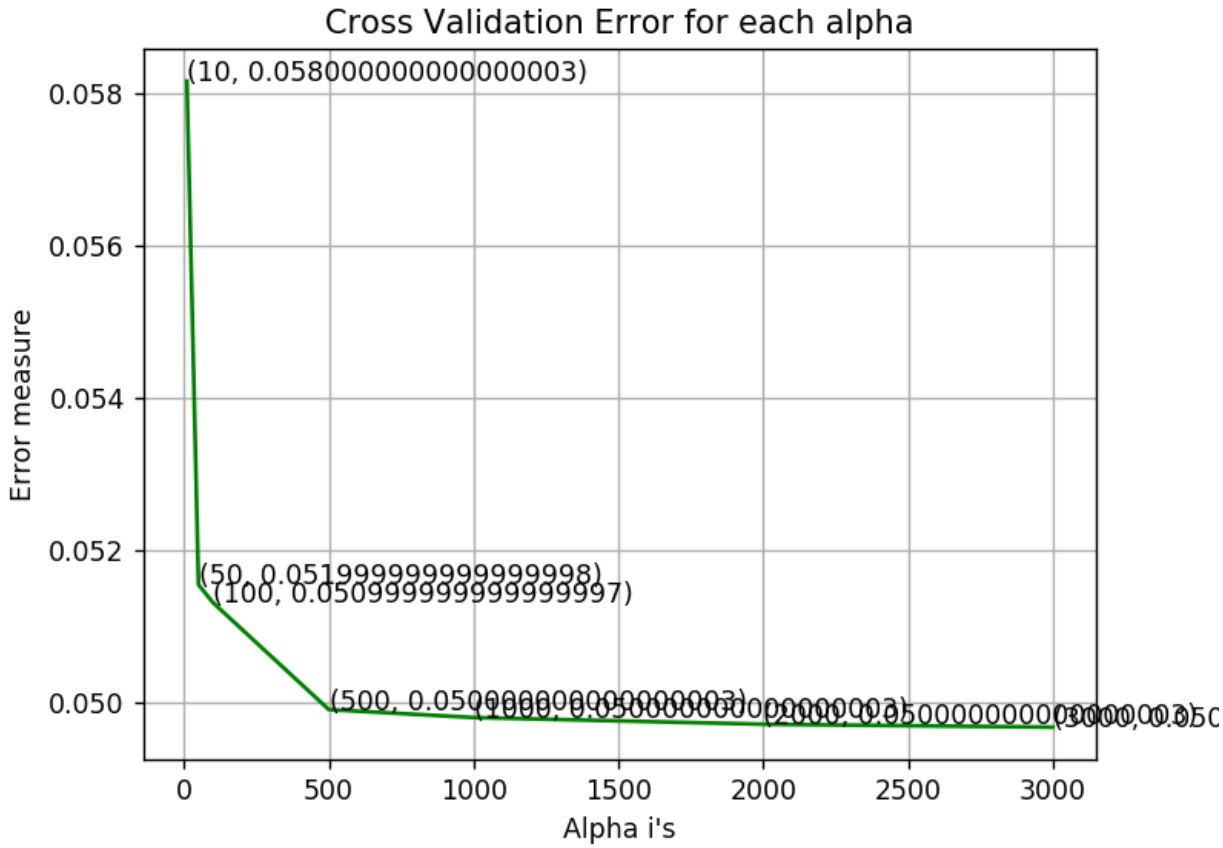
4.4.3 Random Forest Classifier

In [ ]:

```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None
# class_weight=None)

# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight])    Fit the SVM model according to the given training data.
# predict(X)    Perform classification on samples in X.
# predict_proba (X)     Perform classification on samples in X.

# some of attributes of  RandomForestClassifier()
# feature_importances_ : array of shape = [n_features]
# The feature importances (the higher, the more important the feature).

# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/r
# --------------------------------

alpha=[10,50,100,500,1000,2000,3000]
cv_log_error_array=[]
for i in alpha:
    r_cfl=RandomForestClassifier(n_estimators=i,random_state=42,n_jobs=-1)
    r_cfl.fit(X_train_asm,y_train_asm)
    sig_clf = CalibratedClassifierCV(r_cfl, method="sigmoid")
    sig_clf.fit(X_train_asm, y_train_asm)
    predict_y = sig_clf.predict_proba(X_cv_asm)
    cv_log_error_array.append(log_loss(y_cv_asm, predict_y, labels=r_cfl.classes_, eps=

for i in range(len(cv_log_error_array)):
    print ('log_loss for c = ',alpha[i],'is',cv_log_error_array[i])


best_alpha = np.argmin(cv_log_error_array)
```

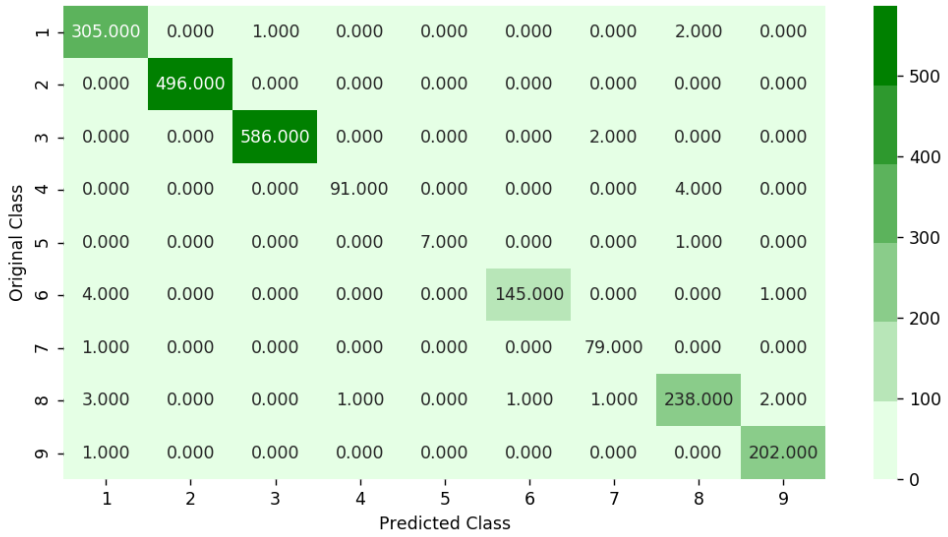
```
fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()

r_cfl=RandomForestClassifier(n_estimators=alpha[best_alpha],random_state=42,n_jobs=-1)
r_cfl.fit(X_train_asm,y_train_asm)
sig_clf = CalibratedClassifierCV(r_cfl, method="sigmoid")
sig_clf.fit(X_train_asm, y_train_asm)
predict_y = sig_clf.predict_proba(X_train_asm)
print ('log loss for train data',(log_loss(y_train_asm, predict_y, labels=sig_clf.class
predict_y = sig_clf.predict_proba(X_cv_asm)
print ('log loss for cv data',(log_loss(y_cv_asm, predict_y, labels=sig_clf.classes_, e
predict_y = sig_clf.predict_proba(X_test_asm)
print ('log loss for test data',(log_loss(y_test_asm, predict_y, labels=sig_clf.classes
plot_confusion_matrix(y_test_asm,sig_clf.predict(X_test_asm))
```

```
log_loss for c =  10 is 0.0581657906023
log_loss for c =  50 is 0.0515443148419
log_loss for c =  100 is 0.0513084973231
log_loss for c =  500 is 0.0499021761479
log_loss for c =  1000 is 0.0497972474298
log_loss for c =  2000 is 0.0497091690815
log_loss for c =  3000 is 0.0496706817633
```

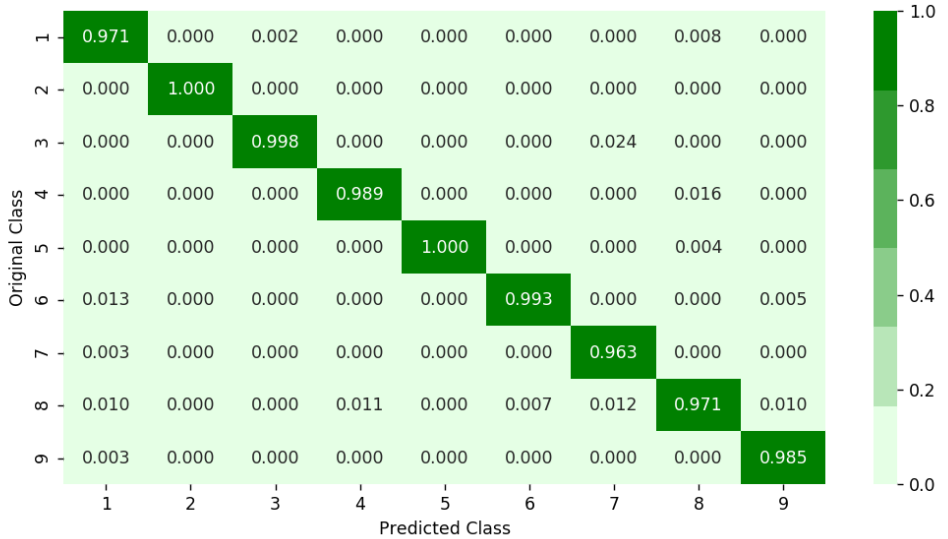



```
log loss for train data 0.0116517052676
log loss for cv data 0.0496706817633
```

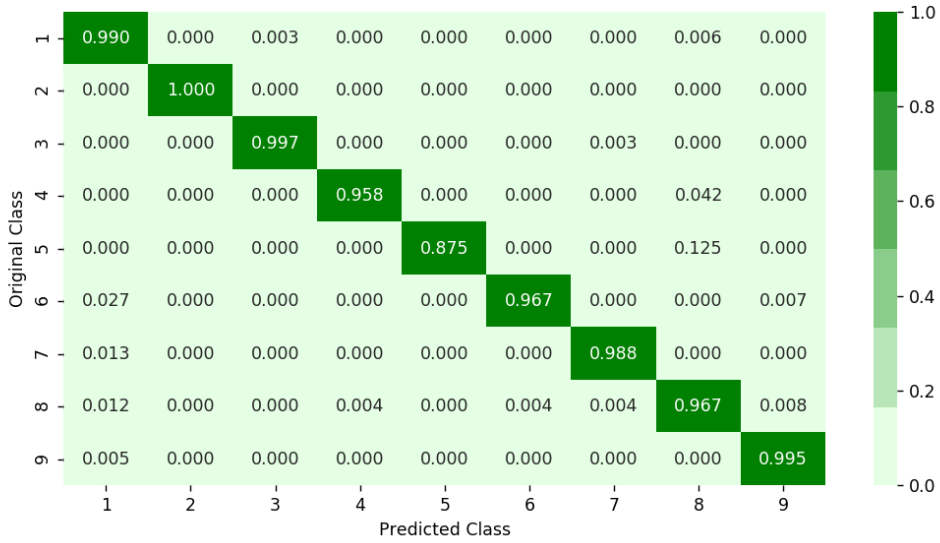
```
log loss for test data 0.0571239496453
Number of misclassified points  1.14995400184
---------------------------------------------------- Confusion matrix --------------------
------------------------------
```




```
---------------------------------------------------- Precision matrix --------------------
------------------------------
```




```
Sum of columns in precision matrix [ 1.  1.  1.  1.  1.  1.  1.  1.  1.]
---------------------------------------------------- Recall matrix ----------------------
--------------------------
```



Sum of rows in precision matrix [1. 1. 1. 1. 1. 1. 1. 1. 1.]

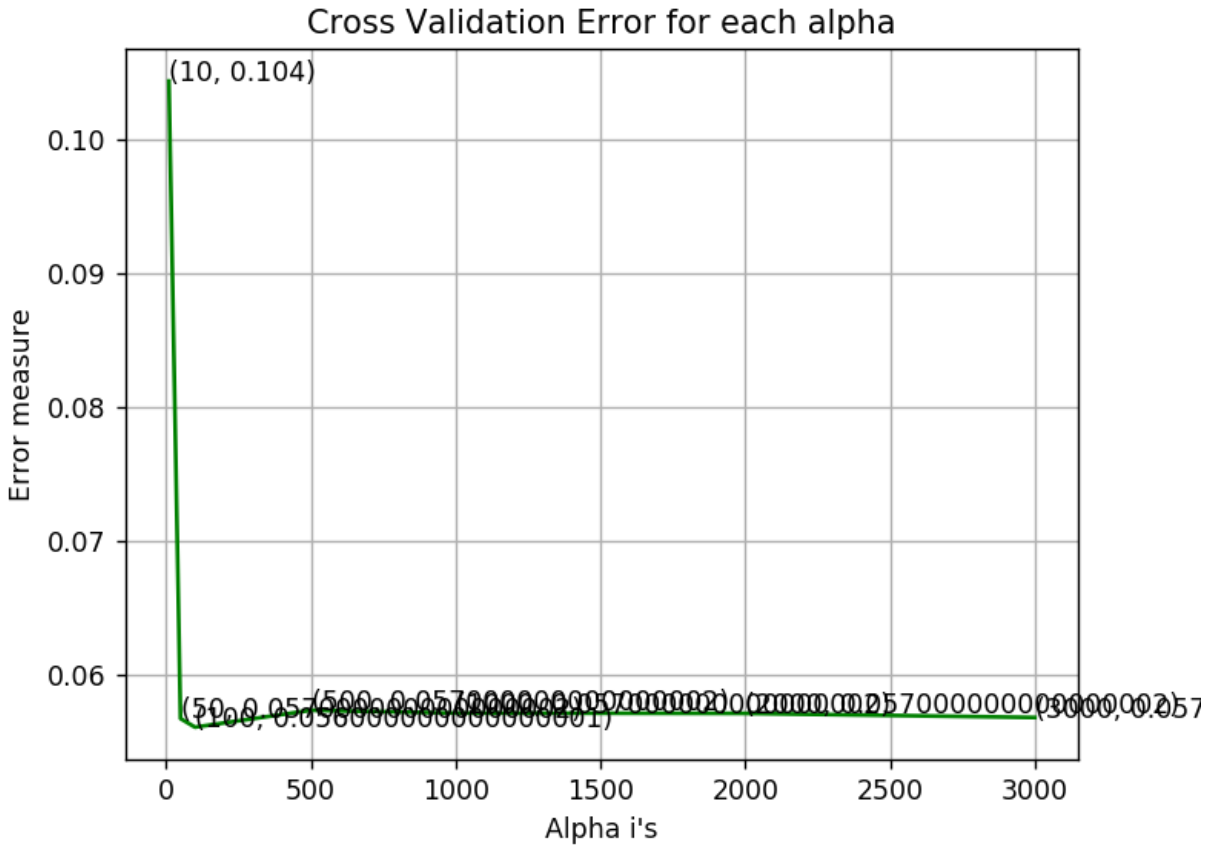
4.4.4 XgBoost Classifier

In [ ]:

```
# Training a hyper-parameter tuned Xg-Boost regressor on our train data

# find more about XGBClassifier function here http://xgboost.readthedocs.io/en/latest/p
# ------------------------
# default paramters
# class xgboost.XGBClassifier(max_depth=3, learning_rate=0.1, n_estimators=100, silent=
# objective='binary:logistic', booster='gbtree', n_jobs=1, nthread=None, gamma=0, min_c
# max_delta_step=0, subsample=1, colsample_bytree=1, colsample_bylevel=1, reg_alpha=0,
# scale_pos_weight=1, base_score=0.5, random_state=0, seed=None, missing=None, **kwargs

# some of methods of RandomForestRegressor()
# fit(X, y, sample_weight=None, eval_set=None, eval_metric=None, early_stopping_rounds=
# get_params([deep])    Get parameters for this estimator.
# predict(data, output_margin=False, ntree_limit=0) : Predict with data. NOTE: This fun
# get_score(importance_type='weight') -> get the feature importance
# -----------------------
# video link2: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/
# -----------------------

alpha=[10,50,100,500,1000,2000,3000]
cv_log_error_array=[]
for i in alpha:
    x_cfl=XGBClassifier(n_estimators=i,nthread=-1)
    x_cfl.fit(X_train_asm,y_train_asm)
    sig_clf = CalibratedClassifierCV(x_cfl, method="sigmoid")
    sig_clf.fit(X_train_asm, y_train_asm)
    predict_y = sig_clf.predict_proba(X_cv_asm)
    cv_log_error_array.append(log_loss(y_cv_asm, predict_y, labels=x_cfl.classes_, eps=

for i in range(len(cv_log_error_array)):
    print ('log_loss for c = ',alpha[i],'is',cv_log_error_array[i])


best_alpha = np.argmin(cv_log_error_array)

fig, ax = plt.subplots()
```

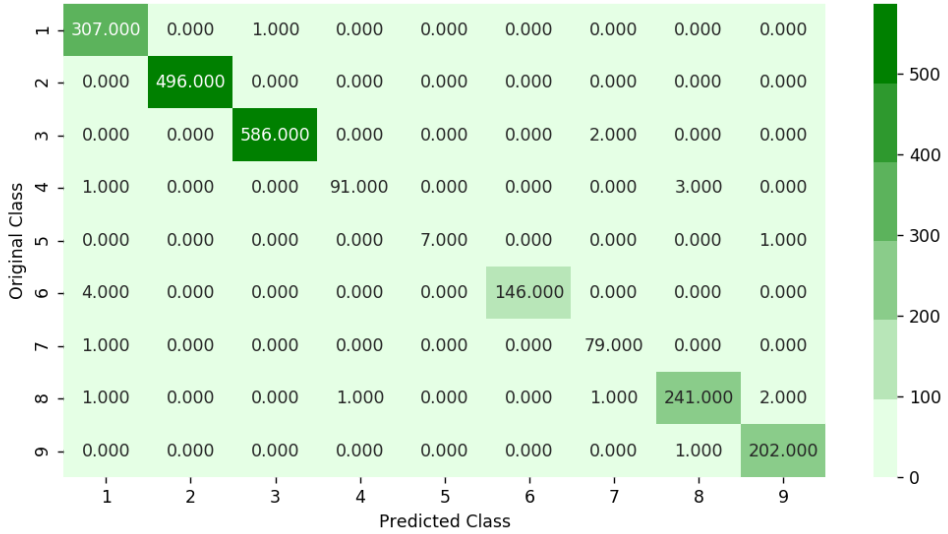
```
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()

x_cfl=XGBClassifier(n_estimators=alpha[best_alpha],nthread=-1)
x_cfl.fit(X_train_asm,y_train_asm)
sig_clf = CalibratedClassifierCV(x_cfl, method="sigmoid")
sig_clf.fit(X_train_asm, y_train_asm)

predict_y = sig_clf.predict_proba(X_train_asm)

print ('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_l
predict_y = sig_clf.predict_proba(X_cv_asm)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss
predict_y = sig_clf.predict_proba(X_test_asm)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_los
plot_confusion_matrix(y_test_asm,sig_clf.predict(X_test_asm))
```

```
log_loss for c =  10 is 0.104344888454
log_loss for c =  50 is 0.0567190635611
log_loss for c =  100 is 0.056075038646
log_loss for c =  500 is 0.057336051683
log_loss for c =  1000 is 0.0571265109903
log_loss for c =  2000 is 0.057103406781
log_loss for c =  3000 is 0.0567993215778
```

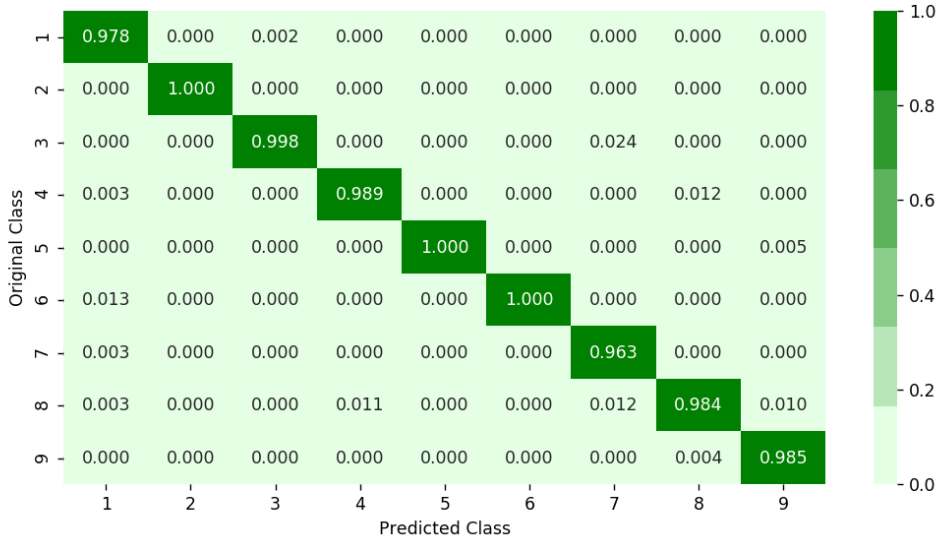



```
For values of best alpha =  100 The train log loss is: 0.0117883742574
```

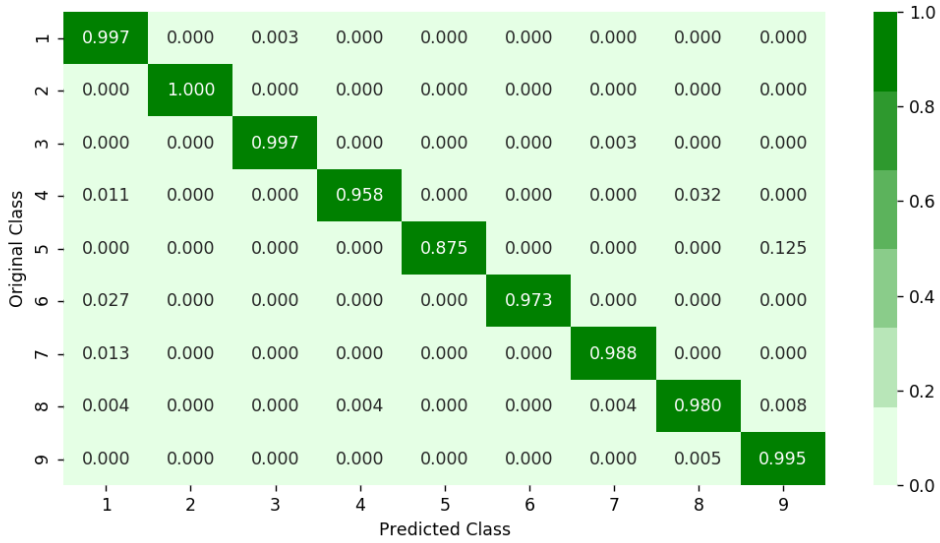
```
For values of best alpha =  100 The cross validation log loss is: 0.056075038646
For values of best alpha =  100 The test log loss is: 0.0491647763845
Number of misclassified points  0.873965041398
---------------------------------------------------- Confusion matrix -------------------
------------------------------
```




```
---------------------------------------------------- Precision matrix -------------------
------------------------------
```




```
Sum of columns in precision matrix [ 1.  1.  1.  1.  1.  1.  1.  1.  1.]
---------------------------------------------------- Recall matrix ----------------------
--------------------------
```




```
Sum of rows in precision matrix [ 1.  1.  1.  1.  1.  1.  1.  1.  1.]
```

4.4.5 Xgboost Classifier with best hyperparameters

In [ ]:
```
x_cfl=XGBClassifier()

prams={
    'learning_rate':[0.01,0.03,0.05,0.1,0.15,0.2],
     'n_estimators':[100,200,500,1000,2000],
     'max_depth':[3,5,10],
    'colsample_bytree':[0.1,0.3,0.5,1],
    'subsample':[0.1,0.3,0.5,1]
}
random_cfl=RandomizedSearchCV(x_cfl,param_distributions=prams,verbose=10,n_jobs=-1,)
random_cfl.fit(X_train_asm,y_train_asm)
```

```
Fitting 3 folds for each of 10 candidates, totalling 30 fits
[Parallel(n_jobs=-1)]: Done   2 tasks      | elapsed:    8.1s
[Parallel(n_jobs=-1)]: Done   9 tasks      | elapsed:   32.8s
[Parallel(n_jobs=-1)]: Done  19 out of  30 | elapsed:  1.1min remaining:   39.3s
[Parallel(n_jobs=-1)]: Done  23 out of  30 | elapsed:  1.3min remaining:   23.0s
[Parallel(n_jobs=-1)]: Done  27 out of  30 | elapsed:  1.4min remaining:    9.2s
[Parallel(n_jobs=-1)]: Done  30 out of  30 | elapsed:  2.3min finished
```

Out[ ]:
```
RandomizedSearchCV(cv=None, error_score='raise',
          estimator=XGBClassifier(base_score=0.5, colsample_bylevel=1, colsample_bytree=
1,
       gamma=0, learning_rate=0.1, max_delta_step=0, max_depth=3,
       min_child_weight=1, missing=None, n_estimators=100, nthread=-1,
       objective='binary:logistic', reg_alpha=0, reg_lambda=1,
       scale_pos_weight=1, seed=0, silent=True, subsample=1),
          fit_params=None, iid=True, n_iter=10, n_jobs=-1,
          param_distributions={'learning_rate': [0.01, 0.03, 0.05, 0.1, 0.15, 0.2], 'n_e
stimators': [100, 200, 500, 1000, 2000], 'max_depth': [3, 5, 10], 'colsample_bytree':
[0.1, 0.3, 0.5, 1], 'subsample': [0.1, 0.3, 0.5, 1]},
          pre_dispatch='2*n_jobs', random_state=None, refit=True,
          return_train_score=True, scoring=None, verbose=10)
```

In [ ]:
```
print (random_cfl.best_params_)
```

```
{'subsample': 1, 'n_estimators': 200, 'max_depth': 5, 'learning_rate': 0.15, 'colsample_
```

```
bytree': 0.5}
```

In [176…

```
# Training a hyper-parameter tuned Xg-Boost regressor on our train data

# find more about XGBClassifier function here http://xgboost.readthedocs.io/en/latest/p
# ------------------------
# default paramters
# class xgboost.XGBClassifier(max_depth=3, learning_rate=0.1, n_estimators=100, silent=
# objective='binary:logistic', booster='gbtree', n_jobs=1, nthread=None, gamma=0, min_c
# max_delta_step=0, subsample=1, colsample_bytree=1, colsample_bylevel=1, reg_alpha=0,
# scale_pos_weight=1, base_score=0.5, random_state=0, seed=None, missing=None, **kwargs

# some of methods of RandomForestRegressor()
# fit(X, y, sample_weight=None, eval_set=None, eval_metric=None, early_stopping_rounds=
# get_params([deep])    Get parameters for this estimator.
# predict(data, output_margin=False, ntree_limit=0) : Predict with data. NOTE: This fun
# get_score(importance_type='weight') -> get the feature importance
# -----------------------
# video link2: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/
# ------------------------

x_cfl=XGBClassifier(n_estimators=200,subsample=0.5,learning_rate=0.15,colsample_bytree=
x_cfl.fit(X_train_asm,y_train_asm)
c_cfl=CalibratedClassifierCV(x_cfl,method='sigmoid')
c_cfl.fit(X_train_asm,y_train_asm)

predict_y = c_cfl.predict_proba(X_train_asm)
print ('train loss',log_loss(y_train_asm, predict_y))
predict_y = c_cfl.predict_proba(X_cv_asm)
print ('cv loss',log_loss(y_cv_asm, predict_y))
predict_y = c_cfl.predict_proba(X_test_asm)
print ('test loss',log_loss(y_test_asm, predict_y))
```

```
[17:30:23] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[17:30:40] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[17:30:53] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[17:31:06] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[17:31:19] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[17:31:32] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
train loss 0.012075110035516543
cv loss 0.01343422215820393
test loss 0.02596457368426434
```

4.5. Machine Learning models on features of both .asm and .bytes files

4.5.1. Merging both asm and byte file features

In [177…
```
result.head()
```

Out[177…

	ID	0	1	2	3	4	5	6
0	01azqd4InC7m9JpocGv5	0.262806	0.005498	0.001567	0.002067	0.002048	0.001835	0.002058
1	01IsoiSMh5gxyDYTl4CB	0.017358	0.011737	0.004033	0.003876	0.005303	0.003873	0.004747
2	01jsnpXSAlgw6aPeDxrU	0.040827	0.013434	0.001429	0.001315	0.005464	0.005280	0.005078
3	01kcPWA9K2BOxQeS5Rju	0.009209	0.001708	0.000404	0.000441	0.000770	0.000354	0.000310
4	01SuzwMJEIXsK7A8dQbl	0.008629	0.001000	0.000168	0.000234	0.000342	0.000232	0.000148

5 rows × 760 columns

In [178…
```
result_asm.head()
```

Out[178…

	ID	HEADER:	.text:	.Pav:	.idata:	.data:	.bss:	.rdata:	.edata:
0	01kcPWA9K2BOxQeS5Rju	0.107345	0.001092	0.0	0.000761	0.000023	0.0	0.000084	0.0
1	1E93CpP60RHFNiT5Qfvn	0.096045	0.001230	0.0	0.000617	0.000019	0.0	0.000000	0.0
2	3ekVow2ajZHbTnBcsDfX	0.096045	0.000627	0.0	0.000300	0.000017	0.0	0.000038	0.0
3	3X2nY7iQaPBIWDrAZqJe	0.096045	0.000333	0.0	0.000258	0.000008	0.0	0.000000	0.0
4	46OZzdsSKDCFV8h7XWxf	0.096045	0.000590	0.0	0.000353	0.000068	0.0	0.000000	0.0

5 rows × 854 columns

In [179…
```
print(result.shape)
print(result_asm.shape)
```

```
(10868, 760)
(10868, 854)
```

In [180…
```
result_x = pd.merge(result,result_asm.drop(['Class'], axis=1),on='ID', how='left')
result_y = result_x['Class']
result_x = result_x.drop(['ID','rtn','.BSS:','.CODE','Class'], axis=1)
result_x.head()
```

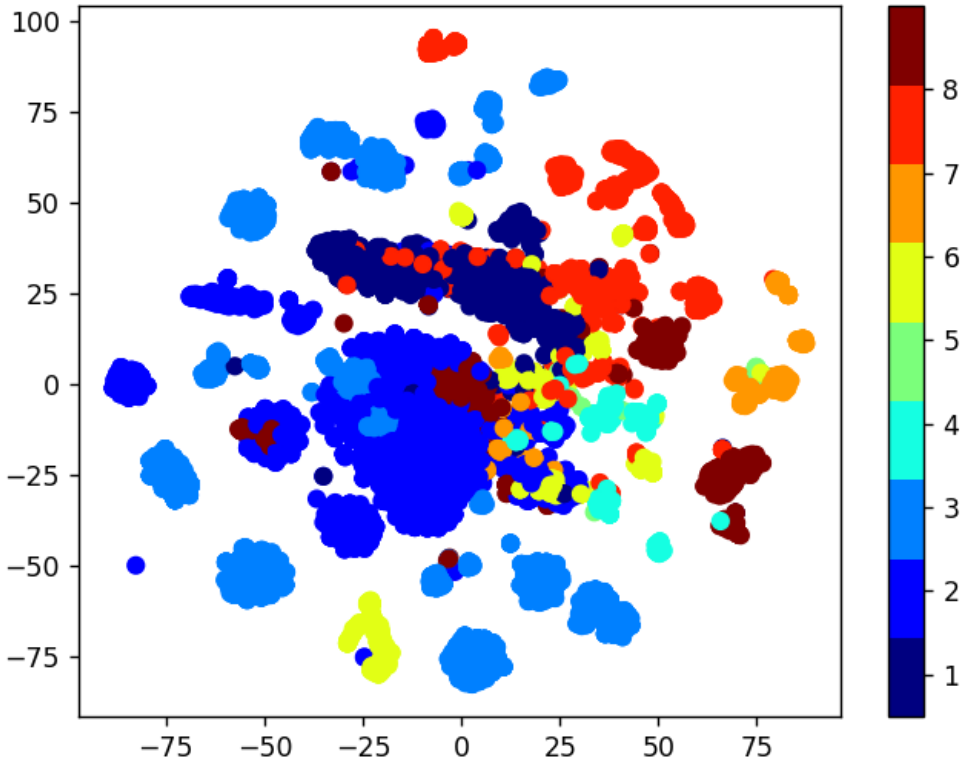
Out[180…

	0	1	2	3	4	5	6	7	8	
0	0.262806	0.005498	0.001567	0.002067	0.002048	0.001835	0.002058	0.002946	0.002638	0.003
1	0.017358	0.011737	0.004033	0.003876	0.005303	0.003873	0.004747	0.006984	0.008267	0.000
2	0.040827	0.013434	0.001429	0.001315	0.005464	0.005280	0.005078	0.002155	0.008104	0.002
3	0.009209	0.001708	0.000404	0.000441	0.000770	0.000354	0.000310	0.000481	0.000959	0.000
4	0.008629	0.001000	0.000168	0.000234	0.000342	0.000232	0.000148	0.000229	0.000376	0.000

5 rows × 1607 columns

4.5.2. Multivariate Analysis on final fearures

In [ ]:
```
xtsne=TSNE(perplexity=50)
results=xtsne.fit_transform(result_x, axis=1))
vis_x = results[:, 0]
vis_y = results[:, 1]
plt.scatter(vis_x, vis_y, c=result_y, cmap=plt.cm.get_cmap("jet", 9))
plt.colorbar(ticks=range(9))
plt.clim(0.5, 9)
plt.show()
```



4.5.3. Train and Test split

In [145…
```
X_train, X_test_merge, y_train, y_test_merge = train_test_split(result_x, result_y,stra
X_train_merge, X_cv_merge, y_train_merge, y_cv_merge = train_test_split(X_train, y_trai
```

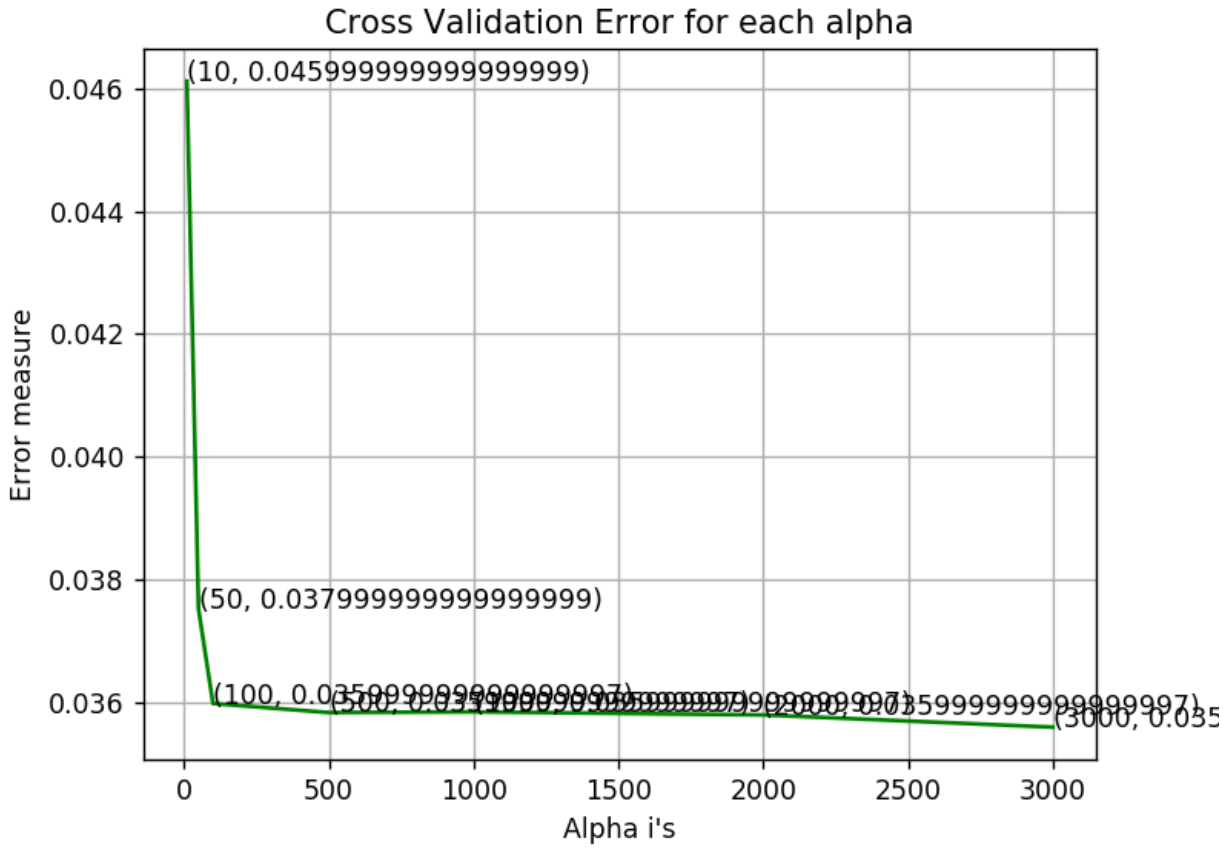
4.5.4. Random Forest Classifier on final features

In [ ]:
```
# ------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=
```

```
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None
# class_weight=None)

# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight])    Fit the SVM model according to the given training data.
# predict(X)    Perform classification on samples in X.
# predict_proba (X)     Perform classification on samples in X.

# some of attributes of  RandomForestClassifier()
# feature_importances_ : array of shape = [n_features]
# The feature importances (the higher, the more important the feature).

# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/r
# --------------------------------

alpha=[10,50,100,500,1000,2000,3000]
cv_log_error_array=[]
from sklearn.ensemble import RandomForestClassifier
for i in alpha:
    r_cfl=RandomForestClassifier(n_estimators=i,random_state=42,n_jobs=-1)
    r_cfl.fit(X_train_merge,y_train_merge)
    sig_clf = CalibratedClassifierCV(r_cfl, method="sigmoid")
    sig_clf.fit(X_train_merge, y_train_merge)
    predict_y = sig_clf.predict_proba(X_cv_merge)
    cv_log_error_array.append(log_loss(y_cv_merge, predict_y, labels=r_cfl.classes_, ep

for i in range(len(cv_log_error_array)):
    print ('log_loss for c = ',alpha[i],'is',cv_log_error_array[i])


best_alpha = np.argmin(cv_log_error_array)

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


r_cfl=RandomForestClassifier(n_estimators=alpha[best_alpha],random_state=42,n_jobs=-1)
r_cfl.fit(X_train_merge,y_train_merge)
sig_clf = CalibratedClassifierCV(r_cfl, method="sigmoid")
sig_clf.fit(X_train_merge, y_train_merge)

predict_y = sig_clf.predict_proba(X_train_merge)
print ('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_l
predict_y = sig_clf.predict_proba(X_cv_merge)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss
predict_y = sig_clf.predict_proba(X_test_merge)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_los
```

```
log_loss for c =  10 is 0.0461221662017
log_loss for c =  50 is 0.0375229563452
log_loss for c =  100 is 0.0359765822455
log_loss for c =  500 is 0.0358291883873
```

```
log_loss for c =  1000 is 0.0358403093496
log_loss for c =  2000 is 0.0357908022178
log_loss for c =  3000 is 0.0355909487962
```




```
For values of best alpha =  3000 The train log loss is: 0.0166267614753
For values of best alpha =  3000 The cross validation log loss is: 0.0355909487962
For values of best alpha =  3000 The test log loss is: 0.0401141303589
```

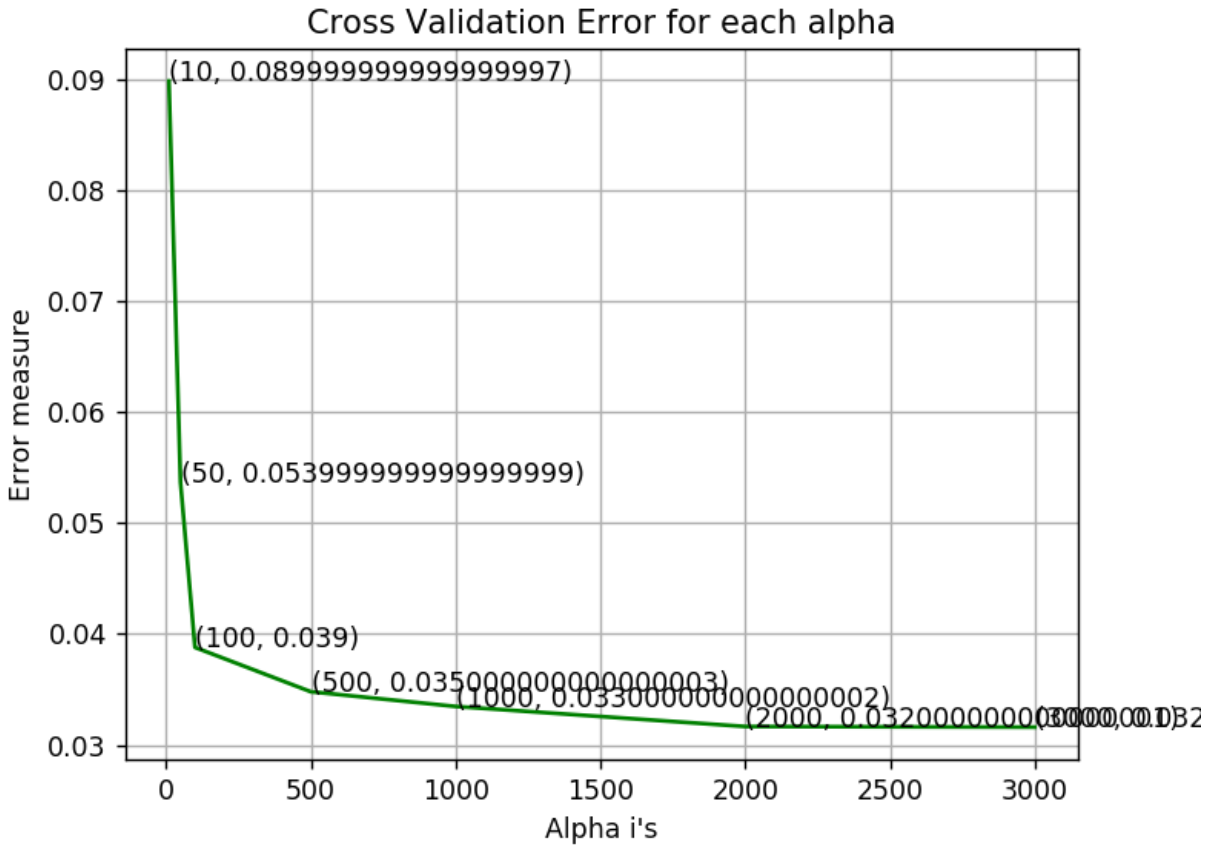
4.5.5. XgBoost Classifier on final features

In [ ]:

```
# Training a hyper-parameter tuned Xg-Boost regressor on our train data

# find more about XGBClassifier function here http://xgboost.readthedocs.io/en/latest/p
# ------------------------
# default paramters
# class xgboost.XGBClassifier(max_depth=3, learning_rate=0.1, n_estimators=100, silent=
# objective='binary:logistic', booster='gbtree', n_jobs=1, nthread=None, gamma=0, min_c
# max_delta_step=0, subsample=1, colsample_bytree=1, colsample_bylevel=1, reg_alpha=0,
# scale_pos_weight=1, base_score=0.5, random_state=0, seed=None, missing=None, **kwargs

# some of methods of RandomForestRegressor()
# fit(X, y, sample_weight=None, eval_set=None, eval_metric=None, early_stopping_rounds=
# get_params([deep])    Get parameters for this estimator.
# predict(data, output_margin=False, ntree_limit=0) : Predict with data. NOTE: This fun
# get_score(importance_type='weight') -> get the feature importance
# -----------------------
# video link2: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/
# -----------------------

alpha=[10,50,100,500,1000,2000,3000]
cv_log_error_array=[]
```

```
for i in alpha:
    x_cfl=XGBClassifier(n_estimators=i)
    x_cfl.fit(X_train_merge,y_train_merge)
    sig_clf = CalibratedClassifierCV(x_cfl, method="sigmoid")
    sig_clf.fit(X_train_merge, y_train_merge)
    predict_y = sig_clf.predict_proba(X_cv_merge)
    cv_log_error_array.append(log_loss(y_cv_merge, predict_y, labels=x_cfl.classes_, ep

for i in range(len(cv_log_error_array)):
    print ('log_loss for c = ',alpha[i],'is',cv_log_error_array[i])


best_alpha = np.argmin(cv_log_error_array)

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()

x_cfl=XGBClassifier(n_estimators=3000,nthread=-1)
x_cfl.fit(X_train_merge,y_train_merge,verbose=True)
sig_clf = CalibratedClassifierCV(x_cfl, method="sigmoid")
sig_clf.fit(X_train_merge, y_train_merge)

predict_y = sig_clf.predict_proba(X_train_merge)
print ('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_l
predict_y = sig_clf.predict_proba(X_cv_merge)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss
predict_y = sig_clf.predict_proba(X_test_merge)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_los
```

```
log_loss for c =  10 is 0.0898979446265
log_loss for c =  50 is 0.0536946658041
log_loss for c =  100 is 0.0387968186177
log_loss for c =  500 is 0.0347960327293
log_loss for c =  1000 is 0.0334668083237
log_loss for c =  2000 is 0.0316569078846
log_loss for c =  3000 is 0.0315972694477
```




```
For values of best alpha =  3000 The train log loss is: 0.0111918809342
For values of best alpha =  3000 The cross validation log loss is: 0.0315972694477
For values of best alpha =  3000 The test log loss is: 0.0323978515915
```

4.5.5. XgBoost Classifier on final features with best hyper parameters using Random search

In [ ]:
```
x_cfl=XGBClassifier()

prams={
    'learning_rate':[0.01,0.03,0.05,0.1,0.15,0.2],
     'n_estimators':[100,200,500,1000,2000],
     'max_depth':[3,5,10],
    'colsample_bytree':[0.1,0.3,0.5,1],
    'subsample':[0.1,0.3,0.5,1]
}
random_cfl=RandomizedSearchCV(x_cfl,cv=3,param_distributions=prams,verbose=10,n_jobs=-1
random_cfl.fit(X_train_merge, y_train_merge)
```

In [ ]:
```
print (random_cfl.best_params_)
```

In [184…
```
print(X_train_merge)
```

```
             0         1         2         3         4         5         6  \
10788  0.053591  0.140907  0.051447  0.049009  0.006894  0.006455  0.006681
3563   0.019035  0.022562  0.008735  0.011115  0.091996  0.007669  0.006763
9057   0.019440  0.020104  0.006573  0.005835  0.262155  0.006141  0.005888
341    0.008819  0.001881  0.000416  0.000497  0.000597  0.000312  0.000321
```

```
6699   0.010120  0.000949  0.000236  0.000285  0.000316  0.000199  0.000198
...         ...       ...       ...       ...       ...       ...       ...
2948   0.032439  0.050067  0.018680  0.018361  0.021367  0.020204  0.019991
5063   0.019021  0.018282  0.007268  0.006787  0.007841  0.006572  0.096015
1910   0.023272  0.003617  0.000680  0.000927  0.000964  0.000527  0.000471
5480   0.003701  0.005955  0.001711  0.001742  0.001929  0.001796  0.001626
7702   0.009236  0.010714  0.001087  0.020502  0.006127  0.000111  0.000076

              7         8         9  ...     p 790     p 791     p 792  \
10788  0.010055  0.009622  0.000317  ...  0.062241  0.062241  0.057328
3563   0.012717  0.012817  0.000273  ...  0.062241  0.062241  0.057328
9057   0.011070  0.011068  0.000189  ...  0.062241  0.062241  0.057328
341    0.000594  0.001931  0.000746  ...  0.062241  0.062241  0.057328
6699   0.000416  0.000327  0.000444  ...  0.062241  0.062241  0.057328
...         ...       ...       ...  ...       ...       ...       ...
2948   0.034057  0.030931  0.029914  ...  0.062241  0.062241  0.057328
5063   0.013444  0.010947  0.000251  ...  0.062241  0.062241  0.057328
1910   0.001094  0.001364  0.001124  ...  0.062241  0.062241  0.057328
5480   0.002861  0.002778  0.003402  ...  0.062241  0.062241  0.057328
7702   0.000078  0.007816  0.000120  ...  0.062241  0.062241  0.057328

          p 793     p 794     p 795     p 796     p 797     p 798     p 799
10788  0.057328  0.057328  0.056236  0.056236  0.056236  0.056782  0.056782
3563   0.057328  0.057328  0.056236  0.056236  0.056236  0.056782  0.056782
9057   0.057328  0.057328  0.056236  0.056236  0.056236  0.056782  0.056782
341    0.057328  0.057328  0.056236  0.056236  0.056236  0.056782  0.056782
6699   0.057328  0.057328  0.056236  0.056236  0.056236  0.056782  0.056782
...         ...       ...       ...       ...       ...       ...       ...
2948   0.057328  0.057328  0.056236  0.056236  0.056236  0.056782  0.056782
5063   0.057328  0.057328  0.056236  0.056236  0.056236  0.056782  0.056782
1910   0.057328  0.057328  0.056236  0.056236  0.056236  0.056782  0.056782
5480   0.057328  0.057328  0.056236  0.056236  0.056236  0.056782  0.056782
7702   0.057328  0.057328  0.056236  0.056236  0.056236  0.056782  0.056782

[6955 rows x 1607 columns]
```

In [ ]:

```
# find more about XGBClassifier function here http://xgboost.readthedocs.io/en/latest/p
# -------------------------
# default paramters
# class xgboost.XGBClassifier(max_depth=3, learning_rate=0.1, n_estimators=100, silent=
# objective='binary:logistic', booster='gbtree', n_jobs=1, nthread=None, gamma=0, min_c
# max_delta_step=0, subsample=1, colsample_bytree=1, colsample_bylevel=1, reg_alpha=0,
# scale_pos_weight=1, base_score=0.5, random_state=0, seed=None, missing=None, **kwargs

# some of methods of RandomForestRegressor()
# fit(X, y, sample_weight=None, eval_set=None, eval_metric=None, early_stopping_rounds=
# get_params([deep])    Get parameters for this estimator.
# predict(data, output_margin=False, ntree_limit=0) : Predict with data. NOTE: This fun
# get_score(importance_type='weight') -> get the feature importance
# -----------------------
# video link2: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/
# -----------------------

x_cfl=XGBClassifier(n_estimators=1000,max_depth=10,learning_rate=0.15,colsample_bytree=
x_cfl.fit(X_train_merge,y_train_merge,verbose=True)
sig_clf = CalibratedClassifierCV(x_cfl, method="sigmoid")
sig_clf.fit(X_train_merge, y_train_merge)

predict_y = sig_clf.predict_proba(X_train_merge)
print ('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_l
predict_y = sig_clf.predict_proba(X_cv_merge)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss
```

```
predict_y = sig_clf.predict_proba(X_test_merge)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_los
plot_confusion_matrix(y_test_asm,sig_clf.predict(X_test_merge))
```

```
For values of best alpha =  3000 The train log loss is: 0.0121922832297
For values of best alpha =  3000 The cross validation log loss is: 0.0344955487471
For values of best alpha =  3000 The test log loss is: 0.0317041132442
```

In [181…

```
###Assignment XGBClassifier with byte bi gram features and pixel intensity features of
import warnings
warnings.filterwarnings('ignore')

x_cfl=XGBClassifier(n_estimators=2000,nthread=-1)
x_cfl.fit(X_train_merge,y_train_merge,verbose=True)
sig_clf = CalibratedClassifierCV(x_cfl, method="sigmoid")
sig_clf.fit(X_train_merge, y_train_merge)

print("SCORES=========================================================================
predict_y = sig_clf.predict_proba(X_train_merge)
print ("The train log loss is:",log_loss(y_train_merge, predict_y))
predict_y = sig_clf.predict_proba(X_cv_merge)
print("The cross validation log loss is:",log_loss(y_cv_merge, predict_y))
predict_y = sig_clf.predict_proba(X_test_merge)
print("The test log loss is:",log_loss(y_test_merge, predict_y))
# plot_confusion_matrix(y_test_asm,sig_clf.predict(X_test_merge))
```

```
[17:31:46] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[17:35:16] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[17:38:05] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[17:40:53] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[17:43:42] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
[17:46:30] WARNING: ../src/learner.cc:1061: Starting in XGBoost 1.3.0, the default evalu
ation metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlog
loss'. Explicitly set eval_metric if you'd like to restore the old behavior.
SCORES=========================================================================
==============================
The train log loss is: 0.012461212503547403
The cross validation log loss is: 0.02556270620405006
The test log loss is: 0.03048635666193773
```

STEPS DONE IN THIS ASSIGHMENT :

1. **Created Bigrams of byte files.**
2. **Picked top 500 bigrams using feature importance of RandomForest.**
3. **Combined unigrams and bigrams to get a log loss of 0.037 dropped from 0.078. (Model used XGBOOST)**
4. **Asmfiles - converted asm files to image files and extracted top 800 pixel values and normalized them.**

5. **Combined these pixel values with unigrams to get a log loss of 0.026 dropped from 0.048. (Model used XGBOOST)**
6. **Combined all features of bytes and asm files to get a log loss of 0.030 dropped from 0.032. (Model used XGBOOST)**

5. Assignments

1. **Add bi-grams on byte files and improve the log-loss**
2. **Watch the video (video) and include pixel intensity features to improve the logloss**

```
1. you need to donwload the train from kaggle, which is of size ~17GB,
after extracting it will occupy ~128GB data your dirve

2. if you are having computation power limitations, you can try using
google colab, with GPU option enabled (you can search for how to enable
GPU in colab) or you can work with the Google Cloud, check this tutorials
by one of our student: https://www.youtube.com/channel/UCRH_z-
oM0LROvHPe_KYR4Wg (we suggest you to use GCP over Colab)

3. To Extract the .7z file in google cloud, once after you upload the
file into server, in your ipython notebook create a new cell and write
thess commands
  a. !sudo apt-get install p7zip
  b. !7z x file_name.7z -o path/where/you/want/to/extract

https://askubuntu.com/a/341637
```