Personalized cancer diagnosis

1. Business Problem

1.1. Description

Source: https://www.kaggle.com/c/msk-redefining-cancer-treatment/

Data: Memorial Sloan Kettering Cancer Center (MSKCC)

Download training_variants.zip and training_text.zip from Kaggle.

Context:

Source: https://www.kaggle.com/c/msk-redefining-cancer-treatment/discussion/35336#198462

Problem statement :

Classify the given genetic variations/mutations based on evidence from text-based clinical literature.

1.2. Source/Useful Links

Some articles and reference blogs about the problem statement

1. https://www.forbes.com/sites/matthewherper/2017/06/03/a-new-cancer-drug-helped-almost-everyone-who-took-it-almost-heres-what-it-teaches-us/#2a44ee2f6b25
2. https://www.youtube.com/watch?v=UwbuW7oK8rk
3. https://www.youtube.com/watch?v=qxXRKVompI8

1.3. Real-world/Business objectives and constraints.

- No low-latency requirement.
- Interpretability is important.
- Errors can be very costly.
- Probability of a data-point belonging to each class is needed.

2. Machine Learning Problem Formulation

2.1. Data

2.1.1. Data Overview

- Source: https://www.kaggle.com/c/msk-redefining-cancer-treatment/data
- We have two data files: one conatins the information about the genetic mutations and the other contains the clinical evidence (text) that human experts/pathologists use to classify the genetic mutations.
- Both these data files are have a common column called ID
- Data file's information:
  - training_variants (ID , Gene, Variations, Class)
  - training_text (ID, Text)

2.1.2. Example Data Point

training_variants

```
ID,Gene,Variation,Class
0,FAM58A,Truncating Mutations,1
1,CBL,W802*,2
2,CBL,Q249E,2
...
```

training_text

```
ID,Text
0||Cyclin-dependent kinases (CDKs) regulate a variety of fundamental cellular processes. CDK10 stands out as one of the last orphan CDKs for which no activating cyclin has been identified and no kinase activity revealed. Previous work has shown that CDK10 silencing increases ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2)-driven activation of the MAPK pathway, which confers tamoxifen resistance to breast cancer cells. The precise mechanisms by which CDK10 modulates ETS2 activity, and more generally the functions of CDK10, remain elusive. Here we demonstrate that CDK10 is a cyclin-dependent kinase by identifying cyclin M as an activating cyclin. Cyclin M, an orphan cyclin, is the product of FAM58A, whose mutations cause STAR syndrome, a human developmental anomaly whose features include toe syndactyly, telecanthus, and anogenital and renal malformations. We show that STAR syndrome-associated cyclin M mutants are unable to interact with CDK10. Cyclin M silencing phenocopies CDK10 silencing in increasing c-Raf and in conferring tamoxifen resistance to breast cancer cells. CDK10/cyclin M phosphorylates ETS2 in vitro, and in cells it positively controls ETS2 degradation by the proteasome. ETS2 protein levels are increased in cells derived from a STAR patient, and this increase is attributable to decreased cyclin M levels. Altogether, our results reveal an additional regulatory mechanism for ETS2, which plays key roles in cancer and development. They also shed light on the molecular mechanisms underlying STAR syndrome.Cyclin-dependent kinases (CDKs) play a pivotal role in the control of a number of fundamental cellular processes (1). The human genome contains 21 genes encoding proteins that can be considered as members of the CDK family owing to their sequence similarity with bona fide CDKs, those known to be activated by cyclins (2). Although discovered almost 20 y ago (3, 4), CDK10 remains one of the two CDKs without an identified cyclin partner. This knowledge gap has largely impeded the exploration of its biological functions. CDK10 can act as a positive cell cycle regulator in some cells (5, 6) or as a tumor suppressor in others (7, 8). CDK10 interacts with the ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2) transcription factor and inhibits its transcriptional activity through an unknown mechanism (9). CDK10 knockdown derepresses ETS2, which increases the expression of the c-Raf protein kinase, activates the MAPK pathway, and induces resistance of MCF7 cells to tamoxifen (6). ...
```

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

2.2.1. Type of Machine Learning Problem

There are nine different classes a genetic mutation can be classified into => Multi class classification problem

2.2.2. Performance Metric

Source: https://www.kaggle.com/c/msk-redefining-cancer-treatment#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:

- Interpretability
- Class probabilities are needed.
- Penalize the errors in class probabilites => Metric is Log-loss.
- No Latency constraints.

2.3. Train, CV and Test Datasets

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

3. Exploratory Data Analysis

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import re
import time
import warnings
import numpy as np
from nltk.corpus import stopwords
from sklearn.decomposition import TruncatedSVD
from sklearn.preprocessing import normalize
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.manifold import TSNE
import seaborn as sns
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix
from sklearn.metrics.classification import accuracy_score, log_loss
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import SGDClassifier
from imblearn.over_sampling import SMOTE
from collections import Counter
from scipy.sparse import hstack
from sklearn.multiclass import OneVsRestClassifier
from sklearn.svm import SVC
#from sklearn.cross_validation import StratifiedKFold
from collections import Counter, defaultdict
from sklearn.calibration import CalibratedClassifierCV
from sklearn.naive_bayes import MultinomialNB
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
import math
from sklearn.metrics import normalized_mutual_info_score
from sklearn.ensemble import RandomForestClassifier
warnings.filterwarnings("ignore")

from mlxtend.classifier import StackingClassifier

from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
```

3.1. Reading Data

3.1.1. Reading Gene and Variation Data

In [2]:

```
data = pd.read_csv('training_variants')
print('Number of data points : ', data.shape[0])
print('Number of features : ', data.shape[1])
print('Features : ', data.columns.values)
data.head()
```

```
Number of data points :  3321
Number of features :  4
Features :  ['ID' 'Gene' 'Variation' 'Class']
```

Out[2]:

	ID	Gene	Variation	Class

	ID	Gene	Variation	Class
0	0	FAM58A	Truncating Mutations	1
1	1	CBL	W802*	2
2	2	CBL	Q249E	2
3	3	CBL	N454D	3
4	4	CBL	L399V	4

training/training_variants is a comma separated file containing the description of the genetic mutations used for training.
Fields are

- **ID :** the id of the row used to link the mutation to the clinical evidence
- **Gene :** the gene where this genetic mutation is located
- **Variation :** the aminoacid change for this mutations
- **Class :** 1-9 the class this genetic mutation has been classified on

3.1.2. Reading Text Data

In [3]:

```
# note the seprator in this file
data_text =pd.read_csv("training_text",sep="\|\|",engine="python",names=["ID","TEXT"],skiprows=1)
print('Number of data points : ', data_text.shape[0])
print('Number of features : ', data_text.shape[1])
print('Features : ', data_text.columns.values)
data_text.head()
```

```
Number of data points :  3321
Number of features :  2
Features :  ['ID' 'TEXT']
```

Out[3]:

	ID	TEXT
0	0	Cyclin-dependent kinases (CDKs) regulate a var...
1	1	Abstract Background Non-small cell lung canc...
2	2	Abstract Background Non-small cell lung canc...
3	3	Recent evidence has demonstrated that acquired...
4	4	Oncogenic mutations in the monomeric Casitas B...

3.1.3. Preprocessing of text

In [4]:

```
# loading stop words from nltk library
stop_words = set(stopwords.words('english'))


def nlp_preprocessing(total_text, index, column):
    if type(total_text) is not int:
        string = ""
        # replace every special char with space
        total_text = re.sub('[^a-zA-Z0-9\n]', ' ', total_text)
        # replace multiple spaces with single space
        total_text = re.sub('\s+',' ', total_text)
        # converting all the chars into lower-case.
        total_text = total_text.lower()

        for word in total_text.split():
        # if the word is a not a stop word then retain that word from the data
            if not word in stop_words:
                string += word + " "
        data_text[column][index] = string
```

In [5]:

```
#text processing stage.
start_time = time.clock()
for index, row in data_text.iterrows():
    if type(row['TEXT']) is str:
        nlp_preprocessing(row['TEXT'], index, 'TEXT')
    else:
        print("there is no text description for id:",index)
print('Time took for preprocessing the text :',time.clock() - start_time, "seconds")
```

```
there is no text description for id: 1109
there is no text description for id: 1277
there is no text description for id: 1407
there is no text description for id: 1639
there is no text description for id: 2755
Time took for preprocessing the text : 141.639947 seconds
```

In [6]:

```
#merging both gene_variations and text data based on ID
result = pd.merge(data, data_text,on='ID', how='left')
result.head()
```

Out[6]:

	ID	Gene	Variation	Class	TEXT
0	0	FAM58A	Truncating Mutations	1	cyclin dependent kinases cdks regulate variety...
1	1	CBL	W802*	2	abstract background non small cell lung cancer...
2	2	CBL	Q249E	2	abstract background non small cell lung cancer...
3	3	CBL	N454D	3	recent evidence demonstrated acquired uniparen...
4	4	CBL	L399V	4	oncogenic mutations monomeric casitas b lineag...

In [7]:

```
result[result.isnull().any(axis=1)]
```

Out[7]:

	ID	Gene	Variation	Class	TEXT
1109	1109	FANCA	S1088F	1	NaN
1277	1277	ARID5B	Truncating Mutations	1	NaN
1407	1407	FGFR3	K508M	6	NaN
1639	1639	FLT1	Amplification	6	NaN
2755	2755	BRAF	G596C	7	NaN

In [8]:

```
result.loc[result['TEXT'].isnull(),'TEXT'] = result['Gene'] +' '+result['Variation']
```

In [9]:

```
result[result['ID']==1109]
```

Out[9]:

	ID	Gene	Variation	Class	TEXT

	ID	Gene	Variation	Class	TEXT
1109	1109	FANCA	S1088F	1	FANCA S1088F

3.1.4. Test, Train and Cross Validation Split

3.1.4.1. Splitting data into train, test and cross validation (64:20:16)

In [10]:

```
y_true = result['Class'].values
result.Gene      = result.Gene.str.replace('\s+', '_')
result.Variation = result.Variation.str.replace('\s+', '_')
# split the data into test and train by maintaining same distribution of output varaible 'y_true' [stratify=y_true]
X_1, X_test, y_1, y_test = train_test_split(result, y_true, stratify=y_true, test_size=0.2)
# split the train data into train and cross validation by maintaining same distribution of output varaible 'y_train' [stratify=y_train]
X_train, X_cv, y_train, y_cv = train_test_split(X_1, y_1, stratify=y_1, test_size=0.2)
```

We split the data into train, test and cross validation data sets, preserving the ratio of class distribution in the original data set

In [11]:

```
print('Number of data points in train data:', X_train.shape[0])
print('Number of data points in test data:', X_test.shape[0])
print('Number of data points in cross validation data:', X_cv.shape[0])
```

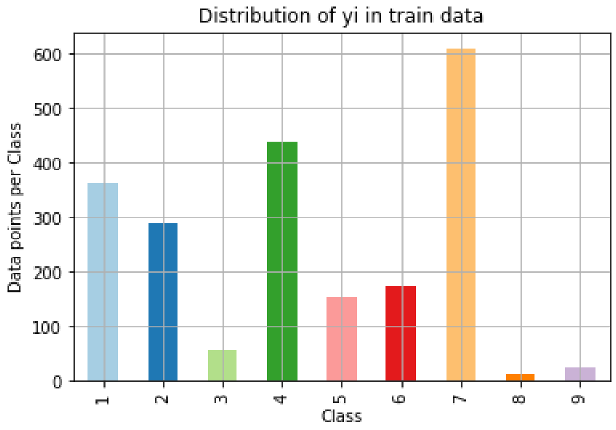
```
Number of data points in train data: 2124
Number of data points in test data: 665
Number of data points in cross validation data: 532
```

3.1.4.2. Distribution of y_i's in Train, Test and Cross Validation datasets

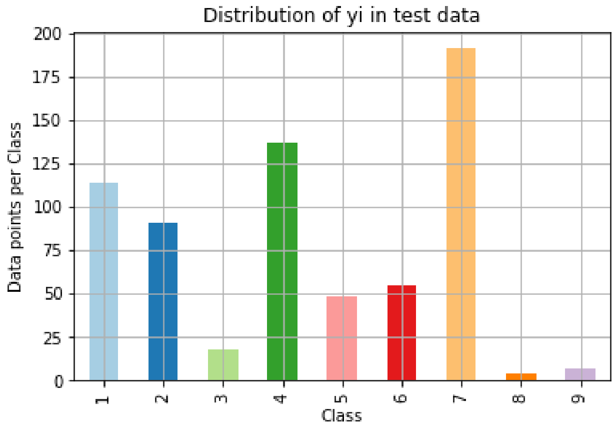
In [12]:

```
#https://stackoverflow.com/questions/50802475/valueerror-invalid-rgba-argument-rgbkymc
# it returns a dict, keys as class labels and values as the number of data points in that class
train_class_distribution = X_train['Class'].value_counts().sort_index()
test_class_distribution = X_test['Class'].value_counts().sort_index()
cv_class_distribution = X_cv['Class'].value_counts().sort_index()

paired_colors = plt.cm.Paired(range(len(train_class_distribution)))
train_class_distribution.plot(kind='bar',color=paired_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], '(', np.round((train_class_distribution.values[i]/X_train.shape[0]*100), 3), '%)')


print('-'*80)
paired_colors = plt.cm.Paired(range(len(test_class_distribution)))
test_class_distribution.plot(kind='bar',color=paired_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], '(', np.rou
nd((test_class_distribution.values[i]/X_test.shape[0]*100), 3), '%)')

print('-'*80)
paired_colors = plt.cm.Paired(range(len(cv_class_distribution)))
cv_class_distribution.plot(kind='bar',color=paired_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], '(', np.round
((cv_class_distribution.values[i]/X_cv.shape[0]*100), 3), '%)')
```

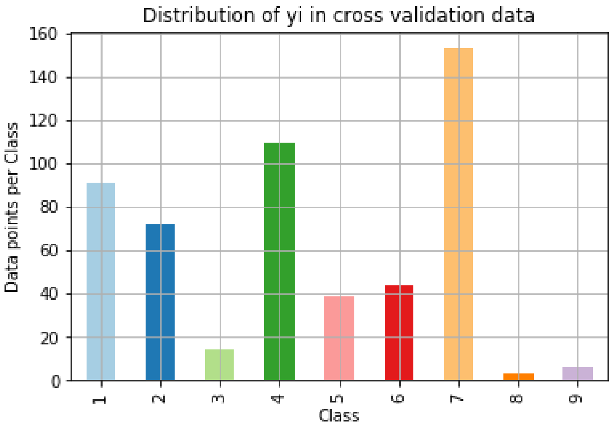


```
Number of data points in class 7 : 609 ( 28.672 %)
Number of data points in class 4 : 439 ( 20.669 %)
Number of data points in class 1 : 363 ( 17.09 %)
Number of data points in class 2 : 289 ( 13.606 %)
Number of data points in class 6 : 176 ( 8.286 %)
Number of data points in class 5 : 155 ( 7.298 %)
Number of data points in class 3 : 57 ( 2.684 %)
Number of data points in class 9 : 24 ( 1.13 %)
Number of data points in class 8 : 12 ( 0.565 %)
--------------------------------------------------------------------------------
```



```
Number of data points in class 7 : 191 ( 28.722 %)
Number of data points in class 4 : 137 ( 20.602 %)
Number of data points in class 1 : 114 ( 17.143 %)
Number of data points in class 2 : 91 ( 13.684 %)
Number of data points in class 6 : 55 ( 8.271 %)
Number of data points in class 5 : 48 ( 7.218 %)
Number of data points in class 3 : 18 ( 2.707 %)
Number of data points in class 9 : 7 ( 1.053 %)
Number of data points in class 8 : 4 ( 0.602 %)
```

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



```
Number of data points in class 7 : 153 ( 28.759 %)
Number of data points in class 4 : 110 ( 20.677 %)
Number of data points in class 1 : 91 ( 17.105 %)
Number of data points in class 2 : 72 ( 13.534 %)
Number of data points in class 6 : 44 ( 8.271 %)
Number of data points in class 5 : 39 ( 7.331 %)
Number of data points in class 3 : 14 ( 2.632 %)
Number of data points in class 9 : 6 ( 1.128 %)
Number of data points in class 8 : 3 ( 0.564 %)
```

3.2 Prediction using a 'Random' Model

In a 'Random' Model, we generate the NINE class probabilites randomly such that they sum to 1.

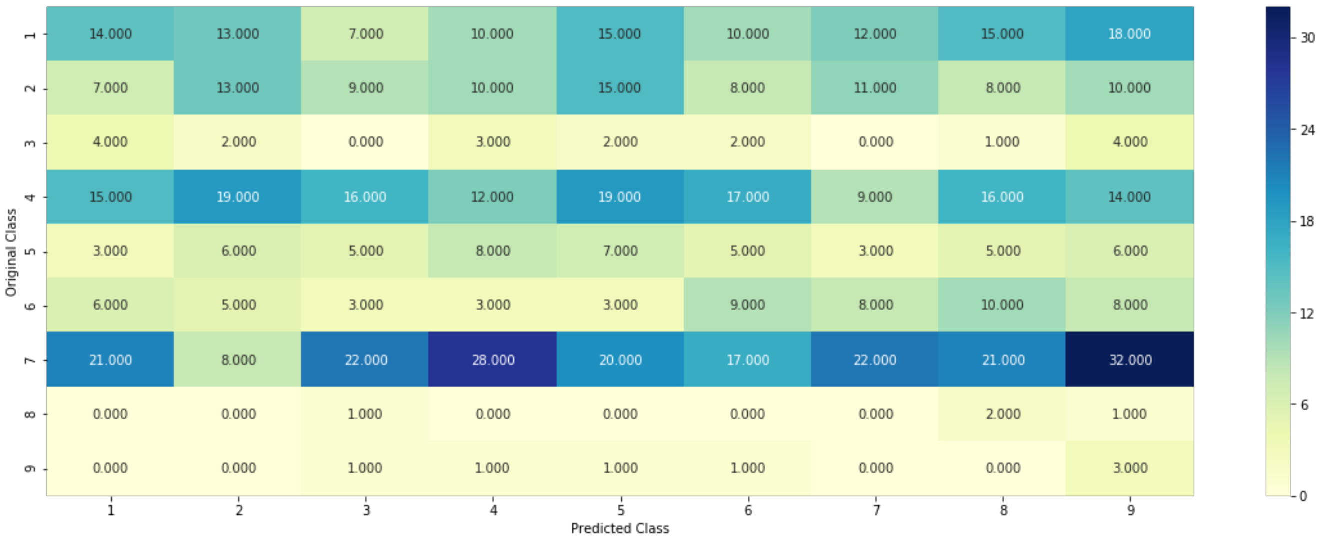
In [13]:

```
# This function plots the confusion matrices given y_i, y_i_hat.
def plot_confusion_matrix(test_y, predict_y):
    C = confusion_matrix(test_y, predict_y)
    # C = 9,9 matrix, each cell (i,j) represents number of points of class i are predicted class j
    
    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 two 
diamensional array
    # 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 two 
diamensional array
    # 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]
    # representing A in heatmap format
    print("-"*20, "Confusion matrix", "-"*20)
    plt.figure(figsize=(20,7))
    sns.heatmap(C, annot=True, cmap="YlGnBu", fmt=".3f", xticklabels=labels, yticklabels=labels)
    plt.xlabel('Predicted Class')
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.show()

    print("-"*20, "Precision matrix (Columm Sum=1)", "-"*20)
    plt.figure(figsize=(20,7))
    sns.heatmap(B, annot=True, cmap="YlGnBu", fmt=".3f", xticklabels=labels, yticklabels=labels)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.show()

    # representing B in heatmap format
    print("-"*20, "Recall matrix (Row sum=1)", "-"*20)
    plt.figure(figsize=(20,7))
    sns.heatmap(A, annot=True, cmap="YlGnBu", fmt=".3f", xticklabels=labels, yticklabels=labels)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.show()
```

In [14]:

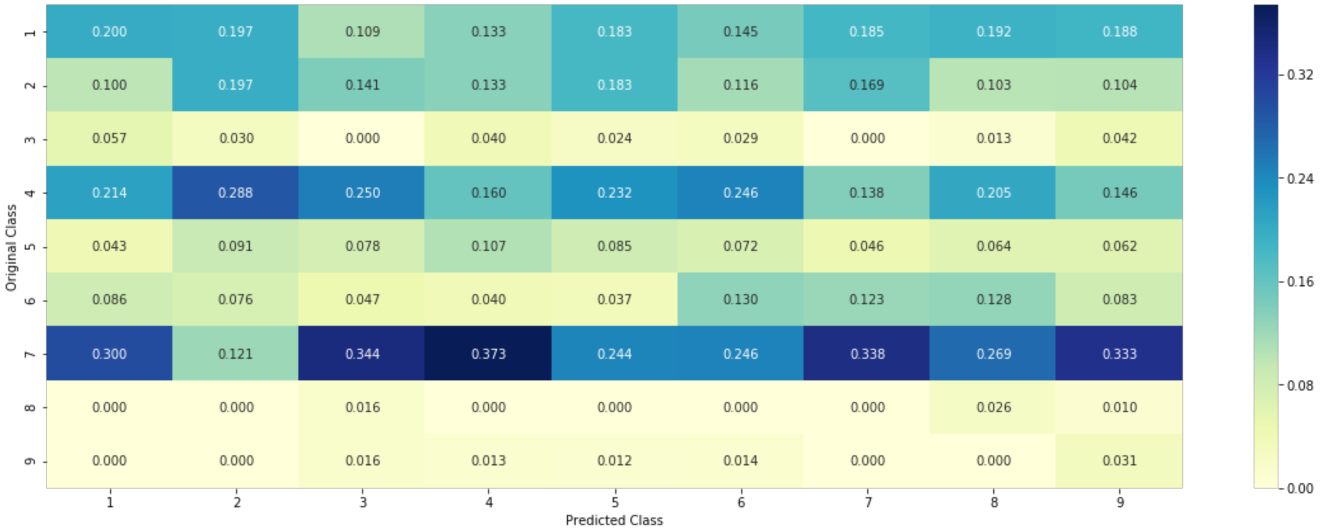
```
# we need to generate 9 numbers and the sum of numbers should be 1
# one solution is to genarate 9 numbers and divide each of the numbers by their sum
# ref: https://stackoverflow.com/a/18662466/4084039
test_data_len = X_test.shape[0]
cv_data_len = X_cv.shape[0]

# 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_y, eps=1e-
15))


# 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=1e-15))

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 2.5358430631813267
Log loss on Test Data using Random Model 2.5182132107572777
-------------------- Confusion matrix --------------------
```

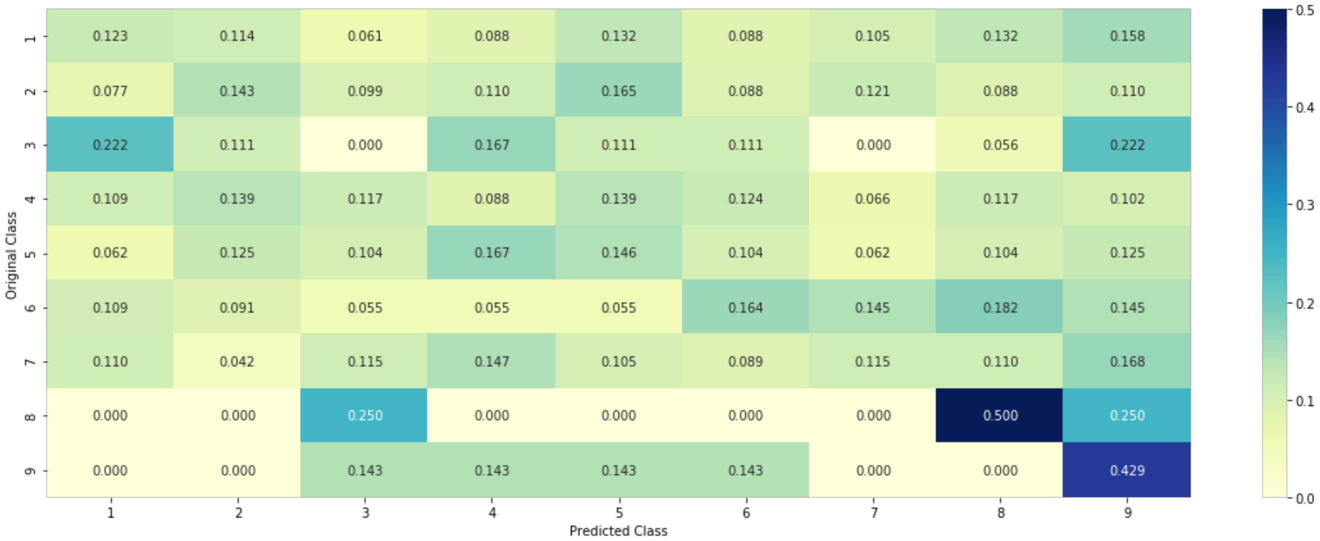


	1	2	3	4	5	6	7	8	9
1	14.000	13.000	7.000	10.000	15.000	10.000	12.000	15.000	18.000
2	7.000	13.000	9.000	10.000	15.000	8.000	11.000	8.000	10.000
3	4.000	2.000	0.000	3.000	2.000	2.000	0.000	1.000	4.000
4	15.000	19.000	16.000	12.000	19.000	17.000	9.000	16.000	14.000
5	3.000	6.000	5.000	8.000	7.000	5.000	3.000	5.000	6.000
6	6.000	5.000	3.000	3.000	3.000	9.000	8.000	10.000	8.000
7	21.000	8.000	22.000	28.000	20.000	17.000	22.000	21.000	32.000
8	0.000	0.000	1.000	0.000	0.000	0.000	0.000	2.000	1.000
9	0.000	0.000	1.000	1.000	1.000	1.000	0.000	0.000	3.000

```
-------------------- Precision matrix (Columm Sum=1) --------------------
```



```
-------------------- Recall matrix (Row sum=1) --------------------
```



3.3 Univariate Analysis

In [15]:

```
# code for response coding with Laplace smoothing.
# alpha : used for laplace smoothing
# feature: ['gene', 'variation']
# df: ['X_train', 'X_test', 'X_cv']
# algorithm
# ----------
# Consider all unique values and the number of occurances of given feature in train data dataframe
# build a vector (1*9) , the first element = (number of times it occured in class1 + 10*alpha / nu
mber of time it occurred in total data+90*alpha)
# gv_dict is like a look up table, for every gene it store a (1*9) representation of it
# for a value of feature in df:
# if it is in train data:
# we add the vector that was stored in 'gv_dict' look up table to 'gv_fea'
# if it is not there is train:
# we add [1/9, 1/9, 1/9, 1/9,1/9, 1/9, 1/9, 1/9, 1/9] to 'gv_fea'
# return 'gv_fea'
# ----------------------

# get_gv_fea_dict: Get Gene varaition Feature Dict
def get_gv_fea_dict(alpha, feature, df):
    # value_count: it contains a dict like
    # print(X_train['Gene'].value_counts())
    # output:
    #         {BRCA1      174
    #          TP53       106
    #          EGFR        86
    #         EGFR        86
    #         BRCA2       75
    #         PTEN        69
    #         KIT         61
    #         BRAF        60
    #         ERBB2       47
    #         PDGFRA      46
    #         ...}
    # print(X_train['Variation'].value_counts())
    # output:
    # {
    # Truncating_Mutations                     63
    # Deletion                                 43
    # Amplification                            43
    # Fusions                                  22
    # Overexpression                            3
    # E17K                                      3
    # Q61L                                      3
    # S222D                                     2
    # P130S                                     2
    # ...
    # }
    value_count = X_train[feature].value_counts()

    # gv_dict : Gene Variation Dict, which contains the probability array for each gene/variation
    gv_dict = dict()

    # denominator will contain the number of time that particular feature occured in whole data
    for i, denominator in value_count.items():
        # vec will contain (p(yi==1/Gi) probability of gene/variation belongs to perticular class
        # vec is 9 diamensional vector
        vec = []
        for k in range(1,10):
            # print(X_train.loc[(X_train['Class']==1) & (X_train['Gene']=='BRCA1')])
            #         ID   Gene             Variation  Class
            # 2470  2470  BRCA1                S1715C      1
            # 2486  2486  BRCA1                S1841R      1
            # 2614  2614  BRCA1                   M1R      1
            # 2432  2432  BRCA1                L1657P      1
            # 2567  2567  BRCA1                T1685A      1
            # 2583  2583  BRCA1                E1660G      1
            # 2634  2634  BRCA1                W1718L      1
            # cls_cnt.shape[0] will return the number of rows

            cls_cnt = X_train.loc[(X_train['Class']==k) & (X_train[feature]==i)]

            # cls_cnt.shape[0](numerator) will contain the number of time that particular feature o
ccured in whole data
            vec.append((cls_cnt.shape[0] + alpha*10)/ (denominator + 90*alpha))

        # we are adding the gene/variation to the dict as key and vec as value
        gv_dict[i]=vec
    return gv_dict

# Get Gene variation feature
def get_gv_feature(alpha, feature, df):
    # print(gv_dict)
    #     {'BRCA1': [0.20075757575757575, 0.03787878787878788, 0.068181818181818177,
0.13636363636363635, 0.25, 0.19318181818181818, 0.03787878787878788, 0.03787878787878788,
0.03787878787878788],
    #      'TP53': [0.32142857142857145, 0.061224489795918366, 0.061224489795918366,
0.27040816326530615, 0.061224489795918366, 0.066326530612244902, 0.051020408163265307, 0.051020408
163265307, 0.056122448979591837],
    #      'EGFR': [0.056818181818181816, 0.21590909090909091, 0.0625, 0.068181818181818177,
0.068181818181818177, 0.0625, 0.34659090909090912, 0.0625, 0.056818181818181816],
    #      'BRCA2': [0.13333333333333333, 0.060606060606060608, 0.060606060606060608,
0.078787878787878782, 0.13939393939393939, 0.34545454545454546, 0.060606060606060608,
0.060606060606060608, 0.060606060606060608],
    #      'PTEN': [0.069182389937106917, 0.062893081761006289, 0.069182389937106917,
0.46540880503144655, 0.075471698113207544, 0.062893081761006289, 0.069182389937106917, 0.062893081
761006289, 0.062893081761006289],
    #      'KIT': [0.066225165562913912, 0.25165562913907286, 0.072847682119205295,
0.072847682119205295, 0.066225165562913912, 0.066225165562913912, 0.27152317880794702,
0.066225165562913912, 0.066225165562913912],
    #      'BRAF': [0.066666666666666666, 0.17999999999999999, 0.073333333333333334,
0.073333333333333334, 0.093333333333333338, 0.080000000000000002, 0.29999999999999999,
0.066666666666666666, 0.066666666666666666],
    #      ...
    #     }
    gv_dict = get_gv_fea_dict(alpha, feature, df)
    # value_count is similar in get_gv_fea_dict
    value_count = X_train[feature].value_counts()

    # gv_fea: Gene_variation feature, it will contain the feature for each feature value in the data
    gv_fea = []
    # for every feature values in the given data frame we will check if it is there in the train data then we will add the feature to gv_fea
    # if not we will add [1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9] to gv_fea
    for index, row in df.iterrows():
        if row[feature] in dict(value_count).keys():
            gv_fea.append(gv_dict[row[feature]])
        else:
            gv_fea.append([1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9])
#             gv_fea.append([-1,-1,-1,-1,-1,-1,-1,-1,-1])
    return gv_fea
```

when we caculate the probability of a feature belongs to any particular class, we apply laplace smoothing

- (numerator + 10*alpha) / (denominator + 90*alpha)

3.2.1 Univariate Analysis on Gene Feature

Q1. Gene, What type of feature it is ?

Ans. Gene is a categorical variable

Q2. How many categories are there and How they are distributed?

In [16]:

```
unique_genes = X_train['Gene'].value_counts()
print('Number of Unique Genes :', unique_genes.shape[0])
# the top 10 genes that occured most
print(unique_genes.head(10))
```

```
Number of Unique Genes : 230
BRCA1    161
TP53     111
EGFR      93
PTEN      82
BRCA2     78
BRAF      65
KIT       57
ERBB2     44
ALK       41
FLT3      37
Name: Gene, dtype: int64
```

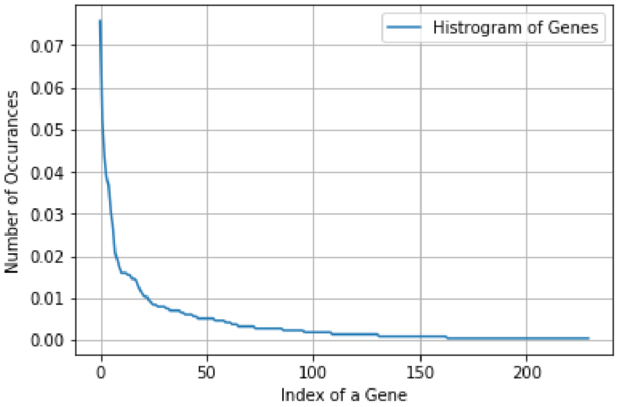
In [17]:

```
print("Ans: There are", unique_genes.shape[0] ,"different categories of genes in the train data, and they are distibuted as follows",)
```

```
Ans: There are 230 different categories of genes in the train data, and they are distibuted as follows
```

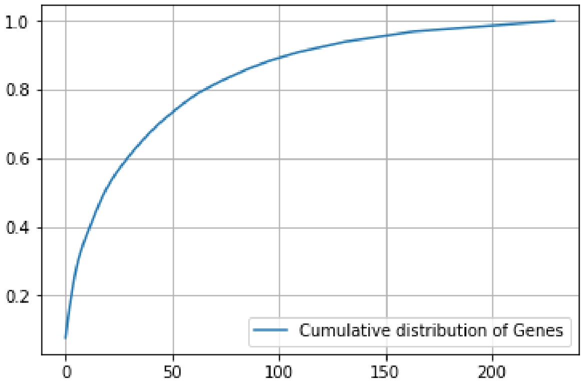
In [18]:

```
s = sum(unique_genes.values);
h = unique_genes.values/s;
plt.plot(h, label="Histrogram of Genes")
plt.xlabel('Index of a Gene')
plt.ylabel('Number of Occurances')
plt.legend()
plt.grid()
plt.show()
```



In [19]:

```
c = np.cumsum(h)
plt.plot(c,label='Cumulative distribution of Genes')
plt.grid()
plt.legend()
plt.show()
```



Q3. How to featurize this Gene feature ?

Ans.there are two ways we can featurize this variable check out this video: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/handling-categorical-and-numerical-features/

1. One hot Encoding
2. Response coding

We will choose the appropriate featurization based on the ML model we use. For this problem of multi-class classification with categorical features, one-hot encoding is better for Logistic regression while response coding is better for Random Forests.

In [20]:

```
#response-coding of the Gene feature
# alpha is used for laplace smoothing
alpha = 1
# train gene feature
train_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", X_train))
# test gene feature
test_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", X_test))
# cross validation gene feature
cv_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", X_cv))
```

In [21]:

```
print("train_gene_feature_responseCoding is converted feature using respone coding method. The sha
pe of gene feature:", train_gene_feature_responseCoding.shape)
```

```
train_gene_feature_responseCoding is converted feature using respone coding method. The shape of g
```

```
ene feature: (2124, 9)
```

In [22]:

```
# one-hot encoding of Gene feature.
gene_vectorizer = CountVectorizer(ngram_range=(1,2))
train_gene_feature_onehotCoding = gene_vectorizer.fit_transform(X_train['Gene'])
test_gene_feature_onehotCoding = gene_vectorizer.transform(X_test['Gene'])
cv_gene_feature_onehotCoding = gene_vectorizer.transform(X_cv['Gene'])
```

In [146]:

```
# tfidf encoding of Gene feature.
gene_tfidf_vectorizer = TfidfVectorizer(ngram_range=(1,1))
train_gene_feature_tfidf = gene_tfidf_vectorizer.fit_transform(X_train['Gene'])
test_gene_feature_tfidf = gene_tfidf_vectorizer.transform(X_test['Gene'])
cv_gene_feature_tfidf = gene_tfidf_vectorizer.transform(X_cv['Gene'])
```

In [112]:

```
print("train_gene_feature_onehotCoding is converted feature using one-hot encoding method. The sha
pe of gene feature:", train_gene_feature_tfidf.shape)
```

```
train_gene_feature_onehotCoding is converted feature using one-hot encoding method. The shape of g
ene feature: (2124, 230)
```

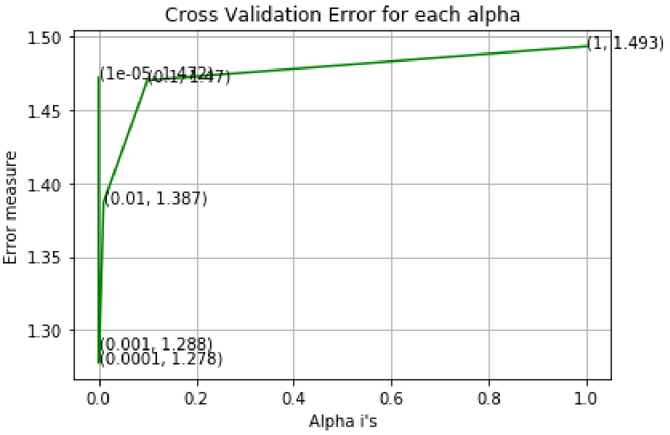
Q4. How good is this gene feature in predicting y_i?

There are many ways to estimate how good a feature is, in predicting y_i. One of the good methods is to build a proper ML model using just this feature. In this case, we will build a logistic regression model using only Gene feature (one hot encoded) to predict y_i.

In [25]:

```
alpha = [10 ** x for x in range(-5, 1)] # hyperparam for SGD classifier.

# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss=’hinge’, penalty=’l2’, alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate=’optimal’, eta0
=0.0, power_t=0.5,
# 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 Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link:
#------------------------------


cv_log_error_array=[]
for i in alpha:
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_gene_feature_onehotCoding, y_train)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_gene_feature_onehotCoding, y_train)
    predict_y = sig_clf.predict_proba(cv_gene_feature_onehotCoding)
    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
    print('For values of alpha = ', i, "The log loss is:",log_loss(y_cv, predict_y, labels=clf.clas
ses_, eps=1e-15))

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()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_gene_feature_onehotCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_gene_feature_onehotCoding, y_train)

predict_y = sig_clf.predict_proba(train_gene_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train,
predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_gene_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_gene_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.4718910639355252
For values of alpha =  0.0001 The log loss is: 1.2775050083514932
For values of alpha =  0.001 The log loss is: 1.2876432742746475
For values of alpha =  0.01 The log loss is: 1.3870437053419553
For values of alpha =  0.1 The log loss is: 1.4701829768717998
For values of alpha =  1 The log loss is: 1.493404590983701
```



```
For values of best alpha =  0.0001 The train log loss is: 1.039603279111539
For values of best alpha =  0.0001 The cross validation log loss is: 1.2775050083514932
For values of best alpha =  0.0001 The test log loss is: 1.1978393112378956
```

Q5. Is the Gene feature stable across all the data sets (Test, Train, Cross validation)?

Ans. Yes, it is. Otherwise, the CV and Test errors would be significantly more than train error.

In [26]:

```
print("Q6. How many data points in Test and CV datasets are covered by the ", unique_genes.shape[0
], " genes in train dataset?")

test_coverage=X_test[X_test['Gene'].isin(list(set(X_train['Gene'])))].shape[0]
cv_coverage=X_cv[X_cv['Gene'].isin(list(set(X_train['Gene'])))].shape[0]

print('Ans\n1. In test data',test_coverage, 'out of',X_test.shape[0], ":",(test_coverage/X_test.sha
pe[0])*100)
print('2. In cross validation data',cv_coverage, 'out of ',X_cv.shape[0],":" ,(cv_coverage/X_cv.sha
pe[0])*100)
```

```
Q6. How many data points in Test and CV datasets are covered by the  230  genes in train dataset?
Ans
1. In test data 645 out of 665 : 96.99248120300751
2. In cross validation data 509 out of  532 : 95.67669172932331
```

3.2.2 Univariate Analysis on Variation Feature

Q7. Variation, What type of feature is it ?

Ans. Variation is a categorical variable

Q8. How many categories are there?

In [27]:

```
unique_variations = X_train['Variation'].value_counts()
print('Number of Unique Variations :', unique_variations.shape[0])
# the top 10 variations that occured most
print(unique_variations.head(10))
```

```
Number of Unique Variations : 1931
Truncating_Mutations     57
Deletion                 50
Amplification            45
Fusions                  16
Overexpression            5
G12V                      4
Q61H                      3
Q61R                      3
T58I                      3
E17K                      2
Name: Variation, dtype: int64
```

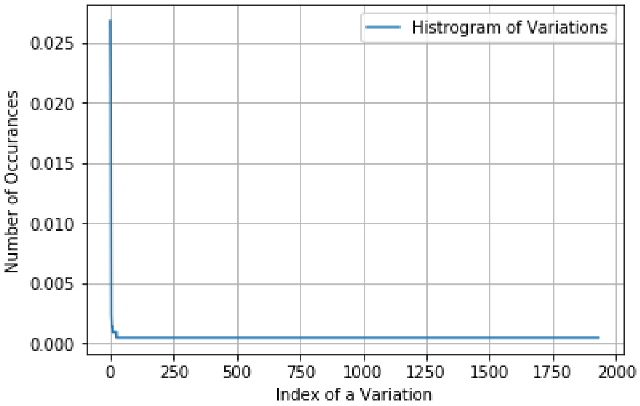
In [28]:

```
print("Ans: There are", unique_variations.shape[0] ,"different categories of variations in the train data, and they are distibuted as follows",)
```

```
Ans: There are 1931 different categories of variations in the train data, and they are distibuted
as follows
```

In [29]:

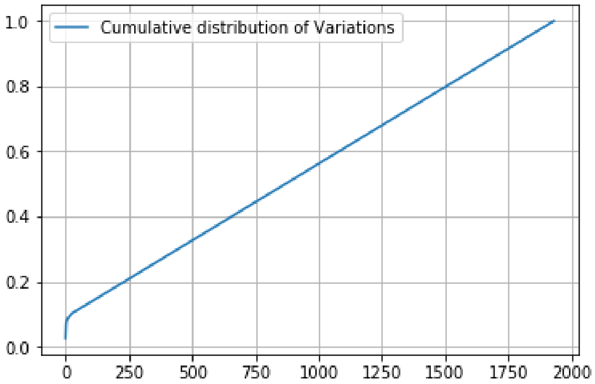
```
s = sum(unique_variations.values);
h = unique_variations.values/s;
plt.plot(h, label="Histrogram of Variations")
plt.xlabel('Index of a Variation')
plt.ylabel('Number of Occurances')
plt.legend()
plt.grid()
plt.show()
```



In [30]:

```
c = np.cumsum(h)
print(c)
plt.plot(c,label='Cumulative distribution of Variations')
plt.grid()
plt.legend()
plt.show()
```

```
[0.02683616 0.05037665 0.07156309 ... 0.99905838 0.99952919 1.        ]
```



Q9. How to featurize this Variation feature ?

Ans. There are two ways we can featurize this variable check out this video: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/handling-categorical-and-numerical-features/

1. One hot Encoding
2. Response coding

We will be using both these methods to featurize the Variation Feature

In [31]:

```
# alpha is used for laplace smoothing
alpha = 1
# train gene feature
train_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", X_train))
# test gene feature
test_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", X_test))
# cross validation gene feature
cv_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", X_cv))
```

In [32]:

```
print("train_variation_feature_responseCoding is a converted feature using the response coding met
hod. The shape of Variation feature:", train_variation_feature_responseCoding.shape)
```

```
train_variation_feature_responseCoding is a converted feature using the response coding method. Th
e shape of Variation feature: (2124, 9)
```

In [33]:

```
# one-hot encoding of variation feature.
variation_vectorizer = CountVectorizer(ngram_range=(1,2))
train_variation_feature_onehotCoding = variation_vectorizer.fit_transform(X_train['Variation'])
test_variation_feature_onehotCoding = variation_vectorizer.transform(X_test['Variation'])
cv_variation_feature_onehotCoding = variation_vectorizer.transform(X_cv['Variation'])
```

In [148]:

```
# tfidf encoding of variation feature.
variation_tfidf_vectorizer = TfidfVectorizer(ngram_range=(1,1))
train_variation_feature_tfidf = variation_tfidf_vectorizer.fit_transform(X_train['Variation'])
test_variation_feature_tfidf = variation_tfidf_vectorizer.transform(X_test['Variation'])
cv_variation_feature_tfidf = variation_tfidf_vectorizer.transform(X_cv['Variation'])
```

In [114]:

```
print("train_variation_feature_onehotEncoded is converted feature using the onne-hot encoding meth
od. The shape of Variation feature:", train_variation_feature_tfidf.shape)
```

```
train_variation_feature_onehotEncoded is converted feature using the onne-hot encoding method. The
shape of Variation feature: (2124, 2066)
```

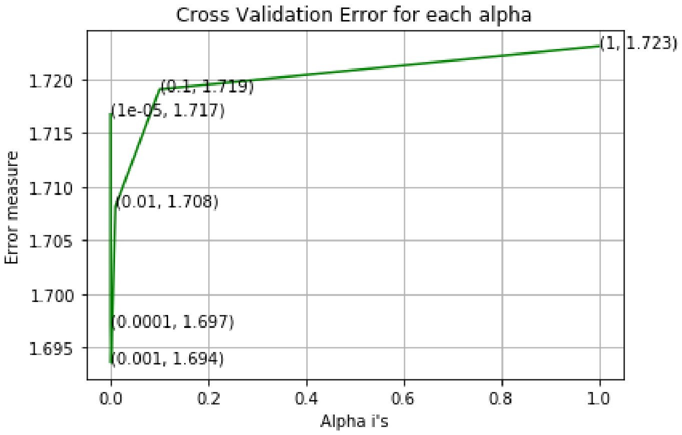
Q10. How good is this Variation feature in predicting y_i?

Let's build a model just like the earlier!

In [36]:

```
alpha = [10 ** x for x in range(-5, 1)]

# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0
=0.0, power_t=0.5,
# 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 Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link:
#------------------------------


cv_log_error_array=[]
for i in alpha:
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_variation_feature_onehotCoding, y_train)

    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_variation_feature_onehotCoding, y_train)
    predict_y = sig_clf.predict_proba(cv_variation_feature_onehotCoding)

    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
    print('For values of alpha = ', i, "The log loss is:",log_loss(y_cv, predict_y, labels=clf.clas
ses_, eps=1e-15))

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()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_variation_feature_onehotCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_variation_feature_onehotCoding, y_train)

predict_y = sig_clf.predict_proba(train_variation_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train,
predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_variation_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_variation_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.7167196126504518
For values of alpha =  0.0001 The log loss is: 1.6970744784336635
For values of alpha =  0.0001 The log loss is: 1.6970744784550655
For values of alpha =  0.001 The log loss is: 1.6935926369802465
For values of alpha =  0.01 The log loss is: 1.7081638298695538
For values of alpha =  0.1 The log loss is: 1.7190330123402875
For values of alpha =  1 The log loss is: 1.7230468914787866
```



```
For values of best alpha =  0.001 The train log loss is: 1.0530749177305776
For values of best alpha =  0.001 The cross validation log loss is: 1.6935926369802465
For values of best alpha =  0.001 The test log loss is: 1.7037824676287987
```

Q11. Is the Variation feature stable across all the data sets (Test, Train, Cross validation)?

Ans. Not sure! But lets be very sure using the below analysis.

In [37]:

```python
print("Q12. How many data points are covered by total ", unique_variations.shape[0], " genes in te
st and cross validation data sets?")
test_coverage=X_test[X_test['Variation'].isin(list(set(X_train['Variation'])))].shape[0]
cv_coverage=X_cv[X_cv['Variation'].isin(list(set(X_train['Variation'])))].shape[0]
print('Ans\n1. In test data',test_coverage, 'out of',X_test.shape[0], ":",(test_coverage/X_test.sha
pe[0])*100)
print('2. In cross validation data',cv_coverage, 'out of ',X_cv.shape[0],":" ,(cv_coverage/X_cv.sha
pe[0])*100)
```

```
Q12. How many data points are covered by total  1931  genes in test and cross validation data
sets?
Ans
1. In test data 66 out of 665 : 9.924812030075188
2. In cross validation data 59 out of  532 : 11.090225563909774
```

3.2.3 Univariate Analysis on Text Feature

1. How many unique words are present in train data?
2. How are word frequencies distributed?
3. How to featurize text field?
4. Is the text feature useful in predicitng y_i?
5. Is the text feature stable across train, test and CV datasets?

In [38]:

```python
# cls_text is a data frame
# for every row in data fram consider the 'TEXT'
# split the words by space
# make a dict with those words
# increment its count whenever we see that word

def extract_dictionary_paddle(cls_text):
    dictionary = defaultdict(int)
    for index, row in cls_text.iterrows():
        for word in row['TEXT'].split():
            dictionary[word] +=1
    return dictionary
```

In [39]:

```
import math
#https://stackoverflow.com/a/1602964
def get_text_responsecoding(df):
    text_feature_responseCoding = np.zeros((df.shape[0],9))
    for i in range(0,9):
        row_index = 0
        for index, row in df.iterrows():
            sum_prob = 0
            for word in row['TEXT'].split():
                sum_prob += math.log(((dict_list[i].get(word,0)+10 )/(total_dict.get(word,0)+90)))
            text_feature_responseCoding[row_index][i] = math.exp(sum_prob/len(row['TEXT'].split()))
            row_index += 1
    return text_feature_responseCoding
```

In [40]:

```
# building a CountVectorizer with all the words that occured minimum 3 times in train data
text_vectorizer = CountVectorizer(min_df=3,ngram_range=(1,2))
train_text_feature_onehotCoding = text_vectorizer.fit_transform(X_train['TEXT'])
# getting all the feature names (words)
train_text_features= text_vectorizer.get_feature_names()

# train_text_feature_onehotCoding.sum(axis=0).A1 will sum every row and returns (1*number of featu
res) vector
train_text_fea_counts = train_text_feature_onehotCoding.sum(axis=0).A1
#print(train_text_fea_counts)
# zip(list(text_features),text_fea_counts) will zip a word with its number of times it occured
text_fea_dict = dict(zip(list(train_text_features),train_text_fea_counts))
#print(text_fea_dict)

print("Total number of unique words in train data :", len(train_text_features))
```

```
Total number of unique words in train data : 766917
```

In [184]:

```
# tfidf encoding of text feature.
text_tfidf_vectorizer = TfidfVectorizer(max_features=1250,ngram_range=(1,1))
train_text_feature_tfidf = text_tfidf_vectorizer.fit_transform(X_train['TEXT'])
test_text_feature_tfidf = text_tfidf_vectorizer.transform(X_test['TEXT'])
cv_text_feature_tfidf = text_tfidf_vectorizer.transform(X_cv['TEXT'])
```

In [42]:

```
dict_list = []
# dict_list =[] contains 9 dictoinaries each corresponds to a class
for i in range(1,10):
    cls_text = X_train[X_train['Class']==i]
    # build a word dict based on the words in that class
    dict_list.append(extract_dictionary_paddle(cls_text))
    # append it to dict_list

# dict_list[i] is build on i'th  class text data
# total_dict is buid on whole training text data
total_dict = extract_dictionary_paddle(X_train)


confuse_array = []
for i in train_text_features:
    ratios = []
    max_val = -1
    for j in range(0,9):
        ratios.append((dict_list[j][i]+10 )/(total_dict[i]+90))
    confuse_array.append(ratios)
confuse_array = np.array(confuse_array)
```

In [43]:

```
#response coding of text features
```

```
#response coding of text features
train_text_feature_responseCoding  = get_text_responsecoding(X_train)
test_text_feature_responseCoding  = get_text_responsecoding(X_test)
cv_text_feature_responseCoding  = get_text_responsecoding(X_cv)
```

In [44]:

```
# https://stackoverflow.com/a/16202486
# we convert each row values such that they sum to 1
train_text_feature_responseCoding =
(train_text_feature_responseCoding.T/train_text_feature_responseCoding.sum(axis=1)).T
test_text_feature_responseCoding =
(test_text_feature_responseCoding.T/test_text_feature_responseCoding.sum(axis=1)).T
cv_text_feature_responseCoding = (cv_text_feature_responseCoding.T/cv_text_feature_responseCoding.
sum(axis=1)).T
```

In [45]:

```
# don't forget to normalize every feature
train_text_feature_onehotCoding = normalize(train_text_feature_onehotCoding, axis=0)

# we use the same vectorizer that was trained on train data
test_text_feature_onehotCoding = text_vectorizer.transform(X_test['TEXT'])
# don't forget to normalize every feature
test_text_feature_onehotCoding = normalize(test_text_feature_onehotCoding, axis=0)

# we use the same vectorizer that was trained on train data
cv_text_feature_onehotCoding = text_vectorizer.transform(X_cv['TEXT'])
# don't forget to normalize every feature
cv_text_feature_onehotCoding = normalize(cv_text_feature_onehotCoding, axis=0)
```

In [185]:

```
#normalizing tfidf features
train_text_feature_tfidf = normalize(train_text_feature_tfidf, axis=0)
cv_text_feature_tfidf = normalize(cv_text_feature_tfidf, axis=0)
test_text_feature_tfidf = normalize(test_text_feature_tfidf, axis=0)
```

In [47]:

```
#https://stackoverflow.com/a/2258273/4084039
sorted_text_fea_dict = dict(sorted(text_fea_dict.items(), key=lambda x: x[1] , reverse=True))
sorted_text_occur = np.array(list(sorted_text_fea_dict.values()))
```

In [48]:

```
# Number of words for a given frequency.
print(Counter(sorted_text_occur))
```

```
Counter({3: 135740, 4: 98253, 5: 76880, 6: 61024, 7: 44774, 8: 40209, 9: 33562, 10: 29975, 12: 227
05, 11: 20082, 13: 18149, 14: 14261, 15: 11250, 16: 9697, 18: 8497, 17: 7070, 20: 6799, 21: 6524,
19: 6113, 24: 5286, 22: 5240, 29: 3857, 23: 3817, 31: 3720, 38: 3603, 25: 3349, 26: 3160, 27: 2995
, 28: 2881, 30: 2802, 32: 2291, 33: 2274, 51: 2270, 45: 2010, 36: 1894, 34: 1819, 35: 1782, 40: 16
64, 39: 1660, 60: 1626, 37: 1503, 42: 1454, 50: 1405, 41: 1254, 44: 1226, 43: 1169, 46: 1116, 48:
1099, 52: 998, 47: 963, 49: 888, 54: 864, 55: 806, 53: 791, 56: 783, 58: 755, 62: 748, 61: 728, 63
: 701, 57: 668, 59: 630, 64: 573, 66: 539, 65: 523, 72: 512, 70: 508, 67: 497, 68: 476, 90: 468, 7
6: 441, 69: 438, 71: 430, 73: 421, 75: 415, 74: 394, 77: 384, 78: 363, 80: 348, 79: 329, 91: 318,
84: 316, 81: 313, 88: 296, 82: 286, 85: 285, 93: 279, 83: 277, 102: 274, 100: 272, 92: 272, 89: 26
4, 86: 264, 87: 261, 96: 259, 95: 246, 97: 245, 94: 239, 99: 224, 98: 217, 104: 210, 120: 204, 103
: 203, 107: 200, 101: 198, 108: 194, 105: 190, 114: 177, 106: 177, 115: 175, 110: 163, 112: 162, 1
16: 159, 113: 154, 131: 153, 126: 153, 109: 152, 117: 150, 123: 148, 111: 147, 124: 146, 121: 144,
119: 142, 129: 141, 135: 139, 122: 139, 144: 130, 118: 129, 125: 127, 128: 125, 134: 120, 127: 120
, 136: 119, 138: 118, 140: 117, 149: 113, 143: 113, 137: 112, 150: 111, 145: 111, 132: 109, 130: 1
07, 133: 106, 139: 105, 154: 100, 153: 98, 141: 97, 155: 96, 168: 94, 142: 92, 147: 91, 146: 90, 1
52: 88, 148: 88, 160: 86, 156: 86, 163: 84, 180: 83, 165: 83, 162: 78, 158: 78, 167: 77, 183: 76,
166: 75, 164: 75, 159: 75, 161: 74, 157: 74, 188: 73, 176: 73, 170: 72, 151: 67, 195: 65, 190: 65,
172: 64, 171: 63, 184: 62, 182: 61, 177: 61, 173: 61, 189: 60, 179: 60, 174: 60, 225: 59, 187: 59,
185: 59, 186: 58, 175: 58, 214: 56, 206: 56, 197: 56, 169: 56, 181: 55, 178: 55, 205: 54, 191: 54,
224: 51, 204: 51, 212: 50, 203: 50, 200: 50, 221: 49, 219: 49, 213: 49, 209: 49, 211: 48, 207: 48,
202: 48, 199: 48, 196: 48, 226: 47, 227: 46, 201: 46, 222: 45, 193: 45, 240: 44, 210: 44, 198: 44,
242: 43, 237: 43, 220: 43, 238: 42, 228: 42, 232: 41, 192: 41, 255: 40, 234: 40, 217: 40, 208: 40,
281: 39, 245: 39, 268: 38, 266: 38, 230: 38, 223: 38, 288: 37, 276: 37, 250: 37, 241: 37, 236: 37,
194: 37, 264: 35, 233: 35, 282: 34, 280: 34, 253: 34, 244: 34, 239: 34, 229: 34, 247: 33, 246: 33,
243: 33, 218: 33, 216: 33, 296: 32, 286: 32, 273: 32, 263: 32, 258: 32, 315: 31, 261: 31, 257: 31,
249: 31, 272: 30, 259: 30, 252: 30, 231: 30, 317: 29, 291: 29, 287: 29, 265: 29, 260: 28, 235: 28,
215: 28, 302: 27, 279: 27, 267: 27, 254: 27, 321: 26, 305: 26, 293: 26, 290: 26, 278: 26, 277: 26,
275: 26, 274: 26, 251: 26, 329: 25, 320: 25, 262: 25, 319: 24, 316: 24, 298: 24, 285: 24, 283: 24,
256: 24, 343: 23, 334: 23, 312: 23, 300: 23, 349: 22, 322: 22, 284: 22, 270: 22, 371: 21, 365: 21,
356: 21, 341: 21, 330: 21, 325: 21, 309: 21, 303: 21, 425: 20, 404: 20, 380: 20, 357: 20, 346: 20,
345: 20, 340: 20, 336: 20, 331: 20, 318: 20, 314: 20, 308: 20, 304: 20, 301: 20, 294: 20, 269: 20,
248: 20, 384: 19, 377: 19, 376: 19, 360: 19, 344: 19, 338: 19, 328: 19, 327: 19, 306: 19, 297: 19,
289: 19, 408: 18, 385: 18, 369: 18, 353: 18, 337: 18, 323: 18, 292: 18, 436: 17, 429: 17, 392: 17,
391: 17, 370: 17, 361: 17, 295: 17, 271: 17, 426: 16, 407: 16, 398: 16, 378: 16, 372: 16, 367: 16,
364: 16, 362: 16, 348: 16, 313: 16, 311: 16, 310: 16, 561: 15, 459: 15, 433: 15, 400: 15, 394: 15,
382: 15, 335: 15, 326: 15, 520: 14, 486: 14, 484: 14, 452: 14, 447: 14, 439: 14, 428: 14, 416: 14,
403: 14, 375: 14, 366: 14, 363: 14, 358: 14, 352: 14, 342: 14, 332: 14, 299: 14, 533: 13, 510: 13,
507: 13, 505: 13, 488: 13, 467: 13, 441: 13, 440: 13, 417: 13, 414: 13, 413: 13, 412: 13, 397: 13,
389: 13, 383: 13, 379: 13, 374: 13, 354: 13, 350: 13, 347: 13, 339: 13, 333: 13, 324: 13, 564: 12,
526: 12, 501: 12, 495: 12, 494: 12, 491: 12, 483: 12, 474: 12, 443: 12, 427: 12, 415: 12, 411: 12,
406: 12, 396: 12, 373: 12, 1091: 11, 682: 11, 620: 11, 555: 11, 552: 11, 509: 11, 470: 11, 466: 11
, 462: 11, 458: 11, 456: 11, 450: 11, 446: 11, 438: 11, 431: 11, 401: 11, 387: 11, 355: 11, 307: 1
1, 775: 10, 742: 10, 627: 10, 598: 10, 570: 10, 538: 10, 512: 10, 508: 10, 502: 10, 490: 10, 478:
10, 455: 10, 453: 10, 449: 10, 423: 10, 422: 10, 409: 10, 402: 10, 399: 10, 368: 10, 359: 10, 714:
9, 680: 9, 677: 9, 671: 9, 639: 9, 616: 9, 613: 9, 574: 9, 559: 9, 548: 9, 547: 9, 543: 9, 542: 9,
537: 9, 535: 9, 534: 9, 531: 9, 529: 9, 522: 9, 511: 9, 498: 9, 480: 9, 477: 9, 465: 9, 442: 9,
437: 9, 430: 9, 421: 9, 418: 9, 405: 9, 393: 9, 386: 9, 351: 9, 806: 8, 801: 8, 660: 8, 650: 8,
649: 8, 648: 8, 646: 8, 638: 8, 610: 8, 600: 8, 599: 8, 594: 8, 591: 8, 586: 8, 584: 8, 556: 8,
530: 8, 518: 8, 516: 8, 489: 8, 481: 8, 476: 8, 464: 8, 461: 8, 460: 8, 454: 8, 445: 8, 434: 8,
432: 8, 420: 8, 410: 8, 395: 8, 1125: 7, 962: 7, 944: 7, 787: 7, 779: 7, 769: 7, 765: 7, 746: 7,
715: 7, 704: 7, 690: 7, 664: 7, 663: 7, 657: 7, 643: 7, 633: 7, 626: 7, 624: 7, 622: 7, 608: 7,
580: 7, 568: 7, 566: 7, 558: 7, 551: 7, 549: 7, 546: 7, 545: 7, 541: 7, 540: 7, 536: 7, 524: 7,
517: 7, 514: 7, 500: 7, 493: 7, 469: 7, 448: 7, 444: 7, 435: 7, 424: 7, 419: 7, 390: 7, 381: 7,
2556: 6, 1279: 6, 1217: 6, 1016: 6, 982: 6, 971: 6, 956: 6, 933: 6, 930: 6, 920: 6, 917: 6, 864: 6,
863: 6, 850: 6, 825: 6, 821: 6, 810: 6, 803: 6, 771: 6, 764: 6, 763: 6, 755: 6, 753: 6, 749: 6,
738: 6, 737: 6, 734: 6, 733: 6, 722: 6, 717: 6, 716: 6, 701: 6, 691: 6, 679: 6, 673: 6, 669: 6,
659: 6, 654: 6, 641: 6, 636: 6, 631: 6, 625: 6, 618: 6, 617: 6, 612: 6, 603: 6, 592: 6, 589: 6,
588: 6, 585: 6, 569: 6, 554: 6, 528: 6, 515: 6, 513: 6, 485: 6, 479: 6, 473: 6, 468: 6, 457: 6,
451: 6, 388: 6, 1350: 5, 1305: 5, 1265: 5, 1130: 5, 1128: 5, 1102: 5, 1098: 5, 1070: 5, 1050: 5, 10
08: 5, 1007: 5, 1003: 5, 985: 5, 969: 5, 950: 5, 934: 5, 905: 5, 900: 5, 897: 5, 895: 5, 878: 5, 86
7: 5, 846: 5, 843: 5, 822: 5, 817: 5, 813: 5, 802: 5, 797: 5, 793: 5, 791: 5, 785: 5, 784: 5, 761:
5, 754: 5, 744: 5, 740: 5, 736: 5, 728: 5, 727: 5, 726: 5, 724: 5, 712: 5, 711: 5, 702: 5, 699: 5,
696: 5, 689: 5, 686: 5, 684: 5, 683: 5, 672: 5, 670: 5, 667: 5, 665: 5, 658: 5, 656: 5, 640: 5,
637: 5, 634: 5, 630: 5, 629: 5, 628: 5, 623: 5, 615: 5, 606: 5, 602: 5, 597: 5, 596: 5, 590: 5,
583: 5, 576: 5, 572: 5, 567: 5, 563: 5, 562: 5, 560: 5, 557: 5, 550: 5, 544: 5, 532: 5, 527: 5,
523: 5, 506: 5, 504: 5, 497: 5, 496: 5, 492: 5, 482: 5, 475: 5, 472: 5, 3401: 4, 2061: 4, 1951: 4,
1927: 4, 1654: 4, 1596: 4, 1569: 4, 1547: 4, 1524: 4, 1459: 4, 1445: 4, 1396: 4, 1384: 4, 1346: 4,
1285: 4, 1266: 4, 1263: 4, 1255: 4, 1241: 4, 1216: 4, 1193: 4, 1151: 4, 1117: 4, 1103: 4, 1097: 4,
1088: 4, 1064: 4, 1031: 4, 1025: 4, 1023: 4, 1018: 4, 1006: 4, 984: 4, 983: 4, 980: 4, 977: 4, 966:
4, 964: 4, 960: 4, 955: 4, 946: 4, 942: 4, 901: 4, 898: 4, 896: 4, 889: 4, 884: 4, 881: 4, 865: 4,
859: 4, 857: 4, 855: 4, 854: 4, 848: 4, 842: 4, 840: 4, 839: 4, 835: 4, 833: 4, 831: 4, 828: 4,
814: 4, 807: 4, 800: 4, 796: 4, 786: 4, 778: 4, 774: 4, 770: 4, 762: 4, 760: 4, 758: 4, 757: 4,
756: 4, 741: 4, 735: 4, 731: 4, 729: 4, 725: 4, 723: 4, 720: 4, 718: 4, 713: 4, 708: 4, 707: 4,
706: 4, 705: 4, 703: 4, 700: 4, 697: 4, 688: 4, 687: 4, 678: 4, 676: 4, 675: 4, 668: 4, 666: 4,
662: 4, 661: 4, 652: 4, 651: 4, 644: 4, 642: 4, 635: 4, 632: 4, 621: 4, 619: 4, 614: 4, 593: 4,
587: 4, 577: 4, 575: 4, 573: 4, 571: 4, 565: 4, 553: 4, 539: 4, 525: 4, 519: 4, 499: 4, 463: 4,
3067: 3, 2628: 3, 2546: 3, 2523: 3, 2522: 3, 2194: 3, 2134: 3, 2083: 3, 2073: 3, 2037: 3, 2024: 3,
1999: 3, 1983: 3, 1966: 3, 1932: 3, 1879: 3, 1837: 3, 1818: 3, 1817: 3, 1807: 3, 1771: 3, 1762: 3,
1747: 3, 1731: 3, 1718: 3, 1692: 3, 1658: 3, 1653: 3, 1647: 3, 1627: 3, 1615: 3, 1611: 3, 1607: 3,
1605: 3, 1602: 3, 1586: 3, 1582: 3, 1575: 3, 1566: 3, 1562: 3, 1557: 3, 1535: 3, 1531: 3, 1502: 3,
1495: 3, 1494: 3, 1470: 3, 1457: 3, 1443: 3, 1422: 3, 1416: 3, 1402: 3, 1382: 3, 1380: 3, 1375: 3,
1374: 3, 1373: 3, 1371: 3, 1367: 3, 1366: 3, 1351: 3, 1314: 3, 1313: 3, 1304: 3, 1294: 3, 1290: 3,
1287: 3, 1277: 3, 1271: 3, 1262: 3, 1260: 3, 1248: 3, 1247: 3, 1245: 3, 1242: 3, 1233: 3, 1229: 3,
1211: 3, 1209: 3, 1201: 3, 1200: 3, 1191: 3, 1187: 3, 1186: 3, 1182: 3, 1179: 3, 1177: 3, 1175: 3,
1174: 3, 1160: 3, 1156: 3, 1152: 3, 1143: 3, 1132: 3, 1126: 3, 1120: 3, 1118: 3, 1114: 3, 1110: 3,
1106: 3, 1104: 3, 1093: 3, 1087: 3, 1084: 3, 1069: 3, 1060: 3, 1053: 3, 1051: 3, 1045: 3, 1037: 3,
1036: 3, 1035: 3, 1034: 3, 1033: 3, 1032: 3, 1029: 3, 1028: 3, 1015: 3, 1011: 3, 1009: 3, 1005: 3,
1000: 3, 995: 3, 994: 3, 991: 3, 979: 3, 975: 3, 973: 3, 967: 3, 965: 3, 953: 3, 951: 3, 949: 3,
948: 3, 947: 3, 941: 3, 940: 3, 938: 3, 937: 3, 928: 3, 923: 3, 918: 3, 916: 3, 912: 3, 911: 3,
902: 3, 894: 3, 893: 3, 887: 3, 880: 3, 876: 3, 875: 3, 874: 3, 873: 3, 870: 3, 862: 3, 861: 3,
858: 3, 851: 3, 849: 3, 837: 3, 832: 3, 826: 3, 816: 3, 812: 3, 804: 3, 783: 3, 781: 3, 777: 3,
776: 3, 772: 3, 768: 3, 766: 3, 752: 3, 751: 3, 750: 3, 747: 3, 739: 3, 732: 3, 730: 3, 710: 3,
709: 3, 692: 3, 611: 3, 609: 3, 605: 3, 604: 3, 601: 3, 595: 3, 578: 3, 521: 3, 471: 3, 9914: 2,
8110: 2, 7827: 2, 7041: 2, 6720: 2, 6715: 2, 6350: 2, 6252: 2, 5923: 2, 5378: 2, 5371: 2, 5076: 2,
4835: 2, 4784: 2, 4723: 2, 4554: 2, 4484: 2, 4128: 2, 4056: 2, 4035: 2, 3864: 2, 3787: 2, 3753: 2,
3735: 2, 3671: 2, 3607: 2, 3598: 2, 3587: 2, 3543: 2, 3449: 2, 3443: 2, 3429: 2, 3397: 2, 3379: 2,
3354: 2, 3317: 2, 3304: 2, 3283: 2, 3266: 2, 3242: 2, 3231: 2, 3219: 2, 3217: 2, 3185: 2, 3175: 2,
3051: 2, 3044: 2, 3016: 2, 2995: 2, 2986: 2, 2912: 2, 2868: 2, 2867: 2, 2853: 2, 2824: 2, 2788: 2,
2786: 2, 2731: 2, 2728: 2, 2720: 2, 2714: 2, 2711: 2, 2688: 2, 2670: 2, 2665: 2, 2639: 2, 2617: 2,
2614: 2, 2600: 2, 2557: 2, 2528: 2, 2524: 2, 2520: 2, 2514: 2, 2451: 2, 2416: 2, 2410: 2, 2388: 2,
2375: 2, 2370: 2, 2342: 2, 2341: 2, 2337: 2, 2334: 2, 2333: 2, 2312: 2, 2308: 2, 2283: 2, 2280: 2,
2274: 2, 2267: 2, 2252: 2, 2229: 2, 2220: 2, 2216: 2, 2214: 2, 2187: 2, 2184: 2, 2176: 2, 2167: 2,
2146: 2, 2141: 2, 2136: 2, 2135: 2, 2131: 2, 2130: 2, 2129: 2, 2116: 2, 2099: 2, 2095: 2, 2093: 2,
2078: 2, 2074: 2, 2069: 2, 2054: 2, 2041: 2, 2031: 2, 2019: 2, 2015: 2, 2013: 2, 2012: 2, 2005: 2,
2002: 2, 1996: 2, 1986: 2, 1982: 2, 1976: 2, 1974: 2, 1971: 2, 1962: 2, 1950: 2, 1948: 2, 1939: 2,
1931: 2, 1930: 2, 1924: 2, 1921: 2, 1915: 2, 1907: 2, 1904: 2, 1902: 2, 1899: 2, 1874: 2, 1866: 2,
1853: 2, 1849: 2, 1843: 2, 1839: 2, 1833: 2, 1830: 2, 1816: 2, 1800: 2, 1788: 2, 1784: 2, 1776: 2,
1770: 2, 1767: 2, 1759: 2, 1755: 2, 1748: 2, 1742: 2, 1737: 2, 1727: 2, 1726: 2, 1723: 2, 1720: 2,
1710: 2, 1697: 2, 1680: 2, 1674: 2, 1656: 2, 1655: 2, 1651: 2, 1646: 2, 1645: 2, 1644: 2, 1633: 2,
1623: 2, 1621: 2, 1619: 2, 1609: 2, 1608: 2, 1606: 2, 1601: 2, 1600: 2, 1592: 2, 1587: 2, 1584: 2,
1580: 2, 1559: 2, 1554: 2, 1550: 2, 1546: 2, 1544: 2, 1542: 2, 1538: 2, 1537: 2, 1532: 2, 1530: 2,
1528: 2, 1517: 2, 1514: 2, 1512: 2, 1509: 2, 1508: 2, 1507: 2, 1501: 2, 1500: 2, 1480: 2, 1479: 2,
1478: 2, 1477: 2, 1474: 2, 1473: 2, 1468: 2, 1461: 2, 1438: 2, 1429: 2, 1428: 2, 1427: 2, 1423: 2,
1420: 2, 1412: 2, 1404: 2, 1391: 2, 1387: 2, 1385: 2, 1379: 2, 1368: 2, 1365: 2, 1364: 2, 1361: 2,
1345: 2, 1336: 2, 1333: 2, 1332: 2, 1331: 2, 1330: 2, 1329: 2, 1328: 2, 1324: 2, 1320: 2, 1306: 2,
1301: 2, 1299: 2, 1298: 2, 1293: 2, 1289: 2, 1282: 2, 1281: 2, 1278: 2, 1275: 2, 1272: 2, 1269: 2,
1268: 2, 1267: 2, 1264: 2, 1257: 2, 1256: 2, 1253: 2, 1249: 2, 1240: 2, 1238: 2, 1237: 2, 1235: 2,
1228: 2, 1226: 2, 1225: 2, 1224: 2, 1220: 2, 1215: 2, 1214: 2, 1213: 2, 1212: 2, 1208: 2, 1205: 2,
1204: 2, 1197: 2, 1196: 2, 1194: 2, 1188: 2, 1185: 2, 1183: 2, 1181: 2, 1172: 2, 1166: 2, 1162: 2,
1161: 2, 1158: 2, 1155: 2, 1154: 2, 1149: 2, 1146: 2, 1144: 2, 1141: 2, 1137: 2, 1136: 2, 1131: 2,
1124: 2, 1108: 2, 1107: 2, 1105: 2, 1095: 2, 1094: 2, 1092: 2, 1085: 2, 1080: 2, 1079: 2, 1073: 2,
1062: 2, 1061: 2, 1057: 2, 1056: 2, 1052: 2, 1049: 2, 1046: 2, 1042: 2, 1041: 2, 1040: 2, 1030: 2,
1026: 2, 1024: 2, 1022: 2, 1019: 2, 1017: 2, 1002: 2, 997: 2, 993: 2, 976: 2, 970: 2, 968: 2, 963:
2, 961: 2, 958: 2, 945: 2, 939: 2, 936: 2, 935: 2, 929: 2, 927: 2, 924: 2, 919: 2, 915: 2, 914: 2,
913: 2, 909: 2, 907: 2, 886: 2, 885: 2, 883: 2, 872: 2, 860: 2, 856: 2, 852: 2, 847: 2, 841: 2,
838: 2, 829: 2, 819: 2, 818: 2, 815: 2, 809: 2, 808: 2, 805: 2, 799: 2, 794: 2, 792: 2, 790: 2,
782: 2, 780: 2, 773: 2, 767: 2, 748: 2, 745: 2, 721: 2, 695: 2, 694: 2, 693: 2, 685: 2, 681: 2,
674: 2, 655: 2, 647: 2, 645: 2, 607: 2, 582: 2, 581: 2, 487: 2, 151768: 1, 118635: 1, 79966: 1,
67037: 1, 66927: 1, 66875: 1, 66842: 1, 66278: 1, 63346: 1, 62666: 1, 55167: 1, 53734: 1, 48371: 1
, 48218: 1, 46235: 1, 46222: 1, 42871: 1, 42456: 1, 42195: 1, 41501: 1, 41164: 1, 40114: 1, 39686:
1, 39509: 1, 38769: 1, 38018: 1, 37484: 1, 36517: 1, 36240: 1, 36086: 1, 35701: 1, 34076: 1, 33911
: 1, 33030: 1, 32925: 1, 32870: 1, 31467: 1, 31460: 1, 29063: 1, 28711: 1, 28121: 1, 26399: 1, 261
87: 1, 25730: 1, 25406: 1, 25357: 1, 24134: 1, 24075: 1, 24019: 1, 23963: 1, 23873: 1, 23612: 1, 2
3327: 1, 22934: 1, 22874: 1, 22022: 1, 21906: 1, 21599: 1, 21426: 1, 21137: 1, 20998: 1, 20769: 1,
20379: 1, 20370: 1, 20216: 1, 19969: 1, 19809: 1, 19301: 1, 19176: 1, 19130: 1, 19084: 1, 19080: 1
, 18719: 1, 18585: 1, 18450: 1, 18440: 1, 18259: 1, 18196: 1, 18051: 1, 18015: 1, 17956: 1, 17888:
1, 17771: 1, 17621: 1, 17579: 1, 17566: 1, 17442: 1, 17393: 1, 17247: 1, 17142: 1, 17086: 1, 16946
: 1, 16864: 1, 16811: 1, 16743: 1, 16709: 1, 16620: 1, 16605: 1, 16364: 1, 16311: 1, 16054: 1, 158
03: 1, 15570: 1, 15534: 1, 15487: 1, 15455: 1, 15438: 1, 15281: 1, 15183: 1, 15160: 1, 15055: 1, 1
4803: 1, 14802: 1, 14635: 1, 14625: 1, 14612: 1, 14578: 1, 14536: 1, 14439: 1, 14378: 1, 14277: 1,
14193: 1, 14190: 1, 14100: 1, 14066: 1, 13572: 1, 13542: 1, 13492: 1, 13443: 1, 13379: 1, 13179: 1
, 13162: 1, 13144: 1, 13064: 1, 13063: 1, 13028: 1, 12927: 1, 12818: 1, 12787: 1, 12751: 1, 12723:
1, 12672: 1, 12637: 1, 12594: 1, 12527: 1, 12510: 1, 12509: 1, 12486: 1, 12466: 1, 12399: 1, 12335
: 1, 12219: 1, 12176: 1, 12161: 1, 12142: 1, 12121: 1, 12061: 1, 12057: 1, 11972: 1, 11956: 1, 119
36: 1, 11917: 1, 11916: 1, 11808: 1, 11805: 1, 11796: 1, 11708: 1, 11679: 1, 11635: 1, 11590: 1, 1
1589: 1, 11585: 1, 11576: 1, 11541: 1, 11493: 1, 11475: 1, 11417: 1, 11190: 1, 11082: 1, 11069: 1,
10996: 1, 10962: 1, 10917: 1, 10879: 1, 10850: 1, 10794: 1, 10744: 1, 10738: 1, 10717: 1, 10713: 1
, 10602: 1, 10532: 1, 10516: 1, 10474: 1, 10409: 1, 10345: 1, 10292: 1, 10245: 1, 10161: 1, 10132:
1, 10087: 1, 10075: 1, 10073: 1, 9989: 1, 9870: 1, 9853: 1, 9832: 1, 9830: 1, 9806: 1, 9745: 1, 970
9: 1, 9570: 1, 9569: 1, 9555: 1, 9545: 1, 9514: 1, 9513: 1, 9463: 1, 9360: 1, 9334: 1, 9324: 1, 931
0: 1, 9309: 1, 9298: 1, 9285: 1, 9217: 1, 9215: 1, 9207: 1, 9076: 1, 9067: 1, 9063: 1, 9057: 1, 905
6: 1, 9046: 1, 9037: 1, 8999: 1, 8975: 1, 8947: 1, 8943: 1, 8936: 1, 8885: 1, 8838: 1, 8834: 1, 880
9: 1, 8786: 1, 8696: 1, 8661: 1, 8600: 1, 8525: 1, 8518: 1, 8515: 1, 8488: 1, 8487: 1, 8354: 1, 834
0: 1, 8316: 1, 8283: 1, 8264: 1, 8188: 1, 8168: 1, 8163: 1, 8081: 1, 8077: 1, 8071: 1, 8058: 1, 805
3: 1, 8052: 1, 8048: 1, 7991: 1, 7964: 1, 7944: 1, 7939: 1, 7915: 1, 7898: 1, 7893: 1, 7883: 1, 787
1: 1, 7864: 1, 7835: 1, 7795: 1, 7788: 1, 7783: 1, 7771: 1, 7767: 1, 7761: 1, 7752: 1, 7750: 1, 771
8: 1, 7693: 1, 7659: 1, 7617: 1, 7603: 1, 7523: 1, 7511: 1, 7507: 1, 7462: 1, 7417: 1, 7379: 1, 737
2: 1, 7341: 1, 7335: 1, 7307: 1, 7288: 1, 7254: 1, 7243: 1, 7240: 1, 7221: 1, 7208: 1, 7186: 1, 717
7: 1, 7169: 1, 7162: 1, 7161: 1, 7142: 1, 7132: 1, 7109: 1, 7067: 1, 7062: 1, 7026: 1, 7013: 1, 700
5: 1, 6998: 1, 6968: 1, 6954: 1, 6941: 1, 6930: 1, 6928: 1, 6918: 1, 6899: 1, 6892: 1, 6887: 1, 688
3: 1, 6863: 1, 6845: 1, 6821: 1, 6798: 1, 6796: 1, 6782: 1, 6740: 1, 6726: 1, 6699: 1, 6694: 1, 668
8: 1, 6685: 1, 6670: 1, 6648: 1, 6631: 1, 6629: 1, 6621: 1, 6620: 1, 6617: 1, 6615: 1, 6600: 1, 657
7: 1, 6570: 1, 6542: 1, 6541: 1, 6508: 1, 6492: 1, 6471: 1, 6449: 1, 6423: 1, 6418: 1, 6415: 1, 640
8: 1, 6405: 1, 6376: 1, 6354: 1, 6339: 1, 6336: 1, 6324: 1, 6288: 1, 6277: 1, 6274: 1, 6268: 1, 624
7: 1, 6241: 1, 6239: 1, 6234: 1, 6208: 1, 6206: 1, 6160: 1, 6152: 1, 6119: 1, 6113: 1, 6107: 1, 609
9: 1, 6092: 1, 6074: 1, 6069: 1, 6066: 1, 6054: 1, 6045: 1, 6035: 1, 6004: 1, 6001: 1, 5989: 1, 597
6: 1, 5975: 1, 5949: 1, 5945: 1, 5937: 1, 5917: 1, 5908: 1, 5901: 1, 5884: 1, 5879: 1, 5847: 1, 584
3: 1, 5817: 1, 5811: 1, 5805: 1, 5789: 1, 5770: 1, 5761: 1, 5746: 1, 5728: 1, 5700: 1, 5694: 1, 569
1: 1, 5677: 1, 5671: 1, 5670: 1, 5661: 1, 5658: 1, 5653: 1, 5627: 1, 5623: 1, 5584: 1, 5577: 1, 555
9: 1, 5551: 1, 5547: 1, 5542: 1, 5530: 1, 5529: 1, 5521: 1, 5510: 1, 5501: 1, 5493: 1, 5464: 1, 546
0: 1, 5452: 1, 5447: 1, 5446: 1, 5431: 1, 5405: 1, 5370: 1, 5358: 1, 5338: 1, 5332: 1, 5329: 1, 532
7: 1, 5305: 1, 5286: 1, 5260: 1, 5227: 1, 5203: 1, 5195: 1, 5178: 1, 5139: 1, 5123: 1, 5121: 1, 509
5: 1, 5092: 1, 5074: 1, 5070: 1, 5063: 1, 5049: 1, 5046: 1, 5044: 1, 5038: 1, 5030: 1, 5028: 1, 501
9: 1, 4996: 1, 4993: 1, 4970: 1, 4964: 1, 4959: 1, 4946: 1, 4926: 1, 4918: 1, 4895: 1, 4892: 1, 488
9: 1, 4888: 1, 4882: 1, 4870: 1, 4864: 1, 4862: 1, 4859: 1, 4858: 1, 4857: 1, 4849: 1, 4841: 1, 483
8: 1, 4832: 1, 4830: 1, 4827: 1, 4817: 1, 4805: 1, 4802: 1, 4794: 1, 4788: 1, 4783: 1, 4773: 1, 476
8: 1, 4748: 1, 4705: 1, 4702: 1, 4701: 1, 4685: 1, 4681: 1, 4667: 1, 4654: 1, 4634: 1, 4633: 1, 462
```

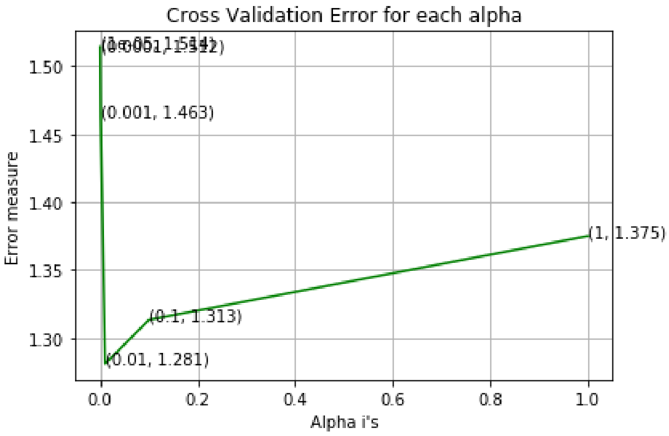
```
4: 1, 4621: 1, 4618: 1, 4616: 1, 4593: 1, 4571: 1, 4565: 1, 4564: 1, 4557: 1, 4556: 1, 4544: 1, 453
7: 1, 4536: 1, 4535: 1, 4518: 1, 4513: 1, 4512: 1, 4507: 1, 4500: 1, 4498: 1, 4495: 1, 4490: 1, 448
8: 1, 4487: 1, 4469: 1, 4458: 1, 4455: 1, 4447: 1, 4441: 1, 4431: 1, 4430: 1, 4421: 1, 4418: 1, 441
5: 1, 4409: 1, 4387: 1, 4384: 1, 4382: 1, 4370: 1, 4363: 1, 4357: 1, 4348: 1, 4321: 1, 4317: 1, 430
5: 1, 4293: 1, 4274: 1, 4268: 1, 4260: 1, 4258: 1, 4256: 1, 4253: 1, 4252: 1, 4248: 1, 4240: 1, 423
6: 1, 4235: 1, 4227: 1, 4222: 1, 4220: 1, 4219: 1, 4209: 1, 4203: 1, 4189: 1, 4188: 1, 4182: 1, 417
9: 1, 4176: 1, 4164: 1, 4160: 1, 4157: 1, 4152: 1, 4148: 1, 4143: 1, 4136: 1, 4134: 1, 4122: 1, 411
8: 1, 4114: 1, 4113: 1, 4111: 1, 4109: 1, 4108: 1, 4100: 1, 4098: 1, 4088: 1, 4072: 1, 4061: 1, 405
9: 1, 4047: 1, 4038: 1, 4016: 1, 4008: 1, 4006: 1, 3991: 1, 3988: 1, 3984: 1, 3981: 1, 3980: 1, 397
8: 1, 3976: 1, 3959: 1, 3956: 1, 3939: 1, 3936: 1, 3929: 1, 3926: 1, 3921: 1, 3920: 1, 3916: 1, 391
0: 1, 3909: 1, 3904: 1, 3891: 1, 3887: 1, 3885: 1, 3878: 1, 3871: 1, 3868: 1, 3857: 1, 3844: 1, 383
2: 1, 3821: 1, 3803: 1, 3802: 1, 3785: 1, 3774: 1, 3773: 1, 3768: 1, 3743: 1, 3740: 1, 3738: 1, 372
6: 1, 3723: 1, 3722: 1, 3721: 1, 3718: 1, 3714: 1, 3713: 1, 3712: 1, 3708: 1, 3699: 1, 3698: 1, 369
4: 1, 3691: 1, 3688: 1, 3686: 1, 3684: 1, 3682: 1, 3675: 1, 3665: 1, 3663: 1, 3660: 1, 3658: 1, 364
0: 1, 3639: 1, 3638: 1, 3626: 1, 3616: 1, 3605: 1, 3603: 1, 3600: 1, 3590: 1, 3589: 1, 3583: 1, 358
2: 1, 3580: 1, 3578: 1, 3574: 1, 3572: 1, 3571: 1, 3566: 1, 3564: 1, 3560: 1, 3555: 1, 3546: 1, 354
0: 1, 3539: 1, 3537: 1, 3534: 1, 3531: 1, 3526: 1, 3513: 1, 3510: 1, 3506: 1, 3496: 1, 3495: 1, 349
3: 1, 3485: 1, 3475: 1, 3470: 1, 3469: 1, 3461: 1, 3457: 1, 3454: 1, 3450: 1, 3446: 1, 3442: 1, 343
1: 1, 3426: 1, 3421: 1, 3416: 1, 3414: 1, 3406: 1, 3405: 1, 3404: 1, 3400: 1, 3399: 1, 3385: 1, 338
3: 1, 3374: 1, 3373: 1, 3364: 1, 3360: 1, 3358: 1, 3353: 1, 3352: 1, 3350: 1, 3349: 1, 3347: 1, 334
4: 1, 3339: 1, 3334: 1, 3323: 1, 3313: 1, 3303: 1, 3297: 1, 3296: 1, 3294: 1, 3285: 1, 3279: 1, 327
6: 1, 3273: 1, 3271: 1, 3265: 1, 3262: 1, 3260: 1, 3258: 1, 3255: 1, 3247: 1, 3245: 1, 3240: 1, 323
6: 1, 3234: 1, 3232: 1, 3227: 1, 3226: 1, 3223: 1, 3222: 1, 3221: 1, 3212: 1, 3211: 1, 3209: 1, 320
8: 1, 3203: 1, 3199: 1, 3197: 1, 3196: 1, 3195: 1, 3194: 1, 3183: 1, 3180: 1, 3179: 1, 3174: 1, 316
8: 1, 3161: 1, 3157: 1, 3152: 1, 3150: 1, 3149: 1, 3146: 1, 3143: 1, 3138: 1, 3137: 1, 3133: 1, 313
0: 1, 3127: 1, 3123: 1, 3115: 1, 3100: 1, 3088: 1, 3086: 1, 3085: 1, 3081: 1, 3079: 1, 3077: 1, 307
6: 1, 3066: 1, 3063: 1, 3058: 1, 3057: 1, 3052: 1, 3049: 1, 3041: 1, 3038: 1, 3035: 1, 3033: 1, 303
2: 1, 3028: 1, 3025: 1, 3023: 1, 3014: 1, 3013: 1, 3012: 1, 3008: 1, 3005: 1, 3001: 1, 3000: 1, 299
7: 1, 2994: 1, 2988: 1, 2981: 1, 2978: 1, 2967: 1, 2956: 1, 2955: 1, 2953: 1, 2941: 1, 2937: 1, 293
3: 1, 2932: 1, 2929: 1, 2921: 1, 2918: 1, 2910: 1, 2908: 1, 2903: 1, 2899: 1, 2898: 1, 2894: 1, 288
6: 1, 2885: 1, 2883: 1, 2873: 1, 2869: 1, 2866: 1, 2863: 1, 2862: 1, 2847: 1, 2842: 1, 2840: 1, 283
8: 1, 2834: 1, 2830: 1, 2829: 1, 2826: 1, 2813: 1, 2811: 1, 2810: 1, 2809: 1, 2807: 1, 2806: 1, 280
4: 1, 2800: 1, 2799: 1, 2787: 1, 2781: 1, 2779: 1, 2778: 1, 2777: 1, 2767: 1, 2766: 1, 2763: 1, 274
5: 1, 2725: 1, 2724: 1, 2723: 1, 2722: 1, 2718: 1, 2716: 1, 2703: 1, 2700: 1, 2699: 1, 2695: 1, 269
2: 1, 2691: 1, 2689: 1, 2687: 1, 2684: 1, 2673: 1, 2664: 1, 2662: 1, 2651: 1, 2650: 1, 2648: 1, 264
6: 1, 2644: 1, 2635: 1, 2631: 1, 2623: 1, 2622: 1, 2619: 1, 2616: 1, 2608: 1, 2606: 1, 2604: 1, 260
1: 1, 2598: 1, 2590: 1, 2588: 1, 2581: 1, 2579: 1, 2578: 1, 2576: 1, 2574: 1, 2572: 1, 2571: 1, 256
6: 1, 2565: 1, 2560: 1, 2558: 1, 2552: 1, 2550: 1, 2547: 1, 2530: 1, 2529: 1, 2527: 1, 2516: 1, 251
5: 1, 2513: 1, 2512: 1, 2509: 1, 2508: 1, 2505: 1, 2503: 1, 2499: 1, 2498: 1, 2496: 1, 2494: 1, 249
3: 1, 2490: 1, 2489: 1, 2484: 1, 2477: 1, 2476: 1, 2475: 1, 2473: 1, 2470: 1, 2469: 1, 2468: 1, 246
7: 1, 2466: 1, 2464: 1, 2459: 1, 2455: 1, 2453: 1, 2452: 1, 2443: 1, 2440: 1, 2438: 1, 2437: 1, 243
0: 1, 2425: 1, 2422: 1, 2421: 1, 2418: 1, 2414: 1, 2411: 1, 2408: 1, 2407: 1, 2406: 1, 2405: 1, 240
2: 1, 2400: 1, 2399: 1, 2397: 1, 2394: 1, 2393: 1, 2389: 1, 2386: 1, 2385: 1, 2384: 1, 2381: 1, 237
9: 1, 2378: 1, 2373: 1, 2371: 1, 2369: 1, 2368: 1, 2360: 1, 2355: 1, 2349: 1, 2348: 1, 2345: 1, 234
4: 1, 2339: 1, 2338: 1, 2336: 1, 2328: 1, 2325: 1, 2324: 1, 2320: 1, 2315: 1, 2307: 1, 2297: 1, 229
0: 1, 2288: 1, 2287: 1, 2284: 1, 2277: 1, 2275: 1, 2262: 1, 2259: 1, 2257: 1, 2255: 1, 2254: 1, 225
3: 1, 2250: 1, 2247: 1, 2246: 1, 2244: 1, 2238: 1, 2235: 1, 2228: 1, 2226: 1, 2225: 1, 2224: 1, 222
3: 1, 2221: 1, 2219: 1, 2217: 1, 2212: 1, 2210: 1, 2207: 1, 2205: 1, 2204: 1, 2203: 1, 2201: 1, 219
5: 1, 2193: 1, 2190: 1, 2188: 1, 2181: 1, 2180: 1, 2177: 1, 2171: 1, 2170: 1, 2163: 1, 2160: 1, 215
3: 1, 2150: 1, 2144: 1, 2142: 1, 2140: 1, 2139: 1, 2138: 1, 2137: 1, 2127: 1, 2125: 1, 2124: 1, 212
2: 1, 2120: 1, 2117: 1, 2108: 1, 2105: 1, 2101: 1, 2100: 1, 2097: 1, 2096: 1, 2094: 1, 2085: 1, 208
2: 1, 2076: 1, 2065: 1, 2064: 1, 2062: 1, 2060: 1, 2056: 1, 2051: 1, 2047: 1, 2045: 1, 2044: 1, 203
9: 1, 2038: 1, 2034: 1, 2029: 1, 2027: 1, 2026: 1, 2025: 1, 2023: 1, 2014: 1, 2009: 1, 2008: 1, 200
7: 1, 2006: 1, 1995: 1, 1988: 1, 1985: 1, 1984: 1, 1980: 1, 1977: 1, 1975: 1, 1973: 1, 1972: 1, 197
0: 1, 1967: 1, 1964: 1, 1961: 1, 1956: 1, 1945: 1, 1944: 1, 1938: 1, 1935: 1, 1934: 1, 1933: 1, 192
2: 1, 1920: 1, 1919: 1, 1917: 1, 1910: 1, 1906: 1, 1900: 1, 1898: 1, 1897: 1, 1894: 1, 1893: 1, 189
2: 1, 1891: 1, 1890: 1, 1889: 1, 1888: 1, 1886: 1, 1882: 1, 1881: 1, 1878: 1, 1877: 1, 1876: 1, 187
5: 1, 1873: 1, 1872: 1, 1870: 1, 1869: 1, 1868: 1, 1864: 1, 1858: 1, 1851: 1, 1850: 1, 1847: 1, 184
5: 1, 1842: 1, 1841: 1, 1838: 1, 1836: 1, 1835: 1, 1834: 1, 1832: 1, 1827: 1, 1820: 1, 1814: 1, 181
2: 1, 1811: 1, 1808: 1, 1805: 1, 1804: 1, 1802: 1, 1801: 1, 1797: 1, 1795: 1, 1793: 1, 1791: 1, 178
9: 1, 1785: 1, 1783: 1, 1782: 1, 1781: 1, 1778: 1, 1777: 1, 1775: 1, 1774: 1, 1772: 1, 1766: 1, 176
5: 1, 1764: 1, 1763: 1, 1760: 1, 1751: 1, 1749: 1, 1746: 1, 1745: 1, 1740: 1, 1736: 1, 1735: 1, 173
0: 1, 1728: 1, 1725: 1, 1724: 1, 1721: 1, 1719: 1, 1716: 1, 1713: 1, 1712: 1, 1707: 1, 1705: 1, 170
4: 1, 1701: 1, 1698: 1, 1694: 1, 1690: 1, 1689: 1, 1687: 1, 1686: 1, 1684: 1, 1683: 1, 1682: 1, 167
9: 1, 1677: 1, 1673: 1, 1672: 1, 1670: 1, 1668: 1, 1667: 1, 1666: 1, 1665: 1, 1660: 1, 1649: 1, 164
1: 1, 1640: 1, 1636: 1, 1631: 1, 1626: 1, 1620: 1, 1618: 1, 1617: 1, 1614: 1, 1613: 1, 1610: 1, 160
3: 1, 1599: 1, 1597: 1, 1595: 1, 1593: 1, 1591: 1, 1588: 1, 1585: 1, 1583: 1, 1578: 1, 1574: 1, 157
3: 1, 1567: 1, 1565: 1, 1564: 1, 1563: 1, 1558: 1, 1551: 1, 1545: 1, 1540: 1, 1536: 1, 1534: 1, 153
3: 1, 1529: 1, 1527: 1, 1525: 1, 1520: 1, 1519: 1, 1518: 1, 1510: 1, 1505: 1, 1504: 1, 1503: 1, 149
7: 1, 1496: 1, 1493: 1, 1492: 1, 1489: 1, 1486: 1, 1485: 1, 1484: 1, 1483: 1, 1482: 1, 1481: 1, 147
2: 1, 1469: 1, 1467: 1, 1466: 1, 1463: 1, 1462: 1, 1460: 1, 1458: 1, 1451: 1, 1449: 1, 1448: 1, 144
7: 1, 1446: 1, 1444: 1, 1442: 1, 1440: 1, 1439: 1, 1437: 1, 1434: 1, 1432: 1, 1431: 1, 1426: 1, 142
5: 1, 1424: 1, 1421: 1, 1418: 1, 1417: 1, 1415: 1, 1411: 1, 1409: 1, 1407: 1, 1406: 1, 1403: 1, 140
1: 1, 1400: 1, 1399: 1, 1397: 1, 1395: 1, 1394: 1, 1381: 1, 1377: 1, 1376: 1, 1363: 1, 1362: 1, 136
0: 1, 1359: 1, 1356: 1, 1355: 1, 1354: 1, 1349: 1, 1348: 1, 1344: 1, 1343: 1, 1339: 1, 1335: 1, 133
4: 1, 1323: 1, 1321: 1, 1319: 1, 1318: 1, 1317: 1, 1312: 1, 1310: 1, 1307: 1, 1303: 1, 1302: 1, 130
0: 1, 1297: 1, 1296: 1, 1292: 1, 1291: 1, 1288: 1, 1286: 1, 1280: 1, 1276: 1, 1274: 1, 1273: 1, 126
```

```
0: 1, 1297: 1, 1296: 1, 1292: 1, 1291: 1, 1288: 1, 1286: 1, 1280: 1, 1276: 1, 1274: 1, 1273: 1, 126
1: 1, 1259: 1, 1258: 1, 1254: 1, 1252: 1, 1251: 1, 1250: 1, 1246: 1, 1244: 1, 1239: 1, 1234: 1, 123
2: 1, 1231: 1, 1223: 1, 1222: 1, 1210: 1, 1207: 1, 1206: 1, 1202: 1, 1199: 1, 1195: 1, 1190: 1, 118
4: 1, 1178: 1, 1176: 1, 1173: 1, 1171: 1, 1170: 1, 1169: 1, 1167: 1, 1164: 1, 1163: 1, 1159: 1, 115
7: 1, 1153: 1, 1150: 1, 1148: 1, 1147: 1, 1145: 1, 1142: 1, 1140: 1, 1139: 1, 1138: 1, 1135: 1, 113
4: 1, 1133: 1, 1127: 1, 1122: 1, 1121: 1, 1116: 1, 1115: 1, 1112: 1, 1109: 1, 1100: 1, 1099: 1, 109
6: 1, 1090: 1, 1086: 1, 1083: 1, 1082: 1, 1078: 1, 1076: 1, 1075: 1, 1074: 1, 1068: 1, 1067: 1, 106
5: 1, 1063: 1, 1055: 1, 1054: 1, 1047: 1, 1044: 1, 1038: 1, 1021: 1, 1020: 1, 1014: 1, 1013: 1, 101
2: 1, 1010: 1, 1004: 1, 999: 1, 998: 1, 989: 1, 988: 1, 987: 1, 981: 1, 978: 1, 974: 1, 959: 1,
957: 1, 954: 1, 952: 1, 932: 1, 931: 1, 926: 1, 925: 1, 921: 1, 910: 1, 906: 1, 904: 1, 903: 1,
899: 1, 892: 1, 891: 1, 890: 1, 888: 1, 882: 1, 879: 1, 871: 1, 869: 1, 868: 1, 853: 1, 844: 1,
836: 1, 834: 1, 830: 1, 827: 1, 824: 1, 820: 1, 811: 1, 798: 1, 795: 1, 789: 1, 788: 1, 759: 1,
743: 1, 698: 1, 653: 1, 579: 1, 503: 1})
```

In [49]:

```
# Train a Logistic regression+Calibration model using text features which are one-hot encoded
alpha = [10 ** x for x in range(-5, 1)]

# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0.5,
# 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 Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link:
#------------------------------


cv_log_error_array=[]
for i in alpha:
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_text_feature_onehotCoding, y_train)

    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_text_feature_onehotCoding, y_train)
    predict_y = sig_clf.predict_proba(cv_text_feature_onehotCoding)
    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
    print('For values of alpha = ', i, "The log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))

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()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_text_feature_onehotCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_text_feature_onehotCoding, y_train)

predict_y = sig_clf.predict_proba(train_text_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_text_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_text_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.5143069505709046
For values of alpha =  0.0001 The log loss is: 1.5117748280810117
For values of alpha =  0.001 The log loss is: 1.4631832833821021
For values of alpha =  0.01 The log loss is: 1.2810406538068342
For values of alpha =  0.1 The log loss is: 1.313435977576093
For values of alpha =  1 The log loss is: 1.375070302009839
```



```
For values of best alpha =  0.01 The train log loss is: 0.8717855736127258
For values of best alpha =  0.01 The cross validation log loss is: 1.2810406538068342
For values of best alpha =  0.01 The test log loss is: 1.1678793551365518
```

Q. Is the Text feature stable across all the data sets (Test, Train, Cross validation)?

Ans. Yes, it seems like!

In [50]:

```python
def get_intersec_text(df):
    df_text_vec = CountVectorizer(min_df=3)
    df_text_fea = df_text_vec.fit_transform(df['TEXT'])
    df_text_features = df_text_vec.get_feature_names()

    df_text_fea_counts = df_text_fea.sum(axis=0).A1
    df_text_fea_dict = dict(zip(list(df_text_features),df_text_fea_counts))
    len1 = len(set(df_text_features))
    len2 = len(set(train_text_features) & set(df_text_features))
    return len1,len2
```

In [51]:

```python
len1,len2 = get_intersec_text(X_test)
print(np.round((len2/len1)*100, 3), "% of word of test data appeared in train data")
len1,len2 = get_intersec_text(X_cv)
print(np.round((len2/len1)*100, 3), "% of word of Cross Validation appeared in train data")
```

```
96.978 % of word of test data appeared in train data
97.305 % of word of Cross Validation appeared in train data
```

4. Machine Learning Models

In [52]:

```python
#Data preparation for ML models.

#Misc. functionns for ML models


def predict_and_plot_confusion_matrix(train_x, train_y,test_x, test_y, clf):
    clf.fit(train_x, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x, train_y)
    pred_y = sig_clf.predict(test_x)
    pred_y = sig_clf.predict(test_x)

    # for calculating log_loss we willl provide the array of probabilities belongs to each class
    print("Log loss :",log_loss(test_y, sig_clf.predict_proba(test_x)))
    # calculating the number of data points that are misclassified
    global mis_per
    mis_per=(np.count_nonzero((pred_y- test_y))/test_y.shape[0])*100
    print("Number of mis-classified points :", np.count_nonzero((pred_y- test_y))/test_y.shape[0])
    plot_confusion_matrix(test_y, pred_y)
```

In [53]:

```
def report_log_loss(train_x, train_y, test_x, test_y,  clf):
    clf.fit(train_x, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x, train_y)
    sig_clf_probs = sig_clf.predict_proba(test_x)
    return log_loss(test_y, sig_clf_probs, eps=1e-15)
```

In [54]:

```
# this function will be used just for naive bayes
# for the given indices, we will print the name of the features
# and we will check whether the feature present in the test point text or not
def get_impfeature_names(indices, text, gene, var, no_features):
    gene_count_vec = CountVectorizer(ngram_range=(1,2))
    var_count_vec = CountVectorizer(ngram_range=(1,2))
    text_count_vec = CountVectorizer(min_df=3,ngram_range=(1,2))

    gene_vec = gene_count_vec.fit(X_train['Gene'])
    var_vec  = var_count_vec.fit(X_train['Variation'])
    text_vec = text_count_vec.fit(X_train['TEXT'])

    fea1_len = len(gene_vec.get_feature_names())
    fea2_len = len(var_count_vec.get_feature_names())

    word_present = 0
    for i,v in enumerate(indices):
        if (v < fea1_len):
            word = gene_vec.get_feature_names()[v]
            yes_no = True if word == gene else False
            if yes_no:
                word_present += 1
                print(i, "Gene feature [{}] present in test data point [{}]".format(word,yes_no))
        elif (v < fea1_len+fea2_len):
            word = var_vec.get_feature_names()[v-(fea1_len)]
            yes_no = True if word == var else False
            if yes_no:
                word_present += 1
                print(i, "variation feature [{}] present in test data point [{}]".format(word,yes_r
o))
        else:
            word = text_vec.get_feature_names()[v-(fea1_len+fea2_len)]
            yes_no = True if word in text.split() else False
            if yes_no:
                word_present += 1
                print(i, "Text feature [{}] present in test data point [{}]".format(word,yes_no))

    print("Out of the top ",no_features," features ", word_present, "are present in query point")
```

Stacking the three types of features

In [186]:

```
# merging gene, variance and text features

# building train, test and cross validation data sets
# a = [[1, 2],
#      [3, 4]]
# b = [[4, 5],
#      [6, 7]]
# hstack(a, b) = [[1, 2, 4, 5],
# hstack(a, b) = [[1, 2, 4, 5],
#                 [ 3, 4, 6, 7]]

#train_gene_var_onehotCoding =
hstack((train_gene_feature_onehotCoding,train_variation_feature_onehotCoding))
#test_gene_var_onehotCoding =
hstack((test_gene_feature_onehotCoding,test_variation_feature_onehotCoding))
#cv_gene_var_onehotCoding =
hstack((cv_gene_feature_onehotCoding,cv_variation_feature_onehotCoding))
#
#train_x_onehotCoding = hstack((train_gene_var_onehotCoding,
train_text_feature_onehotCoding)).tocsr()
#train_y = np.array(list(X_train['Class']))
#
#test_x_onehotCoding = hstack((test_gene_var_onehotCoding,
test_text_feature_onehotCoding)).tocsr()
#test_y = np.array(list(X_test['Class']))
#
#cv_x_onehotCoding = hstack((cv_gene_var_onehotCoding, cv_text_feature_onehotCoding)).tocsr()
#cv_y = np.array(list(X_cv['Class']))
#
#
#train_gene_var_responseCoding =
np.hstack((train_gene_feature_responseCoding,train_variation_feature_responseCoding))
#test_gene_var_responseCoding =
np.hstack((test_gene_feature_responseCoding,test_variation_feature_responseCoding))
#cv_gene_var_responseCoding =
np.hstack((cv_gene_feature_responseCoding,cv_variation_feature_responseCoding))
#
#train_x_responseCoding = np.hstack((train_gene_var_responseCoding,
train_text_feature_responseCoding))
#test_x_responseCoding = np.hstack((test_gene_var_responseCoding,
test_text_feature_responseCoding))
#cv_x_responseCoding = np.hstack((cv_gene_var_responseCoding, cv_text_feature_responseCoding))

train_gene_var_tfidf=hstack((train_gene_feature_tfidf,train_variation_feature_tfidf))
test_gene_var_tfidf=hstack((test_gene_feature_tfidf,test_variation_feature_tfidf))
cv_gene_var_tfidf=hstack((cv_gene_feature_tfidf,cv_variation_feature_tfidf))

train_x_tfidf=hstack((train_gene_var_tfidf,train_text_feature_tfidf))
test_x_tfidf=hstack((test_gene_var_tfidf,test_text_feature_tfidf))
cv_x_tfidf=hstack((cv_gene_var_tfidf,cv_text_feature_tfidf))
```

In [56]:

```
print("One hot encoding features :")
print("(number of data points * number of features) in train data = ", train_x_onehotCoding.shape)
print("(number of data points * number of features) in test data = ", test_x_onehotCoding.shape)
print("(number of data points * number of features) in cross validation data =", cv_x_onehotCoding
.shape)
```

```
One hot encoding features :
(number of data points * number of features) in train data =  (2124, 769211)
(number of data points * number of features) in test data =  (665, 769211)
(number of data points * number of features) in cross validation data = (532, 769211)
```

In [57]:

```
print(" Response encoding features :")
print("(number of data points * number of features) in train data = ", train_x_responseCoding.shape)
print("(number of data points * number of features) in test data = ", test_x_responseCoding.shape)
print("(number of data points * number of features) in cross validation data =",
cv_x_responseCoding.shape)
```

```
 Response encoding features :
(number of data points * number of features) in train data =  (2124, 27)
(number of data points * number of features) in test data =  (665, 27)
(number of data points * number of features) in cross validation data = (532, 27)
```

In [187]:

```
print(" tfidf encoding features :")
print("(number of data points * number of features) in train data = ", train_x_tfidf.shape)
print("(number of data points * number of features) in train data = ", train_x_tfidf.shape)
print("(number of data points * number of features) in test data = ", test_x_tfidf.shape)
print("(number of data points * number of features) in cross validation data =", cv_x_tfidf.shape)
```

```
 tfidf encoding features :
(number of data points * number of features) in train data =  (2124, 3444)
(number of data points * number of features) in test data =  (665, 3444)
(number of data points * number of features) in cross validation data = (532, 3444)
```

In [59]:

```
#groupDF = [X_train, X_cv, X_test]
#for i in groupDF:
#    i.drop(['Class'], axis=1, inplace=True)
```

4.1. Base Line Model

4.1.1. Naive Bayes

4.1.1.1. Hyper parameter tuning

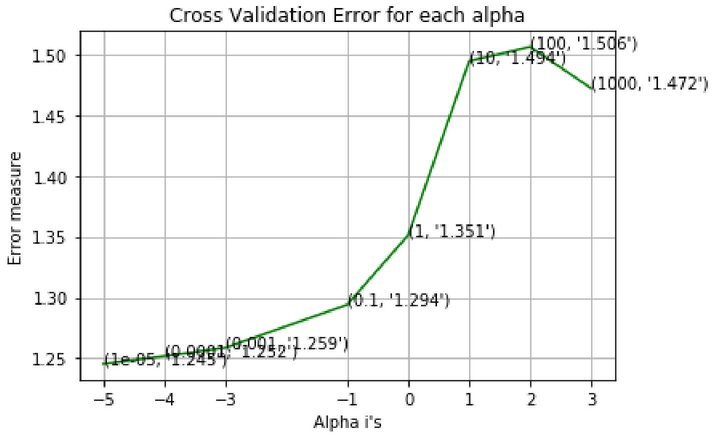
In [60]:

```
from prettytable import PrettyTable

x = PrettyTable()

x.field_names = ["ML Model", "Train Log Loss","CV Log Loss", "Test Log Loss","Misclassification %"]
```

In [61]:

```
# find more about Multinomial Naive base function here http://scikit-
learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html
# -------------------------
# default paramters
# sklearn.naive_bayes.MultinomialNB(alpha=1.0, fit_prior=True, class_prior=None)

# some of methods of MultinomialNB()
# fit(X, y[, sample_weight]) Fit Naive Bayes classifier according to X, y
# predict(X) Perform classification on an array of test vectors X.
# predict_log_proba(X) Return log-probability estimates for the test vector X.
# -----------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/naive-bayes-
algorithm-1/
# -----------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# 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: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/naive-bayes-
algorithm-1/
# -----------------------


alpha = [0.00001, 0.0001, 0.001, 0.1, 1, 10, 100,1000]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = MultinomialNB(alpha=i)
    clf.fit(train_x_tfidf, y_train)
    clf.fit(train_x_tfidf, y_train)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_tfidf, y_train)
    sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
    cv_log_error_array.append(log_loss(y_cv, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(y_cv, sig_clf_probs))

fig, ax = plt.subplots()
ax.plot(np.log10(alpha), cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],str(txt)), (np.log10(alpha[i]),cv_log_error_array[i]))
plt.grid()
plt.xticks(np.log10(alpha))
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = MultinomialNB(alpha=alpha[best_alpha])
clf.fit(train_x_tfidf, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, y_train)


predict_y = sig_clf.predict_proba(train_x_tfidf)
train_ll=log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train,
predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_x_tfidf)
cv_ll=log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_tfidf)
test_ll=log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-05
Log Loss : 1.245401013356264
for alpha = 0.0001
Log Loss : 1.2518259595269317
for alpha = 0.001
Log Loss : 1.258637606741886
for alpha = 0.1
Log Loss : 1.2940837840410697
for alpha = 1
Log Loss : 1.3512928304426302
for alpha = 10
Log Loss : 1.4943723990295623
for alpha = 100
Log Loss : 1.5062717285108203
for alpha = 1000
Log Loss : 1.4721096315440183
```



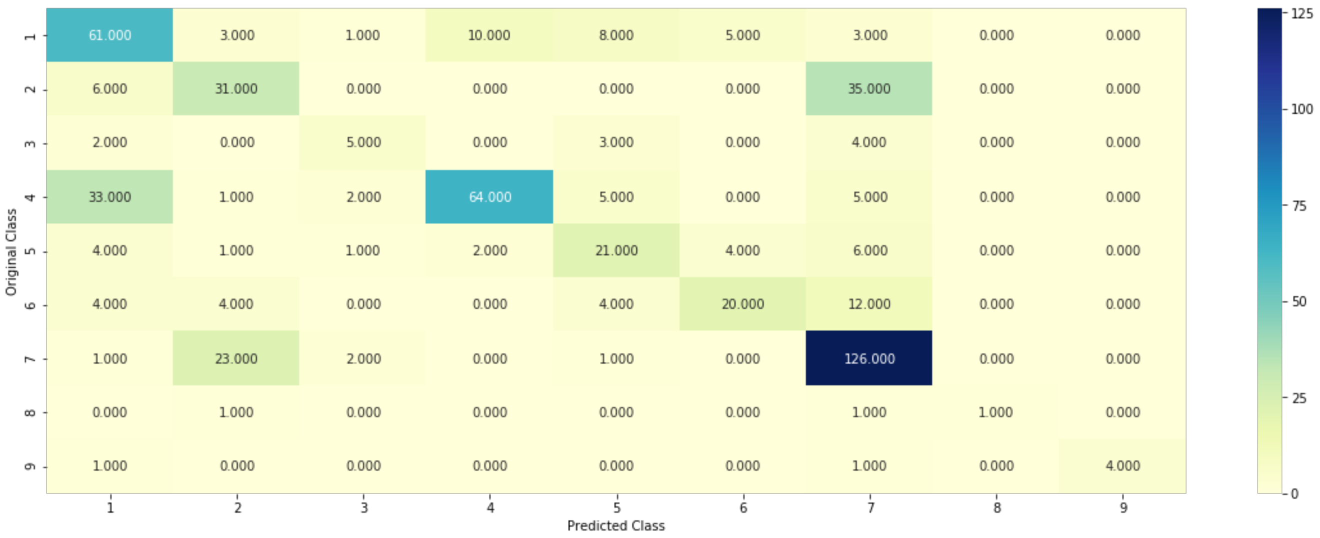
```
For values of best alpha =  1e-05 The train log loss is: 0.6141817542885196
For values of best alpha =  1e-05 The cross validation log loss is: 1.245401013356264
For values of best alpha =  1e-05 The cross validation log loss is: 1.245401013356264
For values of best alpha =  1e-05 The test log loss is: 1.2111163418820183
```

4.1.1.2. Testing the model with best hyper paramters

In [62]:

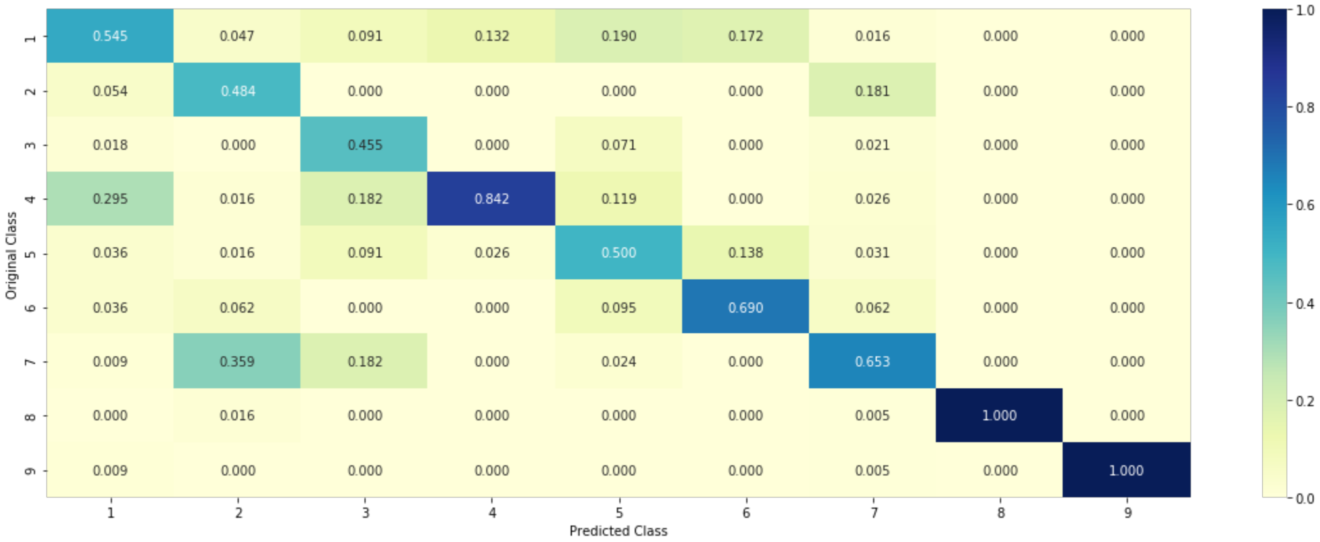
```python
# find more about Multinomial Naive base function here http://scikit-
learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html
# -------------------------
# default paramters
# sklearn.naive_bayes.MultinomialNB(alpha=1.0, fit_prior=True, class_prior=None)

# some of methods of MultinomialNB()
# fit(X, y[, sample_weight]) Fit Naive Bayes classifier according to X, y
# predict(X) Perform classification on an array of test vectors X.
# predict_log_proba(X) Return log-probability estimates for the test vector X.
# -----------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/naive-bayes-
algorithm-1/
# -----------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# 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
# ----------------------------

clf = MultinomialNB(alpha=alpha[best_alpha])
clf.fit(train_x_tfidf, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, y_train)
sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
# to avoid rounding error while multiplying probabilites we use log-probability estimates
print("Log Loss :",log_loss(y_cv, sig_clf_probs))
mis_per=(np.count_nonzero((sig_clf.predict(cv_x_tfidf)- y_cv))/y_cv.shape[0])*100
print("Number of missclassified point :", np.count_nonzero((sig_clf.predict(cv_x_tfidf)- y_cv))/y_
cv.shape[0])
x.add_row(["Naive Bayes(TFIDF)",train_ll,cv_ll,test_ll,mis_per])
plot_confusion_matrix(y_cv, sig_clf.predict(cv_x_tfidf.toarray()))
```

```
Log Loss : 1.245401013356264
Number of missclassified point : 0.37406015037593987
-------------------- Confusion matrix --------------------
```

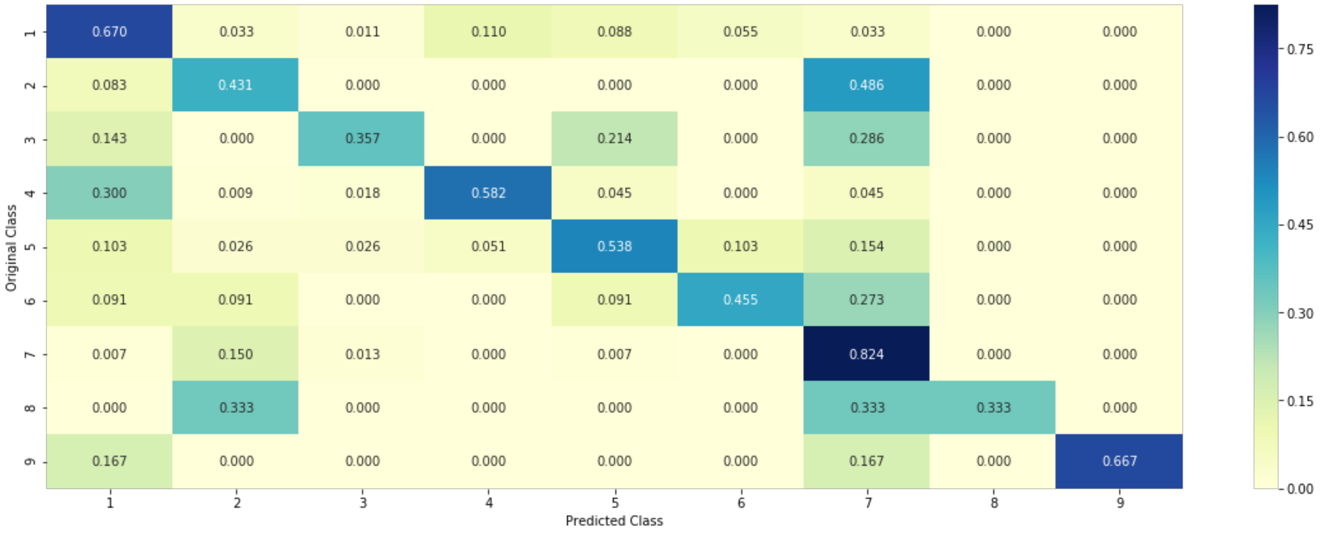


	1	2	3	4	5	6	7	8	9
1	61.000	3.000	1.000	10.000	8.000	5.000	3.000	0.000	0.000
2	6.000	31.000	0.000	0.000	0.000	0.000	35.000	0.000	0.000
3	2.000	0.000	5.000	0.000	3.000	0.000	4.000	0.000	0.000
4	33.000	1.000	2.000	64.000	5.000	0.000	5.000	0.000	0.000
5	4.000	1.000	1.000	2.000	21.000	4.000	6.000	0.000	0.000
6	4.000	4.000	0.000	0.000	4.000	20.000	12.000	0.000	0.000
7	1.000	23.000	2.000	0.000	1.000	0.000	126.000	0.000	0.000
8	0.000	1.000	0.000	0.000	0.000	0.000	1.000	1.000	0.000
9	1.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	4.000

```
-------------------- Precision matrix (Columm Sum=1) --------------------
```



```
-------------------- Recall matrix (Row sum=1) --------------------
```



4.1.1.3. Feature Importance, Correctly classified point

In [63]:

```
test_point_index = 1
no_feature = 100
predicted_cls = sig_clf.predict(test_x_tfidf.tocsr()[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(test_x_tfidf.tocsr()
[test_point_index]),4))
print("Actual Class :", y_test[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], X_test['TEXT'].iloc[test_point_index],X_test['Gene'].iloc[test_poi
nt_index],X_test['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0589 0.0547 0.0147 0.0824 0.0418 0.0404 0.7001 0.0045 0.0024]]
Actual Class : 7
--------------------------------------------------
Out of the top  100  features  0 are present in query point
```

4.1.1.4. Feature Importance, Incorrectly classified point

In [64]:

```
test_point_index = 100
no_feature = 100
predicted_cls = sig_clf.predict(test_x_tfidf.tocsr()[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(test_x_tfidf.tocsr()
[test_point_index]),4))
print("Actual Class :", y_test[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], X_test['TEXT'].iloc[test_point_index],X_test['Gene'].iloc[test_poi
nt_index],X_test['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[0.6457 0.0575 0.0155 0.0889 0.044  0.0425 0.0986 0.0048 0.0026]]
Actual Class : 1
--------------------------------------------------
Out of the top  100  features  0 are present in query point
```

4.2. K Nearest Neighbour Classification

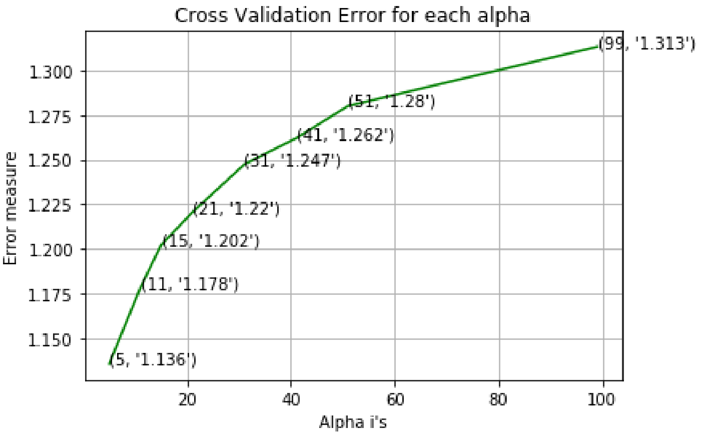
4.2.1. Hyper parameter tuning

In [65]:

```
# find more about KNeighborsClassifier() here http://scikit-
learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
# -------------------------
# default parameter
# KNeighborsClassifier(n_neighbors=5, weights=’uniform’, algorithm=’auto’, leaf_size=30, p=2,
# 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-nearest-ne
ighbors-geometric-intuition-with-a-toy-example-1/
#--------------------------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method=’sigmoid’, cv=3)
#
# 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 = [5, 11, 15, 21, 31, 41, 51, 99]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = KNeighborsClassifier(n_neighbors=i)
    clf.fit(train_x_tfidf, y_train)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_tfidf, y_train)
    sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
    cv_log_error_array.append(log_loss(y_cv, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(y_cv, sig_clf_probs))

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

```
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],str(txt)), (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()


best_alpha = np.argmin(cv_log_error_array)
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_x_tfidf, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, y_train)

predict_y = sig_clf.predict_proba(train_x_tfidf)
train_ll=log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train,
predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_x_tfidf)
cv_ll=log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_tfidf)
test_ll=log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 5
Log Loss : 1.135750482766513
for alpha = 11
Log Loss : 1.178231428538192
for alpha = 15
Log Loss : 1.2022566993612505
for alpha = 21
Log Loss : 1.2203849319429523
for alpha = 31
Log Loss : 1.2472904249925698
for alpha = 41
Log Loss : 1.2619254722221698
for alpha = 51
Log Loss : 1.2799913128119973
for alpha = 99
Log Loss : 1.3130887392595583
```



```
For values of best alpha =  5 The train log loss is: 0.9110478758175219
For values of best alpha =  5 The cross validation log loss is: 1.135750482766513
For values of best alpha =  5 The test log loss is: 1.0775710518751482
```

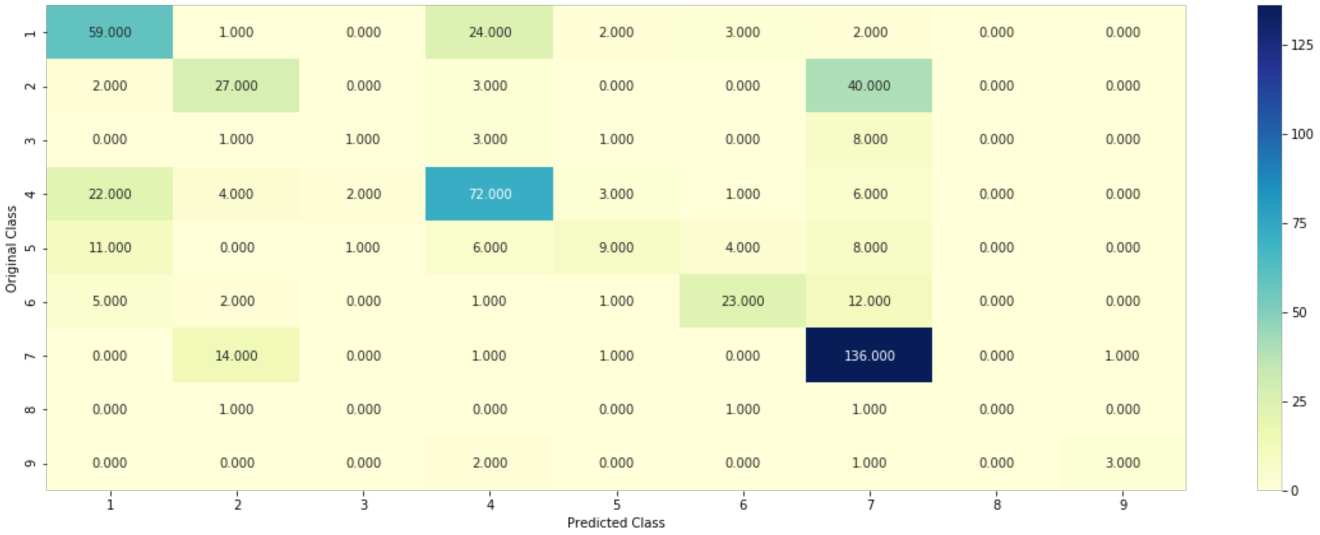
4.2.2. Testing the model with best hyper paramters

In [66]:

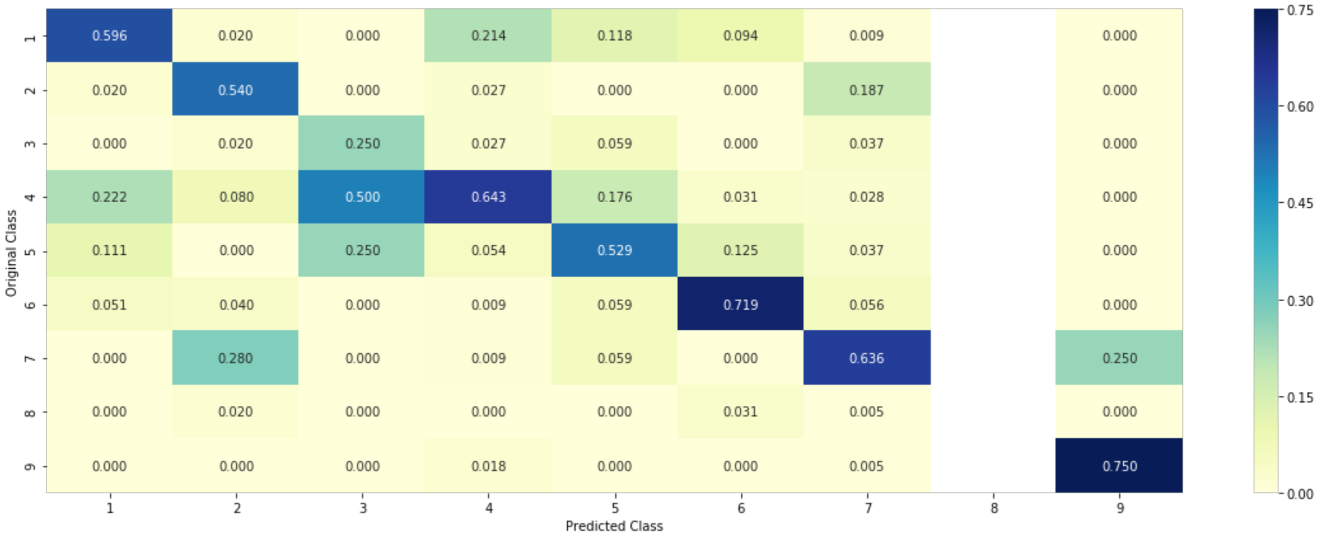
```
# find more about KNeighborsClassifier() here http://scikit-
```

```
learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
# -------------------------
# default parameter
# KNeighborsClassifier(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2,
# 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-nearest-ne
ighbors-geometric-intuition-with-a-toy-example-1/
#-------------------------------------
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
predict_and_plot_confusion_matrix(train_x_tfidf, y_train, cv_x_tfidf, y_cv, clf)
x.add_row(["KNN(TFIDF)",train_ll,cv_ll,test_ll,mis_per])
```

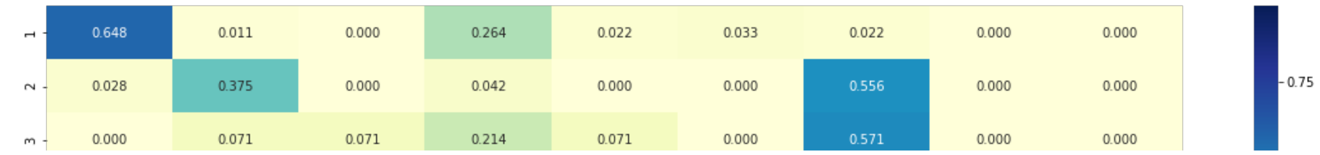
```
Log loss : 1.135750482766513
Number of mis-classified points : 0.37969924812030076
-------------------- Confusion matrix --------------------
```

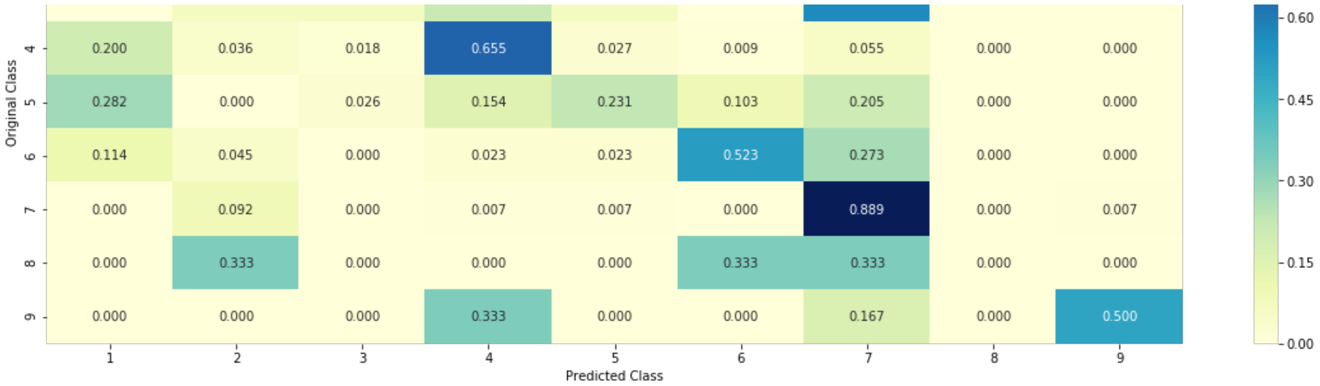


```
-------------------- Precision matrix (Columm Sum=1) --------------------
```



```
-------------------- Recall matrix (Row sum=1) --------------------
```





4.2.3.Sample Query point -1

In [67]:

```
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_x_tfidf, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, y_train)

test_point_index = 1
predicted_cls = sig_clf.predict(test_x_tfidf.tocsr()[0].reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Actual Class :", y_test[test_point_index])
neighbors = clf.kneighbors(test_x_tfidf.tocsr()[test_point_index].reshape(1, -1), alpha[best_alpha]
)
print("The ",alpha[best_alpha]," nearest neighbours of the test points belongs to classes",y_train
[neighbors[1][0]])
print("Fequency of nearest points :",Counter(y_train[neighbors[1][0]]))
```

```
Predicted Class : 4
Actual Class : 7
The  5  nearest neighbours of the test points belongs to classes [7 7 2 7 7]
Fequency of nearest points : Counter({7: 4, 2: 1})
```

4.2.4. Sample Query Point-2

In [68]:

```
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_x_tfidf, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, y_train)

test_point_index = 100

predicted_cls = sig_clf.predict(test_x_tfidf.tocsr()[test_point_index].reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Actual Class :", y_test[test_point_index])
neighbors = clf.kneighbors(test_x_tfidf.tocsr()[test_point_index].reshape(1, -1), alpha[best_alpha]
)
print("the k value for knn is",alpha[best_alpha],"and the nearest neighbours of the test points be
longs to classes",y_train[neighbors[1][0]])
print("Fequency of nearest points :",Counter(y_train[neighbors[1][0]]))
```

```
Predicted Class : 2
Actual Class : 1
the k value for knn is 5 and the nearest neighbours of the test points belongs to classes [1 2 7 2
7]
Fequency of nearest points : Counter({2: 2, 7: 2, 1: 1})
```

4.3. Logistic Regression

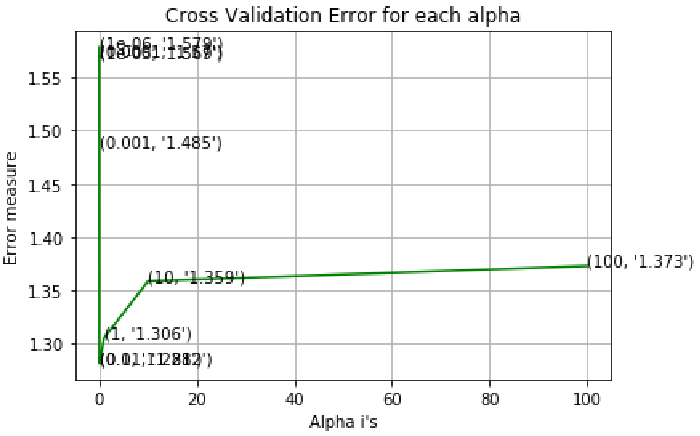
4.3.1. With Class balancing

4.3.1.1. Hyper paramter tuning

In [102]:

```
# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0
=0.0, power_t=0.5,
# 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 Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/geometric-in
tuition-1/
#------------------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# 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 = [10 ** x for x in range(-6, 3)]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = SGDClassifier(class_weight='balanced', alpha=i, penalty='l2', loss='log', random_state=42
)
    clf.fit(train_x_onehotCoding, y_train)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, y_train)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(y_cv, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(y_cv, sig_clf_probs))

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],str(txt)), (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()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
dom_state=42)
clf.fit(train_x_onehotCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, y_train)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
train_ll=log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train,
predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
cv_ll=log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_onehotCoding)
test_ll=log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.5785169808056114
for alpha = 1e-05
Log Loss : 1.5691834800645295
for alpha = 0.0001
Log Loss : 1.5702423695151086
for alpha = 0.001
Log Loss : 1.4850516296136442
for alpha = 0.01
Log Loss : 1.2819931301433514
for alpha = 0.1
Log Loss : 1.2812676609291282
for alpha = 1
Log Loss : 1.3055326215302245
for alpha = 10
Log Loss : 1.3586130305840243
for alpha = 100
Log Loss : 1.3728137804682896
```



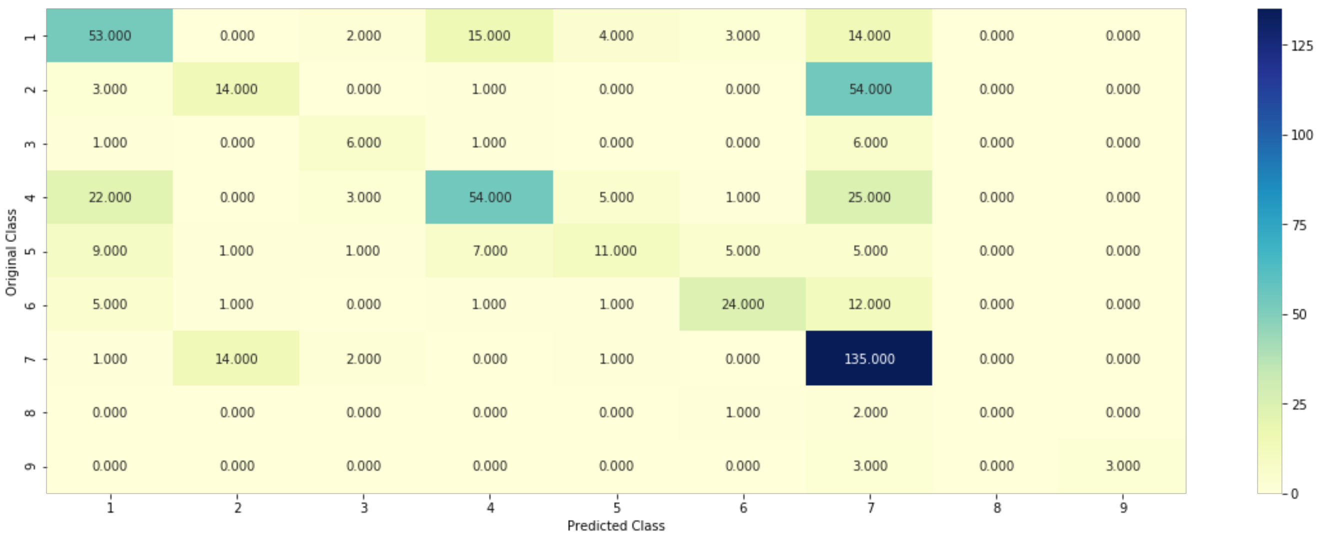
```
For values of best alpha =  0.1 The train log loss is: 0.8577608101533614
For values of best alpha =  0.1 The cross validation log loss is: 1.2812676609291282
For values of best alpha =  0.1 The test log loss is: 1.1527262672860776
```

4.3.1.2. Testing the model with best hyper paramters

In [103]:

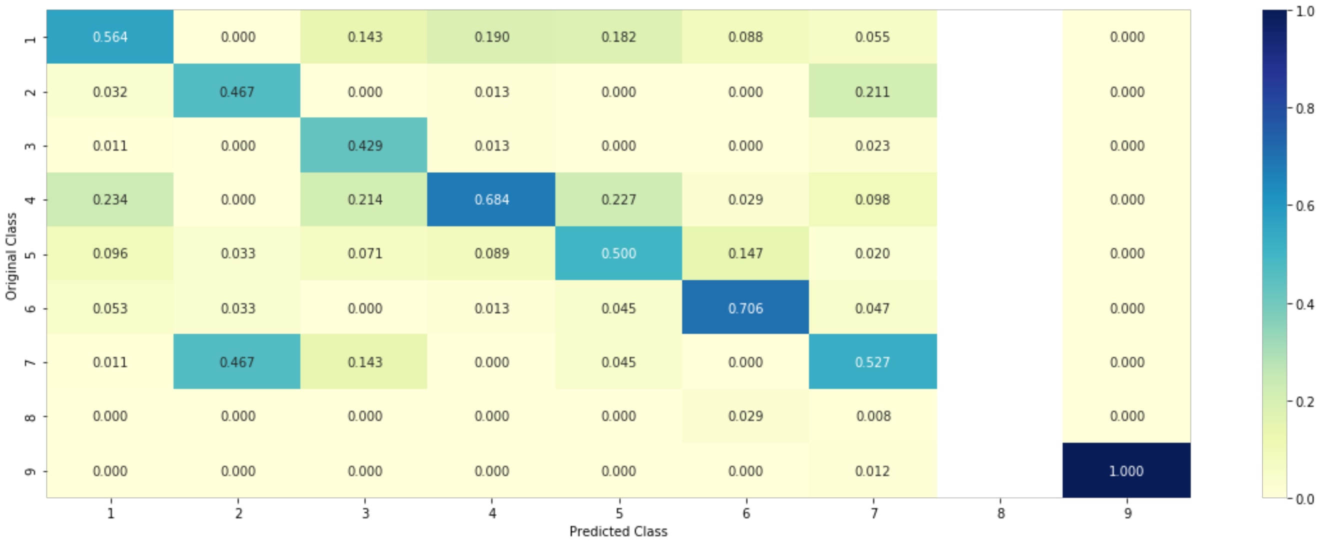
```
# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0
=0.0, power_t=0.5,
# 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 Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/geometric-in
tuition-1/
#------------------------------
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
dom_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, y_train, cv_x_onehotCoding, y_cv, clf)
x.add_row(["LR_Class_Balancing(one hot)",train_ll,cv_ll,test_ll,mis_per])
```

```
Log loss : 1.2812676609291282
Number of mis-classified points : 0.43609022556390975
-------------------- Confusion matrix --------------------
```



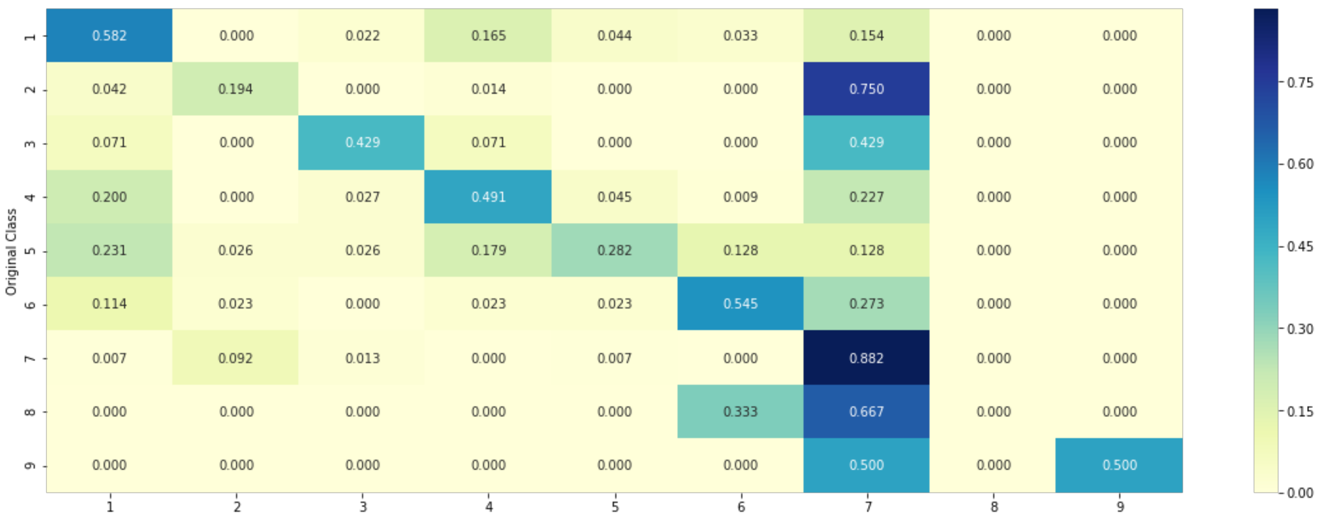
	1	2	3	4	5	6	7	8	9
1	53.000	0.000	2.000	15.000	4.000	3.000	14.000	0.000	0.000
2	3.000	14.000	0.000	1.000	0.000	0.000	54.000	0.000	0.000
3	1.000	0.000	6.000	1.000	0.000	0.000	6.000	0.000	0.000
4	22.000	0.000	3.000	54.000	5.000	1.000	25.000	0.000	0.000
5	9.000	1.000	1.000	7.000	11.000	5.000	5.000	0.000	0.000
6	5.000	1.000	0.000	1.000	1.000	24.000	12.000	0.000	0.000
7	1.000	14.000	2.000	0.000	1.000	0.000	135.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	1.000	2.000	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	3.000	0.000	3.000

```
-------------------- Precision matrix (Columm Sum=1) --------------------
```



	1	2	3	4	5	6	7	8	9
1	0.564	0.000	0.143	0.190	0.182	0.088	0.055		0.000
2	0.032	0.467	0.000	0.013	0.000	0.000	0.211		0.000
3	0.011	0.000	0.429	0.013	0.000	0.000	0.023		0.000
4	0.234	0.000	0.214	0.684	0.227	0.029	0.098		0.000
5	0.096	0.033	0.071	0.089	0.500	0.147	0.020		0.000
6	0.053	0.033	0.000	0.013	0.045	0.706	0.047		0.000
7	0.011	0.467	0.143	0.000	0.045	0.000	0.527		0.000
8	0.000	0.000	0.000	0.000	0.000	0.029	0.008		0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.012		1.000

```
-------------------- Recall matrix (Row sum=1) --------------------
```



	1	2	3	4	5	6	7	8	9
1	0.582	0.000	0.022	0.165	0.044	0.033	0.154	0.000	0.000
2	0.042	0.194	0.000	0.014	0.000	0.000	0.750	0.000	0.000
3	0.071	0.000	0.429	0.071	0.000	0.000	0.429	0.000	0.000
4	0.200	0.000	0.027	0.491	0.045	0.009	0.227	0.000	0.000
5	0.231	0.026	0.026	0.179	0.282	0.128	0.128	0.000	0.000
6	0.114	0.023	0.000	0.023	0.023	0.545	0.273	0.000	0.000
7	0.007	0.092	0.013	0.000	0.007	0.000	0.882	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.333	0.667	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.000	0.500

Predicted Class

4.3.1.3. Feature Importance

In [75]:

```
def get_imp_feature_names(text, indices, removed_ind = []):
    word_present = 0
    tabulte_list = []
    incresingorder_ind = 0
    for i in indices:
        if i < train_gene_feature_onehotCoding.shape[1]:
            tabulte_list.append([incresingorder_ind, "Gene", "Yes"])
        elif i< 18:
            tabulte_list.append([incresingorder_ind,"Variation", "Yes"])
        if ((i > 17) & (i not in removed_ind)) :
            word = train_text_features[i]
            yes_no = True if word in text.split() else False
            if yes_no:
                word_present += 1
            tabulte_list.append([incresingorder_ind,train_text_features[i], yes_no])
        incresingorder_ind += 1
    print(word_present, "most importent features are present in our query point")
    print("-"*50)
    print("The features that are most importent of the ",predicted_cls[0]," class:")
    print (tabulate(tabulte_list, headers=["Index",'Feature name', 'Present or Not']))
```

4.3.1.3.1. Correctly Classified point

In [76]:

```
# from tabulate import tabulate
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
dom_state=42)
clf.fit(train_x_onehotCoding,y_train)
test_point_index = 1
no_feature = 500
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", y_test[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], X_test['TEXT'].iloc[test_point_index],X_test['Gene'].iloc[test_poi
nt_index],X_test['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[4.600e-03 1.640e-02 3.000e-04 8.000e-04 1.600e-03 2.000e-04 9.742
e-01
  1.800e-03 2.000e-04]]
Actual Class : 7
--------------------------------------------------
117 Text feature [constitutive] present in test data point [True]
Out of the top  500  features  1 are present in query point
```

4.3.1.3.2. Incorrectly Classified point

In [77]:

```
test_point_index = 100
no_feature = 500
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", y_test[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], X_test['TEXT'].iloc[test_point_index],X_test['Gene'].iloc[test_poi
nt_index],X_test['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[0.3637 0.1249 0.0171 0.1511 0.055  0.0471 0.2291 0.0066 0.0055]]
Actual Class : 1
--------------------------------------------------
Out of the top  500  features  0 are present in query point
```

4.3.2. Without Class balancing

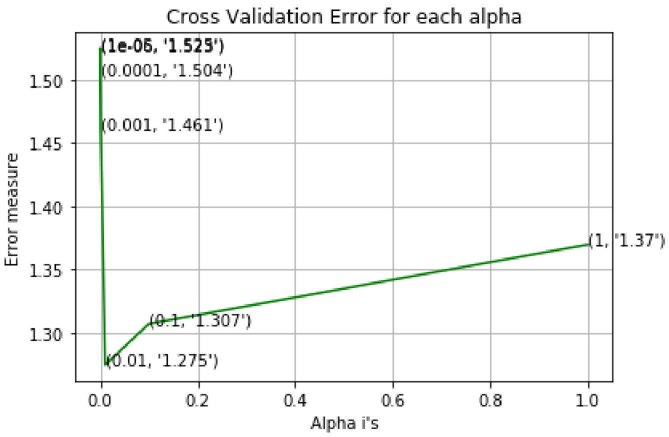
4.3.2.1. Hyper paramter tuning

In [105]:

```
# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0
=0.0, power_t=0.5,
# 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 Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/geometric-in
tuition-1/
#------------------------------



# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# 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 = [10 ** x for x in range(-6, 1)]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_x_onehotCoding, y_train)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, y_train)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(y_cv, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    print("Log Loss :",log_loss(y_cv, sig_clf_probs))

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],str(txt)), (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()
```

```
best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_x_onehotCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, y_train)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
train_ll=log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
cv_ll=log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_onehotCoding)
test_ll=log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.5225157573594494
for alpha = 1e-05
Log Loss : 1.5247006469599462
for alpha = 0.0001
Log Loss : 1.504366862163659
for alpha = 0.001
Log Loss : 1.4609712541867417
for alpha = 0.01
Log Loss : 1.274626138776145
for alpha = 0.1
Log Loss : 1.3071379420420022
for alpha = 1
Log Loss : 1.3696589011726492
```



```
For values of best alpha =  0.01 The train log loss is: 0.8481051995163802
For values of best alpha =  0.01 The cross validation log loss is: 1.274626138776145
For values of best alpha =  0.01 The test log loss is: 1.157109029212211
```

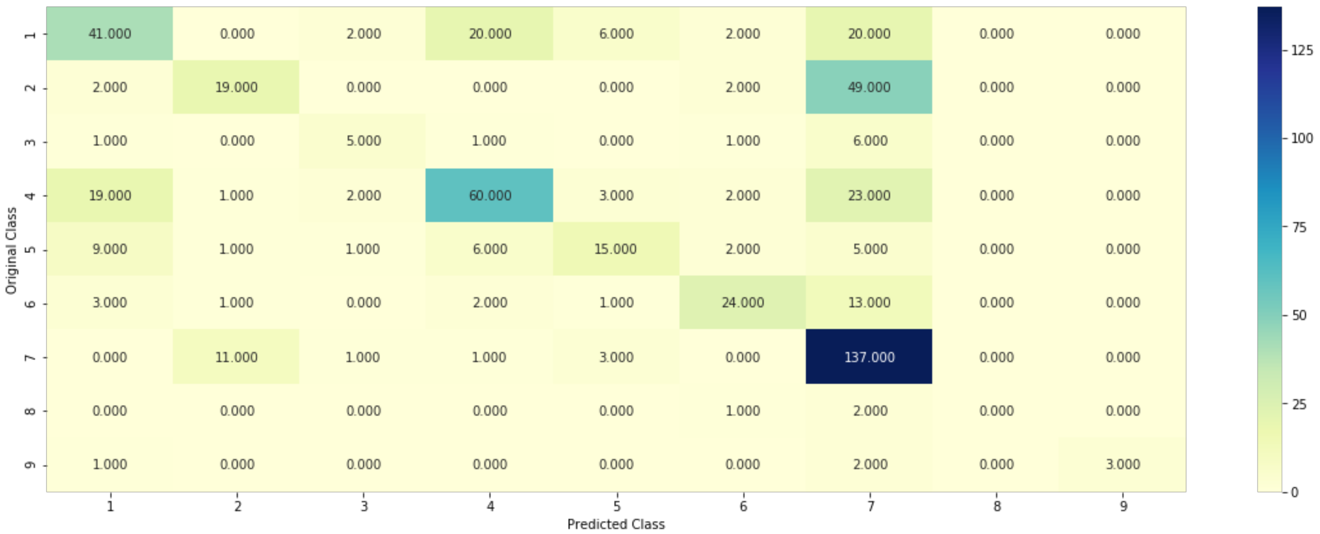
4.3.2.2. Testing model with best hyper parameters

In [106]:

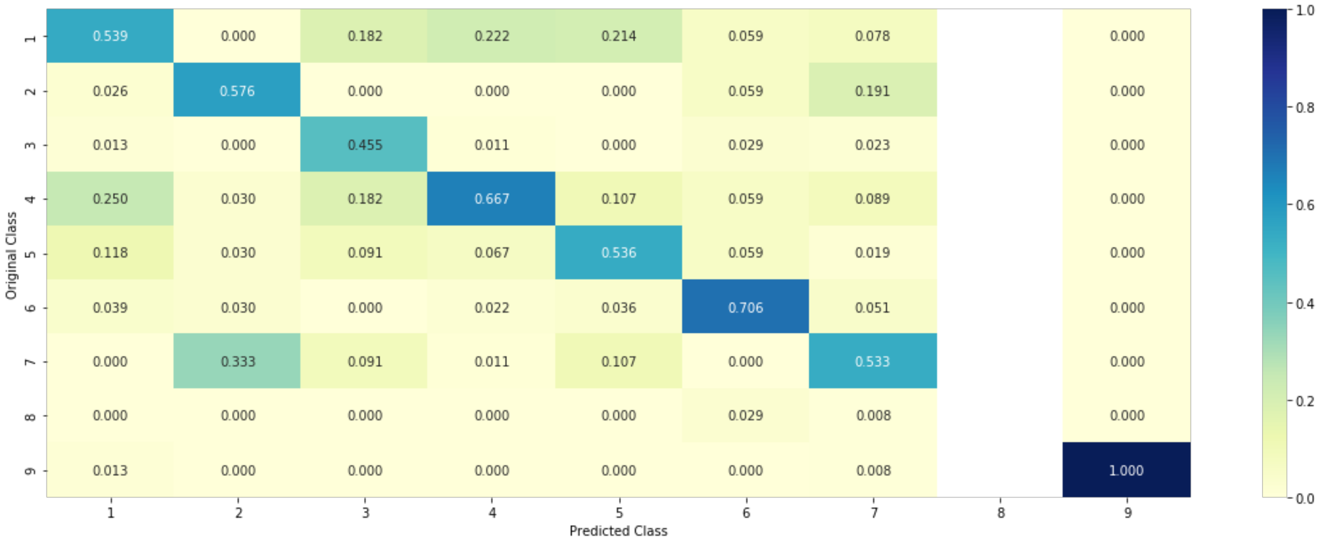
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss=’hinge’, penalty=’l2’, alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate=’optimal’, eta0=0.0, power_t=0.5,
# 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 Gradient Descent.
# predict(X) Predict class labels for samples in X.
```

```
#-------------------------------
# video link:
#------------------------------

clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, y_train, cv_x_onehotCoding, y_cv, clf)
x.add_row(["LR_Without_Class_Balancing(one hot)",train_ll,cv_ll,test_ll,mis_per])
```

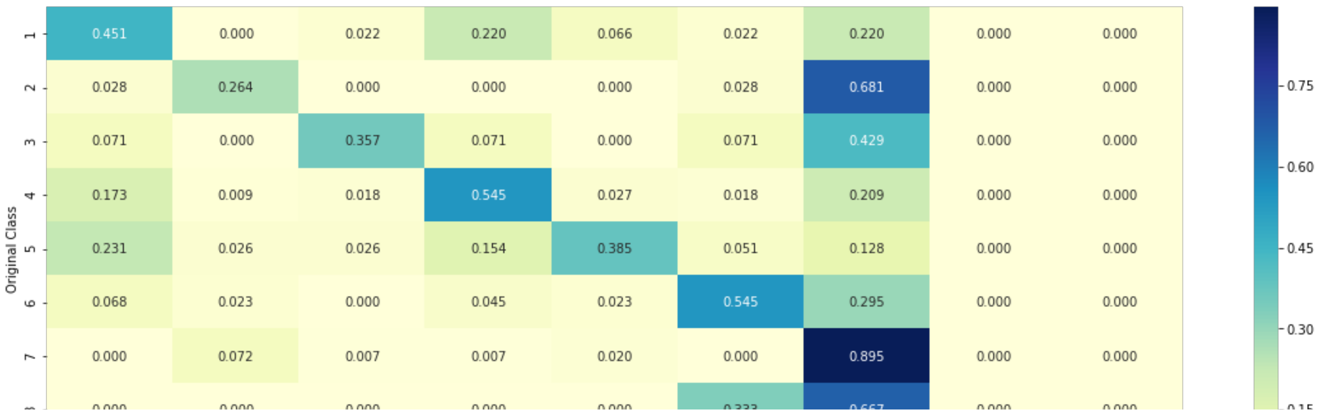
```
Log loss : 1.274626138776145
Number of mis-classified points : 0.42857142857142855
-------------------- Confusion matrix --------------------
```

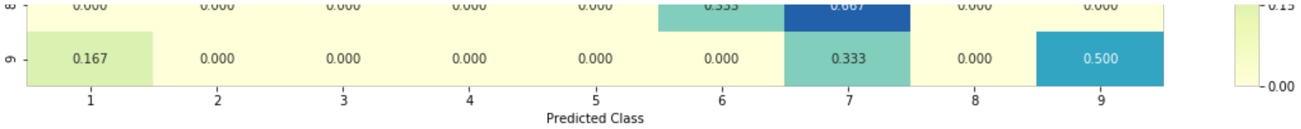


```
-------------------- Precision matrix (Columm Sum=1) --------------------
```



```
-------------------- Recall matrix (Row sum=1) --------------------
```





4.3.2.3. Feature Importance, Correctly Classified point

In [80]:

```
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_x_onehotCoding,y_train)
test_point_index = 1
no_feature = 500
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", y_test[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], X_test['TEXT'].iloc[test_point_index],X_test['Gene'].iloc[test_poi
nt_index],X_test['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[1.400e-02 3.850e-02 5.000e-04 4.400e-03 3.900e-03 1.000e-03 9.366
e-01
  9.000e-04 1.000e-04]]
Actual Class : 7
--------------------------------------------------
226 Text feature [nf] present in test data point [True]
266 Text feature [3t3] present in test data point [True]
Out of the top  500  features  2 are present in query point
```

4.3.2.4. Feature Importance, Inorrectly Classified point

In [81]:

```
test_point_index = 100
no_feature = 500
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", y_test[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], X_test['TEXT'].iloc[test_point_index],X_test['Gene'].iloc[test_poi
nt_index],X_test['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[0.4702 0.109  0.0168 0.1138 0.0506 0.0426 0.1873 0.0055 0.0042]]
Actual Class : 1
--------------------------------------------------
405 Text feature [histologic] present in test data point [True]
Out of the top  500  features  1 are present in query point
```

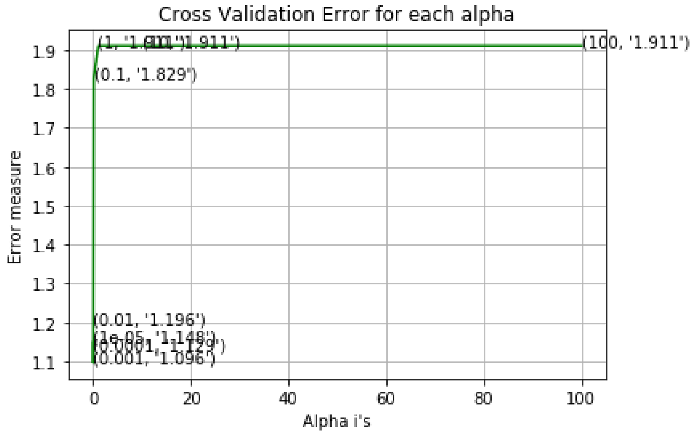
4.4. Linear Support Vector Machines

4.4.1. Hyper paramter tuning

In [82]:

```
# read more about support vector machines with linear kernals here http://scikit-
learn.org/stable/modules/generated/sklearn.svm.SVC.html

# --------------------------------
# default parameters
# SVC(C=1.0, kernel=’rbf’, degree=3, gamma=’auto’, coef0=0.0, shrinking=True, probability=False, tol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape=’ovr’, random_state=None)

# Some of methods of SVM()
# fit(X, y, [sample_weight]) Fit the SVM model according to the given training data.
# predict(X) Perform classification on samples in X.
# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/mathematical-derivation-copy-8/
# --------------------------------



# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method=’sigmoid’, cv=3)
#
# 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 = [10 ** x for x in range(-5, 3)]
cv_log_error_array = []
for i in alpha:
    print("for C =", i)
#     clf = SVC(C=i,kernel='linear',probability=True, class_weight='balanced')
    clf = SGDClassifier( class_weight='balanced', alpha=i, penalty='l2', loss='hinge', random_state=42)
    clf.fit(train_x_tfidf, y_train)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_tfidf, y_train)
    sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
    cv_log_error_array.append(log_loss(y_cv, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    print("Log Loss :",log_loss(y_cv, sig_clf_probs))

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],str(txt)), (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()


best_alpha = np.argmin(cv_log_error_array)
# clf = SVC(C=i,kernel='linear',probability=True, class_weight='balanced')
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='hinge', random_state=42)
clf.fit(train_x_tfidf, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, y_train)

predict_y = sig_clf.predict_proba(train_x_tfidf)
train_ll=log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_x_tfidf)
cv_ll=log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_tfidf)
test_ll=log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for C = 1e-05
Log Loss : 1.148465860963182
for C = 0.0001
Log Loss : 1.129262197787261
for C = 0.001
Log Loss : 1.095803788579623
for C = 0.01
Log Loss : 1.1958044397641217
for C = 0.1
Log Loss : 1.8292089972766103
for C = 1
Log Loss : 1.9114331929676418
for C = 10
Log Loss : 1.9114331977442678
for C = 100
Log Loss : 1.911433271213524
```



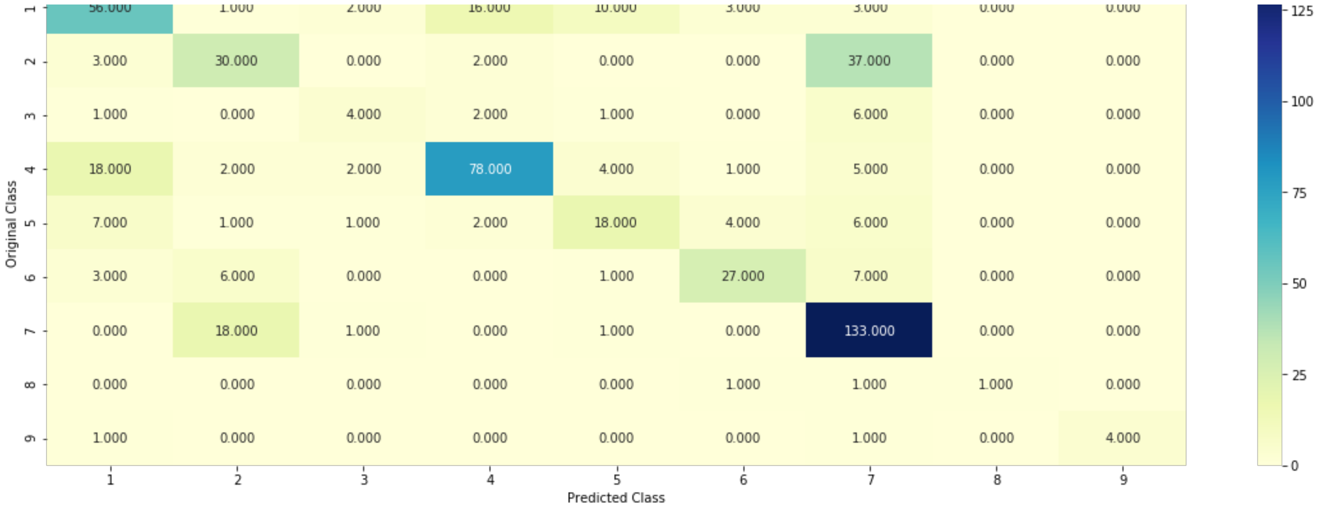
```
For values of best alpha =  0.001 The train log loss is: 0.5476666709175846
For values of best alpha =  0.001 The cross validation log loss is: 1.095803788579623
For values of best alpha =  0.001 The test log loss is: 0.9970739162371267
```

4.4.2. Testing model with best hyper parameters

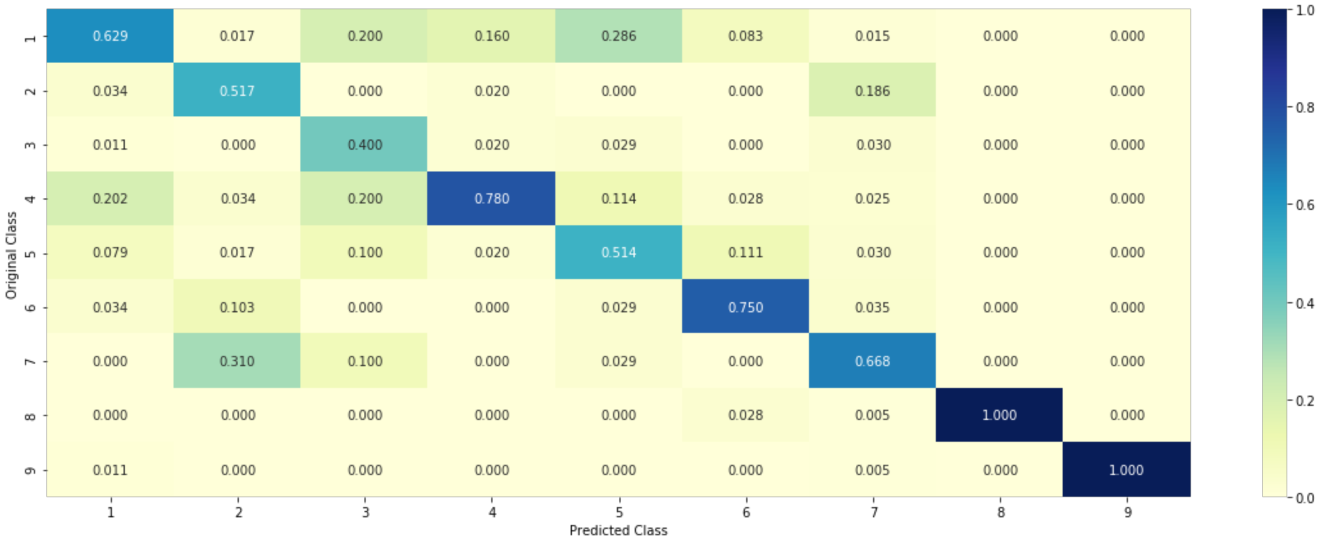
In [83]:

```
# read more about support vector machines with linear kernals here http://scikit-
learn.org/stable/modules/generated/sklearn.svm.SVC.html

# --------------------------------
# default parameters
# SVC(C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking=True, probability=False, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_state=None)

# Some of methods of SVM()
# fit(X, y, [sample_weight]) Fit the SVM model according to the given training data.
# predict(X) Perform classification on samples in X.
# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/mathematical-derivation-copy-8/
# --------------------------------


# clf = SVC(C=alpha[best_alpha],kernel='linear',probability=True, class_weight='balanced')
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='hinge',
random_state=42,class_weight='balanced')
predict_and_plot_confusion_matrix(train_x_tfidf, y_train,cv_x_tfidf,y_cv, clf)
x.add_row(["Linear_SVM(TFIDF)",train_ll,cv_ll,test_ll,mis_per])
```

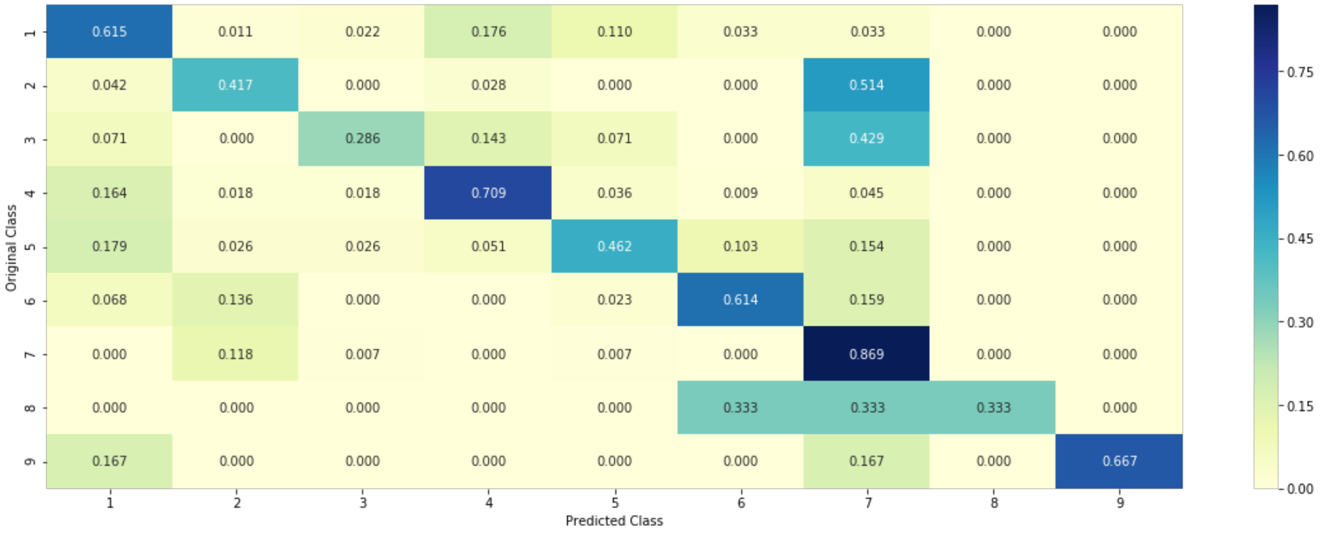
```
Log loss : 1.095803788579623
Number of mis-classified points : 0.34022556390977443
-------------------- Confusion matrix --------------------
```



```
-------------------- Precision matrix (Columm Sum=1) --------------------
```



```
-------------------- Recall matrix (Row sum=1) --------------------
```



4.3.3. Feature Importance

4.3.3.1. For Correctly classified point

In [84]:

```
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='hinge', random_state=42)
clf.fit(train_x_tfidf,y_train)
test_point_index = 1
# test_point_index = 100
no_feature = 500
predicted_cls = sig_clf.predict(test_x_tfidf.tocsr()[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(test_x_tfidf.tocsr()
[test_point_index]),4))
print("Actual Class :", y_test[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], X_test['TEXT'].iloc[test_point_index],X_test['Gene'].iloc[test_poi
nt_index],X_test['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[1.550e-02 8.800e-03 2.500e-03 1.350e-02 5.600e-03 2.800e-03 9.497
e-01
  1.300e-03 3.000e-04]]
Actual Class : 7
--------------------------------------------------
11 Text feature [002] present in test data point [True]
Out of the top  500  features  1 are present in query point
```

4.3.3.2. For Incorrectly classified point

In [85]:

```
test_point_index = 100
no_feature = 500
predicted_cls = sig_clf.predict(test_x_tfidf.tocsr()[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(test_x_tfidf.tocsr()
[test_point_index]),4))
print("Actual Class :", y_test[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], X_test['TEXT'].iloc[test_point_index],X_test['Gene'].iloc[test_poi
nt_index],X_test['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[7.861e-01 4.100e-03 2.100e-03 1.365e-01 5.600e-03 2.500e-03 5.510
e-02
  7.700e-03 3.000e-04]]
Actual Class : 1
--------------------------------------------------
Out of the top  500  features  0 are present in query point
```

4.5 Random Forest Classifier

4.5.1. Hyper paramter tuning (With tfidf Encoding)

In [87]:

```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_
impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None,
verbose=0, warm_start=False,
# 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/random-fores
t-and-their-construction-2/
# --------------------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# 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 = [100,200,500,1000,2000]
max_depth = [5, 10]
cv_log_error_array = []
for i in alpha:
    for j in max_depth:
        print("for n_estimators =", i,"and max depth = ", j)
        clf = RandomForestClassifier(n_estimators=i, criterion='gini', max_depth=j, random_state=42
, n_jobs=-1,class_weight='balanced')
        clf.fit(train_x_tfidf, y_train)
        sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
        sig_clf.fit(train_x_tfidf, y_train)
        sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
        cv_log_error_array.append(log_loss(y_cv, sig_clf_probs, labels=clf.classes_, eps=1e-15))
        print("Log Loss :",log_loss(y_cv, sig_clf_probs))

'''fig, ax = plt.subplots()
features = np.dot(np.array(alpha)[:,None],np.array(max_depth)[None]).ravel()
ax.plot(features, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[int(i/2)],max_depth[int(i%2)],str(txt)),
(features[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()
'''

best_alpha = np.argmin(cv_log_error_array)
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max
_depth[int(best_alpha%2)], random_state=42, n_jobs=-1,class_weight='balanced')
clf.fit(train_x_tfidf, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, y_train)

predict_y = sig_clf.predict_proba(train_x_tfidf)
train_ll=log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best estimator = ', alpha[int(best_alpha/2)], "The train log loss
is:",log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_x_tfidf)
cv_ll=log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best estimator = ', alpha[int(best_alpha/2)], "The cross validation log loss
is:",log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_tfidf)
test_ll=log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best estimator = ', alpha[int(best_alpha/2)], "The test log loss
is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for n_estimators = 100 and max depth =  5
Log Loss : 1.302655926934186
for n_estimators = 100 and max depth =  10
Log Loss : 1.2560688345949846
```

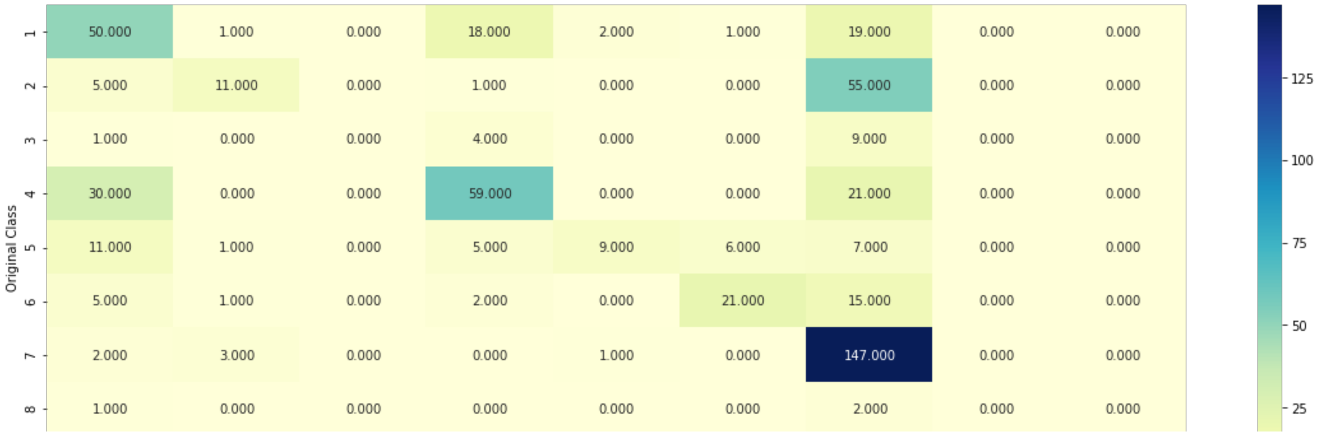
```
for n_estimators = 200 and max depth =  5
Log Loss : 1.2796901438406838
for n_estimators = 200 and max depth =  10
Log Loss : 1.2426942891240191
for n_estimators = 500 and max depth =  5
Log Loss : 1.2549310327900702
for n_estimators = 500 and max depth =  10
Log Loss : 1.2386824693281837
for n_estimators = 1000 and max depth =  5
Log Loss : 1.2461103699655953
for n_estimators = 1000 and max depth =  10
Log Loss : 1.2389164363183003
for n_estimators = 2000 and max depth =  5
Log Loss : 1.2418125372380113
for n_estimators = 2000 and max depth =  10
Log Loss : 1.234955159365341
For values of best estimator =  2000 The train log loss is: 0.653987824666976
For values of best estimator =  2000 The cross validation log loss is: 1.2349551593653405
For values of best estimator =  2000 The test log loss is: 1.1407676316053723
```

4.5.2. Testing model with best hyper parameters (Tfidf Encoding)

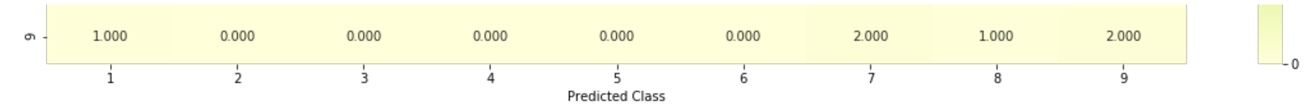
In [88]:

```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_samples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None, verbose=0, warm_start=False,
# 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/random-forest-and-their-construction-2/
# --------------------------------

clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max_depth[int(best_alpha%2)], random_state=42, n_jobs=-1,class_weight='balanced')
predict_and_plot_confusion_matrix(train_x_tfidf, y_train,cv_x_tfidf,y_cv, clf)
x.add_row(["RF(TFIDF)",train_ll,cv_ll,test_ll,mis_per])
```

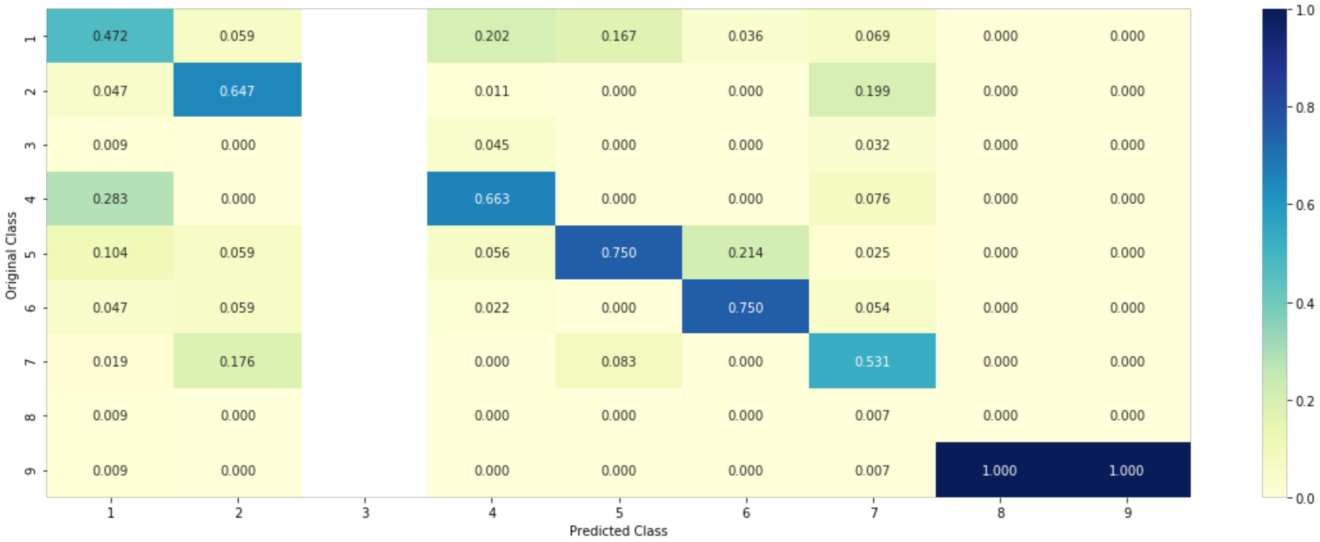
```
Log loss : 1.234955159365341
Number of mis-classified points : 0.43796992481203006
-------------------- Confusion matrix --------------------
```



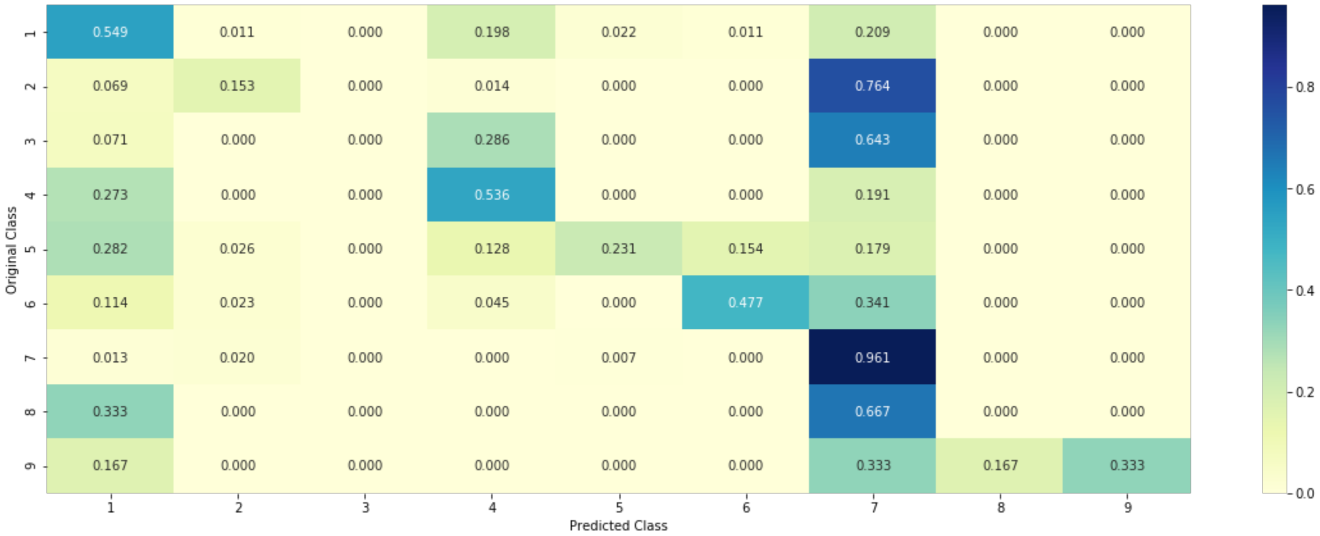
1	50.000	1.000	0.000	18.000	2.000	1.000	19.000	0.000	0.000
2	5.000	11.000	0.000	1.000	0.000	0.000	55.000	0.000	0.000
3	1.000	0.000	0.000	4.000	0.000	0.000	9.000	0.000	0.000
4	30.000	0.000	0.000	59.000	0.000	0.000	21.000	0.000	0.000
5	11.000	1.000	0.000	5.000	9.000	6.000	7.000	0.000	0.000
6	5.000	1.000	0.000	2.000	0.000	21.000	15.000	0.000	0.000
7	2.000	3.000	0.000	0.000	1.000	0.000	147.000	0.000	0.000
8	1.000	0.000	0.000	0.000	0.000	0.000	2.000	0.000	0.000



```
-------------------- Precision matrix (Columm Sum=1) --------------------
```



```
-------------------- Recall matrix (Row sum=1) --------------------
```



4.5.3. Feature Importance

4.5.3.1. Correctly Classified point

In [89]:

```
# test_point_index = 10
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max
_depth[int(best_alpha%2)], random_state=42, n_jobs=-1)
clf.fit(train_x_tfidf, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, y_train)

test_point_index = 1
no_feature = 100
predicted_cls = sig_clf.predict(test_x_tfidf.tocsr()[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(test_x_tfidf.tocsr()
[test_point_index]),4))
print("Actual Class :", y_test[test_point_index])
```

```
indices = np.argsort(-clf.feature_importances_)
print("-"*50)
get_impfeature_names(indices[:no_feature],
X_test['TEXT'].iloc[test_point_index],X_test['Gene'].iloc[test_point_index],X_test['Variation'].ilo
c[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0991 0.2262 0.0207 0.0569 0.0534 0.0441 0.4866 0.0067 0.0063]]
Actual Class : 7
--------------------------------------------------
23 Text feature [002] present in test data point [True]
51 Text feature [10] present in test data point [True]
Out of the top  100  features  2 are present in query point
```

4.5.3.2. Inorrectly Classified point

In [90]:

```
test_point_index = 100
no_feature = 100
predicted_cls = sig_clf.predict(test_x_tfidf.tocsr()[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(test_x_tfidf.tocsr()
[test_point_index]),4))
print("Actuall Class :", y_test[test_point_index])
indices = np.argsort(-clf.feature_importances_)
print("-"*50)
get_impfeature_names(indices[:no_feature],
X_test['TEXT'].iloc[test_point_index],X_test['Gene'].iloc[test_point_index],X_test['Variation'].ilo
c[test_point_index], no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[0.5259 0.0835 0.0202 0.1494 0.0564 0.0463 0.1    0.008  0.0104]]
Actuall Class : 1
--------------------------------------------------
51 Text feature [10] present in test data point [True]
Out of the top  100  features  1 are present in query point
```

4.5.3. Hyper paramter tuning (With Response Coding)

In [91]:

```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_
impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None,
verbose=0, warm_start=False,
# 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/random-fores
t-and-their-construction-2/
# --------------------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# 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 = [10,50,100,200,500,1000]
max_depth = [2,3,5,10]
cv_log_error_array = []
for i in alpha:
    for j in max_depth:
        print("for n_estimators =", i,"and max depth = ", j)
        clf = RandomForestClassifier(n_estimators=i, criterion='gini', max_depth=j, random_state=42
, n_jobs=-1)
        clf.fit(train_x_responseCoding, y_train)
        sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
        sig_clf.fit(train_x_responseCoding, y_train)
        sig_clf_probs = sig_clf.predict_proba(cv_x_responseCoding)
        cv_log_error_array.append(log_loss(y_cv, sig_clf_probs, labels=clf.classes_, eps=1e-15))
        print("Log Loss :",log_loss(y_cv, sig_clf_probs))
'''
fig, ax = plt.subplots()
features = np.dot(np.array(alpha)[:,None],np.array(max_depth)[None]).ravel()
ax.plot(features, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[int(i/4)],max_depth[int(i%4)],str(txt)),
(features[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()
'''

best_alpha = np.argmin(cv_log_error_array)
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/4)], criterion='gini', max_depth=max
_depth[int(best_alpha%4)], random_state=42, n_jobs=-1)
clf.fit(train_x_responseCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, y_train)

predict_y = sig_clf.predict_proba(train_x_responseCoding)
train_ll=log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[int(best_alpha/4)], "The train log loss is:",log_loss(y
_train, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_x_responseCoding)
cv_ll=log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[int(best_alpha/4)], "The cross validation log loss is:"
,log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_responseCoding)
test_ll=log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15)
print('For values of best alpha = ', alpha[int(best_alpha/4)], "The test log loss is:",log_loss(y_
test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for n_estimators = 10 and max depth =  2
Log Loss : 2.188530340508745
for n_estimators = 10 and max depth =  3
Log Loss : 1.748193647254023
for n_estimators = 10 and max depth =  5
Log Loss : 1.632479501054656
for n_estimators = 10 and max depth =  10
Log Loss : 2.201958840078249
for n_estimators = 50 and max depth =  2
Log Loss : 1.87258595336106
for n_estimators = 50 and max depth =  3
Log Loss : 1.5545571255191124
for n_estimators = 50 and max depth =  5
Log Loss : 1.4420123688716178
for n_estimators = 50 and max depth =  10
Log Loss : 1.8962925250887253
for n_estimators = 100 and max depth =  2
Log Loss : 1.6926489169904406
```

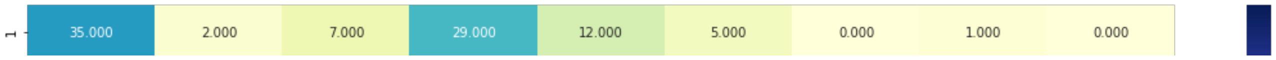
```
for n_estimators = 100 and max depth =  3
Log Loss : 1.592482410598662
for n_estimators = 100 and max depth =  5
Log Loss : 1.3796124081630734
for n_estimators = 100 and max depth =  10
Log Loss : 1.7886967936040354
for n_estimators = 200 and max depth =  2
Log Loss : 1.769698453440723
for n_estimators = 200 and max depth =  3
Log Loss : 1.6278959473741905
for n_estimators = 200 and max depth =  5
Log Loss : 1.4031163407777034
for n_estimators = 200 and max depth =  10
Log Loss : 1.7641361184815787
for n_estimators = 500 and max depth =  2
Log Loss : 1.813643059147104
for n_estimators = 500 and max depth =  3
Log Loss : 1.6599112409597458
for n_estimators = 500 and max depth =  5
Log Loss : 1.4008074801944548
for n_estimators = 500 and max depth =  10
Log Loss : 1.758912668483596
for n_estimators = 1000 and max depth =  2
Log Loss : 1.8301927475252897
for n_estimators = 1000 and max depth =  3
Log Loss : 1.6361184724061835
for n_estimators = 1000 and max depth =  5
Log Loss : 1.381657436749045
for n_estimators = 1000 and max depth =  10
Log Loss : 1.738327583994187
For values of best alpha =  100 The train log loss is: 0.05492990842923589
For values of best alpha =  100 The cross validation log loss is: 1.3796124081630734
For values of best alpha =  100 The test log loss is: 1.3250060963257428
```

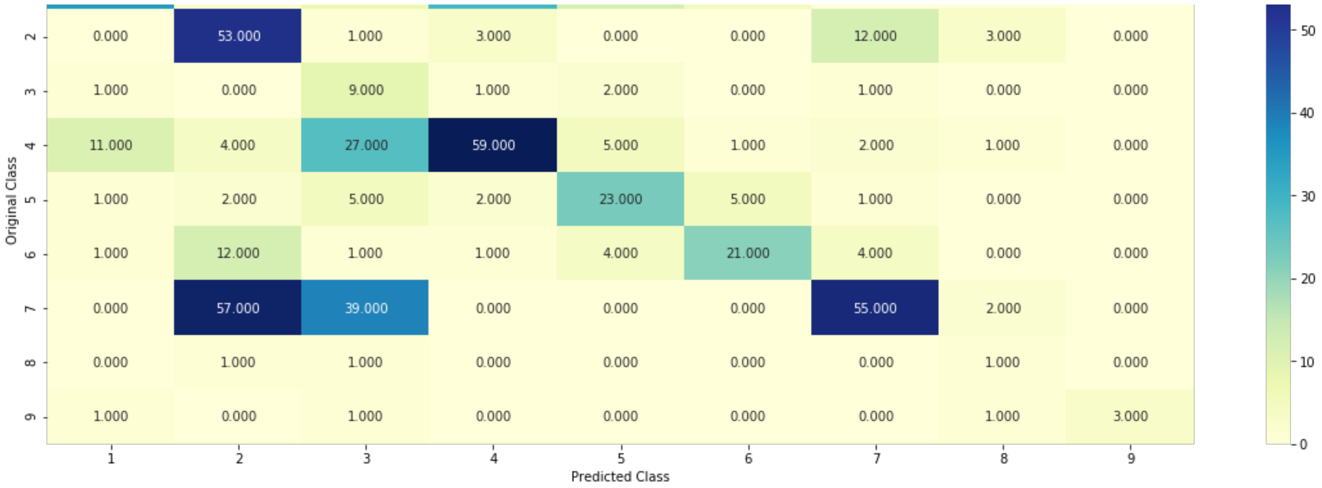
4.5.4. Testing model with best hyper parameters (Response Coding)

In [92]:

```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_
impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None,
verbose=0, warm_start=False,
# 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/random-fores
t-and-their-construction-2/
# --------------------------------

clf = RandomForestClassifier(max_depth=max_depth[int(best_alpha%4)],
n_estimators=alpha[int(best_alpha/4)], criterion='gini', max_features='auto',random_state=42)
predict_and_plot_confusion_matrix(train_x_responseCoding, y_train,cv_x_responseCoding,y_cv, clf)
x.add_row(["RF(Response Coding)",train_ll,cv_ll,test_ll,mis_per])
```

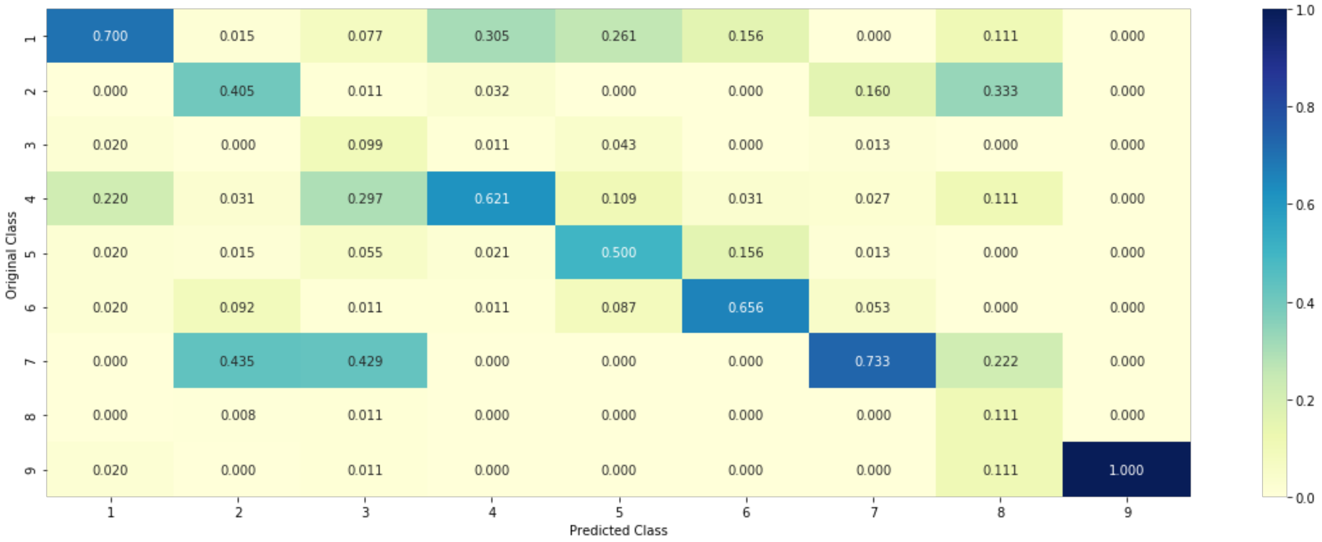
```
Log loss : 1.3796124081630734
Number of mis-classified points : 0.5131578947368421
-------------------- Confusion matrix --------------------
```





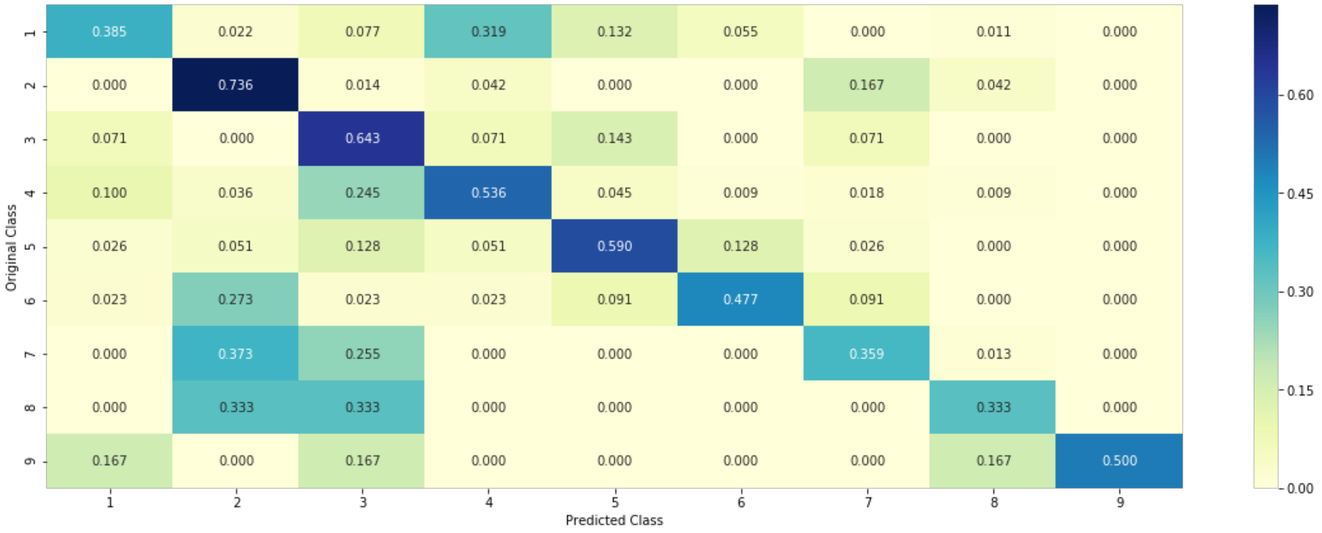
	1	2	3	4	5	6	7	8	9
2	0.000	53.000	1.000	3.000	0.000	0.000	12.000	3.000	0.000
3	1.000	0.000	9.000	1.000	2.000	0.000	1.000	0.000	0.000
4	11.000	4.000	27.000	59.000	5.000	1.000	2.000	1.000	0.000
5	1.000	2.000	5.000	2.000	23.000	5.000	1.000	0.000	0.000
6	1.000	12.000	1.000	1.000	4.000	21.000	4.000	0.000	0.000
7	0.000	57.000	39.000	0.000	0.000	0.000	55.000	2.000	0.000
8	0.000	1.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000
9	1.000	0.000	1.000	0.000	0.000	0.000	0.000	1.000	3.000

```
-------------------- Precision matrix (Columm Sum=1) --------------------
```



	1	2	3	4	5	6	7	8	9
1	0.700	0.015	0.077	0.305	0.261	0.156	0.000	0.111	0.000
2	0.000	0.405	0.011	0.032	0.000	0.000	0.160	0.333	0.000
3	0.020	0.000	0.099	0.011	0.043	0.000	0.013	0.000	0.000
4	0.220	0.031	0.297	0.621	0.109	0.031	0.027	0.111	0.000
5	0.020	0.015	0.055	0.021	0.500	0.156	0.013	0.000	0.000
6	0.020	0.092	0.011	0.011	0.087	0.656	0.053	0.000	0.000
7	0.000	0.435	0.429	0.000	0.000	0.000	0.733	0.222	0.000
8	0.000	0.008	0.011	0.000	0.000	0.000	0.000	0.111	0.000
9	0.020	0.000	0.011	0.000	0.000	0.000	0.000	0.111	1.000

```
-------------------- Recall matrix (Row sum=1) --------------------
```



	1	2	3	4	5	6	7	8	9
1	0.385	0.022	0.077	0.319	0.132	0.055	0.000	0.011	0.000
2	0.000	0.736	0.014	0.042	0.000	0.000	0.167	0.042	0.000
3	0.071	0.000	0.643	0.071	0.143	0.000	0.071	0.000	0.000
4	0.100	0.036	0.245	0.536	0.045	0.009	0.018	0.009	0.000
5	0.026	0.051	0.128	0.051	0.590	0.128	0.026	0.000	0.000
6	0.023	0.273	0.023	0.023	0.091	0.477	0.091	0.000	0.000
7	0.000	0.373	0.255	0.000	0.000	0.000	0.359	0.013	0.000
8	0.000	0.333	0.333	0.000	0.000	0.000	0.000	0.333	0.000
9	0.167	0.000	0.167	0.000	0.000	0.000	0.000	0.167	0.500

4.5.5. Feature Importance

4.5.5.1. Correctly Classified point

In [93]:

```
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/4)], criterion='gini', max_depth=max
_depth[int(best_alpha%4)], random_state=42, n_jobs=-1)
clf.fit(train_x_responseCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, y_train)


test_point_index = 1
no_feature = 27
predicted_cls = sig_clf.predict(test_x_responseCoding[test_point_index].reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_responseCoding[test_point_index].reshape(1,-1)),4))
print("Actual Class :", y_test[test_point_index])
indices = np.argsort(-clf.feature_importances_)
print("-"*50)
for i in indices:
    if i<9:
        print("Gene is important feature")
    elif i<18:
        print("Variation is important feature")
    else:
        print("Text is important feature")
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0179 0.1497 0.3234 0.0169 0.0248 0.0428 0.3851 0.0249 0.0146]]
Actual Class : 7
--------------------------------------------------
Variation is important feature
Variation is important feature
Variation is important feature
Variation is important feature
Gene is important feature
Variation is important feature
Variation is important feature
Text is important feature
Text is important feature
Gene is important feature
Text is important feature
Text is important feature
Text is important feature
Gene is important feature
Gene is important feature
Variation is important feature
Text is important feature
Gene is important feature
Gene is important feature
Variation is important feature
Text is important feature
Text is important feature
Variation is important feature
Gene is important feature
Gene is important feature
Text is important feature
Gene is important feature
```

4.5.5.2. Incorrectly Classified point

In [94]:

```
test_point_index = 100
predicted_cls = sig_clf.predict(test_x_responseCoding[test_point_index].reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_responseCoding[test_point_index].reshape(1,-1)),4))
print("Actual Class :", y_test[test_point_index])
indices = np.argsort(-clf.feature_importances_)
print("-"*50)
for i in indices:
    if i<9:
        print("Gene is important feature")
    elif i<18:
        print("Variation is important feature")
    else:
        print("Text is important feature")
        print('Text is important feature')
```

```
Predicted Class : 1
Predicted Class Probabilities: [[0.3716 0.0422 0.1404 0.1479 0.047  0.0643 0.0142 0.091  0.0814]]
Actual Class : 1
--------------------------------------------------
Variation is important feature
Variation is important feature
Variation is important feature
Variation is important feature
Gene is important feature
Variation is important feature
Variation is important feature
Text is important feature
Text is important feature
Gene is important feature
Text is important feature
Text is important feature
Text is important feature
Gene is important feature
Gene is important feature
Variation is important feature
Text is important feature
Gene is important feature
Gene is important feature
Variation is important feature
Text is important feature
Text is important feature
Variation is important feature
Gene is important feature
Gene is important feature
Text is important feature
Gene is important feature
```

4.7 Stack the models

4.7.1 testing with hyper parameter tuning

In [95]:

```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0.5,
# 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 Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/geometric-intuition-1/
#------------------------------


# read more about support vector machines with linear kernals here http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
# --------------------------------
# default parameters
# SVC(C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking=True, probability=False, tol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', random_state=None)

# Some of methods of SVM()
# fit(X, y, [sample_weight]) Fit the SVM model according to the given training data.
# predict(X) Perform classification on samples in X.
# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/mathematical-derivation-copy-8/
# --------------------------------


# read more about support vector machines with linear kernals here http://scikit-
learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_
impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None,
verbose=0, warm_start=False,
# 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/random-fores
t-and-their-construction-2/
# --------------------------------


clf1 = SGDClassifier(alpha=0.001, penalty='l2', loss='log', class_weight='balanced', random_state=0
)
clf1.fit(train_x_tfidf, y_train)
sig_clf1 = CalibratedClassifierCV(clf1, method="sigmoid")

clf2 = SGDClassifier(alpha=1, penalty='l2', loss='hinge', class_weight='balanced', random_state=0)
clf2.fit(train_x_tfidf, y_train)
sig_clf2 = CalibratedClassifierCV(clf2, method="sigmoid")


clf3 = MultinomialNB(alpha=0.001)
clf3.fit(train_x_tfidf, y_train)
sig_clf3 = CalibratedClassifierCV(clf3, method="sigmoid")

sig_clf1.fit(train_x_tfidf, y_train)
print("Logistic Regression :  Log Loss: %0.2f" % (log_loss(y_cv, sig_clf1.predict_proba(cv_x_tfidf)
)))
sig_clf2.fit(train_x_tfidf, y_train)
print("Support vector machines : Log Loss: %0.2f" % (log_loss(y_cv,
sig_clf2.predict_proba(cv_x_tfidf))))
sig_clf3.fit(train_x_tfidf, y_train)
print("Naive Bayes : Log Loss: %0.2f" % (log_loss(y_cv, sig_clf3.predict_proba(cv_x_tfidf))))
print("-"*50)
alpha = [0.0001,0.001,0.01,0.1,1,10]
best_alpha = 999
for i in alpha:
    lr = LogisticRegression(C=i)
    sclf = StackingClassifier(classifiers=[sig_clf1, sig_clf2, sig_clf3], meta_classifier=lr, use_p
robas=True)
    sclf.fit(train_x_tfidf, y_train)
    print("Stacking Classifer : for the value of alpha: %f Log Loss: %0.3f" % (i, log_loss(y_cv, sc
lf.predict_proba(cv_x_tfidf))))
    log_error =log_loss(y_cv, sclf.predict_proba(cv_x_tfidf))
    if best_alpha > log_error:
        best_alpha = log_error
```

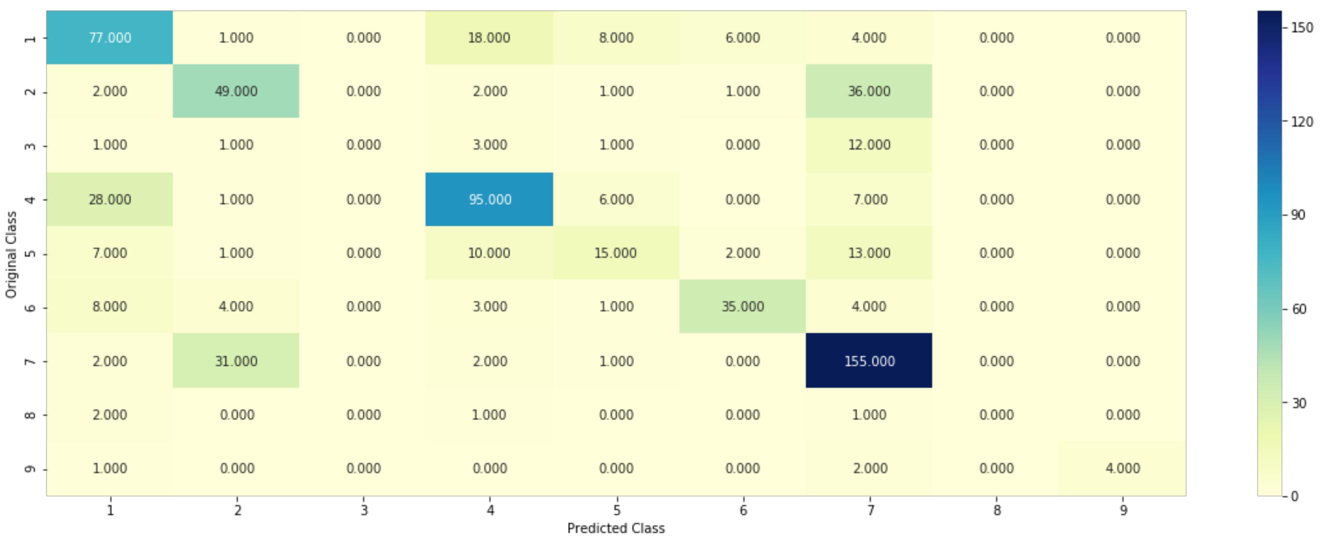
```
Logistic Regression :  Log Loss: 1.08
Support vector machines : Log Loss: 1.91
Naive Bayes : Log Loss: 1.26
--------------------------------------------------
Stacking Classifer : for the value of alpha: 0.000100 Log Loss: 2.178
Stacking Classifer : for the value of alpha: 0.001000 Log Loss: 2.032
Stacking Classifer : for the value of alpha: 0.010000 Log Loss: 1.504
Stacking Classifer : for the value of alpha: 0.100000 Log Loss: 1.179
Stacking Classifer : for the value of alpha: 1.000000 Log Loss: 1.327
Stacking Classifer : for the value of alpha: 10.000000 Log Loss: 1.608
```

4.7.2 testing the model with the best hyper parameters

In [96]:

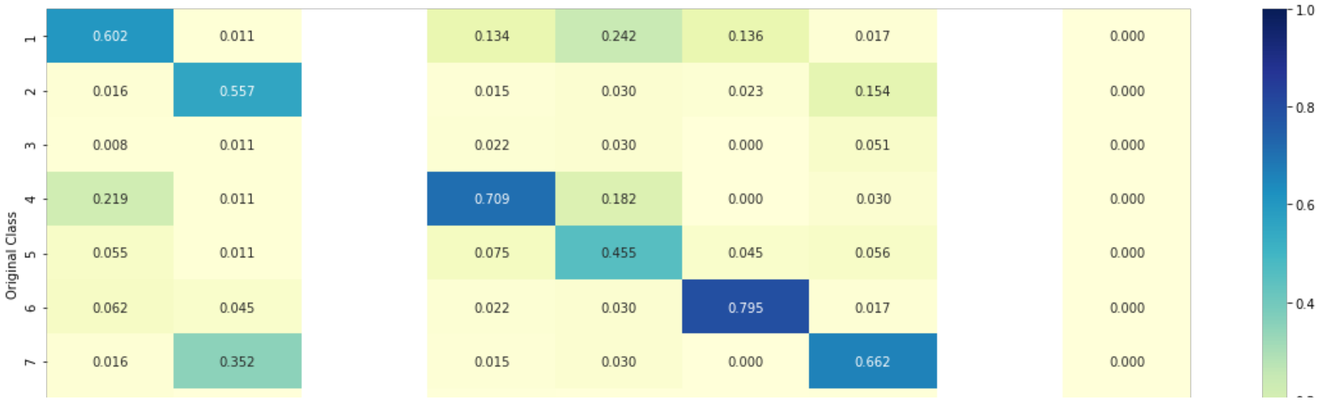
```
lr = LogisticRegression(C=0.1)
sclf = StackingClassifier(classifiers=[sig_clf1, sig_clf2, sig_clf3], meta_classifier=lr, use_proba
s=True)
sclf.fit(train_x_tfidf, y_train)

log_error = log_loss(y_train, sclf.predict_proba(train_x_tfidf))
train_ll=log_error
print("Log loss (train) on the stacking classifier :",log_error)

log_error = log_loss(y_cv, sclf.predict_proba(cv_x_tfidf))
cv_ll=log_error
print("Log loss (CV) on the stacking classifier :",log_error)

log_error = log_loss(y_test, sclf.predict_proba(test_x_tfidf))
test_ll=log_error
print("Log loss (test) on the stacking classifier :",log_error)
mis_per=(np.count_nonzero((sclf.predict(test_x_tfidf)- y_test))/y_test.shape[0]*100)
print("Number of missclassified point :", np.count_nonzero((sclf.predict(test_x_tfidf)- y_test))/y
_test.shape[0])
plot_confusion_matrix(test_y=y_test, predict_y=sclf.predict(test_x_tfidf))
x.add_row(["Stacking(TFIDF)",train_ll,cv_ll,test_ll,mis_per])
```

```
Log loss (train) on the stacking classifier : 0.6457304676713798
Log loss (CV) on the stacking classifier : 1.1787806581187426
Log loss (test) on the stacking classifier : 1.1144747855065302
Number of missclassified point : 0.3533834586466165
-------------------- Confusion matrix --------------------
```

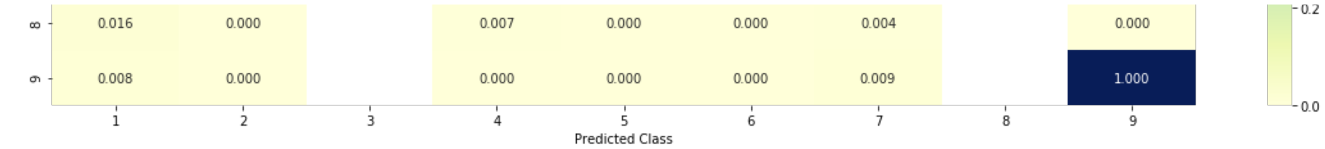


	1	2	3	4	5	6	7	8	9
1	77.000	1.000	0.000	18.000	8.000	6.000	4.000	0.000	0.000
2	2.000	49.000	0.000	2.000	1.000	1.000	36.000	0.000	0.000
3	1.000	1.000	0.000	3.000	1.000	0.000	12.000	0.000	0.000
4	28.000	1.000	0.000	95.000	6.000	0.000	7.000	0.000	0.000
5	7.000	1.000	0.000	10.000	15.000	2.000	13.000	0.000	0.000
6	8.000	4.000	0.000	3.000	1.000	35.000	4.000	0.000	0.000
7	2.000	31.000	0.000	2.000	1.000	0.000	155.000	0.000	0.000
8	2.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000
9	1.000	0.000	0.000	0.000	0.000	0.000	2.000	0.000	4.000

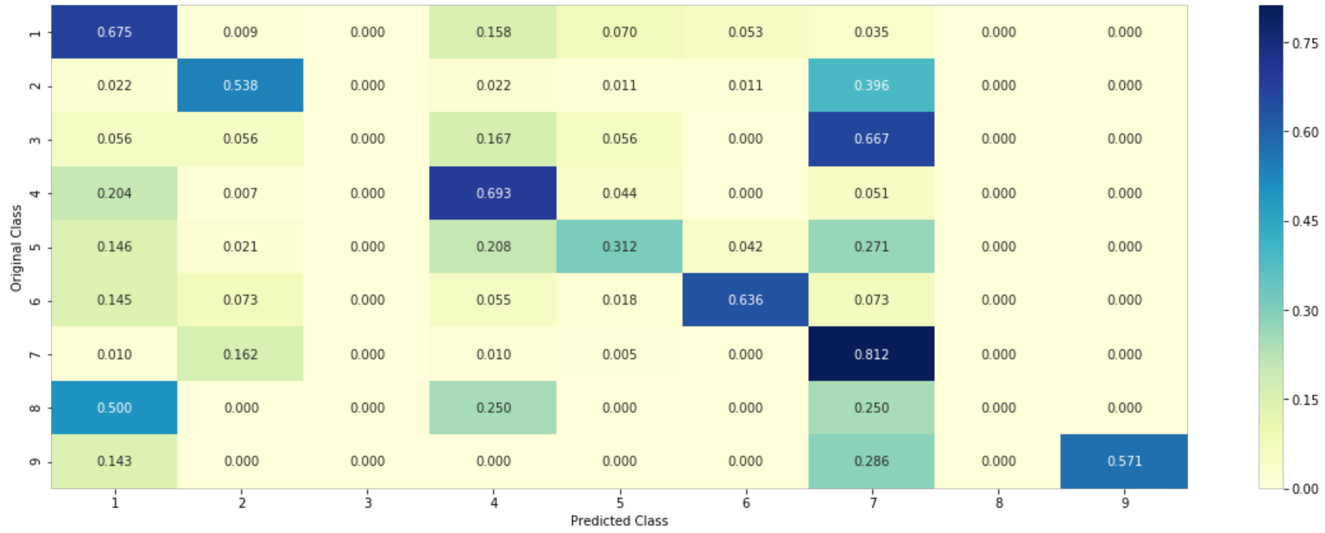
```
-------------------- Precision matrix (Columm Sum=1) --------------------
```



1	0.602	0.011		0.134	0.242	0.136	0.017		0.000
2	0.016	0.557		0.015	0.030	0.023	0.154		0.000
3	0.008	0.011		0.022	0.030	0.000	0.051		0.000
4	0.219	0.011		0.709	0.182	0.000	0.030		0.000
5	0.055	0.011		0.075	0.455	0.045	0.056		0.000
6	0.062	0.045		0.022	0.030	0.795	0.017		0.000
7	0.016	0.352		0.015	0.030	0.000	0.662		0.000



```
-------------------- Recall matrix (Row sum=1) --------------------
```

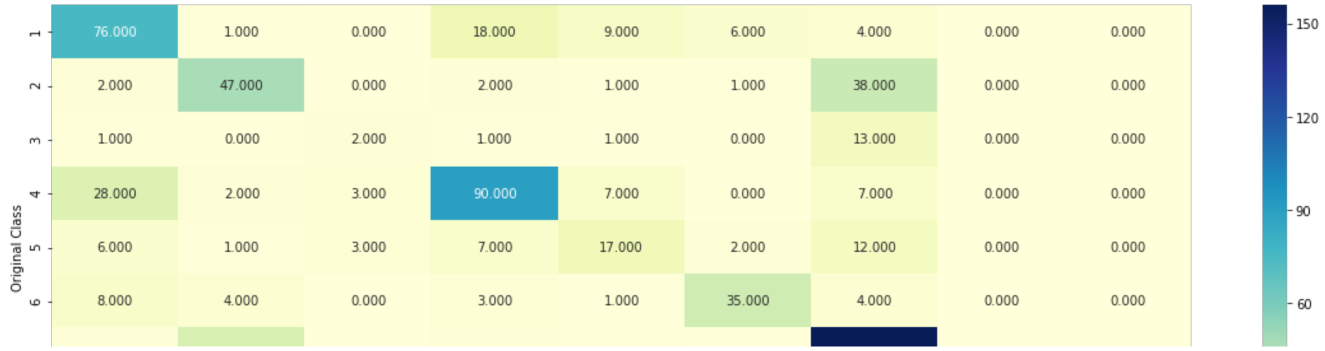


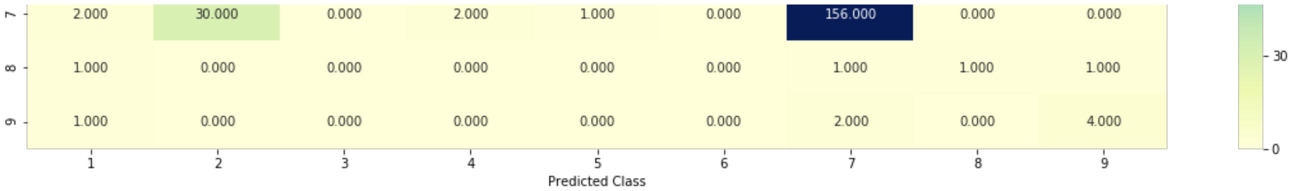
4.7.3 Maximum Voting classifier

In [110]:

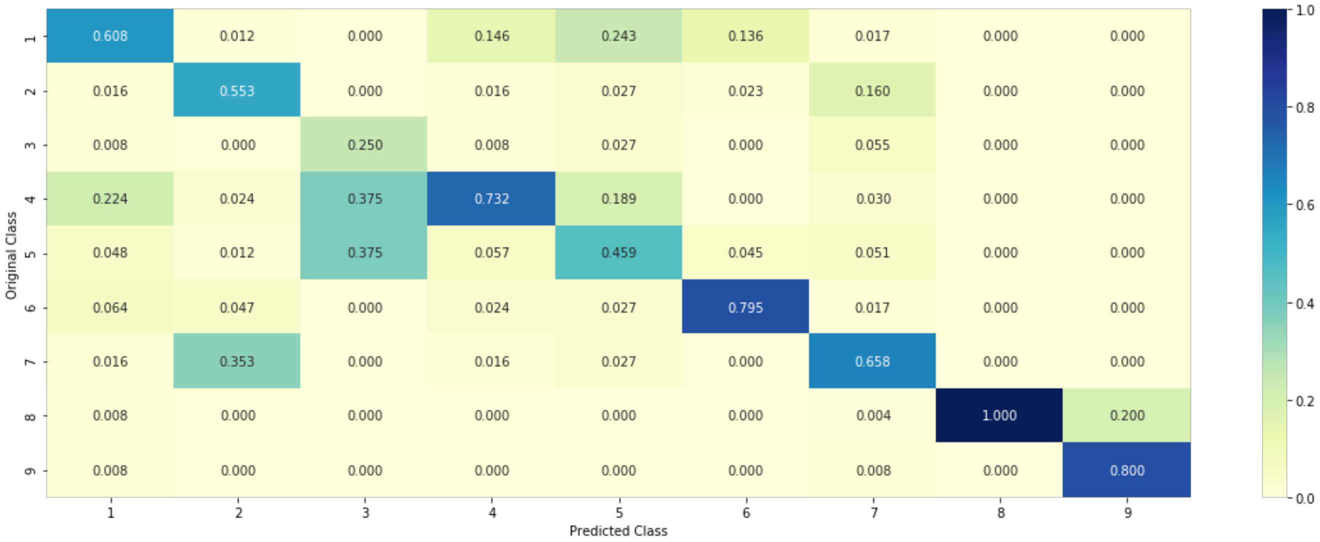
```
#Refer:http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html
from sklearn.ensemble import VotingClassifier
vclf = VotingClassifier(estimators=[('lr', sig_clf1), ('svc', sig_clf2), ('rf', sig_clf3)], voting=
'soft')
vclf.fit(train_x_tfidf, y_train)
print("Log loss (train) on the VotingClassifier :", log_loss(y_train,
vclf.predict_proba(train_x_tfidf)))
train_ll=log_loss(y_train, vclf.predict_proba(train_x_tfidf))
print("Log loss (CV) on the VotingClassifier :", log_loss(y_cv, vclf.predict_proba(cv_x_tfidf)))
cv_ll=log_loss(y_cv, vclf.predict_proba(cv_x_tfidf))
print("Log loss (test) on the VotingClassifier :", log_loss(y_test,
vclf.predict_proba(test_x_tfidf)))
test_ll= log_loss(y_test, vclf.predict_proba(test_x_tfidf))
mis_per=(np.count_nonzero((vclf.predict(test_x_tfidf)- y_test))/y_test.shape[0]*100)
print("Number of missclassified point :", np.count_nonzero((vclf.predict(test_x_tfidf)- y_test))/y
_test.shape[0])
plot_confusion_matrix(test_y=y_test, predict_y=vclf.predict(test_x_tfidf))
x.add_row(["Maximum Voting(TFIDF)",train_ll,cv_ll,test_ll,mis_per])
print(x)
```

```
Log loss (train) on the VotingClassifier : 0.868066769583388
Log loss (CV) on the VotingClassifier : 1.2030467090743675
Log loss (test) on the VotingClassifier : 1.154298306604228
Number of missclassified point : 0.35639097744360904
-------------------- Confusion matrix --------------------
```

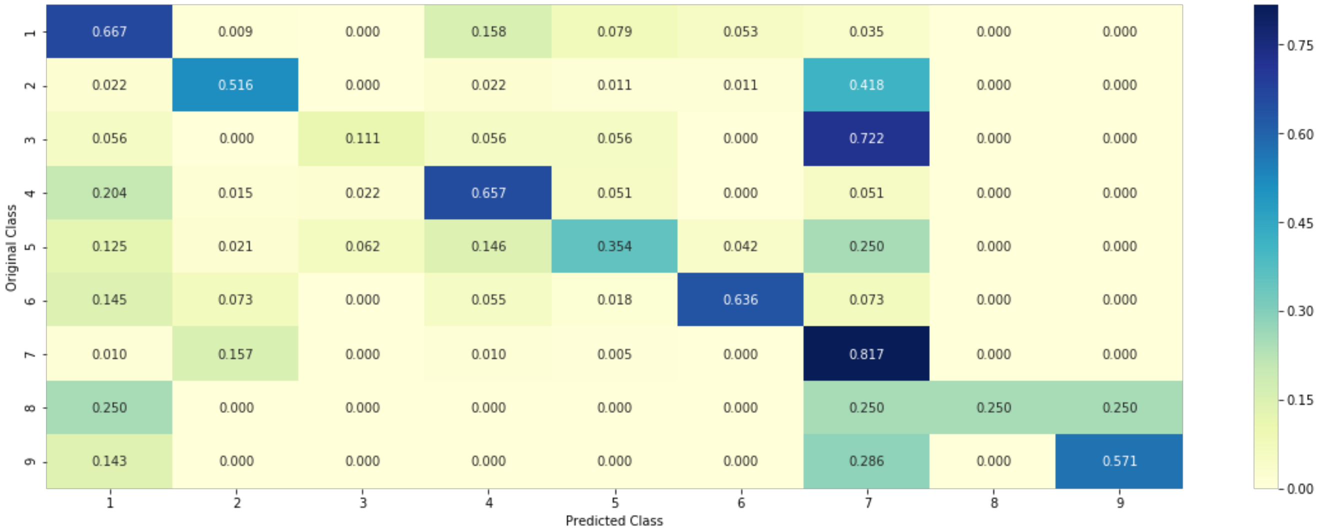




```
-------------------- Precision matrix (Columm Sum=1) --------------------
```



```
-------------------- Recall matrix (Row sum=1) --------------------
```



```
+--------------------------------------+---------------------+---------------------+-----------------
---------------------+
|              ML Model                |   Train Log Loss    |    CV Log Loss      |   Test Log Loss
| Misclassification % |
+--------------------------------------+---------------------+---------------------+-----------------
---------------------+
|          Naive Bayes(TFIDF)          |  0.6141817542885196 | 1.245401013356264   |
1.2111163418820183 |  37.40601503759399   |
|              KNN(TFIDF)              |  0.9110478758175219 | 1.135750482766513   | 1.07757105187514
2 |  37.96992481203007   |
|          Linear_SVM(TFIDF)           |  0.5476666709175846 | 1.095803788579623   |
0.9970739162371267 |  34.02255639097744   |
|              RF(TFIDF)               |  0.653987824666976  | 1.2349551593653405 | 1.14076763160537
3 |  43.796992481203006 |
|         RF(Response Coding)          | 0.05492990842923589 | 1.3796124081630734 |
1.3250060963257428 |  51.31578947368421   |
|            Stacking(TFIDF)           |  0.6457304676713798 | 1.1787806581187426 |
1.1144747855065302 |  35.338345864661655 |
|      LR_Class_Balancing(one hot)      |  0.8577608101533614 | 1.2812676609291282 |
1.1527262672860776 |  43.609022556390975 |
| LR_Without_Class_Balancing(one hot) |  0.8481051995163802 | 1.274626138776145  |
1.157109029212211  |  42.857142857142854 |
|        Maximum Voting(TFIDF)        |  0.868066769583388  | 1.2030467090743675 |
1.154298306604228  |   35.6390977443609  |
+-------------------------------------+---------------------+--------------------+-----------------
---------------------+
```

5. Assignments

1. Apply All the models with tf-idf features (Replace CountVectorizer with tfidfVectorizer and run the same cells)
2. Instead of using all the words in the dataset, use only the top 1000 words based of tf-idf values
3. Apply Logistic regression with CountVectorizer Features, including both unigrams and bigrams
4. Try any of the feature engineering techniques discussed in the course to reduce the CV and test log-loss to a value less than 1.0