Personalized cancer diagnosis

1. Business Problem

1.1. Description

Source: https://www.kaggle.com/c/msk-redefining-cancer-treatment/ (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 (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 (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 (https://www.youtube.com/watch?v=UwbuW7oK8rk)
3. https://www.youtube.com/watch?v=qxXRKVompI8 (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 (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 OR KPI(Key Performance Indicators)

Source: https://www.kaggle.com/c/msk-redefining-cancer-treatment#evaluation (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 [175]:
```
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.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.model_selection 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 [176]:
```
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[176]:

	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 [177]:
```
# 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[177]:

	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...

In [178]:
```
####################################**********************************************
for index,row in data_text.iterrows():
    print(index,row,sep = '\n')
```

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

```
<h3>3.1.3. Preprocessing of text</h3>
```

In [179]:
```
# 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 [180]:
```
#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 : 336.1040944095803 seconds
```

In [181]:
```
#merging both gene_variations and text data based on ID
result = pd.merge(data, data_text,on='ID', how='left')
result.head()
```

Out[181]:

	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 [182]:
```
result[result.isnull().any(axis=1)]
```

Out[182]:

	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 [183]:
```
# result.loc[rows,col] = ....................
result.loc[result['TEXT'].isnull(),'TEXT'] = result['Gene'] +' '+result['Variation']
```

In [184]:
```
result[result['ID']==1277]
```

Out[184]:

	ID	Gene	Variation	Class	TEXT
1277	1277	ARID5B	Truncating Mutations	1	ARID5B Truncating Mutations

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 [185]:
```
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_tru
X_train, test_df, y_train, 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_tra
train_df, cv_df, y_train, y_cv = train_test_split(X_train, y_train, stratify=y_train, test_size=0.2)
```

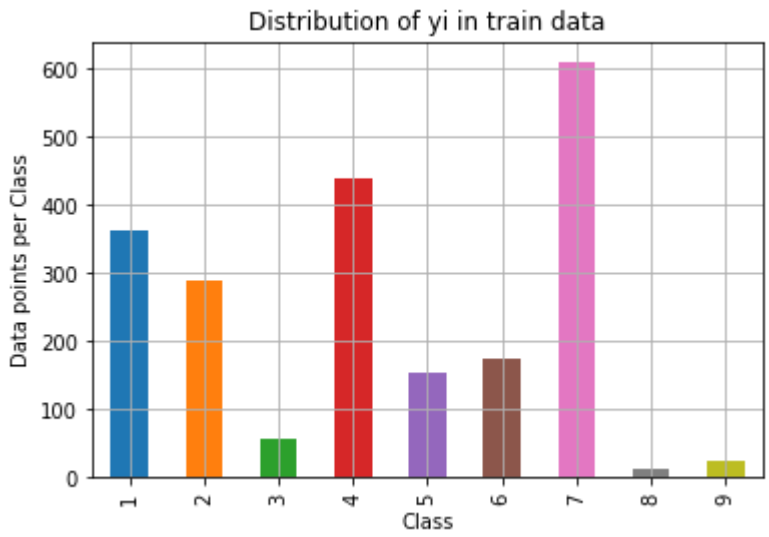
We split the data into train, test and cross validation data sets, preserving the ratio of class distribution in the original data set

In [186]:
```
print('Number of data points in train data:', train_df.shape[0])
print('Number of data points in test data:', test_df.shape[0])
print('Number of data points in cross validation data:', cv_df.shape[0])
```

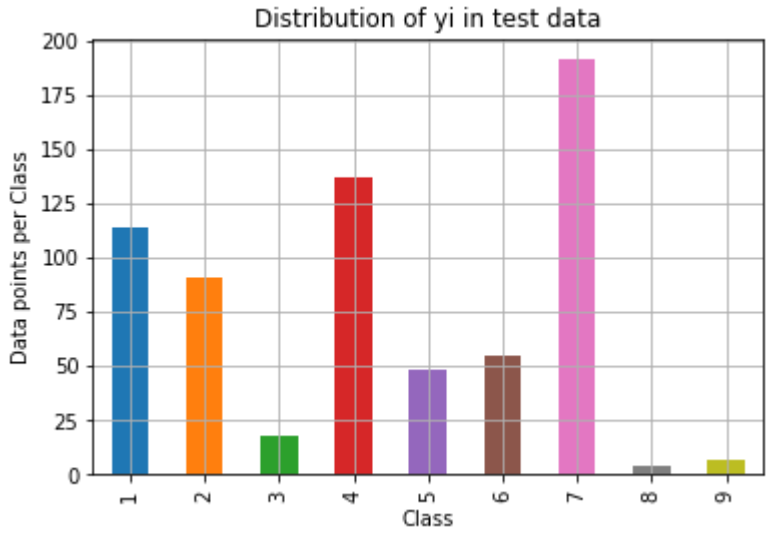
```
Number of data points in train data: 2124
Number of data points in test data: 665
Number of data points in cross validation data: 532
```

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

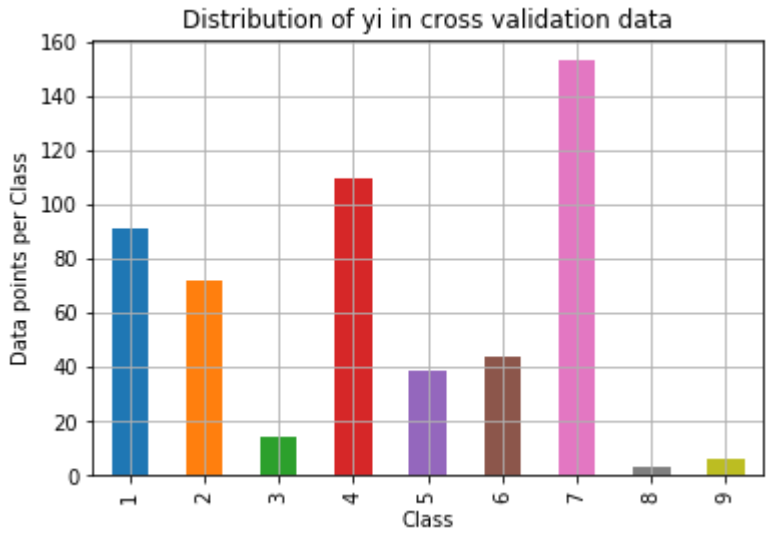
In [187]:
```
# it returns a dict, keys as class labels and values as the number of data points in that class
train_class_distribution = train_df['Class'].value_counts().sortlevel()
test_class_distribution = test_df['Class'].value_counts().sortlevel()
cv_class_distribution = cv_df['Class'].value_counts().sortlevel()

my_colors = 'rgbkymc'
train_class_distribution.plot(kind='bar')
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_cla


print('-'*80)
my_colors = 'rgbkymc'
test_class_distribution.plot(kind='bar')
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.round((test_class

print('-'*80)
my_colors = 'rgbkymc'
cv_class_distribution.plot(kind='bar')
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_dis
```




```
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 %)
```

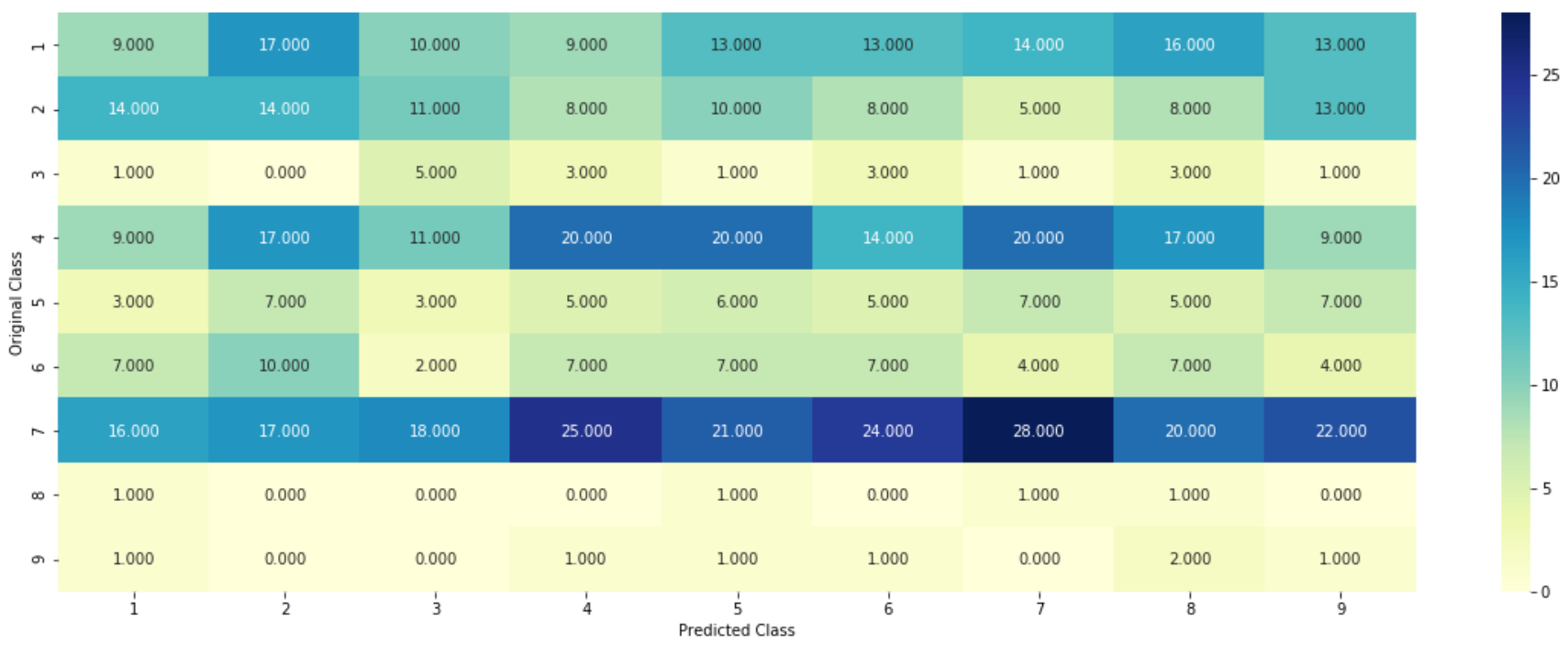
3.2 Prediction using a 'Random' Model

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

In [188]:
```
# 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.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()
```

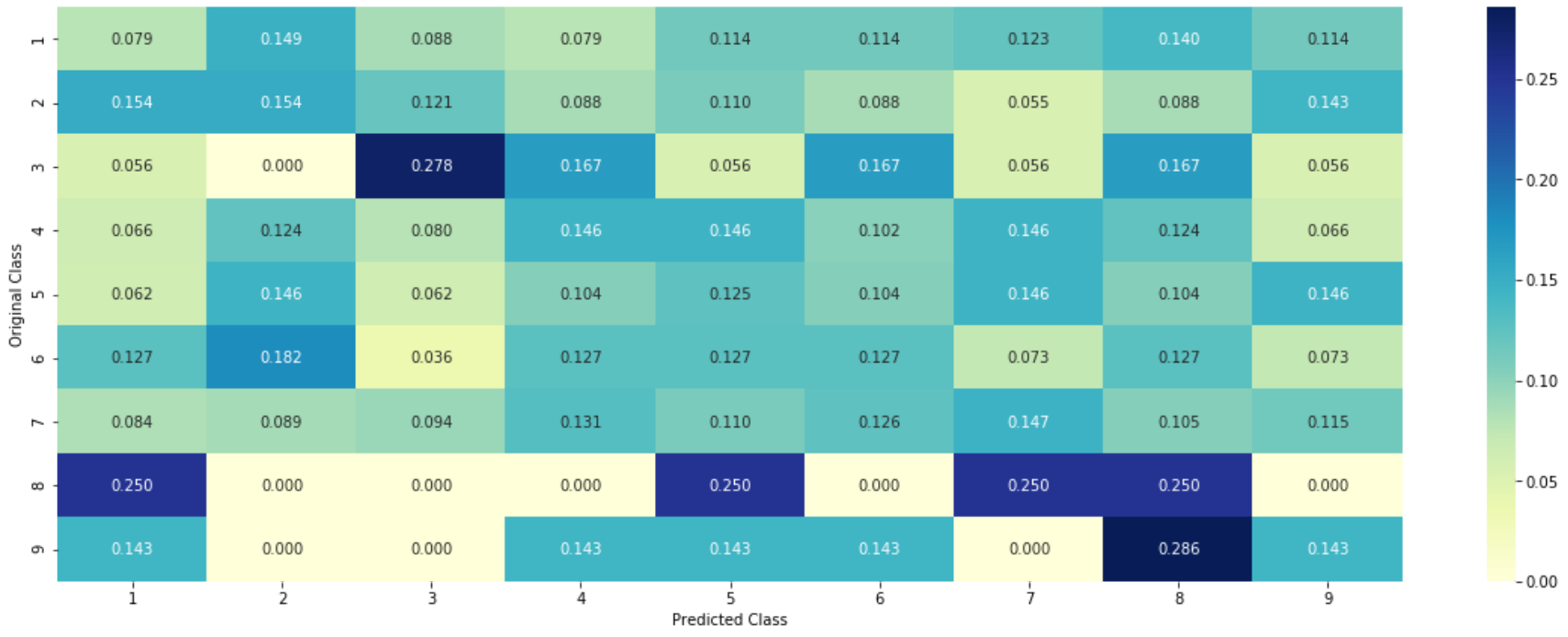
In [189]:
```
# we need to generate 9 numbers and the sum of numbers should be 1
# one solution is to genarate 9 numbers and divide each of the numbers by their sum
# ref: https://stackoverflow.com/a/18662466/4084039
test_data_len = test_df.shape[0]
cv_data_len = cv_df.shape[0]

# 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 2.5620825413140924
Log loss on Test Data using Random Model 2.47569926666256
-------------------- Confusion matrix --------------------
```




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


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



3.3 Univariate Analysis

In [190]:
```
# code for response coding with Laplace smoothing.
# alpha : used for laplace smoothing
# feature: ['gene', 'variation']
# df: ['train_df', 'test_df', 'cv_df']

# algorithm for Response Coding
# ----------
# Consider all unique values and the number of occurances of given feature in train data dataframe
# build a vector (1*9) , the first element in the Vector = (number of times it occured in class1 + 10*alpha / num
# 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(train_df['Gene'].value_counts())
    # output:
    #        {BRCA1      174
    #         TP53       106
    #         EGFR        86
    #         BRCA2       75
    #         PTEN        69
    #         KIT         61
    #         BRAF        60
    #         ERBB2       47
    #         PDGFRA      46
    #         ...}
    # print(train_df['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 = train_df[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(train_df.loc[(train_df['Class']==1) & (train_df['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 = train_df.loc[(train_df['Class']==k) & (train_df[feature]==i)]

            # cls_cnt.shape[0](numerator) will contain the number of time that particular feature occured in whol
            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
    #      'TP53': [0.32142857142857145, 0.061224489795918366, 0.061224489795918366, 0.27040816326530615, 0.0612
    #      'EGFR': [0.056818181818181816, 0.21590909090909091, 0.0625, 0.068181818181818177, 0.0681818181818177
    #      'BRCA2': [0.13333333333333333, 0.060606060606060608, 0.060606060606060608, 0.078787878787878782, 0.13
    #      'PTEN': [0.069182389937106917, 0.062893081761006289, 0.069182389937106917, 0.46540880503144655, 0.0754
    #      'KIT': [0.066225165562913912, 0.25165562913907286, 0.072847682119205295, 0.072847682119205295, 0.0662
    #      'BRAF': [0.066666666666666666, 0.17999999999999999, 0.073333333333333334, 0.073333333333333334, 0.093
    #      ...
    #     }
```

```
    gv_dict = get_gv_fea_dict(alpha, feature, df)
    # value_count is similar in get_gv_fea_dict
    value_count = train_df[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 wil
    # 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?

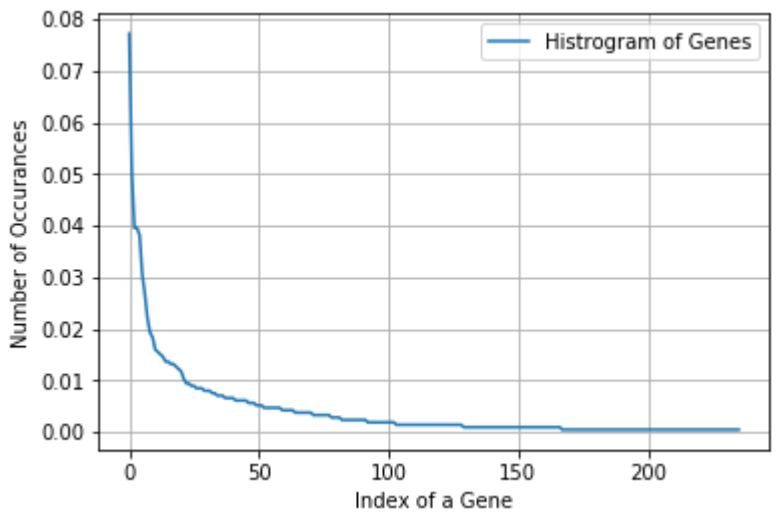
In [191]:
```
unique_genes = train_df['Gene'].value_counts()
print('Number of Unique Genes :', unique_genes.shape[0])
# the top 10 genes that occured most
print(unique_genes.head(10))
```

```
Number of Unique Genes : 236
BRCA1    164
TP53     107
BRCA2     84
EGFR      84
PTEN      81
KIT       65
BRAF      57
ALK       47
ERBB2     41
FGFR2     39
Name: Gene, dtype: int64
```

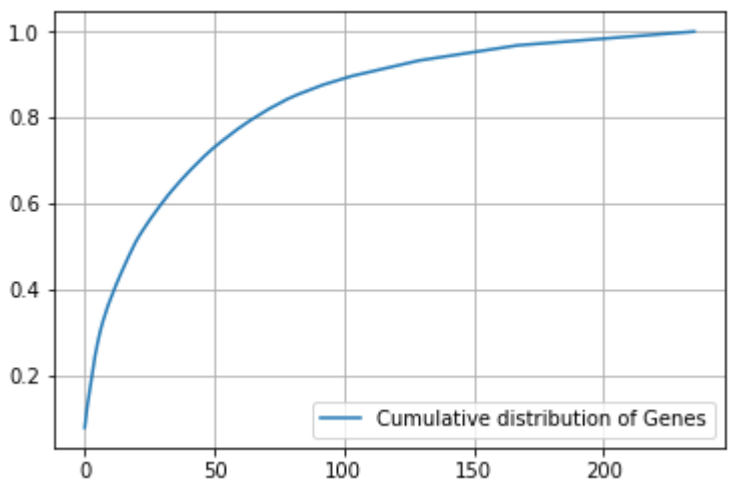
In [192]:
```
print("Ans: There are", unique_genes.shape[0] ,"different categories of genes in the train data, and they are dis
```

```
Ans: There are 236 different categories of genes in the train data, and they are distibuted as follows
```

In [193]:
```
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 [194]:
```
c = np.cumsum(h)
plt.plot(c,label='Cumulative distribution of Genes')
plt.grid()
plt.legend()
plt.show()
#OBSERVATION --x  50th gene value > 75% of the total gene
```



In [195]:
```
## Sample Printing
print(np.cumsum(h)[235])
print(np.cumsum(h)[49])
print(np.cumsum(h)[48])
```

```
0.999999999999996
0.7255178907721278
0.7203389830508472
```

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/ (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 [196]:
```
#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", train_df))
# test gene feature
test_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", test_df))
# cross validation gene feature
cv_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", cv_df))
```

In [197]:
```
print("train_gene_feature_responseCoding is converted feature using respone coding method. The shape of gene feat
```

```
train_gene_feature_responseCoding is converted feature using respone coding method. The shape of gene feature:
(2124, 9)
```

In [198]:
```
# one-hot encoding of Gene feature.
gene_vectorizer = TfidfVectorizer()
train_gene_feature_onehotCoding = gene_vectorizer.fit_transform(train_df['Gene'])
test_gene_feature_onehotCoding = gene_vectorizer.transform(test_df['Gene'])
cv_gene_feature_onehotCoding = gene_vectorizer.transform(cv_df['Gene'])
```

In [199]:
```
# Debugging
unique_variations[0]
```

Out[199]: 63

In [200]:
```
train_df['Gene'].head(50)
```

Out[200]:
```
3255     CASP8
319       ROS1
2692      BRAF
29        TERT
1759      IDH1
1785        AR
437       TP53
2197      PTEN
3296       RET
2042    MAP2K2
2002    MAP2K1
3082    NOTCH1
1119       MET
3162      RAF1
343       CDH1
1852      CTCF
2581     BRCA1
163       EGFR
2078      TET2
3265       RET
3038       KIT
2559     BRCA1
2067      SOX9
837       ABL1
3301     RUNX1
2403       NF1
476       TP53
1386     FGFR1
333       ROS1
2446     BRCA1
1803      ARAF
1000      TSC1
3166      RAF1
73       RAD50
260       EGFR
856       ABL1
486       TP53
1731      MSH2
174       EGFR
2771      BRAF
814      ERCC2
1586     CARM1
2325      JAK2
1954       ATM
2115     GATA3
2610     BRCA1
2460     BRCA1
2562     BRCA1
217       EGFR
2435     BRCA1
Name: Gene, dtype: object
```

In [201]:
```
gene_vectorizer.get_feature_names()
```

```
 'il7r',
 'inpp4b',
 'jak1',
 'jak2',
 'kdm5a',
 'kdm5c',
 'kdm6a',
 'kdr',
 'keap1',
 'kit',
 'kmt2a',
 'kmt2c',
 'kmt2d',
 'knstrn',
 'kras',
 'lats1',
 'lats2',
 'map2k1',
 'map2k2',
```

In [202]:
```
print("train_gene_feature_onehotCoding is converted feature using one-hot encoding method. The shape of gene feat
```

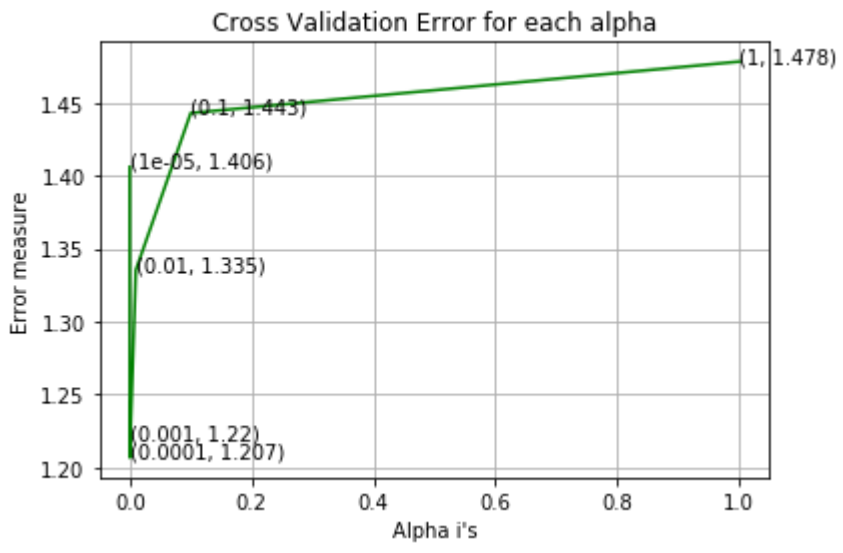
```
train_gene_feature_onehotCoding is converted feature using one-hot encoding method. The shape of gene feature:
(2124, 235)
```

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 [203]:
```
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.SGDCl
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=N
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0
# 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.classes_, eps=1e-1

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, lab
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_loss(y_cv, predi
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, predict_y, label
```

```
For values of alpha =  1e-05 The log loss is: 1.4059030187892425
For values of alpha =  0.0001 The log loss is: 1.20670409748914
For values of alpha =  0.001 The log loss is: 1.2198450093197972
For values of alpha =  0.01 The log loss is: 1.3352572978403012
For values of alpha =  0.1 The log loss is: 1.442853570788702
For values of alpha =  1 The log loss is: 1.4781404112962173
```




```
For values of best alpha =  0.0001 The train log loss is: 1.046639706944138
For values of best alpha =  0.0001 The cross validation log loss is: 1.20670409748914
For values of best alpha =  0.0001 The test log loss is: 1.18087553864529
```

Note: Since in Above case Train Loss == Test Loss == CV Loss ,

- Hence the model is not overfitting and GENE Feature is stable Across All Datasets
- And Since the Train, Test ,And CV logloss < (Logloss of Random Model , This model is Good to Go)

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 [205]:
```
# To Find Overlapping between Train ,Test, And CV Data
print("Q6. How many data points in Test and CV datasets are covered by the ", unique_genes.shape[0], " genes in t

test_coverage=test_df[test_df['Gene'].isin(list(set(train_df['Gene'])))].shape[0]
cv_coverage=cv_df[cv_df['Gene'].isin(list(set(train_df['Gene'])))].shape[0]

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

```
Q6. How many data points in Test and CV datasets are covered by the  236  genes in train dataset?
Ans
1. In test data 647 out of 665 : 97.29323308270676
2. In cross validation data 513 out of  532 : 96.42857142857143
```

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?

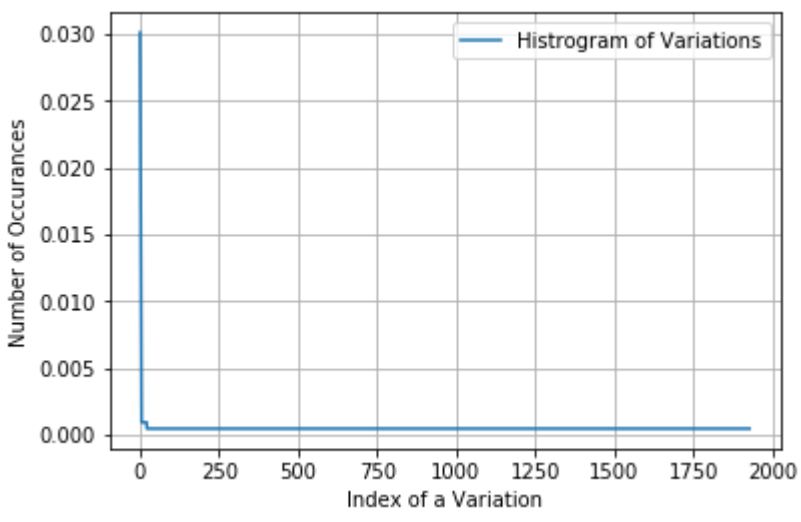
In [206]:
```
unique_variations = train_df['Variation'].value_counts()
print('Number of Unique Variations :', unique_variations.shape[0])
# the top 10 variations that occured most
print(unique_variations.head(10))
```

```
Number of Unique Variations : 1928
Truncating_Mutations    64
Amplification           50
Deletion                45
Fusions                 23
G12V                     2
R841K                    2
Overexpression           2
G12D                     2
E542K                    2
A146V                    2
Name: Variation, dtype: int64
```

In [207]:
```
print("Ans: There are", unique_variations.shape[0] ,"different categories of variations in the train data, and th
```

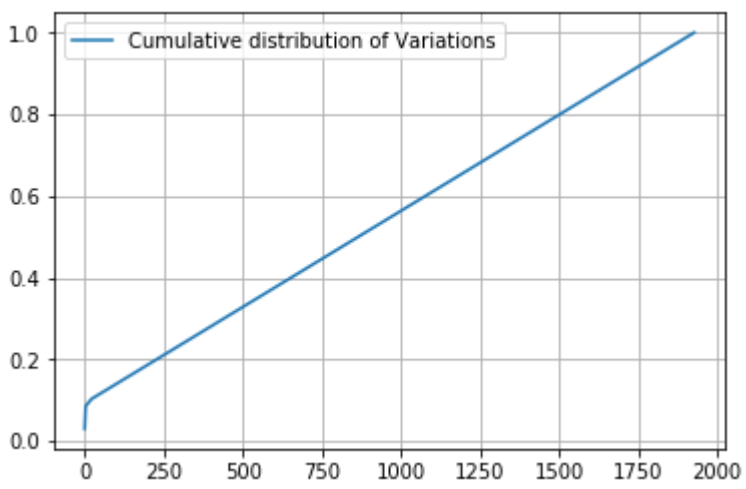
```
Ans: There are 1928 different categories of variations in the train data, and they are distibuted as follows
```

In [208]:
```
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 [209]:
```
c = np.cumsum(h)
print(c)
plt.plot(c,label='Cumulative distribution of Variations')
plt.grid()
plt.legend()
plt.show()
```

```
[0.03013183 0.05367232 0.07485876 ... 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/ (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 [210]:
```
# alpha is used for laplace smoothing
alpha = 1
# train gene feature
train_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", train_df))
# test gene feature
test_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", test_df))
# cross validation gene feature
cv_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", cv_df))
```

In [211]:
```
print("train_variation_feature_responseCoding is a converted feature using the response coding method. The shape
```

```
train_variation_feature_responseCoding is a converted feature using the response coding method. The shape of Va
riation feature: (2124, 9)
```

In [212]:
```
# one-hot encoding of variation feature.
variation_vectorizer = TfidfVectorizer()
train_variation_feature_onehotCoding = variation_vectorizer.fit_transform(train_df['Variation'])
test_variation_feature_onehotCoding = variation_vectorizer.transform(test_df['Variation'])
cv_variation_feature_onehotCoding = variation_vectorizer.transform(cv_df['Variation'])
```

In [213]:
```
print("train_variation_feature_onehotEncoded is converted feature using the onne-hot encoding method. The shape o
```

```
train_variation_feature_onehotEncoded is converted feature using the onne-hot encoding method. The shape of Var
iation feature: (2124, 1965)
```

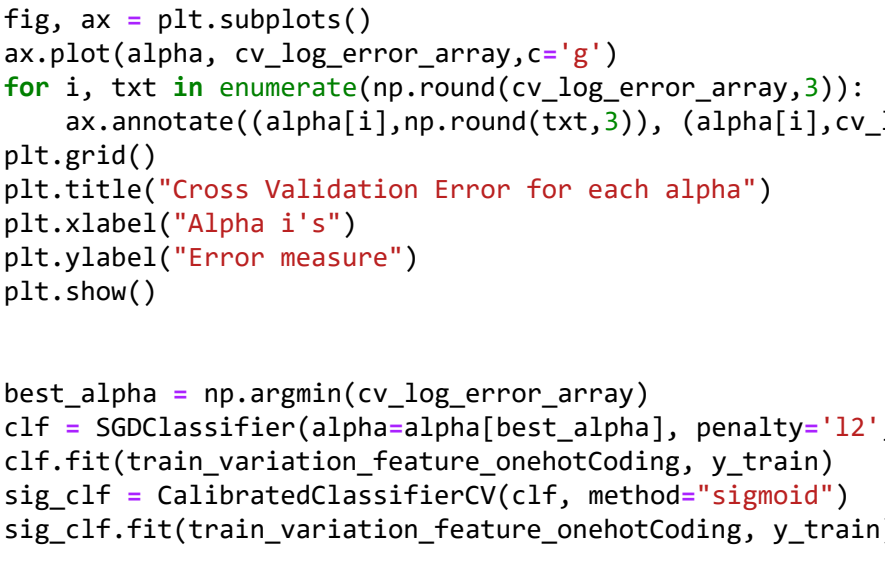
Q10. How good is this Variation feature in predicting y_i?

Let's build a model just like the earlier!

In [214]:
```
alpha = [10 ** x for x in range(-5, 1)]

# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDCl
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=N
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0
# 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.classes_, eps=1e-1
```

```
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, lab
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_loss(y_cv, predi
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, predict_y, label
```

```
For values of alpha =  1e-05 The log loss is: 1.7231701313897025
For values of alpha =  0.0001 The log loss is: 1.7110403616530323
For values of alpha =  0.001 The log loss is: 1.7148138160383877
For values of alpha =  0.01 The log loss is: 1.720647957446547
For values of alpha =  0.1 The log loss is: 1.7262802534224946
For values of alpha =  1 The log loss is: 1.731317410663227
```




```
For values of best alpha =  0.0001 The train log loss is: 0.7042480819939961
For values of best alpha =  0.0001 The cross validation log loss is: 1.7110403616530323
For values of best alpha =  0.0001 The test log loss is: 1.7015737352854592
```

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 [215]:
```
print("Q12. How many data points are covered by total ", unique_variations.shape[0], " genes in test and cross va
test_coverage=test_df[test_df['Variation'].isin(list(set(train_df['Variation'])))].shape[0]
cv_coverage=cv_df[cv_df['Variation'].isin(list(set(train_df['Variation'])))].shape[0]
print('Ans\n1. In test data',test_coverage, 'out of',test_df.shape[0], ":",(test_coverage/test_df.shape[0])*100)
print('2. In cross validation data',cv_coverage, 'out of ',cv_df.shape[0],":" ,(cv_coverage/cv_df.shape[0])*100)
```

```
Q12. How many data points are covered by total  1928  genes in test and cross validation data sets?
Ans
1. In test data 66 out of 665 : 9.924812030075188
2. In cross validation data 51 out of  532 : 9.586466165413533
```

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 [216]:
```
# 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 [217]:
```
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 [218]:
```
# building a CountVectorizer with all the words that occured minimum 3 times in train data
text_vectorizer = TfidfVectorizer(max_features = 1000,min_df=3)
train_text_feature_onehotCoding = text_vectorizer.fit_transform(train_df['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 features) vector
train_text_fea_counts = train_text_feature_onehotCoding.sum(axis=0).A1

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

```
Total number of unique words in train data : 1000
```

In [219]:
```
dict_list = []
# dict_list =[] contains 9 dictoinaries each corresponds to a class
for i in range(1,10):
    cls_text = train_df[train_df['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(train_df)


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 [220]:
```
#response coding of text features
train_text_feature_responseCoding  = get_text_responsecoding(train_df)
test_text_feature_responseCoding  = get_text_responsecoding(test_df)
cv_text_feature_responseCoding  = get_text_responsecoding(cv_df)
```

In [221]:
```
# 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(ax
test_text_feature_responseCoding = (test_text_feature_responseCoding.T/test_text_feature_responseCoding.sum(axis=
cv_text_feature_responseCoding = (cv_text_feature_responseCoding.T/cv_text_feature_responseCoding.sum(axis=1)).T
```

In [222]:
```
# 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(test_df['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(cv_df['TEXT'])
# don't forget to normalize every feature
cv_text_feature_onehotCoding = normalize(cv_text_feature_onehotCoding, axis=0)
```

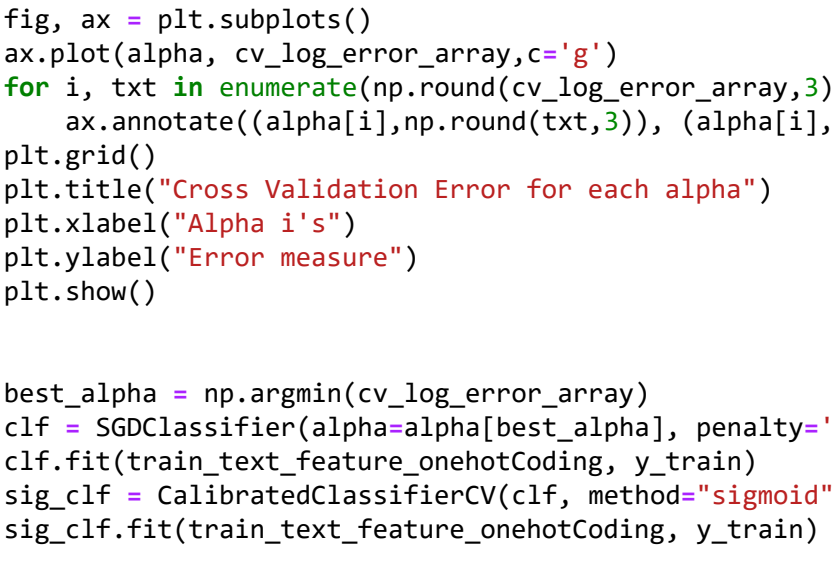
In [223]:
```
#https://stackoverflow.com/a/2258273/4084039
sorted_text_fea_dict = dict(sorted(text_fea_dict.items(), key=lambda x: x[1] , reverse=True))
sorted_text_occur = np.array(list(sorted_text_fea_dict.values()))
```

In [224]:
```
# Number of words for a given frequency.
print(Counter(sorted_text_occur))
```

```
1.112837758058747: 1, 11.101250927192694: 1, 11.083591062965645: 1, 11.062635344785521: 1, 11.05747354950836
8: 1, 11.040845245800002: 1, 11.015866156063053: 1, 11.014795040329918: 1, 11.010332352573133: 1, 11.0088095
05087914: 1, 10.979023380065218: 1, 10.963163403960145: 1, 10.94433888190056: 1, 10.936341257761349: 1, 10.9
21125788709087: 1, 10.915681636837043: 1, 10.91551811795557: 1, 10.906776854387468: 1, 10.903402373548756:
1, 10.863299824194947: 1, 10.861626772842559: 1, 10.861609164923122: 1, 10.839786211020211: 1, 10.8253714445
61275: 1, 10.754986471517679: 1, 10.74206022327933: 1, 10.722788263734145: 1, 10.721725679609827: 1, 10.7102
70864361473: 1, 10.700452544095569: 1, 10.697632727809516: 1, 10.696502885259346: 1, 10.682824772416437: 1,
10.680462830705856: 1, 10.67177486213008: 1, 10.654537217389601: 1, 10.652776078334082: 1, 10.6384827305104
5: 1, 10.62621160136737: 1, 10.625274080419521: 1, 10.60842912386527: 1, 10.590050839775019: 1, 10.580770968
006018: 1, 10.567999739037257: 1, 10.566110895333722: 1, 10.561232457105358: 1, 10.551033629845953: 1, 10.54
108951266666: 1, 10.5405387685527: 1, 10.532723445532145: 1, 10.518065994379922: 1, 10.501819514344081: 1, 1
0.482750183747005: 1, 10.480174663434104: 1, 10.458572201846202: 1, 10.449499068406384: 1, 10.44045363030050
1: 1, 10.433535368762584: 1, 10.43255897229458: 1, 10.415995496534574: 1, 10.415184628500745: 1, 10.41003005
4123727: 1, 10.386075298312354: 1, 10.371643884584365: 1, 10.364867431225637: 1, 10.329159666481285: 1, 10.2
95879431699662: 1, 10.294835885233894: 1, 10.282081402585098: 1, 10.263858112138555: 1, 10.229429757066205:
1, 10.217978909480317: 1, 10.213171699544377: 1, 10.208163277887042: 1, 10.207980451572679: 1, 10.2059276615
9069: 1, 10.205125973250716: 1, 10.198817913599175: 1, 10.181156807238613: 1, 10.173686365000673: 1, 10.1698
93855901512: 1, 10.151968676000015: 1, 10.146078822440876: 1, 10.114296228786145: 1, 10.103185249444387: 1,
10.092487182604591: 1, 10.07514295644835: 1, 10.070931803626438: 1, 10.061933989926365: 1, 10.05191592176976
7: 1, 10.045038635665192: 1, 10.020825678577577: 1, 9.992048094710912: 1, 9.971000952283292: 1, 9.9652108370
```

In [225]:
```
# Train a Logistic regression+Calibration model using text features whicha re on-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.SGDCl
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=N
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0
# 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-1


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, lab
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, predic
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, label
```

```
For values of alpha =  1e-05 The log loss is: 1.148750990636421
For values of alpha =  0.0001 The log loss is: 1.1799131158447351
For values of alpha =  0.001 The log loss is: 1.4273187453204115
For values of alpha =  0.01 The log loss is: 1.8671095458618012
For values of alpha =  0.1 The log loss is: 2.0374555211108967
For values of alpha =  1 The log loss is: 2.0260663619117834
```




```
For values of best alpha =  1e-05 The train log loss is: 0.8048490307170281
For values of best alpha =  1e-05 The cross validation log loss is: 1.148750990636421
For values of best alpha =  1e-05 The test log loss is: 1.0239576039690927
```

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

Ans. Yes, it seems like!

In [226]:
```
def get_intersec_text(df):
    df_text_vec = TfidfVectorizer(max_features = 1000,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 [227]:
```
len1,len2 = get_intersec_text(test_df)
print(np.round((len2/len1)*100, 3), "% of word of test data appeared in train data")
len1,len2 = get_intersec_text(cv_df)
print(np.round((len2/len1)*100, 3), "% of word of Cross Validation appeared in train data")
```

```
95.2 % of word of test data appeared in train data
92.4 % of word of Cross Validation appeared in train data
```

4. Machine Learning Models

In [228]:
```
#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)

    # 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
    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 [229]:
```
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 [230]:
```
# 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 = TfidfVectorizer()
    var_count_vec = TfidfVectorizer()
    text_count_vec = TfidfVectorizer(max_features=1000,min_df=3)

    gene_vec = gene_count_vec.fit(train_df['Gene'])
    var_vec  = var_count_vec.fit(train_df['Variation'])
    text_vec = text_count_vec.fit(train_df['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_no))
        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 [231]:
```
# 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],
#                 [ 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(train_df['Class']))

test_x_onehotCoding = hstack((test_gene_var_onehotCoding, test_text_feature_onehotCoding)).tocsr()
test_y = np.array(list(test_df['Class']))

cv_x_onehotCoding = hstack((cv_gene_var_onehotCoding, cv_text_feature_onehotCoding)).tocsr()
cv_y = np.array(list(cv_df['Class']))


train_gene_var_responseCoding = np.hstack((train_gene_feature_responseCoding,train_variation_feature_responseCodi
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))
```

In [232]:
```
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, 3200)
(number of data points * number of features) in test data =  (665, 3200)
(number of data points * number of features) in cross validation data = (532, 3200)
```

In [233]:
```
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)
```

4.1. Base Line Model

4.1.1. Naive Bayes

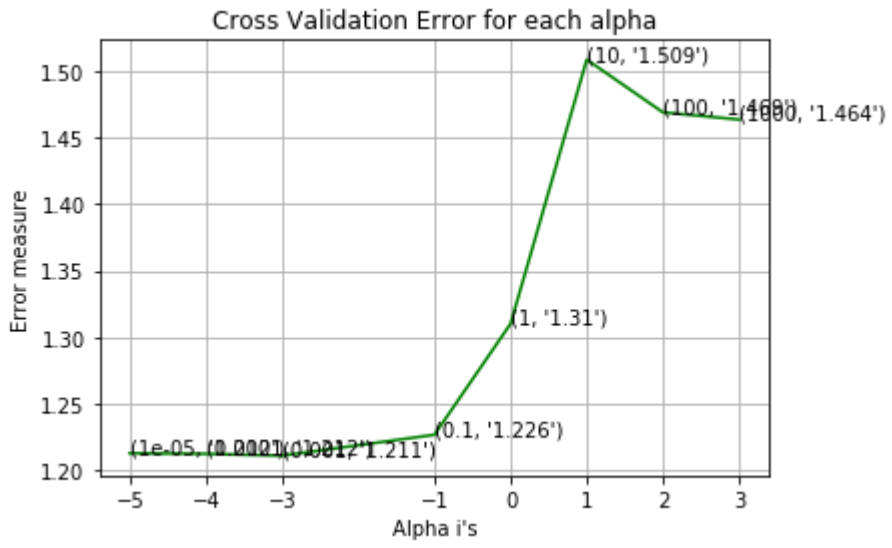
4.1.1.1. Hyper parameter tuning

In [234]:
```
# find more about Multinomial Naive base function here http://scikit-learn.org/stable/modules/generated/sklearn.n
# ------------------------
# 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.calibra
# ----------------------------
# 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_onehotCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, 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(cv_y, 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_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)
```

```
predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, predict_y, lab
predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_cv, predic
predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predict_y, label
```

```
for alpha = 1e-05
Log Loss : 1.2124916854730916
for alpha = 0.0001
Log Loss : 1.2120525269502955
for alpha = 0.001
Log Loss : 1.2106799153730616
for alpha = 0.1
Log Loss : 1.2263786958632021
for alpha = 1
Log Loss : 1.3100019899776518
for alpha = 10
Log Loss : 1.50886623984461
for alpha = 100
Log Loss : 1.469153837892789
for alpha = 1000
Log Loss : 1.4637513879908013
```

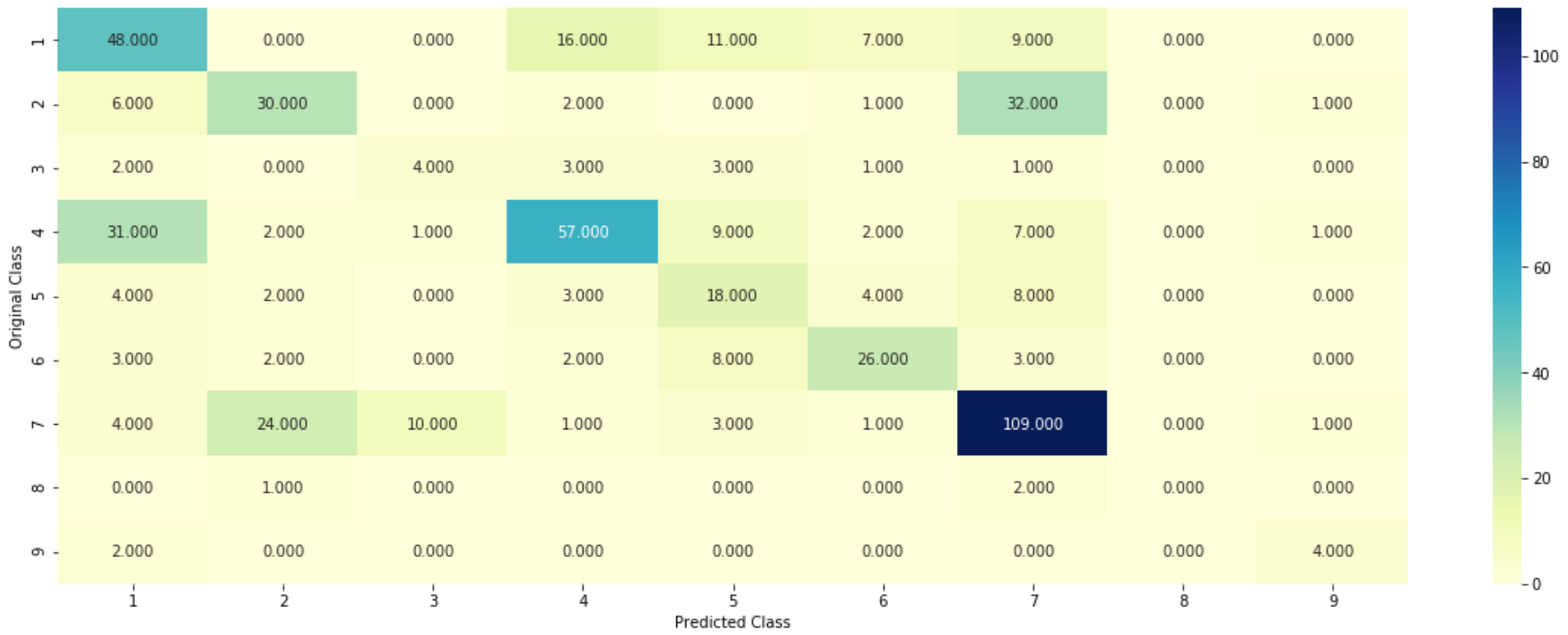



```
For values of best alpha =  0.001 The train log loss is: 0.5261403312693558
For values of best alpha =  0.001 The cross validation log loss is: 1.2106799153730616
For values of best alpha =  0.001 The test log loss is: 1.1428914713058833
```

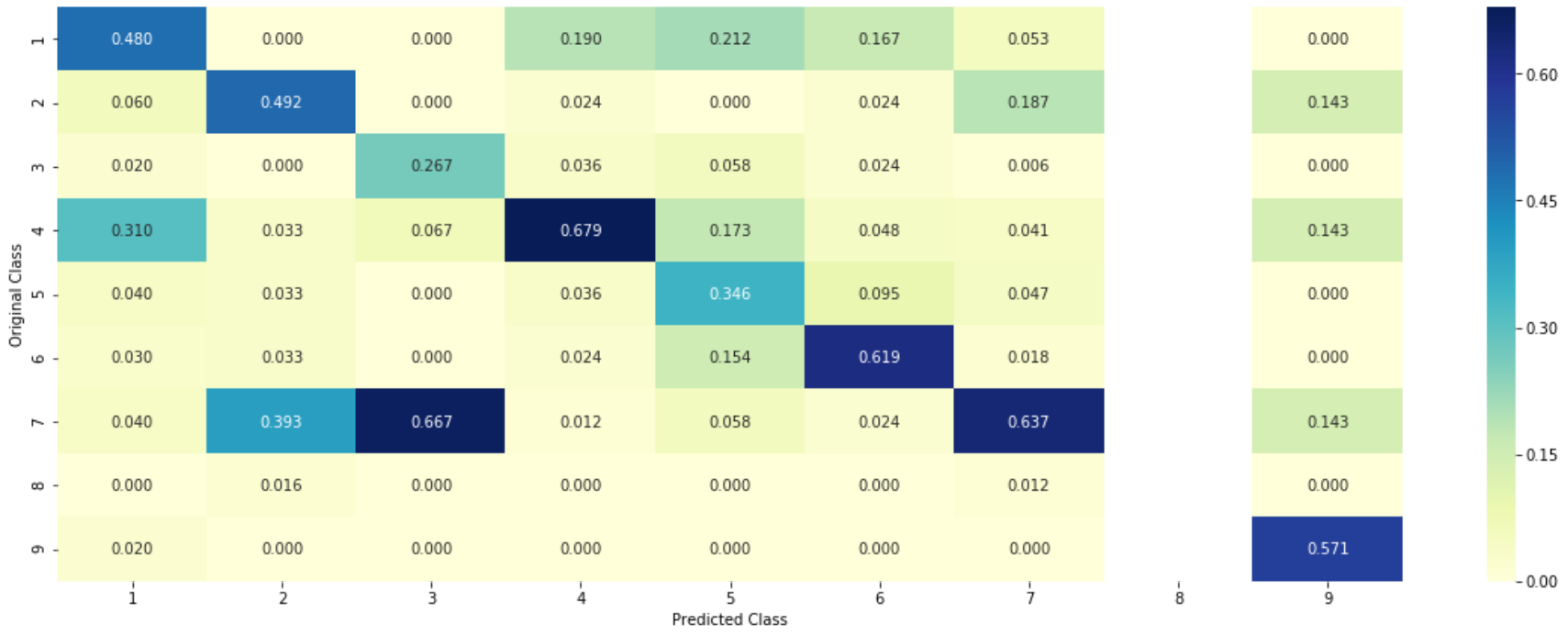
4.1.1.2. Testing the model with best hyper paramters

In [69]:
```
# find more about Multinomial Naive base function here http://scikit-learn.org/stable/modules/generated/sklearn.n
# -------------------------
# 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.calibra
# ----------------------------
# 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_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)
sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
# to avoid rounding error while multiplying probabilites we use log-probability estimates
print("Log Loss :",log_loss(cv_y, sig_clf_probs))
print("Number of missclassified point :", np.count_nonzero((sig_clf.predict(cv_x_onehotCoding)- cv_y))/cv_y.shape
plot_confusion_matrix(cv_y, sig_clf.predict(cv_x_onehotCoding.toarray()))
```

```
Log Loss : 1.3509235125982695
Number of missclassified point : 0.44360902255639095
-------------------- Confusion matrix --------------------
```




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




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

4.1.1.3. Feature Importance, Correctly classified point

In [236]:
```
test_point_index = 1
no_feature = 1000
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 :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],te
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.0563 0.0417 0.0118 0.7382 0.0321 0.0305 0.0845 0.003  0.0019]]
Actual Class : 4
--------------------------------------------------
11 Text feature [activity] present in test data point [True]
12 Text feature [protein] present in test data point [True]
15 Text feature [function] present in test data point [True]
16 Text feature [proteins] present in test data point [True]
17 Text feature [missense] present in test data point [True]
18 Text feature [results] present in test data point [True]
19 Text feature [type] present in test data point [True]
20 Text feature [pten] present in test data point [True]
22 Text feature [shown] present in test data point [True]
23 Text feature [acid] present in test data point [True]
24 Text feature [important] present in test data point [True]
25 Text feature [functional] present in test data point [True]
26 Text feature [wild] present in test data point [True]
27 Text feature [whether] present in test data point [True]
28 Text feature [whereas] present in test data point [True]
29 Text feature [also] present in test data point [True]
30 Text feature [may] present in test data point [True]
31 Text feature [mutations] present in test data point [True]
32 Text feature [amino] present in test data point [True]
33 Text feature [suppressor] present in test data point [True]
34 Text feature [two] present in test data point [True]
35 Text feature [described] present in test data point [True]
36 Text feature [reduced] present in test data point [True]
37 Text feature [although] present in test data point [True]
38 Text feature [30] present in test data point [True]
41 Text feature [indicated] present in test data point [True]
42 Text feature [therefore] present in test data point [True]
43 Text feature [levels] present in test data point [True]
45 Text feature [previously] present in test data point [True]
46 Text feature [determine] present in test data point [True]
47 Text feature [suggesting] present in test data point [True]
48 Text feature [either] present in test data point [True]
49 Text feature [ability] present in test data point [True]
50 Text feature [discussion] present in test data point [True]
51 Text feature [related] present in test data point [True]
52 Text feature [thus] present in test data point [True]
56 Text feature [analysis] present in test data point [True]
57 Text feature [expressed] present in test data point [True]
60 Text feature [suggest] present in test data point [True]
61 Text feature [three] present in test data point [True]
62 Text feature [show] present in test data point [True]
63 Text feature [associated] present in test data point [True]
65 Text feature [vitro] present in test data point [True]
66 Text feature [result] present in test data point [True]
67 Text feature [introduction] present in test data point [True]
68 Text feature [effects] present in test data point [True]
70 Text feature [using] present in test data point [True]
71 Text feature [one] present in test data point [True]
72 Text feature [similar] present in test data point [True]
73 Text feature [effect] present in test data point [True]
75 Text feature [analyzed] present in test data point [True]
76 Text feature [found] present in test data point [True]
78 Text feature [buffer] present in test data point [True]
79 Text feature [mutants] present in test data point [True]
80 Text feature [loss] present in test data point [True]
81 Text feature [assay] present in test data point [True]
83 Text feature [phosphatase] present in test data point [True]
84 Text feature [mutation] present in test data point [True]
85 Text feature [used] present in test data point [True]
87 Text feature [several] present in test data point [True]
88 Text feature [due] present in test data point [True]
89 Text feature [vector] present in test data point [True]
90 Text feature [cells] present in test data point [True]
92 Text feature [generated] present in test data point [True]
93 Text feature [role] present in test data point [True]
96 Text feature [10] present in test data point [True]
97 Text feature [addition] present in test data point [True]
98 Text feature [50] present in test data point [True]
99 Text feature [data] present in test data point [True]
100 Text feature [site] present in test data point [True]
101 Text feature [functions] present in test data point [True]
102 Text feature [mutant] present in test data point [True]
103 Text feature [affect] present in test data point [True]
104 Text feature [stability] present in test data point [True]
105 Text feature [high] present in test data point [True]
```

```
106 Text feature [reported] present in test data point [True]
107 Text feature [figure] present in test data point [True]
109 Text feature [transfected] present in test data point [True]
110 Text feature [tagged] present in test data point [True]
112 Text feature [together] present in test data point [True]
113 Text feature [performed] present in test data point [True]
115 Text feature [see] present in test data point [True]
116 Text feature [transfection] present in test data point [True]
117 Text feature [contribute] present in test data point [True]
118 Text feature [expression] present in test data point [True]
119 Text feature [incubated] present in test data point [True]
120 Text feature [consistent] present in test data point [True]
122 Text feature [vivo] present in test data point [True]
123 Text feature [15] present in test data point [True]
124 Text feature [compared] present in test data point [True]
126 Text feature [well] present in test data point [True]
127 Text feature [table] present in test data point [True]
128 Text feature [control] present in test data point [True]
129 Text feature [binding] present in test data point [True]
130 Text feature [different] present in test data point [True]
132 Text feature [possible] present in test data point [True]
133 Text feature [cellular] present in test data point [True]
134 Text feature [resulting] present in test data point [True]
135 Text feature [according] present in test data point [True]
136 Text feature [page] present in test data point [True]
138 Text feature [fig] present in test data point [True]
140 Text feature [derived] present in test data point [True]
143 Text feature [tested] present in test data point [True]
145 Text feature [materials] present in test data point [True]
146 Text feature [likely] present in test data point [True]
148 Text feature [25] present in test data point [True]
149 Text feature [might] present in test data point [True]
150 Text feature [analyses] present in test data point [True]
155 Text feature [observed] present in test data point [True]
156 Text feature [another] present in test data point [True]
157 Text feature [cell] present in test data point [True]
158 Text feature [respectively] present in test data point [True]
160 Text feature [20] present in test data point [True]
161 Text feature [recent] present in test data point [True]
163 Text feature [directly] present in test data point [True]
164 Text feature [31] present in test data point [True]
165 Text feature [frequently] present in test data point [True]
166 Text feature [terminal] present in test data point [True]
167 Text feature [contrast] present in test data point [True]
170 Text feature [dna] present in test data point [True]
171 Text feature [comparison] present in test data point [True]
172 Text feature [27] present in test data point [True]
174 Text feature [western] present in test data point [True]
180 Text feature [direct] present in test data point [True]
181 Text feature [significant] present in test data point [True]
182 Text feature [washed] present in test data point [True]
184 Text feature [studies] present in test data point [True]
186 Text feature [negative] present in test data point [True]
187 Text feature [methods] present in test data point [True]
188 Text feature [relative] present in test data point [True]
190 Text feature [26] present in test data point [True]
192 Text feature [human] present in test data point [True]
193 Text feature [four] present in test data point [True]
194 Text feature [present] present in test data point [True]
195 Text feature [least] present in test data point [True]
196 Text feature [showed] present in test data point [True]
197 Text feature [phenotype] present in test data point [True]
198 Text feature [provide] present in test data point [True]
199 Text feature [germline] present in test data point [True]
200 Text feature [first] present in test data point [True]
203 Text feature [specific] present in test data point [True]
204 Text feature [sds] present in test data point [True]
205 Text feature [many] present in test data point [True]
206 Text feature [low] present in test data point [True]
207 Text feature [dependent] present in test data point [True]
208 Text feature [detected] present in test data point [True]
209 Text feature [measured] present in test data point [True]
210 Text feature [localization] present in test data point [True]
212 Text feature [42] present in test data point [True]
213 Text feature [mm] present in test data point [True]
214 Text feature [highly] present in test data point [True]
217 Text feature [confirmed] present in test data point [True]
218 Text feature [assays] present in test data point [True]
219 Text feature [24] present in test data point [True]
222 Text feature [domain] present in test data point [True]
225 Text feature [37] present in test data point [True]
227 Text feature [green] present in test data point [True]
228 Text feature [investigated] present in test data point [True]
231 Text feature [28] present in test data point [True]
232 Text feature [study] present in test data point [True]
233 Text feature [considered] present in test data point [True]
234 Text feature [12] present in test data point [True]
236 Text feature [min] present in test data point [True]
238 Text feature [total] present in test data point [True]
239 Text feature [shows] present in test data point [True]
241 Text feature [part] present in test data point [True]
```

```
242 Text feature [affected] present in test data point [True]
243 Text feature [multiple] present in test data point [True]
244 Text feature [tumor] present in test data point [True]
245 Text feature [mutagenesis] present in test data point [True]
246 Text feature [lysates] present in test data point [True]
248 Text feature [relatively] present in test data point [True]
249 Text feature [4a] present in test data point [True]
250 Text feature [distinct] present in test data point [True]
254 Text feature [hypothesis] present in test data point [True]
255 Text feature [s1] present in test data point [True]
257 Text feature [gene] present in test data point [True]
258 Text feature [sequence] present in test data point [True]
259 Text feature [level] present in test data point [True]
260 Text feature [antibodies] present in test data point [True]
261 Text feature [40] present in test data point [True]
263 Text feature [major] present in test data point [True]
264 Text feature [followed] present in test data point [True]
266 Text feature [significantly] present in test data point [True]
267 Text feature [assessed] present in test data point [True]
269 Text feature [29] present in test data point [True]
270 Text feature [lead] present in test data point [True]
271 Text feature [ca] present in test data point [True]
272 Text feature [antibody] present in test data point [True]
274 Text feature [would] present in test data point [True]
275 Text feature [presence] present in test data point [True]
277 Text feature [normal] present in test data point [True]
279 Text feature [interestingly] present in test data point [True]
281 Text feature [essential] present in test data point [True]
283 Text feature [taken] present in test data point [True]
284 Text feature [100] present in test data point [True]
286 Text feature [13] present in test data point [True]
287 Text feature [induced] present in test data point [True]
288 Text feature [revealed] present in test data point [True]
291 Text feature [genetic] present in test data point [True]
297 Text feature [identified] present in test data point [True]
299 Text feature [ml] present in test data point [True]
300 Text feature [increased] present in test data point [True]
301 Text feature [system] present in test data point [True]
302 Text feature [nuclear] present in test data point [True]
304 Text feature [even] present in test data point [True]
305 Text feature [activities] present in test data point [True]
307 Text feature [33] present in test data point [True]
309 Text feature [obtained] present in test data point [True]
314 Text feature [model] present in test data point [True]
316 Text feature [cancer] present in test data point [True]
317 Text feature [decreased] present in test data point [True]
318 Text feature [appears] present in test data point [True]
320 Text feature [among] present in test data point [True]
321 Text feature [since] present in test data point [True]
322 Text feature [3a] present in test data point [True]
323 Text feature [indeed] present in test data point [True]
324 Text feature [still] present in test data point [True]
325 Text feature [distribution] present in test data point [True]
326 Text feature [dominant] present in test data point [True]
327 Text feature [directed] present in test data point [True]
330 Text feature [cannot] present in test data point [True]
331 Text feature [35] present in test data point [True]
333 Text feature [times] present in test data point [True]
334 Text feature [mechanism] present in test data point [True]
337 Text feature [known] present in test data point [True]
339 Text feature [plasmid] present in test data point [True]
340 Text feature [1a] present in test data point [True]
342 Text feature [mediated] present in test data point [True]
343 Text feature [play] present in test data point [True]
346 Text feature [moreover] present in test data point [True]
350 Text feature [cause] present in test data point [True]
351 Text feature [90] present in test data point [True]
352 Text feature [test] present in test data point [True]
354 Text feature [38] present in test data point [True]
355 Text feature [anti] present in test data point [True]
356 Text feature [et] present in test data point [True]
358 Text feature [18] present in test data point [True]
359 Text feature [regulation] present in test data point [True]
360 Text feature [complete] present in test data point [True]
368 Text feature [al] present in test data point [True]
369 Text feature [active] present in test data point [True]
370 Text feature [following] present in test data point [True]
371 Text feature [approximately] present in test data point [True]
372 Text feature [resulted] present in test data point [True]
378 Text feature [impaired] present in test data point [True]
381 Text feature [various] present in test data point [True]
382 Text feature [recently] present in test data point [True]
388 Text feature [assess] present in test data point [True]
390 Text feature [studied] present in test data point [True]
392 Text feature [six] present in test data point [True]
393 Text feature [32] present in test data point [True]
394 Text feature [change] present in test data point [True]
396 Text feature [research] present in test data point [True]
397 Text feature [able] present in test data point [True]
400 Text feature [23] present in test data point [True]
```

```
402 Text feature [express] present in test data point [True]
405 Text feature [mouse] present in test data point [True]
406 Text feature [36] present in test data point [True]
408 Text feature [similarly] present in test data point [True]
410 Text feature [located] present in test data point [True]
411 Text feature [increase] present in test data point [True]
414 Text feature [occur] present in test data point [True]
415 Text feature [common] present in test data point [True]
419 Text feature [reaction] present in test data point [True]
420 Text feature [mechanisms] present in test data point [True]
421 Text feature [demonstrate] present in test data point [True]
422 Text feature [go] present in test data point [True]
424 Text feature [domains] present in test data point [True]
426 Text feature [cancers] present in test data point [True]
428 Text feature [potential] present in test data point [True]
432 Text feature [flag] present in test data point [True]
433 Text feature [targeting] present in test data point [True]
436 Text feature [next] present in test data point [True]
439 Text feature [range] present in test data point [True]
440 Text feature [via] present in test data point [True]
442 Text feature [substrate] present in test data point [True]
448 Text feature [manufacturer] present in test data point [True]
449 Text feature [like] present in test data point [True]
451 Text feature [observations] present in test data point [True]
454 Text feature [pathogenic] present in test data point [True]
461 Text feature [evidence] present in test data point [True]
462 Text feature [contains] present in test data point [True]
465 Text feature [carrying] present in test data point [True]
467 Text feature [demonstrated] present in test data point [True]
468 Text feature [pathways] present in test data point [True]
474 Text feature [types] present in test data point [True]
478 Text feature [right] present in test data point [True]
481 Text feature [line] present in test data point [True]
484 Text feature [subjected] present in test data point [True]
485 Text feature [detection] present in test data point [True]
488 Text feature [process] present in test data point [True]
489 Text feature [wt] present in test data point [True]
491 Text feature [molecular] present in test data point [True]
493 Text feature [series] present in test data point [True]
494 Text feature [presented] present in test data point [True]
495 Text feature [plates] present in test data point [True]
498 Text feature [pcr] present in test data point [True]
499 Text feature [developed] present in test data point [True]
503 Text feature [seen] present in test data point [True]
504 Text feature [11] present in test data point [True]
506 Text feature [heterozygous] present in test data point [True]
509 Text feature [applied] present in test data point [True]
511 Text feature [supplementary] present in test data point [True]
512 Text feature [left] present in test data point [True]
513 Text feature [variants] present in test data point [True]
516 Text feature [nucleus] present in test data point [True]
520 Text feature [stably] present in test data point [True]
521 Text feature [induction] present in test data point [True]
522 Text feature [led] present in test data point [True]
525 Text feature [expressing] present in test data point [True]
528 Text feature [05] present in test data point [True]
533 Text feature [culture] present in test data point [True]
534 Text feature [manner] present in test data point [True]
535 Text feature [controls] present in test data point [True]
536 Text feature [identify] present in test data point [True]
537 Text feature [genomic] present in test data point [True]
538 Text feature [panel] present in test data point [True]
540 Text feature [correlation] present in test data point [True]
541 Text feature [leading] present in test data point [True]
542 Text feature [use] present in test data point [True]
544 Text feature [chain] present in test data point [True]
548 Text feature [group] present in test data point [True]
550 Text feature [lines] present in test data point [True]
554 Text feature [database] present in test data point [True]
556 Text feature [17] present in test data point [True]
558 Text feature [22] present in test data point [True]
566 Text feature [enhanced] present in test data point [True]
569 Text feature [34] present in test data point [True]
571 Text feature [regulated] present in test data point [True]
573 Text feature [positive] present in test data point [True]
577 Text feature [stable] present in test data point [True]
580 Text feature [2a] present in test data point [True]
588 Text feature [age] present in test data point [True]
589 Text feature [subset] present in test data point [True]
591 Text feature [proportion] present in test data point [True]
592 Text feature [s2] present in test data point [True]
593 Text feature [blot] present in test data point [True]
595 Text feature [cultured] present in test data point [True]
597 Text feature [lanes] present in test data point [True]
602 Text feature [risk] present in test data point [True]
603 Text feature [isolated] present in test data point [True]
604 Text feature [proliferation] present in test data point [True]
606 Text feature [39] present in test data point [True]
607 Text feature [alterations] present in test data point [True]
610 Text feature [defects] present in test data point [True]
611 Text feature [fold] present in test data point [True]
```

```
615 Text feature [60] present in test data point [True]
616 Text feature [impact] present in test data point [True]
618 Text feature [www] present in test data point [True]
620 Text feature [mice] present in test data point [True]
629 Text feature [samples] present in test data point [True]
631 Text feature [exons] present in test data point [True]
632 Text feature [signal] present in test data point [True]
633 Text feature [per] present in test data point [True]
634 Text feature [decrease] present in test data point [True]
636 Text feature [14] present in test data point [True]
639 Text feature [sporadic] present in test data point [True]
640 Text feature [somatic] present in test data point [True]
643 Text feature [mutated] present in test data point [True]
644 Text feature [elevated] present in test data point [True]
645 Text feature [clinical] present in test data point [True]
646 Text feature [fraction] present in test data point [True]
648 Text feature [time] present in test data point [True]
650 Text feature [21] present in test data point [True]
651 Text feature [mrna] present in test data point [True]
654 Text feature [tissue] present in test data point [True]
657 Text feature [whole] present in test data point [True]
658 Text feature [sequencing] present in test data point [True]
660 Text feature [basis] present in test data point [True]
662 Text feature [activation] present in test data point [True]
666 Text feature [factor] present in test data point [True]
668 Text feature [novel] present in test data point [True]
673 Text feature [invitrogen] present in test data point [True]
675 Text feature [tumors] present in test data point [True]
676 Text feature [subsequent] present in test data point [True]
677 Text feature [signaling] present in test data point [True]
678 Text feature [pathway] present in test data point [True]
680 Text feature [degradation] present in test data point [True]
682 Text feature [reverse] present in test data point [True]
685 Text feature [state] present in test data point [True]
689 Text feature [tissues] present in test data point [True]
692 Text feature [primary] present in test data point [True]
694 Text feature [terminus] present in test data point [True]
699 Text feature [inhibition] present in test data point [True]
700 Text feature [patient] present in test data point [True]
706 Text feature [concentration] present in test data point [True]
707 Text feature [ubiquitin] present in test data point [True]
710 Text feature [double] present in test data point [True]
711 Text feature [long] present in test data point [True]
713 Text feature [hours] present in test data point [True]
714 Text feature [transcription] present in test data point [True]
718 Text feature [primers] present in test data point [True]
721 Text feature [activated] present in test data point [True]
724 Text feature [induce] present in test data point [True]
725 Text feature [carcinoma] present in test data point [True]
729 Text feature [values] present in test data point [True]
734 Text feature [criteria] present in test data point [True]
737 Text feature [statistical] present in test data point [True]
738 Text feature [000] present in test data point [True]
749 Text feature [patients] present in test data point [True]
752 Text feature [years] present in test data point [True]
753 Text feature [combination] present in test data point [True]
756 Text feature [factors] present in test data point [True]
767 Text feature [sensitivity] present in test data point [True]
769 Text feature [inhibited] present in test data point [True]
770 Text feature [blood] present in test data point [True]
772 Text feature [amplified] present in test data point [True]
773 Text feature [end] present in test data point [True]
774 Text feature [individuals] present in test data point [True]
776 Text feature [treated] present in test data point [True]
777 Text feature [breast] present in test data point [True]
779 Text feature [serum] present in test data point [True]
783 Text feature [overexpression] present in test data point [True]
793 Text feature [activate] present in test data point [True]
794 Text feature [correlated] present in test data point [True]
800 Text feature [kinase] present in test data point [True]
812 Text feature [colon] present in test data point [True]
815 Text feature [mg] present in test data point [True]
819 Text feature [syndrome] present in test data point [True]
823 Text feature [survival] present in test data point [True]
827 Text feature [effective] present in test data point [True]
832 Text feature [response] present in test data point [True]
834 Text feature [targeted] present in test data point [True]
843 Text feature [deleterious] present in test data point [True]
845 Text feature [involving] present in test data point [True]
847 Text feature [malignant] present in test data point [True]
848 Text feature [alternative] present in test data point [True]
853 Text feature [inhibitor] present in test data point [True]
855 Text feature [clinically] present in test data point [True]
857 Text feature [nm] present in test data point [True]
859 Text feature [pi3k] present in test data point [True]
860 Text feature [treatment] present in test data point [True]
868 Text feature [exon] present in test data point [True]
870 Text feature [gain] present in test data point [True]
874 Text feature [akt] present in test data point [True]
877 Text feature [01] present in test data point [True]
882 Text feature [endometrial] present in test data point [True]
```

```
884 Text feature [atp] present in test data point [True]
918 Text feature [cohort] present in test data point [True]
921 Text feature [differentiation] present in test data point [True]
924 Text feature [inhibitors] present in test data point [True]
928 Text feature [harboring] present in test data point [True]
951 Text feature [secondary] present in test data point [True]
954 Text feature [specimens] present in test data point [True]
959 Text feature [lymphoma] present in test data point [True]
960 Text feature [mapk] present in test data point [True]
965 Text feature [drug] present in test data point [True]
976 Text feature [erk] present in test data point [True]
Out of the top  1000  features  449 are present in query point
```

4.1.1.4. Feature Importance, Incorrectly classified point

In [237]:
```
test_point_index = 100
no_feature = 1000
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 :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],te
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.1558 0.0436 0.0119 0.631  0.0337 0.032  0.0869 0.0031 0.002 ]]
Actual Class : 1
--------------------------------------------------
11 Text feature [activity] present in test data point [True]
12 Text feature [protein] present in test data point [True]
15 Text feature [function] present in test data point [True]
16 Text feature [proteins] present in test data point [True]
17 Text feature [missense] present in test data point [True]
18 Text feature [results] present in test data point [True]
19 Text feature [type] present in test data point [True]
21 Text feature [experiments] present in test data point [True]
22 Text feature [shown] present in test data point [True]
23 Text feature [acid] present in test data point [True]
24 Text feature [important] present in test data point [True]
25 Text feature [functional] present in test data point [True]
26 Text feature [wild] present in test data point [True]
27 Text feature [whether] present in test data point [True]
28 Text feature [whereas] present in test data point [True]
29 Text feature [also] present in test data point [True]
30 Text feature [may] present in test data point [True]
31 Text feature [mutations] present in test data point [True]
32 Text feature [amino] present in test data point [True]
33 Text feature [suppressor] present in test data point [True]
34 Text feature [two] present in test data point [True]
36 Text feature [reduced] present in test data point [True]
37 Text feature [although] present in test data point [True]
40 Text feature [determined] present in test data point [True]
41 Text feature [indicated] present in test data point [True]
42 Text feature [therefore] present in test data point [True]
43 Text feature [levels] present in test data point [True]
44 Text feature [indicate] present in test data point [True]
46 Text feature [determine] present in test data point [True]
48 Text feature [either] present in test data point [True]
49 Text feature [ability] present in test data point [True]
50 Text feature [discussion] present in test data point [True]
51 Text feature [related] present in test data point [True]
52 Text feature [thus] present in test data point [True]
53 Text feature [fact] present in test data point [True]
54 Text feature [containing] present in test data point [True]
55 Text feature [mammalian] present in test data point [True]
56 Text feature [analysis] present in test data point [True]
57 Text feature [expressed] present in test data point [True]
61 Text feature [three] present in test data point [True]
63 Text feature [associated] present in test data point [True]
65 Text feature [vitro] present in test data point [True]
66 Text feature [result] present in test data point [True]
68 Text feature [effects] present in test data point [True]
69 Text feature [however] present in test data point [True]
70 Text feature [using] present in test data point [True]
71 Text feature [one] present in test data point [True]
72 Text feature [similar] present in test data point [True]
73 Text feature [effect] present in test data point [True]
76 Text feature [found] present in test data point [True]
77 Text feature [could] present in test data point [True]
79 Text feature [mutants] present in test data point [True]
80 Text feature [loss] present in test data point [True]
81 Text feature [assay] present in test data point [True]
82 Text feature [bind] present in test data point [True]
84 Text feature [mutation] present in test data point [True]
85 Text feature [used] present in test data point [True]
87 Text feature [several] present in test data point [True]
88 Text feature [due] present in test data point [True]
89 Text feature [vector] present in test data point [True]
90 Text feature [cells] present in test data point [True]
92 Text feature [generated] present in test data point [True]
93 Text feature [role] present in test data point [True]
94 Text feature [involved] present in test data point [True]
95 Text feature [lower] present in test data point [True]
97 Text feature [addition] present in test data point [True]
98 Text feature [50] present in test data point [True]
99 Text feature [data] present in test data point [True]
102 Text feature [mutant] present in test data point [True]
103 Text feature [affect] present in test data point [True]
106 Text feature [reported] present in test data point [True]
107 Text feature [figure] present in test data point [True]
109 Text feature [transfected] present in test data point [True]
110 Text feature [tagged] present in test data point [True]
113 Text feature [performed] present in test data point [True]
```

```
114 Text feature [critical] present in test data point [True]
116 Text feature [transfection] present in test data point [True]
118 Text feature [expression] present in test data point [True]
119 Text feature [incubated] present in test data point [True]
120 Text feature [consistent] present in test data point [True]
122 Text feature [vivo] present in test data point [True]
124 Text feature [compared] present in test data point [True]
125 Text feature [within] present in test data point [True]
126 Text feature [well] present in test data point [True]
128 Text feature [control] present in test data point [True]
129 Text feature [binding] present in test data point [True]
130 Text feature [different] present in test data point [True]
131 Text feature [indicating] present in test data point [True]
134 Text feature [resulting] present in test data point [True]
136 Text feature [page] present in test data point [True]
137 Text feature [residues] present in test data point [True]
139 Text feature [examined] present in test data point [True]
140 Text feature [derived] present in test data point [True]
141 Text feature [changes] present in test data point [True]
144 Text feature [including] present in test data point [True]
147 Text feature [note] present in test data point [True]
149 Text feature [might] present in test data point [True]
150 Text feature [analyses] present in test data point [True]
151 Text feature [acids] present in test data point [True]
152 Text feature [observation] present in test data point [True]
155 Text feature [observed] present in test data point [True]
156 Text feature [another] present in test data point [True]
157 Text feature [cell] present in test data point [True]
158 Text feature [respectively] present in test data point [True]
159 Text feature [required] present in test data point [True]
163 Text feature [directly] present in test data point [True]
165 Text feature [frequently] present in test data point [True]
166 Text feature [terminal] present in test data point [True]
169 Text feature [furthermore] present in test data point [True]
170 Text feature [dna] present in test data point [True]
173 Text feature [properties] present in test data point [True]
174 Text feature [western] present in test data point [True]
181 Text feature [significant] present in test data point [True]
183 Text feature [caused] present in test data point [True]
184 Text feature [studies] present in test data point [True]
186 Text feature [negative] present in test data point [True]
189 Text feature [large] present in test data point [True]
191 Text feature [full] present in test data point [True]
192 Text feature [human] present in test data point [True]
193 Text feature [four] present in test data point [True]
195 Text feature [least] present in test data point [True]
196 Text feature [showed] present in test data point [True]
198 Text feature [provide] present in test data point [True]
199 Text feature [germline] present in test data point [True]
200 Text feature [first] present in test data point [True]
201 Text feature [single] present in test data point [True]
203 Text feature [specific] present in test data point [True]
204 Text feature [sds] present in test data point [True]
205 Text feature [many] present in test data point [True]
207 Text feature [dependent] present in test data point [True]
208 Text feature [detected] present in test data point [True]
209 Text feature [measured] present in test data point [True]
210 Text feature [localization] present in test data point [True]
214 Text feature [highly] present in test data point [True]
217 Text feature [confirmed] present in test data point [True]
221 Text feature [based] present in test data point [True]
222 Text feature [domain] present in test data point [True]
230 Text feature [blue] present in test data point [True]
232 Text feature [study] present in test data point [True]
240 Text feature [unable] present in test data point [True]
241 Text feature [part] present in test data point [True]
242 Text feature [affected] present in test data point [True]
244 Text feature [tumor] present in test data point [True]
245 Text feature [mutagenesis] present in test data point [True]
246 Text feature [lysates] present in test data point [True]
247 Text feature [remains] present in test data point [True]
248 Text feature [relatively] present in test data point [True]
252 Text feature [without] present in test data point [True]
254 Text feature [hypothesis] present in test data point [True]
257 Text feature [gene] present in test data point [True]
258 Text feature [sequence] present in test data point [True]
259 Text feature [level] present in test data point [True]
260 Text feature [antibodies] present in test data point [True]
263 Text feature [major] present in test data point [True]
264 Text feature [followed] present in test data point [True]
265 Text feature [association] present in test data point [True]
266 Text feature [significantly] present in test data point [True]
270 Text feature [lead] present in test data point [True]
271 Text feature [ca] present in test data point [True]
272 Text feature [antibody] present in test data point [True]
273 Text feature [findings] present in test data point [True]
274 Text feature [would] present in test data point [True]

275 Text feature [presence] present in test data point [True]
276 Text feature [residue] present in test data point [True]
277 Text feature [normal] present in test data point [True]
```

```
278 Text feature [region] present in test data point [True]
285 Text feature [membrane] present in test data point [True]
287 Text feature [induced] present in test data point [True]
290 Text feature [second] present in test data point [True]
294 Text feature [constructs] present in test data point [True]
297 Text feature [identified] present in test data point [True]
298 Text feature [growth] present in test data point [True]
300 Text feature [increased] present in test data point [True]
302 Text feature [nuclear] present in test data point [True]
303 Text feature [substitution] present in test data point [True]
305 Text feature [activities] present in test data point [True]
310 Text feature [consequences] present in test data point [True]
311 Text feature [carried] present in test data point [True]
312 Text feature [interact] present in test data point [True]
316 Text feature [cancer] present in test data point [True]
317 Text feature [decreased] present in test data point [True]
319 Text feature [members] present in test data point [True]
320 Text feature [among] present in test data point [True]
321 Text feature [since] present in test data point [True]
322 Text feature [3a] present in test data point [True]
325 Text feature [distribution] present in test data point [True]
326 Text feature [dominant] present in test data point [True]
328 Text feature [48] present in test data point [True]
332 Text feature [complex] present in test data point [True]
335 Text feature [alone] present in test data point [True]
336 Text feature [gel] present in test data point [True]
342 Text feature [mediated] present in test data point [True]
344 Text feature [form] present in test data point [True]
345 Text feature [co] present in test data point [True]
347 Text feature [1998] present in test data point [True]
348 Text feature [independent] present in test data point [True]
350 Text feature [cause] present in test data point [True]
351 Text feature [90] present in test data point [True]
353 Text feature [groups] present in test data point [True]
355 Text feature [anti] present in test data point [True]
356 Text feature [et] present in test data point [True]
357 Text feature [3b] present in test data point [True]
359 Text feature [regulation] present in test data point [True]
360 Text feature [complete] present in test data point [True]
361 Text feature [tumorigenesis] present in test data point [True]
367 Text feature [length] present in test data point [True]
368 Text feature [al] present in test data point [True]
369 Text feature [active] present in test data point [True]
374 Text feature [five] present in test data point [True]
378 Text feature [impaired] present in test data point [True]
379 Text feature [sufficient] present in test data point [True]
381 Text feature [various] present in test data point [True]
382 Text feature [recently] present in test data point [True]
384 Text feature [conserved] present in test data point [True]
385 Text feature [number] present in test data point [True]
387 Text feature [families] present in test data point [True]
389 Text feature [2000] present in test data point [True]
391 Text feature [staining] present in test data point [True]
392 Text feature [six] present in test data point [True]
394 Text feature [change] present in test data point [True]
397 Text feature [able] present in test data point [True]
399 Text feature [interaction] present in test data point [True]
404 Text feature [strongly] present in test data point [True]
405 Text feature [mouse] present in test data point [True]
407 Text feature [red] present in test data point [True]
410 Text feature [located] present in test data point [True]
411 Text feature [increase] present in test data point [True]
413 Text feature [top] present in test data point [True]
414 Text feature [occur] present in test data point [True]
415 Text feature [common] present in test data point [True]
416 Text feature [regions] present in test data point [True]
417 Text feature [family] present in test data point [True]
419 Text feature [reaction] present in test data point [True]
420 Text feature [mechanisms] present in test data point [True]
424 Text feature [domains] present in test data point [True]
425 Text feature [development] present in test data point [True]
426 Text feature [cancers] present in test data point [True]
429 Text feature [early] present in test data point [True]
430 Text feature [target] present in test data point [True]
432 Text feature [flag] present in test data point [True]
434 Text feature [deletion] present in test data point [True]
435 Text feature [finding] present in test data point [True]
436 Text feature [next] present in test data point [True]
440 Text feature [via] present in test data point [True]
443 Text feature [image] present in test data point [True]
445 Text feature [occurs] present in test data point [True]
457 Text feature [absence] present in test data point [True]
459 Text feature [inhibit] present in test data point [True]
462 Text feature [contains] present in test data point [True]
464 Text feature [biological] present in test data point [True]
466 Text feature [genes] present in test data point [True]
467 Text feature [demonstrated] present in test data point [True]
469 Text feature [200] present in test data point [True]
471 Text feature [leads] present in test data point [True]
472 Text feature [structure] present in test data point [True]
480 Text feature [formation] present in test data point [True]
```

```
481 Text feature [line] present in test data point [True]
482 Text feature [nature] present in test data point [True]
484 Text feature [subjected] present in test data point [True]
491 Text feature [molecular] present in test data point [True]
492 Text feature [particular] present in test data point [True]
498 Text feature [pcr] present in test data point [True]
501 Text feature [structural] present in test data point [True]
502 Text feature [bound] present in test data point [True]
504 Text feature [11] present in test data point [True]
509 Text feature [applied] present in test data point [True]
516 Text feature [nucleus] present in test data point [True]
522 Text feature [led] present in test data point [True]
523 Text feature [1997] present in test data point [True]
525 Text feature [expressing] present in test data point [True]
527 Text feature [reduction] present in test data point [True]
531 Text feature [reports] present in test data point [True]
534 Text feature [manner] present in test data point [True]
536 Text feature [identify] present in test data point [True]
538 Text feature [panel] present in test data point [True]
539 Text feature [loop] present in test data point [True]
544 Text feature [chain] present in test data point [True]
546 Text feature [1999] present in test data point [True]
547 Text feature [position] present in test data point [True]
548 Text feature [group] present in test data point [True]
551 Text feature [contact] present in test data point [True]
552 Text feature [reporter] present in test data point [True]
559 Text feature [downstream] present in test data point [True]
566 Text feature [enhanced] present in test data point [True]
578 Text feature [require] present in test data point [True]
579 Text feature [binds] present in test data point [True]
580 Text feature [2a] present in test data point [True]
585 Text feature [2b] present in test data point [True]
586 Text feature [new] present in test data point [True]
594 Text feature [upon] present in test data point [True]
597 Text feature [lanes] present in test data point [True]
601 Text feature [case] present in test data point [True]
603 Text feature [isolated] present in test data point [True]
604 Text feature [proliferation] present in test data point [True]
605 Text feature [contained] present in test data point [True]
611 Text feature [fold] present in test data point [True]
615 Text feature [60] present in test data point [True]
623 Text feature [core] present in test data point [True]
624 Text feature [locus] present in test data point [True]
625 Text feature [unknown] present in test data point [True]
629 Text feature [samples] present in test data point [True]
632 Text feature [signal] present in test data point [True]
635 Text feature [apoptosis] present in test data point [True]
638 Text feature [gst] present in test data point [True]
641 Text feature [interactions] present in test data point [True]
642 Text feature [lane] present in test data point [True]
643 Text feature [mutated] present in test data point [True]
645 Text feature [clinical] present in test data point [True]
647 Text feature [ii] present in test data point [True]
649 Text feature [transcriptional] present in test data point [True]
652 Text feature [size] present in test data point [True]
660 Text feature [basis] present in test data point [True]
662 Text feature [activation] present in test data point [True]
663 Text feature [1996] present in test data point [True]
666 Text feature [factor] present in test data point [True]
669 Text feature [complexes] present in test data point [True]
671 Text feature [signals] present in test data point [True]
672 Text feature [set] present in test data point [True]
675 Text feature [tumors] present in test data point [True]
677 Text feature [signaling] present in test data point [True]
678 Text feature [pathway] present in test data point [True]
685 Text feature [state] present in test data point [True]
686 Text feature [phosphorylated] present in test data point [True]
692 Text feature [primary] present in test data point [True]
694 Text feature [terminus] present in test data point [True]
697 Text feature [2001] present in test data point [True]
699 Text feature [inhibition] present in test data point [True]
704 Text feature [allele] present in test data point [True]
711 Text feature [long] present in test data point [True]
714 Text feature [transcription] present in test data point [True]
716 Text feature [luciferase] present in test data point [True]
717 Text feature [side] present in test data point [True]
719 Text feature [promoter] present in test data point [True]
721 Text feature [activated] present in test data point [True]

725 Text feature [carcinoma] present in test data point [True]
726 Text feature [normalized] present in test data point [True]
751 Text feature [promote] present in test data point [True]
756 Text feature [factors] present in test data point [True]
761 Text feature [2002] present in test data point [True]
763 Text feature [smad4] present in test data point [True]
765 Text feature [helix] present in test data point [True]
769 Text feature [inhibited] present in test data point [True]
773 Text feature [end] present in test data point [True]
776 Text feature [treated] present in test data point [True]
777 Text feature [breast] present in test data point [True]
783 Text feature [overexpression] present in test data point [True]
```

```
790 Text feature [colorectal] present in test data point [True]
793 Text feature [activate] present in test data point [True]
800 Text feature [kinase] present in test data point [True]
801 Text feature [2003] present in test data point [True]
803 Text feature [tgf] present in test data point [True]
812 Text feature [colon] present in test data point [True]
828 Text feature [author] present in test data point [True]
829 Text feature [neutral] present in test data point [True]
831 Text feature [aberrant] present in test data point [True]
832 Text feature [response] present in test data point [True]
839 Text feature [confer] present in test data point [True]
842 Text feature [box] present in test data point [True]
848 Text feature [alternative] present in test data point [True]
850 Text feature [hydrophobic] present in test data point [True]
860 Text feature [treatment] present in test data point [True]
865 Text feature [repeats] present in test data point [True]
867 Text feature [stage] present in test data point [True]
903 Text feature [leukemia] present in test data point [True]
904 Text feature [conformation] present in test data point [True]
909 Text feature [lung] present in test data point [True]
912 Text feature [ng] present in test data point [True]
913 Text feature [constitutively] present in test data point [True]
915 Text feature [driven] present in test data point [True]
917 Text feature [bone] present in test data point [True]
919 Text feature [receptor] present in test data point [True]
921 Text feature [differentiation] present in test data point [True]
922 Text feature [basal] present in test data point [True]
924 Text feature [inhibitors] present in test data point [True]
934 Text feature [liver] present in test data point [True]
935 Text feature [p21] present in test data point [True]
937 Text feature [cyclin] present in test data point [True]
939 Text feature [stimulated] present in test data point [True]
940 Text feature [hydrogen] present in test data point [True]
941 Text feature [smad3] present in test data point [True]
942 Text feature [potent] present in test data point [True]
949 Text feature [smad2] present in test data point [True]
950 Text feature [stimulation] present in test data point [True]
951 Text feature [secondary] present in test data point [True]
961 Text feature [myc] present in test data point [True]
968 Text feature [metastatic] present in test data point [True]
969 Text feature [extracellular] present in test data point [True]
Out of the top  1000  features  387 are present in query point
```

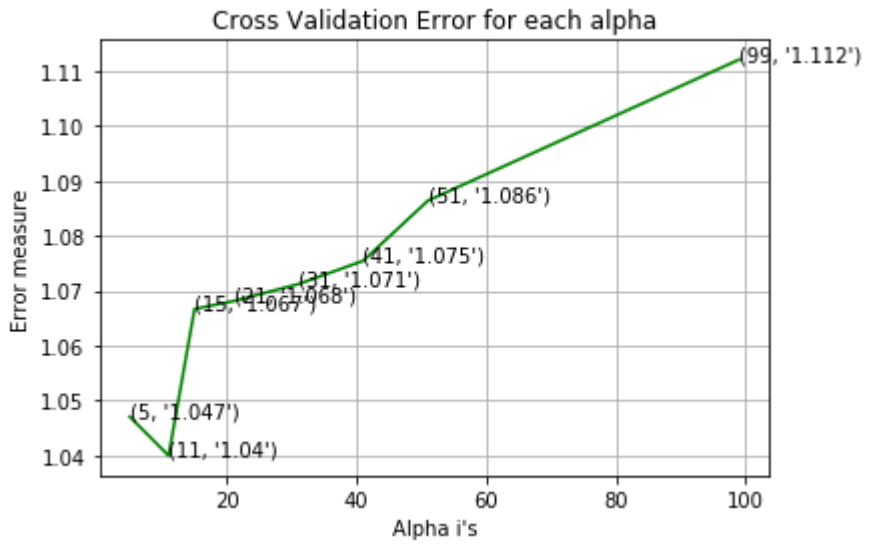
4.2. K Nearest Neighbour Classification

4.2.1. Hyper parameter tuning

In [238]:
```
# find more about KNeighborsClassifier() here http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.
# ------------------------
# 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-neighbors-geometr
#--------------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/generated/sklearn.calibra
# ----------------------------
# 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_responseCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_responseCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_responseCoding)
    cv_log_error_array.append(log_loss(cv_y, 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(cv_y, 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 = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_responseCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, predict_y, lab
predict_y = sig_clf.predict_proba(cv_x_responseCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_cv, predic
predict_y = sig_clf.predict_proba(test_x_responseCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predict_y, label
```

```
for alpha = 5
Log Loss : 1.0470822160366484
for alpha = 11
Log Loss : 1.0399928306605806
for alpha = 15
Log Loss : 1.0666871502746276
for alpha = 21
Log Loss : 1.0681299320767772
for alpha = 31
Log Loss : 1.0712092515441383
for alpha = 41
Log Loss : 1.0754482303366437
for alpha = 51
Log Loss : 1.0864093764174962
for alpha = 99
Log Loss : 1.1120814813174422
```

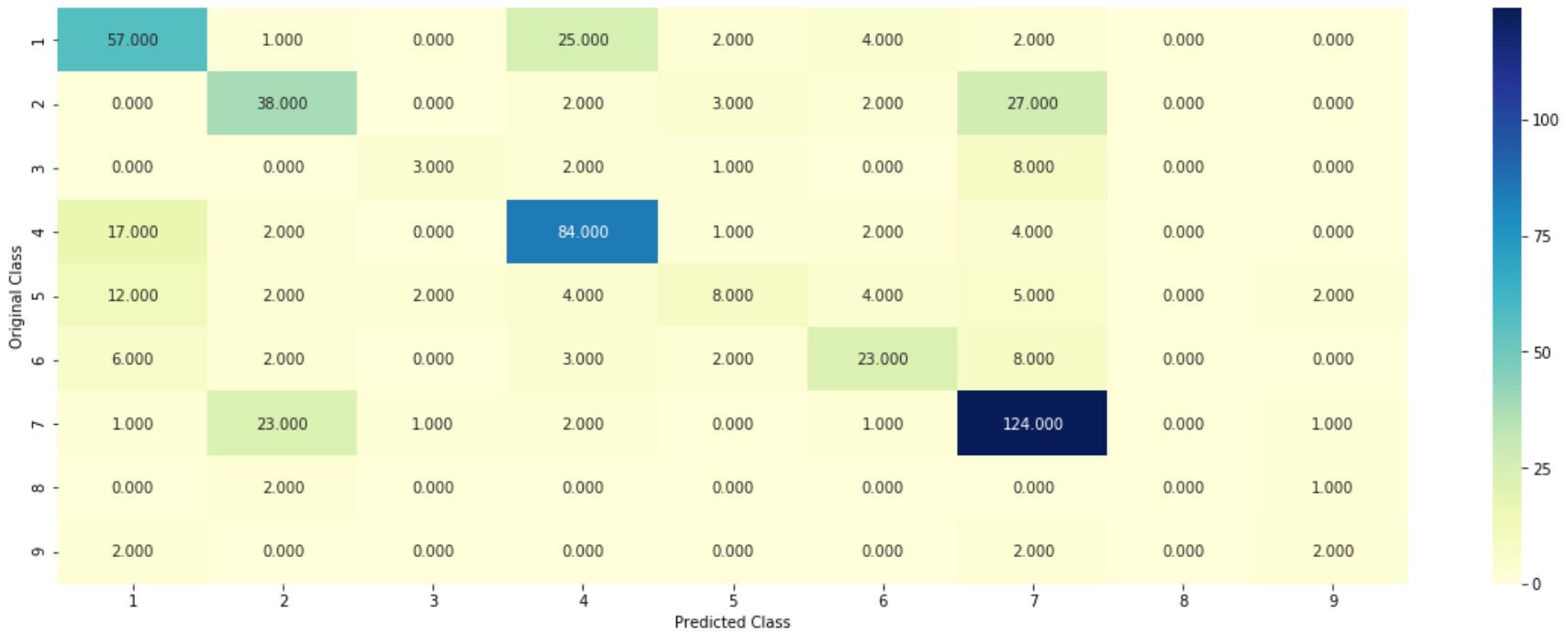



```
For values of best alpha =  11 The train log loss is: 0.6500392723689956
For values of best alpha =  11 The cross validation log loss is: 1.0399928306605806
For values of best alpha =  11 The test log loss is: 0.9835121342822057
```

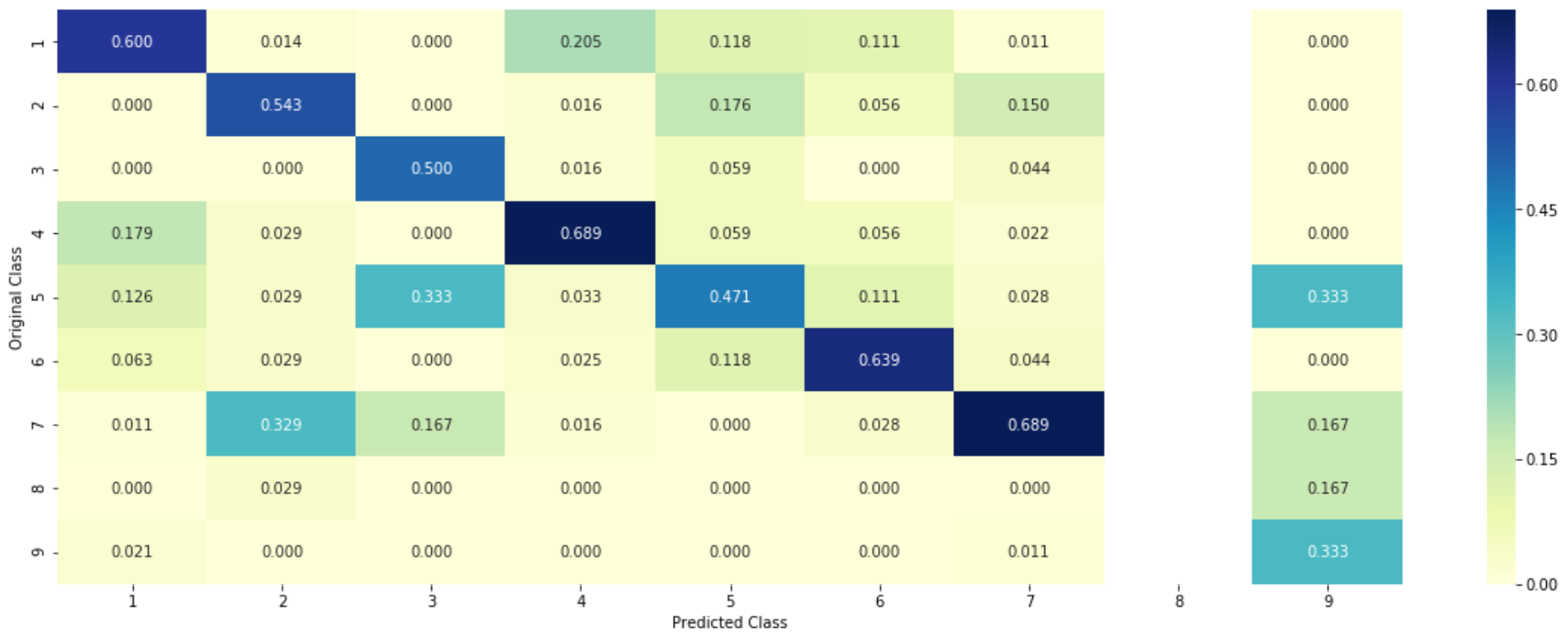
4.2.2. Testing the model with best hyper paramters

In [239]:
```
# find more about KNeighborsClassifier() here http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.
# ------------------------
# 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-neighbors-geometr
#-------------------------------------
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
predict_and_plot_confusion_matrix(train_x_responseCoding, train_y, cv_x_responseCoding, cv_y, clf)
```

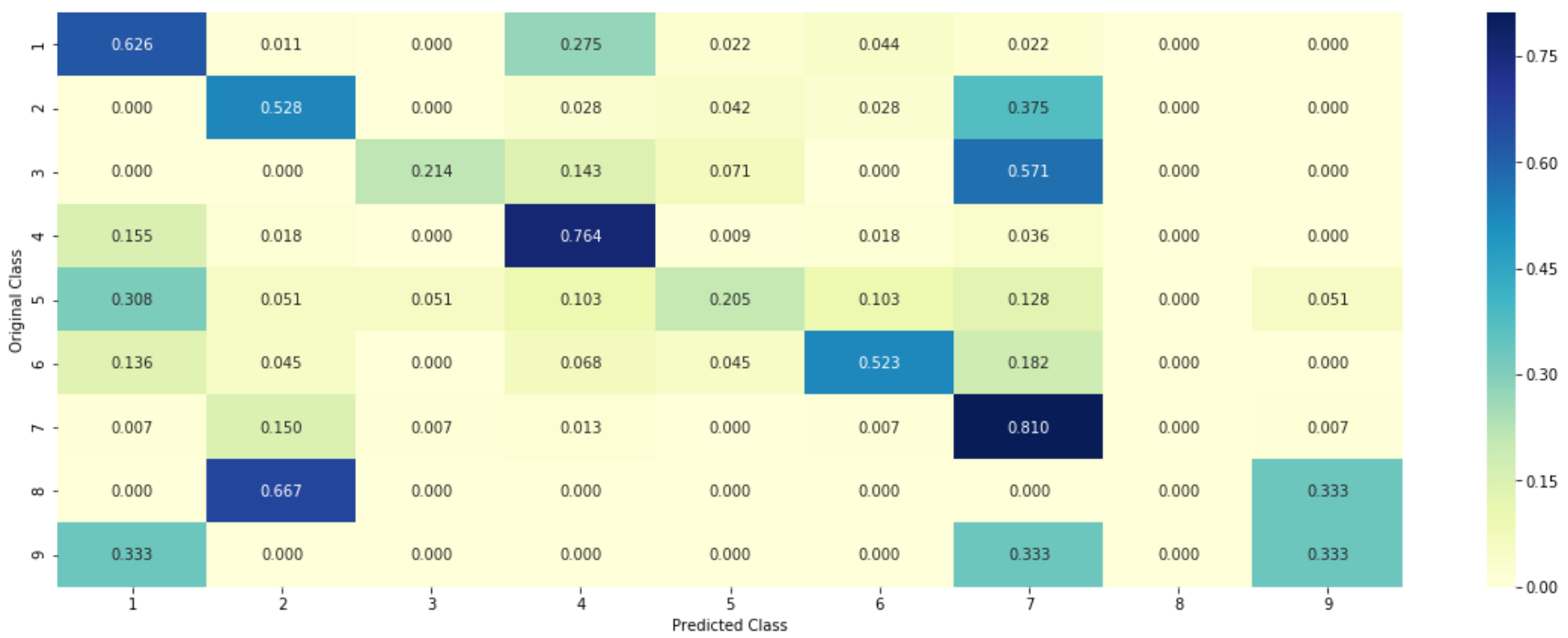
```
Log loss : 1.0399928306605806
Number of mis-classified points : 0.36278195488721804
-------------------- Confusion matrix --------------------
```




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




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



4.2.3.Sample Query point -1

In [240]:
```
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)

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

```
Predicted Class : 1
Actual Class : 4
The  11  nearest neighbours of the test points belongs to classes [4 4 4 4 4 4 4 7 4 4 4]
Fequency of nearest points : Counter({4: 10, 7: 1})
```

4.2.4. Sample Query Point-2

In [242]:
```
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)

test_point_index = 105

predicted_cls = sig_clf.predict(test_x_responseCoding[test_point_index].reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Actual Class :", test_y[test_point_index])
neighbors = clf.kneighbors(test_x_responseCoding[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 belongs to classe
print("Fequency of nearest points :",Counter(train_y[neighbors[1][0]]))
```

```
Predicted Class : 7
Actual Class : 7
the k value for knn is 11 and the nearest neighbours of the test points belongs to classes [7 7 7 7 7 7 7 7 2 7
7]
Fequency of nearest points : Counter({7: 10, 2: 1})
```

4.3. Logistic Regression

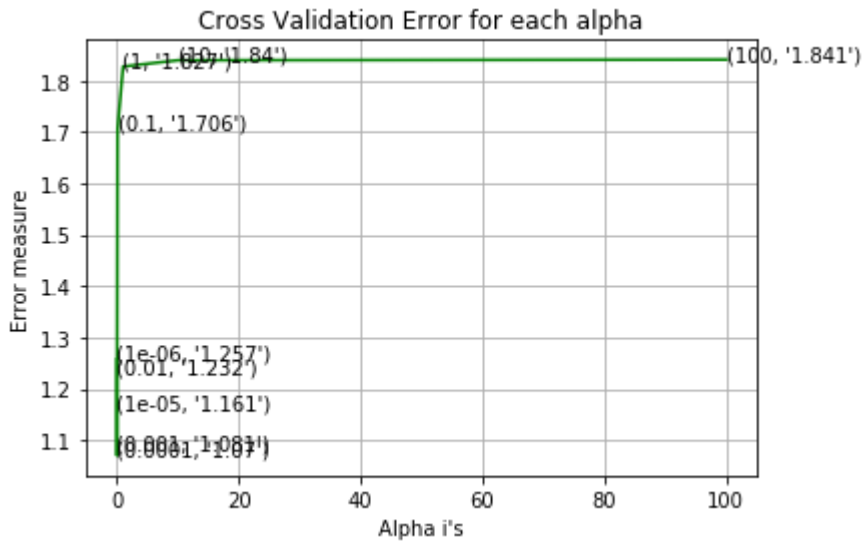
4.3.1. With Class balancing

4.3.1.1. Hyper paramter tuning

In [245]:

```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDCl
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=N
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0
# 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/
#-------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/generated/sklearn.calibra
# ----------------------------
# 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, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, 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(cv_y, 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', random_state=42)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, predict_y, lab
predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_cv, predic
predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predict_y, label
```

```
for alpha = 1e-06
Log Loss : 1.2573299788549024
for alpha = 1e-05
Log Loss : 1.1612089458566022
for alpha = 0.0001
Log Loss : 1.0697848339167542
for alpha = 0.001
Log Loss : 1.0805896866264562
for alpha = 0.01
Log Loss : 1.2315884147204645
for alpha = 0.1
Log Loss : 1.7062598092283299
for alpha = 1
Log Loss : 1.8273471238086016
for alpha = 10
Log Loss : 1.8395211307318156
for alpha = 100
Log Loss : 1.840846253827411
```

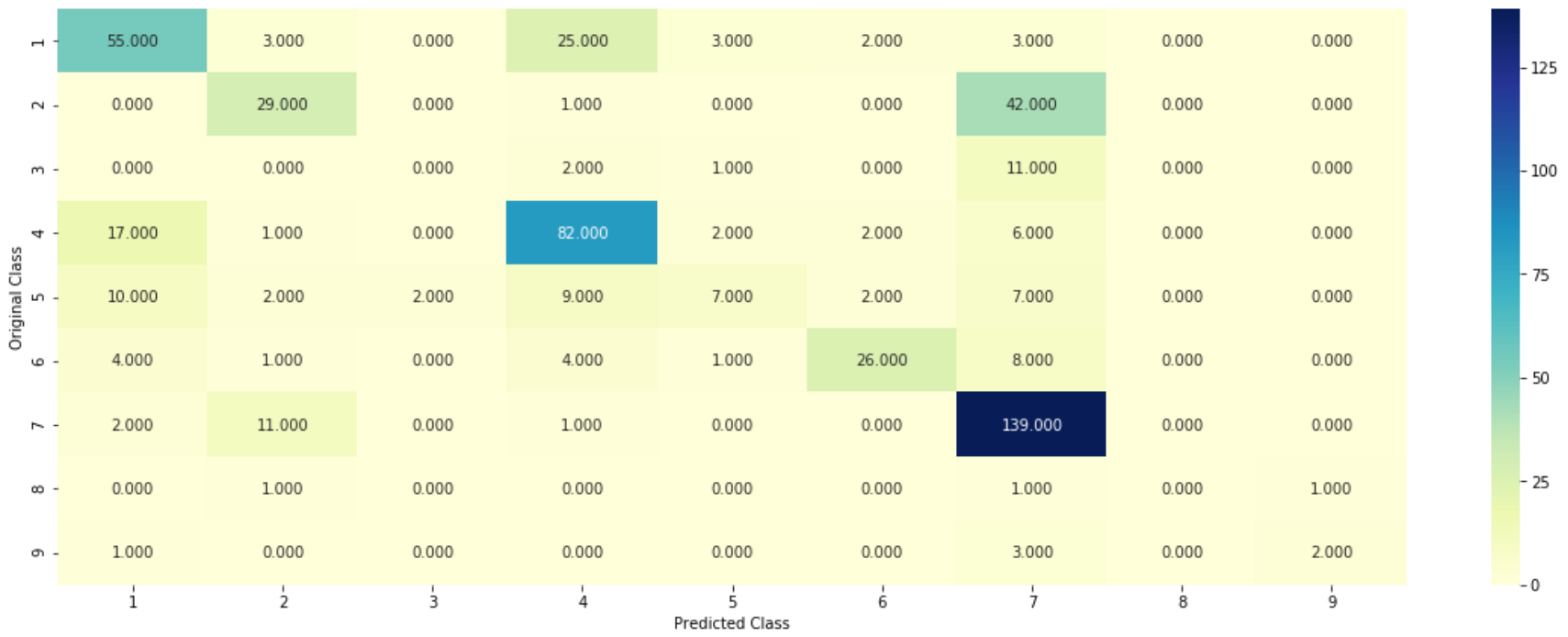



```
For values of best alpha =  0.0001 The train log loss is: 0.4472405130488467
For values of best alpha =  0.0001 The cross validation log loss is: 1.0697848339167542
For values of best alpha =  0.0001 The test log loss is: 0.923840342701383
```

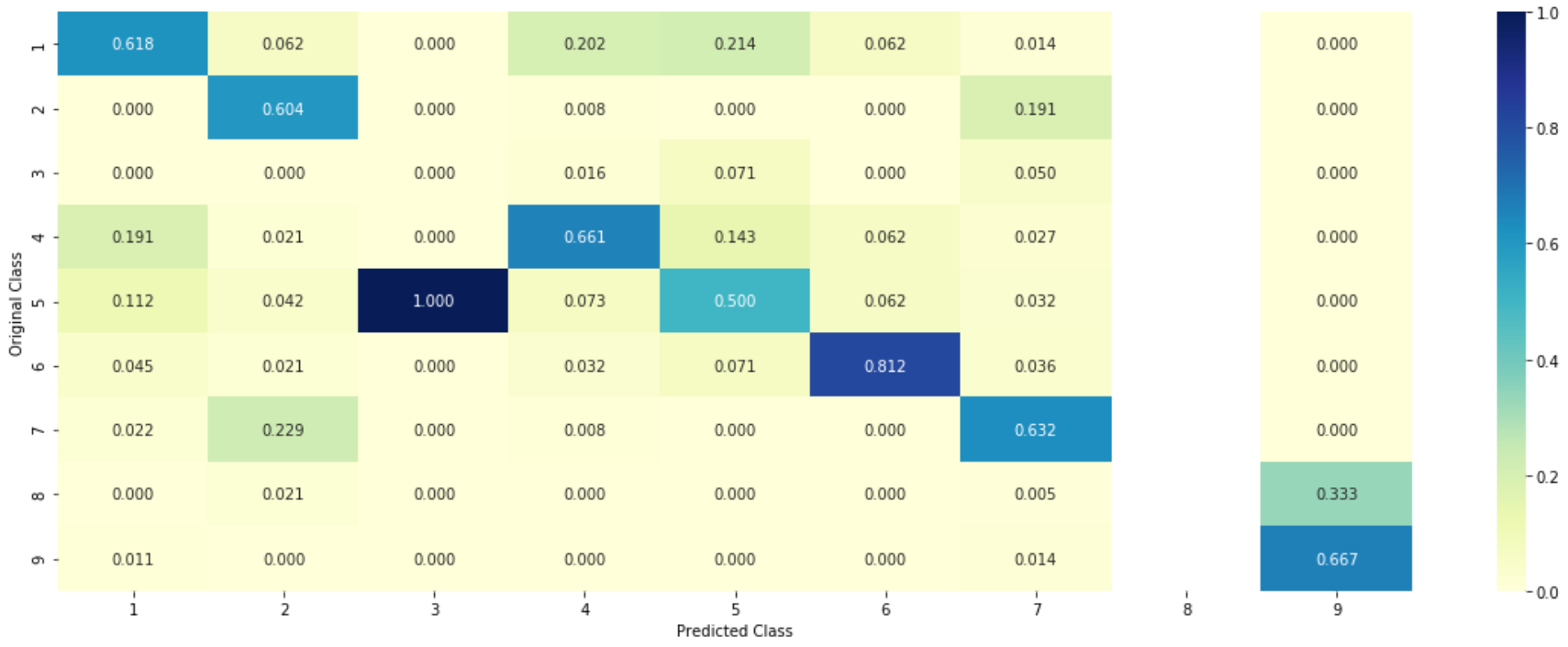
4.3.1.2. Testing the model with best hyper paramters

In [246]:
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDCl
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=N
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0
# 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/
#-------------------------------
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_onehotCoding, cv_y, clf)
```

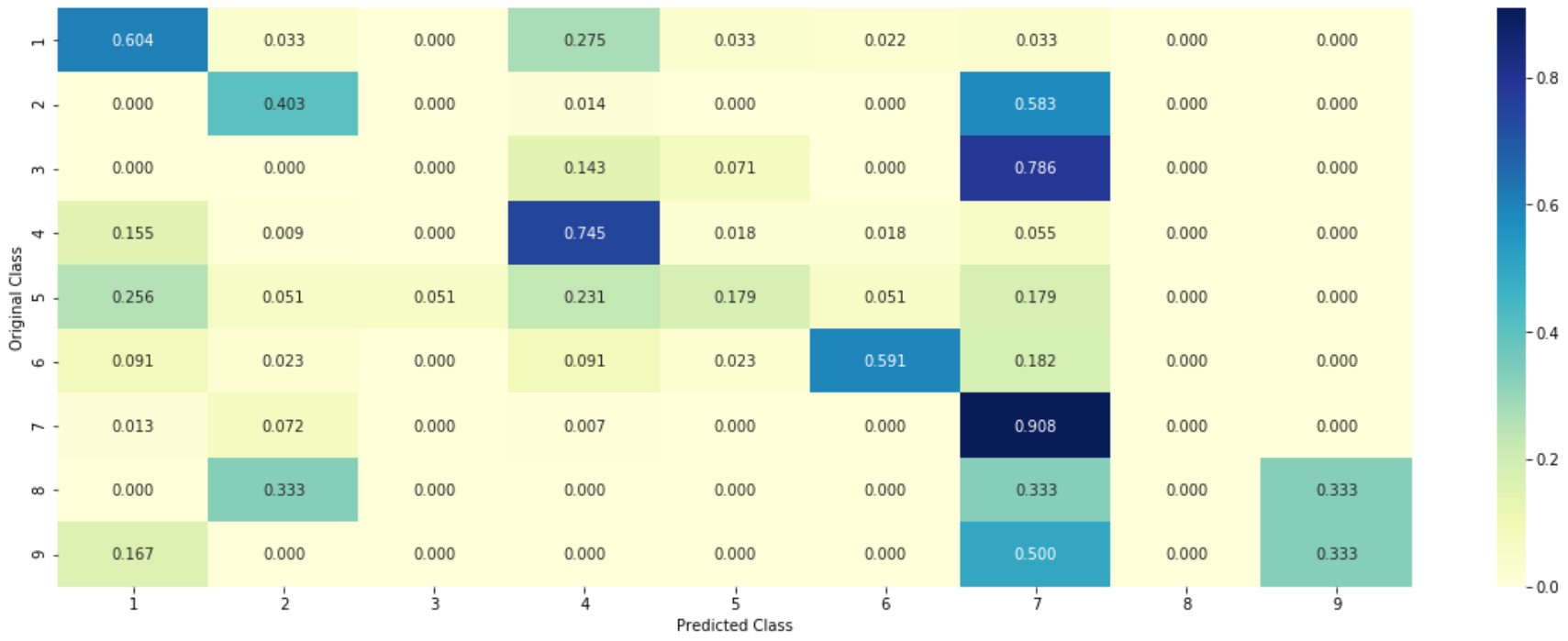
```
Log loss : 1.0697848339167542
Number of mis-classified points : 0.3609022556390977
-------------------- Confusion matrix --------------------
```




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




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



4.3.1.3. Feature Importance

In [247]:
```
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 [248]:
```
# from tabulate import tabulate
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 1
no_feature = 1000
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 :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],te
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.0132 0.0056 0.0125 0.927  0.0069 0.0046 0.0251 0.0028 0.0023]]
Actual Class : 4
--------------------------------------------------
18 Text feature [suppressor] present in test data point [True]
89 Text feature [missense] present in test data point [True]
115 Text feature [right] present in test data point [True]
117 Text feature [suggesting] present in test data point [True]
157 Text feature [lanes] present in test data point [True]
167 Text feature [due] present in test data point [True]
192 Text feature [flag] present in test data point [True]
194 Text feature [dominant] present in test data point [True]
224 Text feature [phenotype] present in test data point [True]
233 Text feature [mice] present in test data point [True]
235 Text feature [reduced] present in test data point [True]
269 Text feature [western] present in test data point [True]
271 Text feature [direct] present in test data point [True]
276 Text feature [play] present in test data point [True]
285 Text feature [mm] present in test data point [True]
301 Text feature [functional] present in test data point [True]
313 Text feature [see] present in test data point [True]
318 Text feature [resulting] present in test data point [True]
321 Text feature [proportion] present in test data point [True]
332 Text feature [show] present in test data point [True]
333 Text feature [comparison] present in test data point [True]
339 Text feature [transfected] present in test data point [True]
350 Text feature [germline] present in test data point [True]
351 Text feature [substrate] present in test data point [True]
352 Text feature [suggest] present in test data point [True]
357 Text feature [protein] present in test data point [True]
373 Text feature [phosphatase] present in test data point [True]
374 Text feature [ca] present in test data point [True]
375 Text feature [high] present in test data point [True]
381 Text feature [age] present in test data point [True]
382 Text feature [risk] present in test data point [True]
391 Text feature [stably] present in test data point [True]
393 Text feature [investigated] present in test data point [True]
394 Text feature [4a] present in test data point [True]
396 Text feature [plasmid] present in test data point [True]
398 Text feature [washed] present in test data point [True]
402 Text feature [supplementary] present in test data point [True]
416 Text feature [left] present in test data point [True]
418 Text feature [tagged] present in test data point [True]
426 Text feature [stability] present in test data point [True]
430 Text feature [38] present in test data point [True]
431 Text feature [analyses] present in test data point [True]
438 Text feature [determine] present in test data point [True]
440 Text feature [decreased] present in test data point [True]
441 Text feature [nucleus] present in test data point [True]
442 Text feature [following] present in test data point [True]
443 Text feature [presented] present in test data point [True]
449 Text feature [29] present in test data point [True]
464 Text feature [considered] present in test data point [True]
466 Text feature [transfection] present in test data point [True]
477 Text feature [loss] present in test data point [True]
478 Text feature [negative] present in test data point [True]
482 Text feature [impaired] present in test data point [True]
484 Text feature [appears] present in test data point [True]
489 Text feature [low] present in test data point [True]
491 Text feature [sds] present in test data point [True]
495 Text feature [nuclear] present in test data point [True]
496 Text feature [detection] present in test data point [True]
499 Text feature [materials] present in test data point [True]
501 Text feature [plates] present in test data point [True]
502 Text feature [dna] present in test data point [True]
505 Text feature [assay] present in test data point [True]
510 Text feature [activity] present in test data point [True]
511 Text feature [role] present in test data point [True]
512 Text feature [reaction] present in test data point [True]
516 Text feature [combination] present in test data point [True]
518 Text feature [indeed] present in test data point [True]
520 Text feature [heterozygous] present in test data point [True]
524 Text feature [mutants] present in test data point [True]
529 Text feature [endometrial] present in test data point [True]
531 Text feature [30] present in test data point [True]
538 Text feature [years] present in test data point [True]
```

```
544 Text feature [therefore] present in test data point [True]
552 Text feature [multiple] present in test data point [True]
557 Text feature [recent] present in test data point [True]
559 Text feature [relatively] present in test data point [True]
563 Text feature [shows] present in test data point [True]
571 Text feature [complete] present in test data point [True]
572 Text feature [first] present in test data point [True]
575 Text feature [contribute] present in test data point [True]
581 Text feature [since] present in test data point [True]
586 Text feature [express] present in test data point [True]
598 Text feature [pten] present in test data point [True]
600 Text feature [proteins] present in test data point [True]
601 Text feature [genetic] present in test data point [True]
603 Text feature [performed] present in test data point [True]
604 Text feature [tissue] present in test data point [True]
608 Text feature [still] present in test data point [True]
610 Text feature [26] present in test data point [True]
614 Text feature [mouse] present in test data point [True]
617 Text feature [levels] present in test data point [True]
618 Text feature [regulation] present in test data point [True]
622 Text feature [whether] present in test data point [True]
623 Text feature [elevated] present in test data point [True]
624 Text feature [type] present in test data point [True]
625 Text feature [vector] present in test data point [True]
630 Text feature [cancers] present in test data point [True]
631 Text feature [important] present in test data point [True]
634 Text feature [frequently] present in test data point [True]
635 Text feature [antibodies] present in test data point [True]
636 Text feature [localization] present in test data point [True]
637 Text feature [induced] present in test data point [True]
639 Text feature [blot] present in test data point [True]
641 Text feature [mechanism] present in test data point [True]
646 Text feature [times] present in test data point [True]
647 Text feature [affected] present in test data point [True]
648 Text feature [wild] present in test data point [True]
653 Text feature [12] present in test data point [True]
655 Text feature [figure] present in test data point [True]
656 Text feature [observed] present in test data point [True]
658 Text feature [able] present in test data point [True]
659 Text feature [function] present in test data point [True]
660 Text feature [buffer] present in test data point [True]
664 Text feature [likely] present in test data point [True]
669 Text feature [lymphoma] present in test data point [True]
671 Text feature [several] present in test data point [True]
677 Text feature [incubated] present in test data point [True]
685 Text feature [deleterious] present in test data point [True]
690 Text feature [antibody] present in test data point [True]
693 Text feature [page] present in test data point [True]
696 Text feature [moreover] present in test data point [True]
697 Text feature [induction] present in test data point [True]
699 Text feature [process] present in test data point [True]
702 Text feature [manufacturer] present in test data point [True]
706 Text feature [seen] present in test data point [True]
707 Text feature [40] present in test data point [True]
709 Text feature [consistent] present in test data point [True]
710 Text feature [relative] present in test data point [True]
711 Text feature [two] present in test data point [True]
715 Text feature [research] present in test data point [True]
721 Text feature [whole] present in test data point [True]
728 Text feature [effects] present in test data point [True]
734 Text feature [together] present in test data point [True]
736 Text feature [syndrome] present in test data point [True]
737 Text feature [colon] present in test data point [True]
739 Text feature [analysis] present in test data point [True]
746 Text feature [assays] present in test data point [True]
749 Text feature [pathogenic] present in test data point [True]
753 Text feature [mrna] present in test data point [True]
755 Text feature [31] present in test data point [True]
756 Text feature [27] present in test data point [True]
757 Text feature [significant] present in test data point [True]
761 Text feature [various] present in test data point [True]
764 Text feature [effect] present in test data point [True]
765 Text feature [lead] present in test data point [True]
766 Text feature [demonstrate] present in test data point [True]
768 Text feature [according] present in test data point [True]
770 Text feature [group] present in test data point [True]
772 Text feature [positive] present in test data point [True]
773 Text feature [control] present in test data point [True]
781 Text feature [level] present in test data point [True]
782 Text feature [distribution] present in test data point [True]
786 Text feature [detected] present in test data point [True]
788 Text feature [wt] present in test data point [True]
792 Text feature [enhanced] present in test data point [True]
793 Text feature [90] present in test data point [True]
798 Text feature [pathways] present in test data point [True]
799 Text feature [led] present in test data point [True]
800 Text feature [thus] present in test data point [True]
802 Text feature [types] present in test data point [True]
804 Text feature [normal] present in test data point [True]
805 Text feature [amplified] present in test data point [True]
806 Text feature [mechanisms] present in test data point [True]
```

```
807 Text feature [molecular] present in test data point [True]
808 Text feature [used] present in test data point [True]
810 Text feature [cellular] present in test data point [True]
811 Text feature [variants] present in test data point [True]
816 Text feature [similarly] present in test data point [True]
818 Text feature [patient] present in test data point [True]
820 Text feature [anti] present in test data point [True]
821 Text feature [shown] present in test data point [True]
822 Text feature [confirmed] present in test data point [True]
825 Text feature [genomic] present in test data point [True]
829 Text feature [least] present in test data point [True]

830 Text feature [fig] present in test data point [True]
833 Text feature [test] present in test data point [True]
834 Text feature [amino] present in test data point [True]
839 Text feature [subjected] present in test data point [True]
841 Text feature [green] present in test data point [True]
842 Text feature [malignant] present in test data point [True]
845 Text feature [42] present in test data point [True]
846 Text feature [studied] present in test data point [True]
848 Text feature [associated] present in test data point [True]
851 Text feature [alterations] present in test data point [True]
852 Text feature [results] present in test data point [True]
856 Text feature [blood] present in test data point [True]
858 Text feature [impact] present in test data point [True]
860 Text feature [sporadic] present in test data point [True]
864 Text feature [cohort] present in test data point [True]
867 Text feature [cannot] present in test data point [True]
868 Text feature [atp] present in test data point [True]
871 Text feature [introduction] present in test data point [True]
873 Text feature [may] present in test data point [True]
874 Text feature [cause] present in test data point [True]
875 Text feature [functions] present in test data point [True]
877 Text feature [28] present in test data point [True]
880 Text feature [whereas] present in test data point [True]
882 Text feature [decrease] present in test data point [True]
883 Text feature [3a] present in test data point [True]
885 Text feature [hours] present in test data point [True]
886 Text feature [min] present in test data point [True]
889 Text feature [via] present in test data point [True]
890 Text feature [methods] present in test data point [True]
895 Text feature [23] present in test data point [True]
900 Text feature [degradation] present in test data point [True]
901 Text feature [major] present in test data point [True]
904 Text feature [described] present in test data point [True]
906 Text feature [s1] present in test data point [True]
908 Text feature [domains] present in test data point [True]
915 Text feature [acid] present in test data point [True]
920 Text feature [breast] present in test data point [True]
934 Text feature [approximately] present in test data point [True]
946 Text feature [table] present in test data point [True]
950 Text feature [long] present in test data point [True]
978 Text feature [part] present in test data point [True]
988 Text feature [proliferation] present in test data point [True]
995 Text feature [37] present in test data point [True]
Out of the top  1000  features  217 are present in query point
```

4.3.1.3.2. Incorrectly Classified point

In [249]:
```
test_point_index = 100
no_feature = 5000
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 :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],te
```

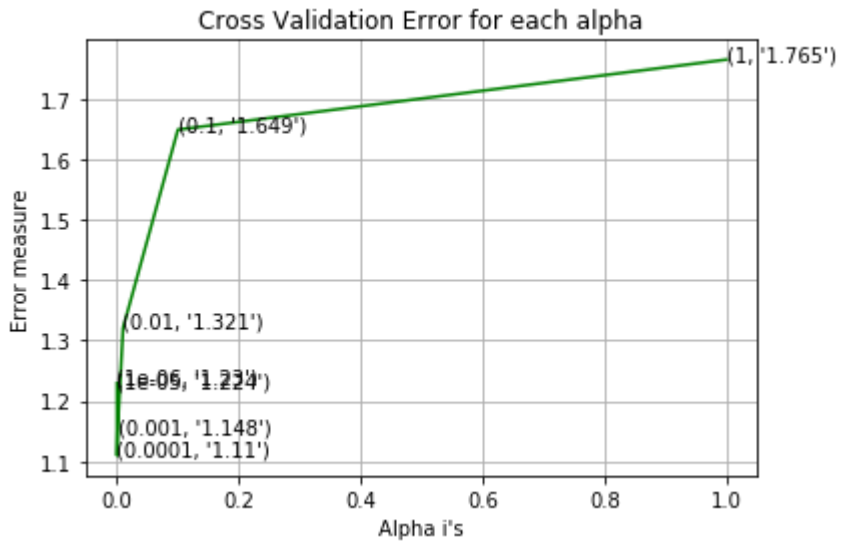
```
Predicted Class : 4
Predicted Class Probabilities: [[0.2081 0.0101 0.0135 0.7173 0.0139 0.0135 0.0169 0.0038 0.0029]]
Actual Class : 1
--------------------------------------------------
18 Text feature [suppressor] present in test data point [True]
89 Text feature [missense] present in test data point [True]
157 Text feature [lanes] present in test data point [True]
167 Text feature [due] present in test data point [True]
179 Text feature [families] present in test data point [True]
186 Text feature [tumorigenesis] present in test data point [True]
192 Text feature [flag] present in test data point [True]
194 Text feature [dominant] present in test data point [True]
200 Text feature [liver] present in test data point [True]
235 Text feature [reduced] present in test data point [True]
258 Text feature [mammalian] present in test data point [True]
259 Text feature [deletion] present in test data point [True]
269 Text feature [western] present in test data point [True]
274 Text feature [smad4] present in test data point [True]
291 Text feature [1998] present in test data point [True]
```

4.3.2. Without Class balancing

4.3.2.1. Hyper paramter tuning

In [250]:
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDCl
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=N
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0
# 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/
#-------------------------------



# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/generated/sklearn.calibra
# ------------------------------
# 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, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    print("Log Loss :",log_loss(cv_y, 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, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, predict_y, lab
predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_cv, predic
predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predict_y, label
```

```
for alpha = 1e-06
Log Loss : 1.229616452548897
for alpha = 1e-05
Log Loss : 1.223546560197277
for alpha = 0.0001
Log Loss : 1.1095533267013222
for alpha = 0.001
Log Loss : 1.1479587491146248
for alpha = 0.01
Log Loss : 1.3211186732304874
for alpha = 0.1
Log Loss : 1.6492446361561401
for alpha = 1
Log Loss : 1.7651410655674775
```

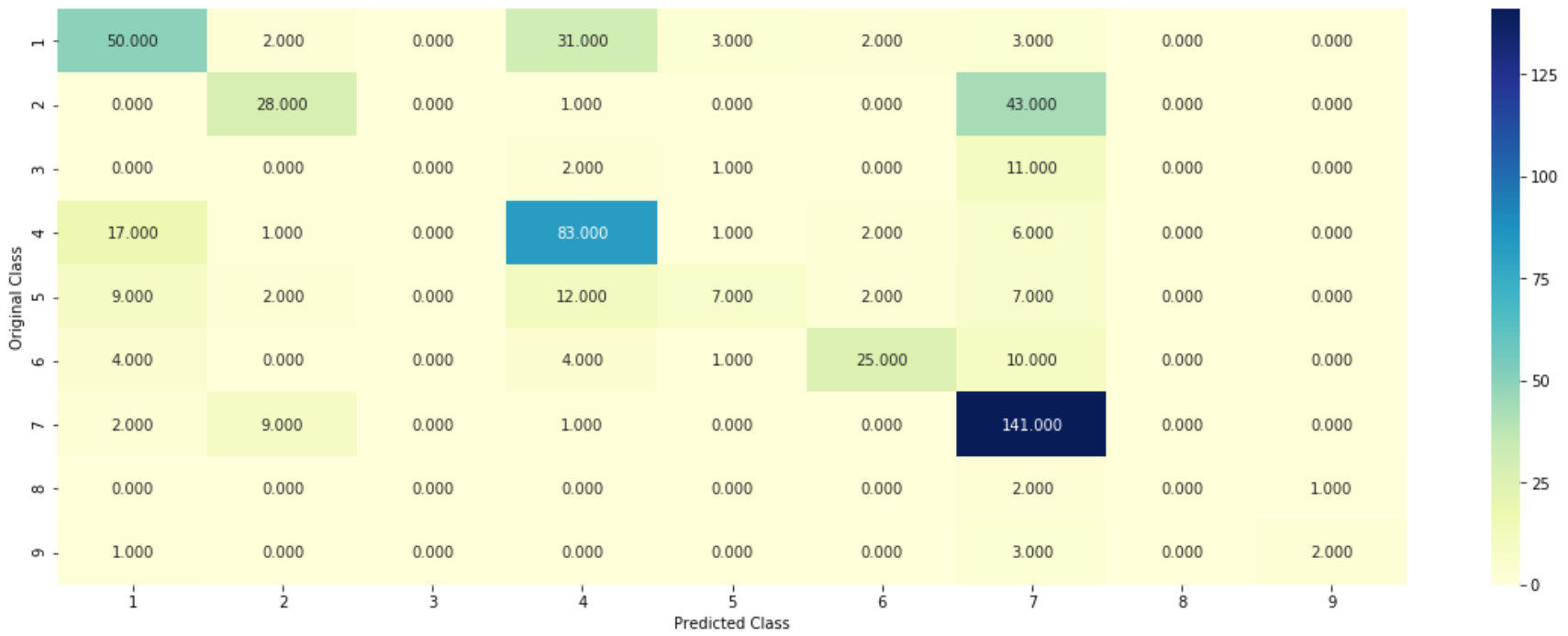



```
For values of best alpha =  0.0001 The train log loss is: 0.44102471844656876
For values of best alpha =  0.0001 The cross validation log loss is: 1.1095533267013222
For values of best alpha =  0.0001 The test log loss is: 0.9523338413522304
```

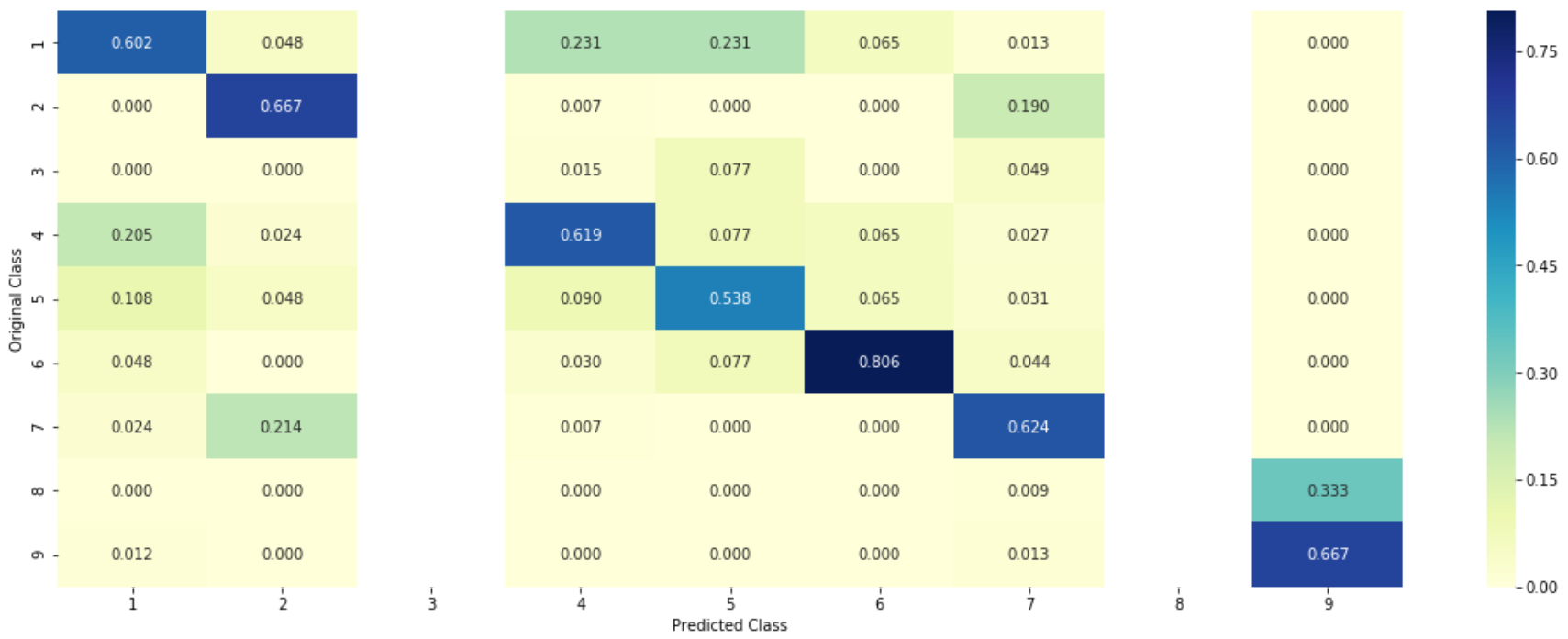
4.3.2.2. Testing model with best hyper parameters

In [251]:
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDCl
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=N
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0
# 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, train_y, cv_x_onehotCoding, cv_y, clf)
```

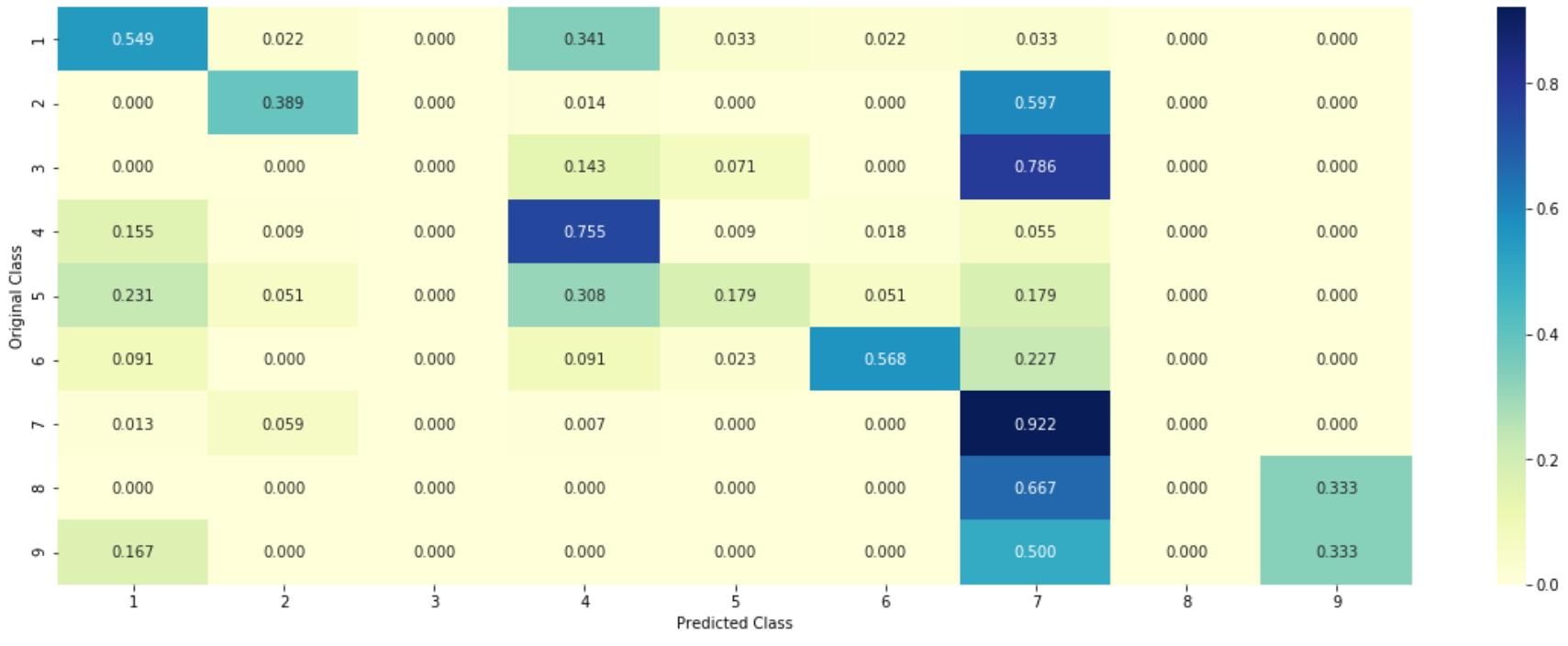
```
Log loss : 1.1095533267013222
Number of mis-classified points : 0.3684210526315789
-------------------- Confusion matrix --------------------
```




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




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



4.3.2.3. Feature Importance, Correctly Classified point

In [252]:
```
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 1
no_feature = 1000
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 :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],te
```

```
Predicted Class : 4
Predicted Class Probabilities: [[1.430e-02 5.800e-03 1.590e-02 9.176e-01 6.400e-03 4.300e-03 3.400e-02
  1.200e-03 6.000e-04]]
Actual Class : 4
--------------------------------------------------
37 Text feature [suppressor] present in test data point [True]
105 Text feature [suggesting] present in test data point [True]
111 Text feature [missense] present in test data point [True]
118 Text feature [right] present in test data point [True]
143 Text feature [lanes] present in test data point [True]
177 Text feature [due] present in test data point [True]
230 Text feature [flag] present in test data point [True]
231 Text feature [dominant] present in test data point [True]
233 Text feature [play] present in test data point [True]
244 Text feature [reduced] present in test data point [True]
267 Text feature [suggest] present in test data point [True]
272 Text feature [western] present in test data point [True]
274 Text feature [see] present in test data point [True]
278 Text feature [mice] present in test data point [True]
280 Text feature [functional] present in test data point [True]
285 Text feature [phenotype] present in test data point [True]
291 Text feature [resulting] present in test data point [True]
295 Text feature [direct] present in test data point [True]
300 Text feature [proportion] present in test data point [True]
329 Text feature [mm] present in test data point [True]
355 Text feature [substrate] present in test data point [True]
358 Text feature [protein] present in test data point [True]
363 Text feature [show] present in test data point [True]
366 Text feature [transfected] present in test data point [True]
368 Text feature [plasmid] present in test data point [True]
369 Text feature [ca] present in test data point [True]
371 Text feature [high] present in test data point [True]
374 Text feature [stably] present in test data point [True]
378 Text feature [4a] present in test data point [True]
379 Text feature [risk] present in test data point [True]
383 Text feature [germline] present in test data point [True]
390 Text feature [comparison] present in test data point [True]
400 Text feature [supplementary] present in test data point [True]
403 Text feature [phosphatase] present in test data point [True]
407 Text feature [analyses] present in test data point [True]
409 Text feature [decreased] present in test data point [True]
411 Text feature [determine] present in test data point [True]
413 Text feature [investigated] present in test data point [True]
415 Text feature [age] present in test data point [True]
418 Text feature [washed] present in test data point [True]
420 Text feature [presented] present in test data point [True]
425 Text feature [left] present in test data point [True]
429 Text feature [stability] present in test data point [True]
433 Text feature [tagged] present in test data point [True]
434 Text feature [following] present in test data point [True]
441 Text feature [38] present in test data point [True]
442 Text feature [nucleus] present in test data point [True]
443 Text feature [29] present in test data point [True]
451 Text feature [plates] present in test data point [True]
455 Text feature [impaired] present in test data point [True]
457 Text feature [indeed] present in test data point [True]
464 Text feature [considered] present in test data point [True]
467 Text feature [role] present in test data point [True]
471 Text feature [transfection] present in test data point [True]
474 Text feature [detection] present in test data point [True]
476 Text feature [assay] present in test data point [True]
483 Text feature [materials] present in test data point [True]
492 Text feature [complete] present in test data point [True]
494 Text feature [activity] present in test data point [True]
500 Text feature [sds] present in test data point [True]
506 Text feature [low] present in test data point [True]
507 Text feature [combination] present in test data point [True]
510 Text feature [reaction] present in test data point [True]
512 Text feature [mutants] present in test data point [True]
513 Text feature [loss] present in test data point [True]
514 Text feature [first] present in test data point [True]
517 Text feature [negative] present in test data point [True]
519 Text feature [dna] present in test data point [True]
524 Text feature [endometrial] present in test data point [True]
525 Text feature [contribute] present in test data point [True]
526 Text feature [recent] present in test data point [True]
531 Text feature [therefore] present in test data point [True]
```

```
535 Text feature [multiple] present in test data point [True]
536 Text feature [heterozygous] present in test data point [True]
538 Text feature [elevated] present in test data point [True]
540 Text feature [relatively] present in test data point [True]
541 Text feature [appears] present in test data point [True]
542 Text feature [nuclear] present in test data point [True]
548 Text feature [still] present in test data point [True]
555 Text feature [express] present in test data point [True]
560 Text feature [since] present in test data point [True]
561 Text feature [years] present in test data point [True]
563 Text feature [levels] present in test data point [True]
565 Text feature [30] present in test data point [True]
566 Text feature [tissue] present in test data point [True]
570 Text feature [antibodies] present in test data point [True]
572 Text feature [26] present in test data point [True]
573 Text feature [performed] present in test data point [True]
575 Text feature [mouse] present in test data point [True]
576 Text feature [regulation] present in test data point [True]
582 Text feature [likely] present in test data point [True]
586 Text feature [shows] present in test data point [True]
591 Text feature [observed] present in test data point [True]
595 Text feature [induced] present in test data point [True]
597 Text feature [type] present in test data point [True]
598 Text feature [vector] present in test data point [True]
600 Text feature [genetic] present in test data point [True]
603 Text feature [function] present in test data point [True]
610 Text feature [able] present in test data point [True]
613 Text feature [blot] present in test data point [True]
617 Text feature [frequently] present in test data point [True]
619 Text feature [40] present in test data point [True]
620 Text feature [wild] present in test data point [True]
623 Text feature [pten] present in test data point [True]
626 Text feature [important] present in test data point [True]
628 Text feature [several] present in test data point [True]
630 Text feature [figure] present in test data point [True]
633 Text feature [whether] present in test data point [True]
634 Text feature [times] present in test data point [True]
636 Text feature [cancers] present in test data point [True]
647 Text feature [moreover] present in test data point [True]
649 Text feature [relative] present in test data point [True]
650 Text feature [12] present in test data point [True]
651 Text feature [induction] present in test data point [True]
652 Text feature [proteins] present in test data point [True]
653 Text feature [mechanism] present in test data point [True]
654 Text feature [antibody] present in test data point [True]
657 Text feature [lead] present in test data point [True]
658 Text feature [localization] present in test data point [True]
661 Text feature [two] present in test data point [True]
663 Text feature [together] present in test data point [True]
665 Text feature [manufacturer] present in test data point [True]
666 Text feature [lymphoma] present in test data point [True]
668 Text feature [level] present in test data point [True]
672 Text feature [mrna] present in test data point [True]
673 Text feature [colon] present in test data point [True]
677 Text feature [positive] present in test data point [True]
680 Text feature [research] present in test data point [True]
687 Text feature [process] present in test data point [True]
690 Text feature [pathogenic] present in test data point [True]
692 Text feature [assays] present in test data point [True]
693 Text feature [various] present in test data point [True]
698 Text feature [led] present in test data point [True]
699 Text feature [detected] present in test data point [True]
700 Text feature [demonstrate] present in test data point [True]
701 Text feature [31] present in test data point [True]
704 Text feature [affected] present in test data point [True]
708 Text feature [seen] present in test data point [True]
709 Text feature [syndrome] present in test data point [True]
710 Text feature [deleterious] present in test data point [True]
712 Text feature [buffer] present in test data point [True]
715 Text feature [significant] present in test data point [True]
718 Text feature [consistent] present in test data point [True]
723 Text feature [page] present in test data point [True]
725 Text feature [types] present in test data point [True]
738 Text feature [analysis] present in test data point [True]
748 Text feature [whole] present in test data point [True]
750 Text feature [incubated] present in test data point [True]
751 Text feature [effects] present in test data point [True]
762 Text feature [control] present in test data point [True]
767 Text feature [enhanced] present in test data point [True]
772 Text feature [thus] present in test data point [True]
775 Text feature [green] present in test data point [True]
776 Text feature [mechanisms] present in test data point [True]
777 Text feature [90] present in test data point [True]
779 Text feature [anti] present in test data point [True]
781 Text feature [similarly] present in test data point [True]
786 Text feature [wt] present in test data point [True]
789 Text feature [effect] present in test data point [True]
792 Text feature [variants] present in test data point [True]
795 Text feature [cellular] present in test data point [True]
796 Text feature [27] present in test data point [True]
804 Text feature [according] present in test data point [True]
```

```
808 Text feature [cohort] present in test data point [True]
810 Text feature [confirmed] present in test data point [True]
812 Text feature [malignant] present in test data point [True]
814 Text feature [amino] present in test data point [True]
815 Text feature [group] present in test data point [True]
817 Text feature [shown] present in test data point [True]
818 Text feature [molecular] present in test data point [True]
825 Text feature [normal] present in test data point [True]
831 Text feature [used] present in test data point [True]
832 Text feature [patient] present in test data point [True]
837 Text feature [may] present in test data point [True]

838 Text feature [impact] present in test data point [True]
840 Text feature [subjected] present in test data point [True]
842 Text feature [test] present in test data point [True]
845 Text feature [hours] present in test data point [True]
847 Text feature [pathways] present in test data point [True]
848 Text feature [least] present in test data point [True]
849 Text feature [amplified] present in test data point [True]
850 Text feature [genomic] present in test data point [True]
851 Text feature [3a] present in test data point [True]
852 Text feature [fig] present in test data point [True]
854 Text feature [introduction] present in test data point [True]
855 Text feature [via] present in test data point [True]
860 Text feature [cannot] present in test data point [True]
861 Text feature [blood] present in test data point [True]
868 Text feature [28] present in test data point [True]
870 Text feature [atp] present in test data point [True]
871 Text feature [42] present in test data point [True]
872 Text feature [associated] present in test data point [True]
873 Text feature [results] present in test data point [True]
881 Text feature [whereas] present in test data point [True]
883 Text feature [23] present in test data point [True]
887 Text feature [domains] present in test data point [True]
889 Text feature [functions] present in test data point [True]
890 Text feature [cause] present in test data point [True]
891 Text feature [methods] present in test data point [True]
892 Text feature [distribution] present in test data point [True]
893 Text feature [decrease] present in test data point [True]
896 Text feature [long] present in test data point [True]
899 Text feature [major] present in test data point [True]
900 Text feature [sporadic] present in test data point [True]
904 Text feature [described] present in test data point [True]
905 Text feature [s1] present in test data point [True]
906 Text feature [breast] present in test data point [True]
907 Text feature [18] present in test data point [True]
911 Text feature [occur] present in test data point [True]
915 Text feature [acid] present in test data point [True]
924 Text feature [developed] present in test data point [True]
927 Text feature [similar] present in test data point [True]
935 Text feature [studied] present in test data point [True]
941 Text feature [60] present in test data point [True]
944 Text feature [alterations] present in test data point [True]
947 Text feature [part] present in test data point [True]
950 Text feature [analyzed] present in test data point [True]
951 Text feature [37] present in test data point [True]
954 Text feature [25] present in test data point [True]
964 Text feature [approximately] present in test data point [True]
966 Text feature [sensitivity] present in test data point [True]
981 Text feature [proliferation] present in test data point [True]
Out of the top  1000  features  222 are present in query point
```

4.3.2.4. Feature Importance, Inorrectly Classified point

In [253]:
```
test_point_index = 100
no_feature = 1000
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 :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],te
```

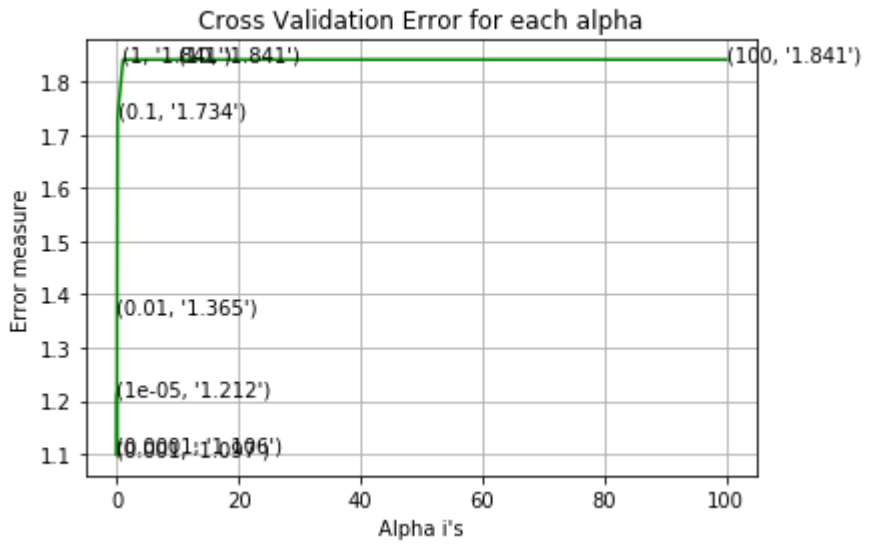
```
Predicted Class : 4
Predicted Class Probabilities: [[0.2121 0.0101 0.0088 0.7237 0.0131 0.0124 0.0169 0.0019 0.0009]]
Actual Class : 1
--------------------------------------------------
37 Text feature [suppressor] present in test data point [True]
111 Text feature [missense] present in test data point [True]
136 Text feature [liver] present in test data point [True]
143 Text feature [lanes] present in test data point [True]
177 Text feature [due] present in test data point [True]
204 Text feature [tumorigenesis] present in test data point [True]
230 Text feature [flag] present in test data point [True]
231 Text feature [dominant] present in test data point [True]
234 Text feature [smad4] present in test data point [True]
244 Text feature [reduced] present in test data point [True]
252 Text feature [families] present in test data point [True]
272 Text feature [western] present in test data point [True]
280 Text feature [functional] present in test data point [True]
284 Text feature [mammalian] present in test data point [True]
291 Text feature [resulting] present in test data point [True]
```

4.4. Linear Support Vector Machines

4.4.1. Hyper paramter tuning

In [255]:
```
# read more about support vector machines with linear kernals here http://scikit-learn.org/stable/modules/generat

# ---------------------------------
# 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-cop
# ---------------------------------



# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/generated/sklearn.calibra
# ----------------------------
# 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_onehotCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    print("Log Loss :",log_loss(cv_y, 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_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, predict_y, lab
predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_cv, predic
predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predict_y, label
```

```
for C = 1e-05
Log Loss : 1.211749346977332
for C = 0.0001
Log Loss : 1.1062660169438554
for C = 0.001
Log Loss : 1.0970708397941669
for C = 0.01
Log Loss : 1.364591779972023
for C = 0.1
Log Loss : 1.7336498713010107
for C = 1
Log Loss : 1.841109933247229
for C = 10
Log Loss : 1.8411098522912799
for C = 100
Log Loss : 1.841109877788849
```

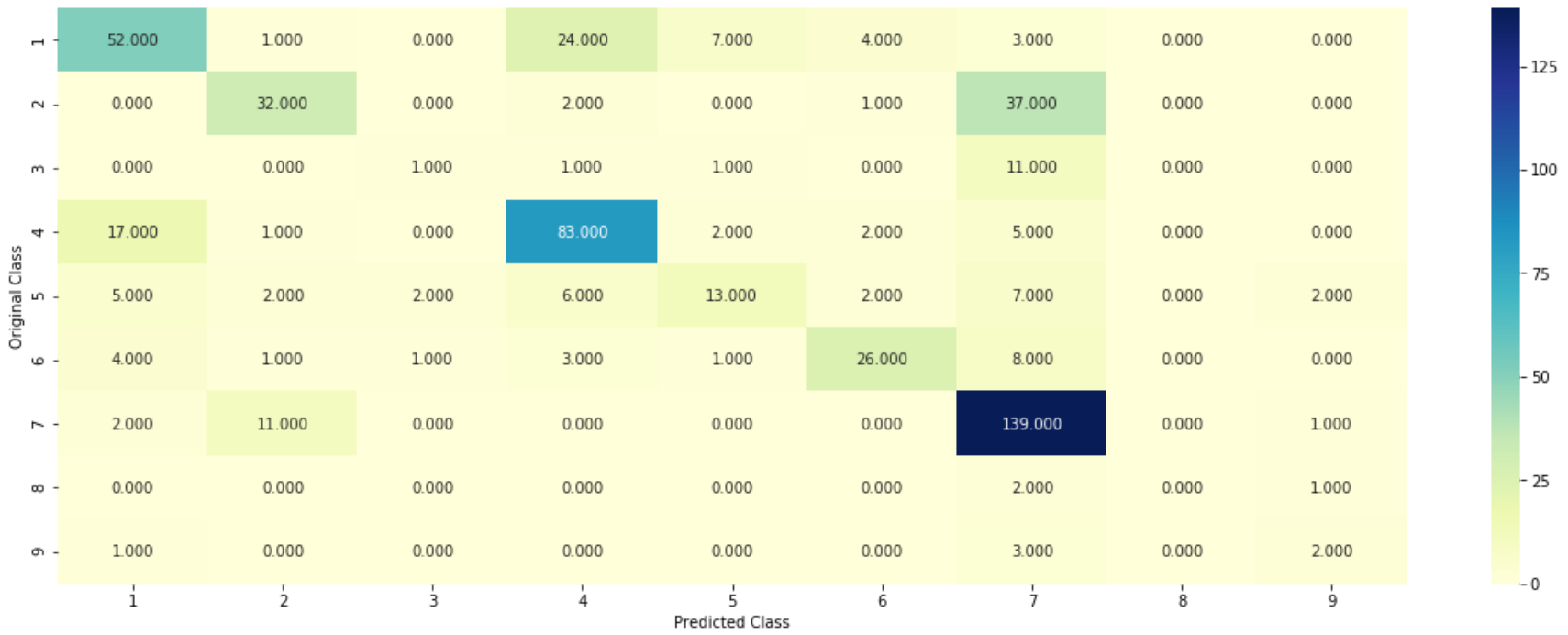



```
For values of best alpha =  0.001 The train log loss is: 0.5852589956860752
For values of best alpha =  0.001 The cross validation log loss is: 1.0970708397941669
For values of best alpha =  0.001 The test log loss is: 1.0043301962481346
```

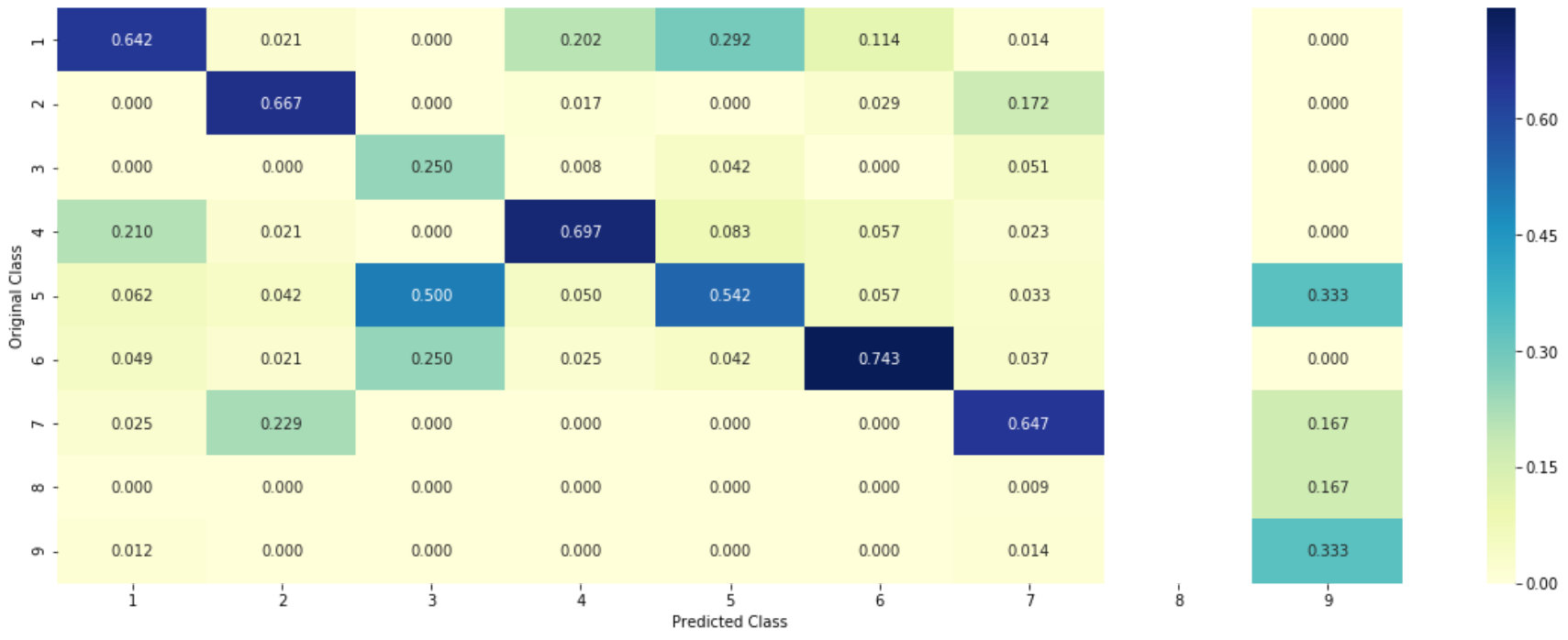
4.4.2. Testing model with best hyper parameters

In [256]:
```
# read more about support vector machines with linear kernals here http://scikit-learn.org/stable/modules/generat

# --------------------------------
# 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-co

# --------------------------------


# 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_onehotCoding, train_y,cv_x_onehotCoding,cv_y, clf)
```

```
Log loss : 1.0970708397941669
Number of mis-classified points : 0.3458646616541353
-------------------- 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 [257]:
```
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='hinge', random_state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 1
# test_point_index = 100
no_feature = 1000
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 :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],te
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.0384 0.0221 0.0122 0.766  0.0199 0.0149 0.1216 0.0023 0.0026]]
Actual Class : 4
--------------------------------------------------
19 Text feature [suppressor] present in test data point [True]
24 Text feature [missense] present in test data point [True]
27 Text feature [suggesting] present in test data point [True]
28 Text feature [right] present in test data point [True]
29 Text feature [direct] present in test data point [True]
31 Text feature [flag] present in test data point [True]
32 Text feature [lanes] present in test data point [True]
34 Text feature [due] present in test data point [True]
35 Text feature [proportion] present in test data point [True]
41 Text feature [show] present in test data point [True]
255 Text feature [germline] present in test data point [True]
256 Text feature [stably] present in test data point [True]
258 Text feature [comparison] present in test data point [True]
259 Text feature [mice] present in test data point [True]
262 Text feature [western] present in test data point [True]
```

4.3.3.2. For Incorrectly classified point

In [258]:
```
test_point_index = 100
no_feature =1000
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 :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],te
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.1645 0.0414 0.0245 0.6386 0.0259 0.0257 0.0719 0.0051 0.0024]]
Actual Class : 1
--------------------------------------------------
19 Text feature [suppressor] present in test data point [True]
24 Text feature [missense] present in test data point [True]
31 Text feature [flag] present in test data point [True]
32 Text feature [lanes] present in test data point [True]
34 Text feature [due] present in test data point [True]
37 Text feature [tumorigenesis] present in test data point [True]
255 Text feature [germline] present in test data point [True]
257 Text feature [1998] present in test data point [True]
260 Text feature [families] present in test data point [True]
261 Text feature [mammalian] present in test data point [True]
262 Text feature [western] present in test data point [True]
344 Text feature [transfected] present in test data point [True]
345 Text feature [smad4] present in test data point [True]
346 Text feature [reduced] present in test data point [True]
347 Text feature [consequences] present in test data point [True]
348 Text feature [dominant] present in test data point [True]
349 Text feature [fact] present in test data point [True]
350 Text feature [unable] present in test data point [True]
351 Text feature [tgf] present in test data point [True]
354 Text feature [functional] present in test data point [True]
363 Text feature [indicate] present in test data point [True]
365 Text feature [resulting] present in test data point [True]
368 Text feature [deletion] present in test data point [True]
375 Text feature [caused] present in test data point [True]
376 Text feature [alone] present in test data point [True]
381 Text feature [mutants] present in test data point [True]
417 Text feature [liver] present in test data point [True]
427 Text feature [determine] present in test data point [True]
428 Text feature [family] present in test data point [True]
429 Text feature [involved] present in test data point [True]
432 Text feature [activity] present in test data point [True]
434 Text feature [nature] present in test data point [True]
451 Text feature [negative] present in test data point [True]
454 Text feature [case] present in test data point [True]
455 Text feature [members] present in test data point [True]
457 Text feature [transfection] present in test data point [True]
461 Text feature [groups] present in test data point [True]
463 Text feature [analyses] present in test data point [True]
464 Text feature [ca] present in test data point [True]
465 Text feature [blue] present in test data point [True]
469 Text feature [able] present in test data point [True]
471 Text feature [various] present in test data point [True]
473 Text feature [tagged] present in test data point [True]
475 Text feature [relatively] present in test data point [True]
479 Text feature [p21] present in test data point [True]
480 Text feature [top] present in test data point [True]
481 Text feature [induced] present in test data point [True]
482 Text feature [association] present in test data point [True]
485 Text feature [staining] present in test data point [True]
486 Text feature [cancers] present in test data point [True]
532 Text feature [nucleus] present in test data point [True]
534 Text feature [large] present in test data point [True]
540 Text feature [colon] present in test data point [True]
545 Text feature [impaired] present in test data point [True]
546 Text feature [since] present in test data point [True]
547 Text feature [complex] present in test data point [True]
549 Text feature [observed] present in test data point [True]
550 Text feature [loss] present in test data point [True]
552 Text feature [wild] present in test data point [True]
554 Text feature [consistent] present in test data point [True]
556 Text feature [interact] present in test data point [True]
558 Text feature [early] present in test data point [True]
562 Text feature [gst] present in test data point [True]
564 Text feature [second] present in test data point [True]
565 Text feature [1999] present in test data point [True]
566 Text feature [dna] present in test data point [True]
567 Text feature [protein] present in test data point [True]
569 Text feature [nuclear] present in test data point [True]
571 Text feature [distribution] present in test data point [True]
572 Text feature [require] present in test data point [True]
575 Text feature [role] present in test data point [True]
576 Text feature [2000] present in test data point [True]
577 Text feature [first] present in test data point [True]
578 Text feature [1996] present in test data point [True]
579 Text feature [reaction] present in test data point [True]
```

```
581 Text feature [associated] present in test data point [True]
583 Text feature [whether] present in test data point [True]
587 Text feature [type] present in test data point [True]
588 Text feature [changes] present in test data point [True]
589 Text feature [decreased] present in test data point [True]
590 Text feature [sds] present in test data point [True]
591 Text feature [therefore] present in test data point [True]
592 Text feature [normal] present in test data point [True]
594 Text feature [figure] present in test data point [True]
596 Text feature [localization] present in test data point [True]
598 Text feature [vector] present in test data point [True]
601 Text feature [frequently] present in test data point [True]
602 Text feature [membrane] present in test data point [True]
605 Text feature [antibodies] present in test data point [True]
606 Text feature [thus] present in test data point [True]
611 Text feature [experiments] present in test data point [True]
612 Text feature [important] present in test data point [True]
613 Text feature [reporter] present in test data point [True]
619 Text feature [mouse] present in test data point [True]
620 Text feature [regions] present in test data point [True]
621 Text feature [proliferation] present in test data point [True]
622 Text feature [two] present in test data point [True]
624 Text feature [complete] present in test data point [True]
625 Text feature [substitution] present in test data point [True]
626 Text feature [lane] present in test data point [True]
627 Text feature [co] present in test data point [True]
638 Text feature [performed] present in test data point [True]
643 Text feature [enhanced] present in test data point [True]
644 Text feature [bind] present in test data point [True]
646 Text feature [significant] present in test data point [True]
648 Text feature [loop] present in test data point [True]
650 Text feature [confirmed] present in test data point [True]
653 Text feature [mutations] present in test data point [True]
657 Text feature [colorectal] present in test data point [True]
659 Text feature [mechanisms] present in test data point [True]
664 Text feature [analysis] present in test data point [True]
725 Text feature [note] present in test data point [True]
727 Text feature [used] present in test data point [True]
734 Text feature [absence] present in test data point [True]
737 Text feature [residue] present in test data point [True]
739 Text feature [proteins] present in test data point [True]
742 Text feature [assay] present in test data point [True]
747 Text feature [lead] present in test data point [True]
748 Text feature [affected] present in test data point [True]
749 Text feature [lower] present in test data point [True]
757 Text feature [reports] present in test data point [True]
758 Text feature [observation] present in test data point [True]
759 Text feature [control] present in test data point [True]
760 Text feature [regulation] present in test data point [True]
761 Text feature [properties] present in test data point [True]
762 Text feature [indicating] present in test data point [True]
763 Text feature [shown] present in test data point [True]
764 Text feature [cause] present in test data point [True]
767 Text feature [based] present in test data point [True]
770 Text feature [antibody] present in test data point [True]
773 Text feature [author] present in test data point [True]
775 Text feature [incubated] present in test data point [True]
776 Text feature [leukemia] present in test data point [True]
778 Text feature [examined] present in test data point [True]
781 Text feature [particular] present in test data point [True]
783 Text feature [leads] present in test data point [True]
787 Text feature [levels] present in test data point [True]
789 Text feature [neutral] present in test data point [True]
791 Text feature [mutation] present in test data point [True]
792 Text feature [whereas] present in test data point [True]
793 Text feature [48] present in test data point [True]
796 Text feature [including] present in test data point [True]
798 Text feature [biological] present in test data point [True]
800 Text feature [90] present in test data point [True]
802 Text feature [bound] present in test data point [True]
803 Text feature [several] present in test data point [True]
805 Text feature [group] present in test data point [True]
807 Text feature [anti] present in test data point [True]
810 Text feature [led] present in test data point [True]
816 Text feature [effect] present in test data point [True]
823 Text feature [least] present in test data point [True]
824 Text feature [inhibited] present in test data point [True]
828 Text feature [finding] present in test data point [True]
838 Text feature [detected] present in test data point [True]
839 Text feature [amino] present in test data point [True]
843 Text feature [inhibition] present in test data point [True]
844 Text feature [2a] present in test data point [True]
846 Text feature [unknown] present in test data point [True]
848 Text feature [carried] present in test data point [True]
852 Text feature [effects] present in test data point [True]
855 Text feature [molecular] present in test data point [True]
856 Text feature [3b] present in test data point [True]
859 Text feature [page] present in test data point [True]
860 Text feature [via] present in test data point [True]
862 Text feature [length] present in test data point [True]
```

```
863 Text feature [cells] present in test data point [True]
864 Text feature [generated] present in test data point [True]
868 Text feature [remains] present in test data point [True]
869 Text feature [secondary] present in test data point [True]
871 Text feature [response] present in test data point [True]
872 Text feature [however] present in test data point [True]
876 Text feature [number] present in test data point [True]
877 Text feature [acid] present in test data point [True]
878 Text feature [could] present in test data point [True]
879 Text feature [six] present in test data point [True]
887 Text feature [breast] present in test data point [True]
889 Text feature [results] present in test data point [True]
893 Text feature [image] present in test data point [True]
896 Text feature [showed] present in test data point [True]
897 Text feature [measured] present in test data point [True]
899 Text feature [60] present in test data point [True]
900 Text feature [well] present in test data point [True]
Out of the top  1000  features  182 are present in query point
```

4.5 Random Forest Classifier

4.5.1. Hyper paramter tuning (With One hot Encoding)

In [259]:
```
# ------------------------------
# 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_decrea
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None, verbose=0, warm_start=Fa
# 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-cor
# ------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/generated/sklearn.calibra
# ----------------------------
# 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)
        clf.fit(train_x_onehotCoding, train_y)
        sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
        sig_clf.fit(train_x_onehotCoding, train_y)
        sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
        cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
        print("Log Loss :",log_loss(cv_y, 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(bes
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best estimator = ', alpha[int(best_alpha/2)], "The train log loss is:",log_loss(y_train, pre
predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best estimator = ', alpha[int(best_alpha/2)], "The cross validation log loss is:",log_loss(y
predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best estimator = ', alpha[int(best_alpha/2)], "The test log loss is:",log_loss(y_test, predi
```

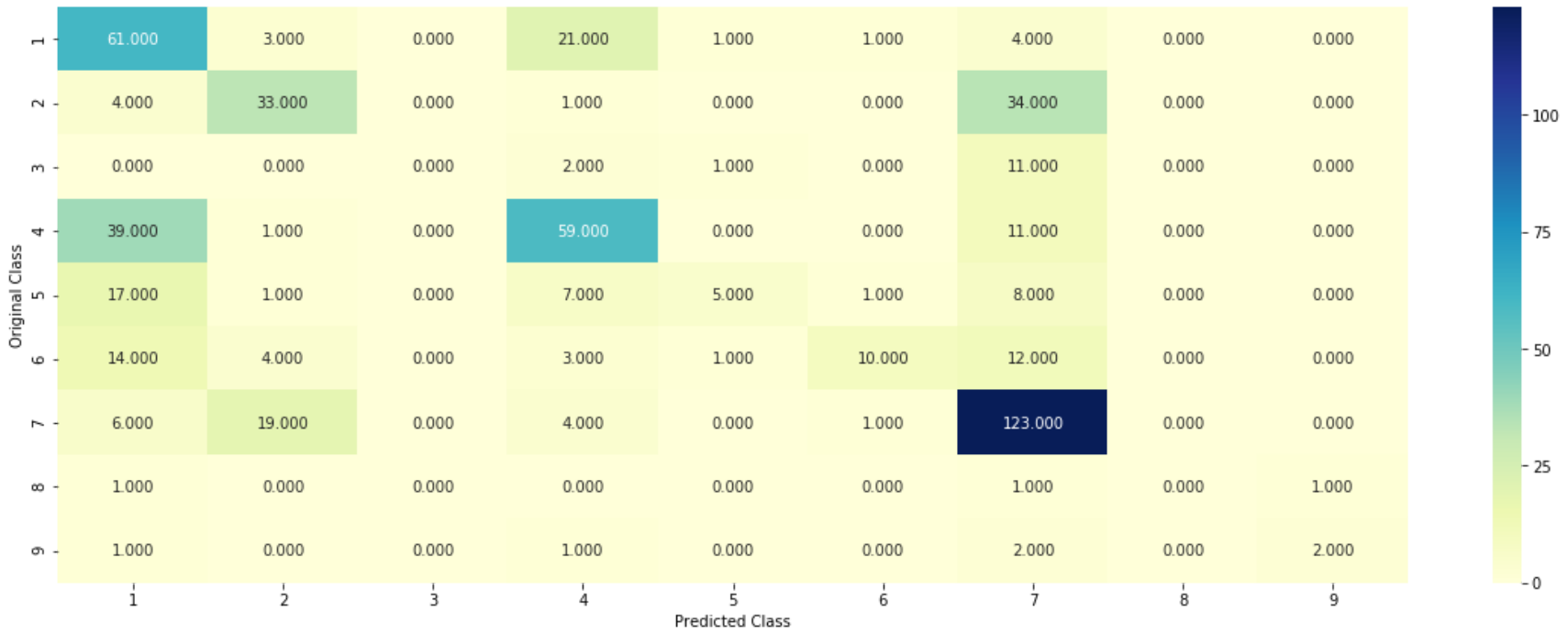
```
for n_estimators = 100 and max depth =  5
Log Loss : 1.2617533227550406
for n_estimators = 100 and max depth =  10
Log Loss : 1.2677898402470533
for n_estimators = 200 and max depth =  5
Log Loss : 1.2507056140571386
for n_estimators = 200 and max depth =  10
Log Loss : 1.2595669384468542
for n_estimators = 500 and max depth =  5
Log Loss : 1.2409268392873736
for n_estimators = 500 and max depth =  10
Log Loss : 1.2500210086674584
for n_estimators = 1000 and max depth =  5
Log Loss : 1.2347724500480501
for n_estimators = 1000 and max depth =  10
```

```
Log Loss : 1.2481507685432582
for n_estimators = 2000 and max depth =  5
Log Loss : 1.2338557657799185
for n_estimators = 2000 and max depth =  10
Log Loss : 1.2490606710097445
For values of best estimator =  2000 The train log loss is: 0.8779870034714718
For values of best estimator =  2000 The cross validation log loss is: 1.2338557657799185
For values of best estimator =  2000 The test log loss is: 1.1629955256895583
```

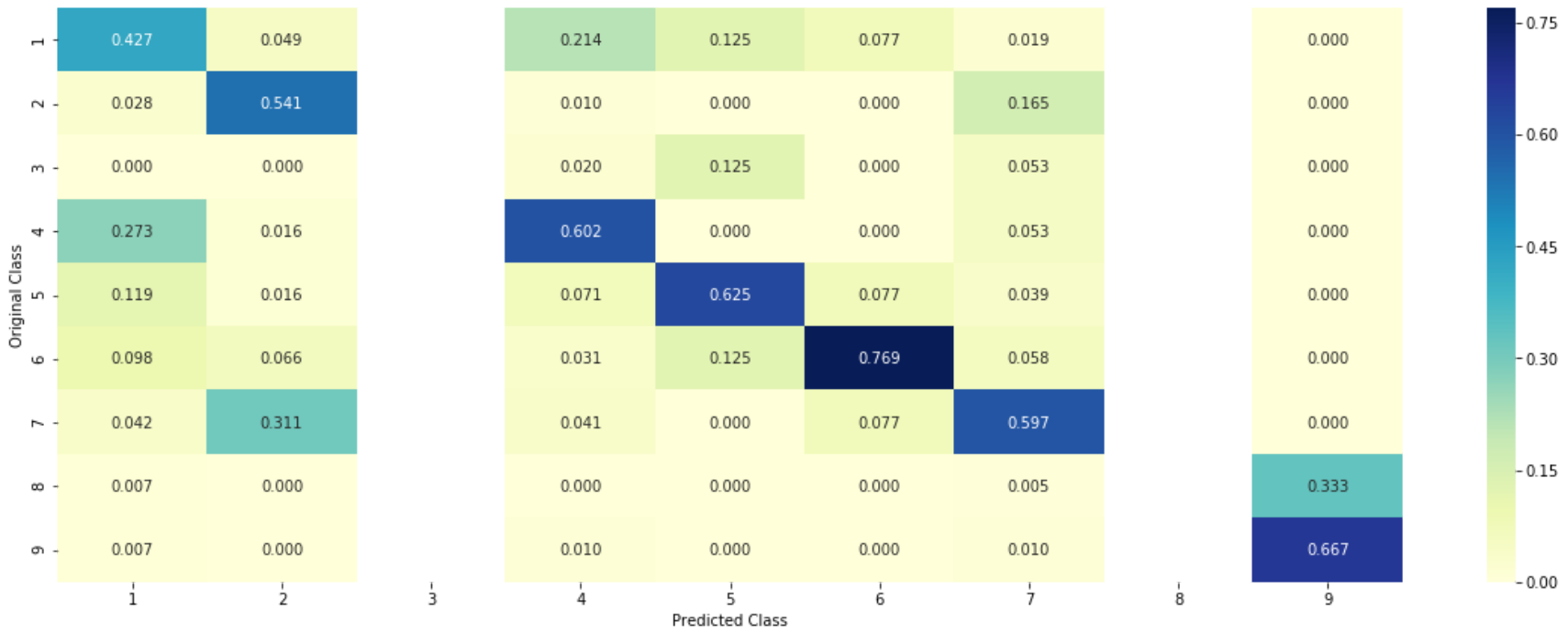
4.5.2. Testing model with best hyper parameters (One Hot Encoding)

In [260]:
```
# ------------------------------
# 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_decrea
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None, verbose=0, warm_start=Fa
# 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-cor
# ------------------------------

clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max_depth[int(bes
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y,cv_x_onehotCoding,cv_y, clf)
```

```
Log loss : 1.2338557657799185
Number of mis-classified points : 0.4492481203007519
-------------------- Confusion matrix --------------------
```



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



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

4.5.3. Feature Importance

4.5.3.1. Correctly Classified point

In [261]:
```
# test_point_index = 10
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max_depth[int(bes
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

test_point_index = 1
no_feature = 1000
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 :", test_y[test_point_index])
indices = np.argsort(-clf.feature_importances_)
print("-"*50)
get_impfeature_names(indices[:no_feature], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.1249 0.0717 0.0296 0.4736 0.0564 0.0493 0.1813 0.0066 0.0065]]
Actual Class : 4
--------------------------------------------------
0 Text feature [kinase] present in test data point [True]
3 Text feature [activation] present in test data point [True]
4 Text feature [suppressor] present in test data point [True]
5 Text feature [inhibitors] present in test data point [True]
6 Text feature [activated] present in test data point [True]
7 Text feature [function] present in test data point [True]
10 Text feature [inhibitor] present in test data point [True]
11 Text feature [missense] present in test data point [True]
13 Text feature [loss] present in test data point [True]
14 Text feature [treatment] present in test data point [True]
15 Text feature [erk] present in test data point [True]
19 Text feature [pathogenic] present in test data point [True]
25 Text feature [pten] present in test data point [True]
27 Text feature [deleterious] present in test data point [True]
28 Text feature [stability] present in test data point [True]
29 Text feature [protein] present in test data point [True]
31 Text feature [cell] present in test data point [True]
33 Text feature [activate] present in test data point [True]
34 Text feature [signaling] present in test data point [True]
35 Text feature [variants] present in test data point [True]
36 Text feature [functions] present in test data point [True]
41 Text feature [cells] present in test data point [True]
45 Text feature [functional] present in test data point [True]
46 Text feature [drug] present in test data point [True]
49 Text feature [inhibition] present in test data point [True]
50 Text feature [treated] present in test data point [True]
52 Text feature [patients] present in test data point [True]
54 Text feature [expression] present in test data point [True]
57 Text feature [clinical] present in test data point [True]
61 Text feature [inhibited] present in test data point [True]
62 Text feature [proteins] present in test data point [True]
64 Text feature [phosphatase] present in test data point [True]
70 Text feature [akt] present in test data point [True]
71 Text feature [resistance] present in test data point [True]
73 Text feature [expressing] present in test data point [True]
74 Text feature [mek] present in test data point [True]
75 Text feature [response] present in test data point [True]
79 Text feature [proliferation] present in test data point [True]
80 Text feature [database] present in test data point [True]
87 Text feature [harboring] present in test data point [True]
89 Text feature [lines] present in test data point [True]
91 Text feature [activity] present in test data point [True]
98 Text feature [nuclear] present in test data point [True]
100 Text feature [serum] present in test data point [True]
104 Text feature [dna] present in test data point [True]
105 Text feature [32] present in test data point [True]
106 Text feature [assays] present in test data point [True]
107 Text feature [tagged] present in test data point [True]
110 Text feature [potential] present in test data point [True]
112 Text feature [use] present in test data point [True]
113 Text feature [gene] present in test data point [True]
114 Text feature [sensitivity] present in test data point [True]
117 Text feature [research] present in test data point [True]
118 Text feature [pathway] present in test data point [True]
121 Text feature [proportion] present in test data point [True]
124 Text feature [21] present in test data point [True]
127 Text feature [tumors] present in test data point [True]
131 Text feature [defects] present in test data point [True]
132 Text feature [mapk] present in test data point [True]
134 Text feature [survival] present in test data point [True]
136 Text feature [presence] present in test data point [True]
137 Text feature [dose] present in test data point [True]
140 Text feature [sequencing] present in test data point [True]
141 Text feature [affect] present in test data point [True]
143 Text feature [effective] present in test data point [True]
145 Text feature [sequence] present in test data point [True]
146 Text feature [patient] present in test data point [True]
148 Text feature [one] present in test data point [True]
149 Text feature [33] present in test data point [True]
```

```
150 Text feature [wild] present in test data point [True]
151 Text feature [results] present in test data point [True]
152 Text feature [risk] present in test data point [True]
154 Text feature [factor] present in test data point [True]
156 Text feature [basis] present in test data point [True]
157 Text feature [31] present in test data point [True]
158 Text feature [terminal] present in test data point [True]
159 Text feature [known] present in test data point [True]
162 Text feature [assay] present in test data point [True]
168 Text feature [affected] present in test data point [True]
169 Text feature [ability] present in test data point [True]
171 Text feature [34] present in test data point [True]
173 Text feature [binding] present in test data point [True]
175 Text feature [multiple] present in test data point [True]
177 Text feature [likely] present in test data point [True]
180 Text feature [active] present in test data point [True]
181 Text feature [evidence] present in test data point [True]
182 Text feature [vector] present in test data point [True]
183 Text feature [studies] present in test data point [True]
184 Text feature [transfected] present in test data point [True]
189 Text feature [specific] present in test data point [True]
190 Text feature [previously] present in test data point [True]
192 Text feature [model] present in test data point [True]
194 Text feature [mutant] present in test data point [True]
195 Text feature [35] present in test data point [True]
198 Text feature [01] present in test data point [True]
199 Text feature [role] present in test data point [True]
200 Text feature [values] present in test data point [True]
203 Text feature [first] present in test data point [True]
204 Text feature [23] present in test data point [True]
205 Text feature [two] present in test data point [True]
207 Text feature [000] present in test data point [True]
210 Text feature [identified] present in test data point [True]
211 Text feature [effects] present in test data point [True]
216 Text feature [clinically] present in test data point [True]
217 Text feature [type] present in test data point [True]
224 Text feature [indicated] present in test data point [True]
226 Text feature [mutants] present in test data point [True]
227 Text feature [18] present in test data point [True]
228 Text feature [29] present in test data point [True]
229 Text feature [ubiquitin] present in test data point [True]
230 Text feature [26] present in test data point [True]
231 Text feature [enhanced] present in test data point [True]
232 Text feature [14] present in test data point [True]
233 Text feature [pi3k] present in test data point [True]
234 Text feature [reduced] present in test data point [True]
235 Text feature [domains] present in test data point [True]
236 Text feature [levels] present in test data point [True]
237 Text feature [germline] present in test data point [True]
238 Text feature [observed] present in test data point [True]
239 Text feature [gain] present in test data point [True]
240 Text feature [introduction] present in test data point [True]
241 Text feature [effect] present in test data point [True]
242 Text feature [figure] present in test data point [True]
246 Text feature [weeks] present in test data point [True]
247 Text feature [mutated] present in test data point [True]
249 Text feature [least] present in test data point [True]
250 Text feature [site] present in test data point [True]
251 Text feature [many] present in test data point [True]
255 Text feature [42] present in test data point [True]
259 Text feature [increased] present in test data point [True]
260 Text feature [individuals] present in test data point [True]
261 Text feature [anti] present in test data point [True]
262 Text feature [39] present in test data point [True]
263 Text feature [human] present in test data point [True]
264 Text feature [breast] present in test data point [True]
269 Text feature [important] present in test data point [True]
272 Text feature [suggesting] present in test data point [True]
274 Text feature [50] present in test data point [True]
276 Text feature [related] present in test data point [True]
281 Text feature [primary] present in test data point [True]
283 Text feature [mechanism] present in test data point [True]
284 Text feature [significant] present in test data point [True]
285 Text feature [15] present in test data point [True]
286 Text feature [36] present in test data point [True]
287 Text feature [transcription] present in test data point [True]
288 Text feature [12] present in test data point [True]

289 Text feature [mutations] present in test data point [True]
290 Text feature [molecular] present in test data point [True]
292 Text feature [11] present in test data point [True]
293 Text feature [possible] present in test data point [True]
295 Text feature [recently] present in test data point [True]
298 Text feature [17] present in test data point [True]
299 Text feature [culture] present in test data point [True]
302 Text feature [significantly] present in test data point [True]
306 Text feature [analysis] present in test data point [True]
308 Text feature [hypothesis] present in test data point [True]
313 Text feature [3a] present in test data point [True]
316 Text feature [60] present in test data point [True]
317 Text feature [developed] present in test data point [True]
```

```
321 Text feature [used] present in test data point [True]
322 Text feature [therefore] present in test data point [True]
324 Text feature [tissue] present in test data point [True]
326 Text feature [37] present in test data point [True]
330 Text feature [expressed] present in test data point [True]
332 Text feature [cancers] present in test data point [True]
333 Text feature [40] present in test data point [True]
334 Text feature [studied] present in test data point [True]
335 Text feature [although] present in test data point [True]
337 Text feature [study] present in test data point [True]
338 Text feature [table] present in test data point [True]
341 Text feature [methods] present in test data point [True]
342 Text feature [control] present in test data point [True]
346 Text feature [fraction] present in test data point [True]
347 Text feature [38] present in test data point [True]
348 Text feature [may] present in test data point [True]
349 Text feature [13] present in test data point [True]
351 Text feature [another] present in test data point [True]
352 Text feature [total] present in test data point [True]
353 Text feature [1a] present in test data point [True]
355 Text feature [identify] present in test data point [True]
360 Text feature [induced] present in test data point [True]
363 Text feature [statistical] present in test data point [True]
364 Text feature [several] present in test data point [True]
365 Text feature [22] present in test data point [True]
367 Text feature [lysates] present in test data point [True]
368 Text feature [performed] present in test data point [True]
369 Text feature [cultured] present in test data point [True]
370 Text feature [incubated] present in test data point [True]
374 Text feature [suggest] present in test data point [True]
375 Text feature [novel] present in test data point [True]
376 Text feature [line] present in test data point [True]
377 Text feature [positive] present in test data point [True]
379 Text feature [demonstrated] present in test data point [True]
382 Text feature [lead] present in test data point [True]
384 Text feature [data] present in test data point [True]
385 Text feature [transfection] present in test data point [True]
387 Text feature [impaired] present in test data point [True]
388 Text feature [controls] present in test data point [True]
390 Text feature [essential] present in test data point [True]
391 Text feature [mouse] present in test data point [True]
392 Text feature [also] present in test data point [True]
393 Text feature [mediated] present in test data point [True]
396 Text feature [level] present in test data point [True]
401 Text feature [detected] present in test data point [True]
405 Text feature [reported] present in test data point [True]
408 Text feature [common] present in test data point [True]
412 Text feature [different] present in test data point [True]
413 Text feature [due] present in test data point [True]
414 Text feature [times] present in test data point [True]
415 Text feature [30] present in test data point [True]
418 Text feature [showed] present in test data point [True]
421 Text feature [hours] present in test data point [True]
422 Text feature [mutagenesis] present in test data point [True]
423 Text feature [mg] present in test data point [True]
424 Text feature [time] present in test data point [True]
426 Text feature [discussion] present in test data point [True]
429 Text feature [al] present in test data point [True]
430 Text feature [tested] present in test data point [True]
431 Text feature [pathways] present in test data point [True]
432 Text feature [acid] present in test data point [True]
434 Text feature [ml] present in test data point [True]
436 Text feature [plasmid] present in test data point [True]
437 Text feature [leading] present in test data point [True]
438 Text feature [occur] present in test data point [True]
439 Text feature [described] present in test data point [True]
441 Text feature [substrate] present in test data point [True]
444 Text feature [dominant] present in test data point [True]
445 Text feature [associated] present in test data point [True]
448 Text feature [page] present in test data point [True]
450 Text feature [among] present in test data point [True]
453 Text feature [obtained] present in test data point [True]
454 Text feature [vitro] present in test data point [True]
456 Text feature [exon] present in test data point [True]
457 Text feature [might] present in test data point [True]
458 Text feature [genetic] present in test data point [True]
460 Text feature [atp] present in test data point [True]
462 Text feature [25] present in test data point [True]
463 Text feature [following] present in test data point [True]
464 Text feature [2a] present in test data point [True]
467 Text feature [24] present in test data point [True]
468 Text feature [highly] present in test data point [True]
471 Text feature [respectively] present in test data point [True]
472 Text feature [present] present in test data point [True]
473 Text feature [contrast] present in test data point [True]
474 Text feature [like] present in test data point [True]
475 Text feature [plates] present in test data point [True]
477 Text feature [domain] present in test data point [True]
478 Text feature [10] present in test data point [True]
479 Text feature [antibody] present in test data point [True]
481 Text feature [well] present in test data point [True]
```

```
482 Text feature [distribution] present in test data point [True]
483 Text feature [group] present in test data point [True]
484 Text feature [somatic] present in test data point [True]
486 Text feature [change] present in test data point [True]
488 Text feature [measured] present in test data point [True]
489 Text feature [localization] present in test data point [True]
490 Text feature [according] present in test data point [True]
493 Text feature [using] present in test data point [True]
494 Text feature [whether] present in test data point [True]
497 Text feature [100] present in test data point [True]
499 Text feature [27] present in test data point [True]
500 Text feature [via] present in test data point [True]
501 Text feature [state] present in test data point [True]
502 Text feature [subsequent] present in test data point [True]
503 Text feature [relative] present in test data point [True]
505 Text feature [regulation] present in test data point [True]
508 Text feature [materials] present in test data point [True]
511 Text feature [28] present in test data point [True]
512 Text feature [dependent] present in test data point [True]
515 Text feature [provide] present in test data point [True]
516 Text feature [low] present in test data point [True]
517 Text feature [detection] present in test data point [True]
518 Text feature [confirmed] present in test data point [True]
519 Text feature [found] present in test data point [True]
521 Text feature [revealed] present in test data point [True]
523 Text feature [types] present in test data point [True]
526 Text feature [4a] present in test data point [True]
528 Text feature [determine] present in test data point [True]
533 Text feature [high] present in test data point [True]
535 Text feature [process] present in test data point [True]
537 Text feature [six] present in test data point [True]
538 Text feature [even] present in test data point [True]
539 Text feature [shown] present in test data point [True]
541 Text feature [vivo] present in test data point [True]
543 Text feature [mutation] present in test data point [True]
545 Text feature [amino] present in test data point [True]
546 Text feature [signal] present in test data point [True]
547 Text feature [considered] present in test data point [True]
549 Text feature [mice] present in test data point [True]
550 Text feature [next] present in test data point [True]
552 Text feature [nucleus] present in test data point [True]
553 Text feature [fold] present in test data point [True]
554 Text feature [samples] present in test data point [True]
556 Text feature [wt] present in test data point [True]
558 Text feature [washed] present in test data point [True]
561 Text feature [compared] present in test data point [True]
562 Text feature [cause] present in test data point [True]
565 Text feature [phenotype] present in test data point [True]
566 Text feature [western] present in test data point [True]
568 Text feature [fusion] present in test data point [True]

571 Text feature [together] present in test data point [True]
572 Text feature [major] present in test data point [True]
574 Text feature [carrying] present in test data point [True]
578 Text feature [mechanisms] present in test data point [True]
579 Text feature [per] present in test data point [True]
581 Text feature [cannot] present in test data point [True]
583 Text feature [system] present in test data point [True]
586 Text feature [complete] present in test data point [True]
587 Text feature [cellular] present in test data point [True]
588 Text feature [degradation] present in test data point [True]
590 Text feature [mm] present in test data point [True]
591 Text feature [either] present in test data point [True]
595 Text feature [genomic] present in test data point [True]
597 Text feature [criteria] present in test data point [True]
602 Text feature [syndrome] present in test data point [True]
605 Text feature [long] present in test data point [True]
607 Text feature [involving] present in test data point [True]
610 Text feature [mrna] present in test data point [True]
611 Text feature [consistent] present in test data point [True]
612 Text feature [investigated] present in test data point [True]
613 Text feature [green] present in test data point [True]
614 Text feature [antibodies] present in test data point [True]
617 Text feature [tumor] present in test data point [True]
619 Text feature [interestingly] present in test data point [True]
621 Text feature [double] present in test data point [True]
623 Text feature [distinct] present in test data point [True]
624 Text feature [directed] present in test data point [True]
626 Text feature [presented] present in test data point [True]
630 Text feature [flag] present in test data point [True]
632 Text feature [negative] present in test data point [True]
633 Text feature [directly] present in test data point [True]
634 Text feature [analyzed] present in test data point [True]
635 Text feature [frequently] present in test data point [True]
636 Text feature [concentration] present in test data point [True]
637 Text feature [located] present in test data point [True]
638 Text feature [panel] present in test data point [True]
645 Text feature [90] present in test data point [True]
647 Text feature [20] present in test data point [True]
649 Text feature [alternative] present in test data point [True]
651 Text feature [cancer] present in test data point [True]
```

```
652 Text feature [thus] present in test data point [True]
657 Text feature [range] present in test data point [True]
660 Text feature [direct] present in test data point [True]
661 Text feature [alterations] present in test data point [True]
662 Text feature [four] present in test data point [True]
665 Text feature [observations] present in test data point [True]
666 Text feature [et] present in test data point [True]
668 Text feature [pcr] present in test data point [True]
672 Text feature [similar] present in test data point [True]
673 Text feature [increase] present in test data point [True]
677 Text feature [malignant] present in test data point [True]
681 Text feature [buffer] present in test data point [True]
682 Text feature [elevated] present in test data point [True]
683 Text feature [shows] present in test data point [True]
687 Text feature [result] present in test data point [True]
694 Text feature [differentiation] present in test data point [True]
696 Text feature [s2] present in test data point [True]
698 Text feature [assess] present in test data point [True]
700 Text feature [three] present in test data point [True]
701 Text feature [end] present in test data point [True]
702 Text feature [stably] present in test data point [True]
704 Text feature [amplified] present in test data point [True]
705 Text feature [whereas] present in test data point [True]
709 Text feature [normal] present in test data point [True]
711 Text feature [cohort] present in test data point [True]
713 Text feature [taken] present in test data point [True]
716 Text feature [generated] present in test data point [True]
717 Text feature [isolated] present in test data point [True]
719 Text feature [similarly] present in test data point [True]
720 Text feature [addition] present in test data point [True]
729 Text feature [specimens] present in test data point [True]
730 Text feature [fig] present in test data point [True]
732 Text feature [terminus] present in test data point [True]
736 Text feature [test] present in test data point [True]
738 Text feature [contains] present in test data point [True]
742 Text feature [go] present in test data point [True]
744 Text feature [relatively] present in test data point [True]
745 Text feature [seen] present in test data point [True]
748 Text feature [exons] present in test data point [True]
749 Text feature [targeting] present in test data point [True]
751 Text feature [would] present in test data point [True]
752 Text feature [various] present in test data point [True]
757 Text feature [resulted] present in test data point [True]
760 Text feature [derived] present in test data point [True]
761 Text feature [since] present in test data point [True]
763 Text feature [targeted] present in test data point [True]
767 Text feature [show] present in test data point [True]
768 Text feature [subset] present in test data point [True]
770 Text feature [applied] present in test data point [True]
774 Text feature [subunit] present in test data point [True]
776 Text feature [impact] present in test data point [True]
781 Text feature [sds] present in test data point [True]
782 Text feature [followed] present in test data point [True]
784 Text feature [www] present in test data point [True]
788 Text feature [factors] present in test data point [True]
790 Text feature [part] present in test data point [True]
791 Text feature [contribute] present in test data point [True]
798 Text feature [05] present in test data point [True]
799 Text feature [primers] present in test data point [True]
805 Text feature [still] present in test data point [True]
808 Text feature [min] present in test data point [True]
809 Text feature [s1] present in test data point [True]
811 Text feature [able] present in test data point [True]
812 Text feature [resulting] present in test data point [True]
814 Text feature [reverse] present in test data point [True]
816 Text feature [approximately] present in test data point [True]
817 Text feature [analyses] present in test data point [True]
822 Text feature [comparison] present in test data point [True]
823 Text feature [supplementary] present in test data point [True]
824 Text feature [appears] present in test data point [True]
827 Text feature [recent] present in test data point [True]
828 Text feature [tissues] present in test data point [True]
830 Text feature [correlation] present in test data point [True]
832 Text feature [activities] present in test data point [True]
833 Text feature [sporadic] present in test data point [True]
834 Text feature [manufacturer] present in test data point [True]
838 Text feature [blot] present in test data point [True]
845 Text feature [moreover] present in test data point [True]
847 Text feature [age] present in test data point [True]
848 Text feature [overexpression] present in test data point [True]
850 Text feature [express] present in test data point [True]
852 Text feature [indeed] present in test data point [True]
853 Text feature [carcinoma] present in test data point [True]
857 Text feature [regulated] present in test data point [True]
858 Text feature [assessed] present in test data point [True]
859 Text feature [led] present in test data point [True]
864 Text feature [heterozygous] present in test data point [True]
865 Text feature [ca] present in test data point [True]
866 Text feature [decreased] present in test data point [True]
873 Text feature [years] present in test data point [True]
882 Text feature [see] present in test data point [True]
```

```
886 Text feature [correlated] present in test data point [True]
887 Text feature [combination] present in test data point [True]
890 Text feature [manner] present in test data point [True]
892 Text feature [right] present in test data point [True]
893 Text feature [lanes] present in test data point [True]
894 Text feature [demonstrate] present in test data point [True]
895 Text feature [chain] present in test data point [True]
901 Text feature [blood] present in test data point [True]
902 Text feature [play] present in test data point [True]
903 Text feature [series] present in test data point [True]
909 Text feature [invitrogen] present in test data point [True]
911 Text feature [reaction] present in test data point [True]
913 Text feature [secondary] present in test data point [True]
916 Text feature [subjected] present in test data point [True]
918 Text feature [stable] present in test data point [True]
921 Text feature [induction] present in test data point [True]
923 Text feature [nm] present in test data point [True]
930 Text feature [endometrial] present in test data point [True]
935 Text feature [colon] present in test data point [True]
946 Text feature [induce] present in test data point [True]
947 Text feature [left] present in test data point [True]
948 Text feature [decrease] present in test data point [True]

962 Text feature [whole] present in test data point [True]
999 Text feature [lymphoma] present in test data point [True]
Out of the top  1000  features  455 are present in query point
```

4.5.3.2. Inorrectly Classified point

In [262]:
```
test_point_index = 100
no_feature = 1000
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("Actuall Class :", test_y[test_point_index])
indices = np.argsort(-clf.feature_importances_)
print("-"*50)
get_impfeature_names(indices[:no_feature], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.2658 0.0453 0.0194 0.4364 0.0531 0.0489 0.1171 0.0073 0.0067]]
Actuall Class : 1
--------------------------------------------------
0 Text feature [kinase] present in test data point [True]
3 Text feature [activation] present in test data point [True]
4 Text feature [suppressor] present in test data point [True]
5 Text feature [inhibitors] present in test data point [True]
6 Text feature [activated] present in test data point [True]
7 Text feature [function] present in test data point [True]
11 Text feature [missense] present in test data point [True]
13 Text feature [loss] present in test data point [True]
14 Text feature [treatment] present in test data point [True]
16 Text feature [transforming] present in test data point [True]
21 Text feature [growth] present in test data point [True]
22 Text feature [receptor] present in test data point [True]
29 Text feature [protein] present in test data point [True]
30 Text feature [constitutively] present in test data point [True]
31 Text feature [cell] present in test data point [True]
32 Text feature [neutral] present in test data point [True]
33 Text feature [activate] present in test data point [True]
34 Text feature [signaling] present in test data point [True]
38 Text feature [extracellular] present in test data point [True]
41 Text feature [cells] present in test data point [True]
45 Text feature [functional] present in test data point [True]
49 Text feature [inhibition] present in test data point [True]
50 Text feature [treated] present in test data point [True]
54 Text feature [expression] present in test data point [True]
57 Text feature [clinical] present in test data point [True]
61 Text feature [inhibited] present in test data point [True]
62 Text feature [proteins] present in test data point [True]
67 Text feature [ligand] present in test data point [True]
73 Text feature [expressing] present in test data point [True]
75 Text feature [response] present in test data point [True]
79 Text feature [proliferation] present in test data point [True]
88 Text feature [conserved] present in test data point [True]
91 Text feature [activity] present in test data point [True]
95 Text feature [phosphorylated] present in test data point [True]
98 Text feature [nuclear] present in test data point [True]
103 Text feature [downstream] present in test data point [True]
104 Text feature [dna] present in test data point [True]
107 Text feature [tagged] present in test data point [True]
111 Text feature [core] present in test data point [True]
113 Text feature [gene] present in test data point [True]
118 Text feature [pathway] present in test data point [True]
127 Text feature [tumors] present in test data point [True]
129 Text feature [stimulation] present in test data point [True]
135 Text feature [mammalian] present in test data point [True]
136 Text feature [presence] present in test data point [True]
137 Text feature [dose] present in test data point [True]
139 Text feature [family] present in test data point [True]
141 Text feature [affect] present in test data point [True]
144 Text feature [leukemia] present in test data point [True]
145 Text feature [sequence] present in test data point [True]
148 Text feature [one] present in test data point [True]
150 Text feature [wild] present in test data point [True]
151 Text feature [results] present in test data point [True]
153 Text feature [membrane] present in test data point [True]
154 Text feature [factor] present in test data point [True]
155 Text feature [lung] present in test data point [True]
156 Text feature [basis] present in test data point [True]
158 Text feature [terminal] present in test data point [True]
162 Text feature [assay] present in test data point [True]
165 Text feature [based] present in test data point [True]
166 Text feature [interaction] present in test data point [True]
167 Text feature [families] present in test data point [True]
168 Text feature [affected] present in test data point [True]
169 Text feature [ability] present in test data point [True]
173 Text feature [binding] present in test data point [True]
174 Text feature [structure] present in test data point [True]
180 Text feature [active] present in test data point [True]
182 Text feature [vector] present in test data point [True]
183 Text feature [studies] present in test data point [True]
184 Text feature [transfected] present in test data point [True]
187 Text feature [receptors] present in test data point [True]
189 Text feature [specific] present in test data point [True]
191 Text feature [interact] present in test data point [True]
194 Text feature [mutant] present in test data point [True]
199 Text feature [role] present in test data point [True]
```

```
201 Text feature [large] present in test data point [True]
203 Text feature [first] present in test data point [True]
205 Text feature [two] present in test data point [True]
208 Text feature [structural] present in test data point [True]
209 Text feature [transcriptional] present in test data point [True]
210 Text feature [identified] present in test data point [True]
211 Text feature [effects] present in test data point [True]
212 Text feature [experiments] present in test data point [True]
214 Text feature [genes] present in test data point [True]
217 Text feature [type] present in test data point [True]
218 Text feature [changes] present in test data point [True]
223 Text feature [independent] present in test data point [True]
224 Text feature [indicated] present in test data point [True]
226 Text feature [mutants] present in test data point [True]
231 Text feature [enhanced] present in test data point [True]
234 Text feature [reduced] present in test data point [True]
235 Text feature [domains] present in test data point [True]
236 Text feature [levels] present in test data point [True]
237 Text feature [germline] present in test data point [True]
238 Text feature [observed] present in test data point [True]
241 Text feature [effect] present in test data point [True]
242 Text feature [figure] present in test data point [True]
245 Text feature [co] present in test data point [True]
247 Text feature [mutated] present in test data point [True]
248 Text feature [interactions] present in test data point [True]
249 Text feature [least] present in test data point [True]
251 Text feature [many] present in test data point [True]
256 Text feature [bind] present in test data point [True]
259 Text feature [increased] present in test data point [True]
261 Text feature [anti] present in test data point [True]
263 Text feature [human] present in test data point [True]
264 Text feature [breast] present in test data point [True]
269 Text feature [important] present in test data point [True]
270 Text feature [reporter] present in test data point [True]
274 Text feature [50] present in test data point [True]
276 Text feature [related] present in test data point [True]
277 Text feature [target] present in test data point [True]
278 Text feature [inhibit] present in test data point [True]
279 Text feature [length] present in test data point [True]
280 Text feature [number] present in test data point [True]
281 Text feature [primary] present in test data point [True]
284 Text feature [significant] present in test data point [True]
287 Text feature [transcription] present in test data point [True]
289 Text feature [mutations] present in test data point [True]
290 Text feature [molecular] present in test data point [True]
292 Text feature [11] present in test data point [True]
294 Text feature [acids] present in test data point [True]
295 Text feature [recently] present in test data point [True]
300 Text feature [single] present in test data point [True]
302 Text feature [significantly] present in test data point [True]
303 Text feature [however] present in test data point [True]
306 Text feature [analysis] present in test data point [True]
307 Text feature [residues] present in test data point [True]
308 Text feature [hypothesis] present in test data point [True]
310 Text feature [including] present in test data point [True]
313 Text feature [3a] present in test data point [True]
314 Text feature [3b] present in test data point [True]
315 Text feature [metastatic] present in test data point [True]
316 Text feature [60] present in test data point [True]
319 Text feature [promoter] present in test data point [True]
320 Text feature [deletion] present in test data point [True]
321 Text feature [used] present in test data point [True]
322 Text feature [therefore] present in test data point [True]
323 Text feature [within] present in test data point [True]
325 Text feature [require] present in test data point [True]
329 Text feature [hydrophobic] present in test data point [True]
330 Text feature [expressed] present in test data point [True]
331 Text feature [involved] present in test data point [True]
332 Text feature [cancers] present in test data point [True]
335 Text feature [although] present in test data point [True]
337 Text feature [study] present in test data point [True]
342 Text feature [control] present in test data point [True]
343 Text feature [absence] present in test data point [True]
348 Text feature [may] present in test data point [True]
351 Text feature [another] present in test data point [True]
354 Text feature [region] present in test data point [True]
355 Text feature [identify] present in test data point [True]
358 Text feature [locus] present in test data point [True]
360 Text feature [induced] present in test data point [True]
361 Text feature [set] present in test data point [True]
364 Text feature [several] present in test data point [True]
367 Text feature [lysates] present in test data point [True]
368 Text feature [performed] present in test data point [True]
370 Text feature [incubated] present in test data point [True]
371 Text feature [containing] present in test data point [True]
376 Text feature [line] present in test data point [True]
379 Text feature [demonstrated] present in test data point [True]
381 Text feature [2b] present in test data point [True]
382 Text feature [lead] present in test data point [True]
384 Text feature [data] present in test data point [True]
385 Text feature [transfection] present in test data point [True]
```

```
387 Text feature [impaired] present in test data point [True]
389 Text feature [could] present in test data point [True]
391 Text feature [mouse] present in test data point [True]
392 Text feature [also] present in test data point [True]
393 Text feature [mediated] present in test data point [True]
394 Text feature [helix] present in test data point [True]
396 Text feature [level] present in test data point [True]
398 Text feature [indicate] present in test data point [True]
401 Text feature [detected] present in test data point [True]
402 Text feature [confer] present in test data point [True]
404 Text feature [determined] present in test data point [True]
405 Text feature [reported] present in test data point [True]
406 Text feature [substitution] present in test data point [True]
408 Text feature [common] present in test data point [True]
409 Text feature [full] present in test data point [True]

412 Text feature [different] present in test data point [True]
413 Text feature [due] present in test data point [True]
417 Text feature [stimulated] present in test data point [True]
418 Text feature [showed] present in test data point [True]
419 Text feature [complex] present in test data point [True]
420 Text feature [members] present in test data point [True]
422 Text feature [mutagenesis] present in test data point [True]
426 Text feature [discussion] present in test data point [True]
429 Text feature [al] present in test data point [True]
432 Text feature [acid] present in test data point [True]
433 Text feature [form] present in test data point [True]
438 Text feature [occur] present in test data point [True]
443 Text feature [basal] present in test data point [True]
444 Text feature [dominant] present in test data point [True]
445 Text feature [associated] present in test data point [True]
447 Text feature [findings] present in test data point [True]
448 Text feature [page] present in test data point [True]
450 Text feature [among] present in test data point [True]
452 Text feature [allele] present in test data point [True]
454 Text feature [vitro] present in test data point [True]
457 Text feature [might] present in test data point [True]
459 Text feature [leads] present in test data point [True]
464 Text feature [2a] present in test data point [True]
465 Text feature [development] present in test data point [True]
468 Text feature [highly] present in test data point [True]
471 Text feature [respectively] present in test data point [True]
477 Text feature [domain] present in test data point [True]
479 Text feature [antibody] present in test data point [True]
480 Text feature [side] present in test data point [True]
481 Text feature [well] present in test data point [True]
482 Text feature [distribution] present in test data point [True]
483 Text feature [group] present in test data point [True]
486 Text feature [change] present in test data point [True]
488 Text feature [measured] present in test data point [True]
489 Text feature [localization] present in test data point [True]
493 Text feature [using] present in test data point [True]
494 Text feature [whether] present in test data point [True]
496 Text feature [lower] present in test data point [True]
500 Text feature [via] present in test data point [True]
501 Text feature [state] present in test data point [True]
505 Text feature [regulation] present in test data point [True]
506 Text feature [alk] present in test data point [True]
510 Text feature [formation] present in test data point [True]
512 Text feature [dependent] present in test data point [True]
513 Text feature [required] present in test data point [True]
515 Text feature [provide] present in test data point [True]
518 Text feature [confirmed] present in test data point [True]
519 Text feature [found] present in test data point [True]
522 Text feature [properties] present in test data point [True]
524 Text feature [indicating] present in test data point [True]
525 Text feature [signals] present in test data point [True]
528 Text feature [determine] present in test data point [True]
532 Text feature [remains] present in test data point [True]
536 Text feature [conformation] present in test data point [True]
537 Text feature [six] present in test data point [True]
539 Text feature [shown] present in test data point [True]
540 Text feature [binds] present in test data point [True]
541 Text feature [vivo] present in test data point [True]
542 Text feature [gel] present in test data point [True]
543 Text feature [mutation] present in test data point [True]
545 Text feature [amino] present in test data point [True]
546 Text feature [signal] present in test data point [True]
550 Text feature [next] present in test data point [True]
551 Text feature [regions] present in test data point [True]
552 Text feature [nucleus] present in test data point [True]
553 Text feature [fold] present in test data point [True]
554 Text feature [samples] present in test data point [True]
560 Text feature [apoptosis] present in test data point [True]
561 Text feature [compared] present in test data point [True]
562 Text feature [cause] present in test data point [True]
566 Text feature [western] present in test data point [True]
568 Text feature [fusion] present in test data point [True]
570 Text feature [fact] present in test data point [True]
572 Text feature [major] present in test data point [True]
576 Text feature [furthermore] present in test data point [True]
```

```
578 Text feature [mechanisms] present in test data point [True]
585 Text feature [occurs] present in test data point [True]
586 Text feature [complete] present in test data point [True]
591 Text feature [either] present in test data point [True]
593 Text feature [groups] present in test data point [True]
596 Text feature [critical] present in test data point [True]
599 Text feature [case] present in test data point [True]
605 Text feature [long] present in test data point [True]
608 Text feature [residue] present in test data point [True]
611 Text feature [consistent] present in test data point [True]
614 Text feature [antibodies] present in test data point [True]
615 Text feature [without] present in test data point [True]
617 Text feature [tumor] present in test data point [True]
620 Text feature [complexes] present in test data point [True]
627 Text feature [constructs] present in test data point [True]
628 Text feature [five] present in test data point [True]
629 Text feature [blue] present in test data point [True]
630 Text feature [flag] present in test data point [True]
631 Text feature [note] present in test data point [True]
632 Text feature [negative] present in test data point [True]
633 Text feature [directly] present in test data point [True]
635 Text feature [frequently] present in test data point [True]
637 Text feature [located] present in test data point [True]
638 Text feature [panel] present in test data point [True]
641 Text feature [bound] present in test data point [True]
642 Text feature [examined] present in test data point [True]
645 Text feature [90] present in test data point [True]
646 Text feature [new] present in test data point [True]
649 Text feature [alternative] present in test data point [True]
650 Text feature [loop] present in test data point [True]
651 Text feature [cancer] present in test data point [True]
652 Text feature [thus] present in test data point [True]
654 Text feature [potent] present in test data point [True]
655 Text feature [reports] present in test data point [True]
658 Text feature [unknown] present in test data point [True]
662 Text feature [four] present in test data point [True]
666 Text feature [et] present in test data point [True]
668 Text feature [pcr] present in test data point [True]
672 Text feature [similar] present in test data point [True]
673 Text feature [increase] present in test data point [True]
679 Text feature [staining] present in test data point [True]
685 Text feature [second] present in test data point [True]
687 Text feature [result] present in test data point [True]
691 Text feature [carried] present in test data point [True]
692 Text feature [1998] present in test data point [True]
694 Text feature [differentiation] present in test data point [True]
699 Text feature [association] present in test data point [True]
700 Text feature [three] present in test data point [True]
701 Text feature [end] present in test data point [True]
705 Text feature [whereas] present in test data point [True]
706 Text feature [liver] present in test data point [True]
709 Text feature [normal] present in test data point [True]
710 Text feature [consequences] present in test data point [True]
716 Text feature [generated] present in test data point [True]
717 Text feature [isolated] present in test data point [True]
718 Text feature [observation] present in test data point [True]
720 Text feature [addition] present in test data point [True]
723 Text feature [promote] present in test data point [True]
726 Text feature [aberrant] present in test data point [True]
732 Text feature [terminus] present in test data point [True]
734 Text feature [ii] present in test data point [True]
737 Text feature [unable] present in test data point [True]
738 Text feature [contains] present in test data point [True]
744 Text feature [relatively] present in test data point [True]
750 Text feature [reduction] present in test data point [True]
751 Text feature [would] present in test data point [True]
752 Text feature [various] present in test data point [True]
753 Text feature [2001] present in test data point [True]

755 Text feature [hydrogen] present in test data point [True]
756 Text feature [tgf] present in test data point [True]
759 Text feature [2000] present in test data point [True]
760 Text feature [derived] present in test data point [True]
761 Text feature [since] present in test data point [True]
765 Text feature [sufficient] present in test data point [True]
766 Text feature [gst] present in test data point [True]
769 Text feature [position] present in test data point [True]
770 Text feature [applied] present in test data point [True]
777 Text feature [p21] present in test data point [True]
781 Text feature [sds] present in test data point [True]
782 Text feature [followed] present in test data point [True]
783 Text feature [upon] present in test data point [True]
788 Text feature [factors] present in test data point [True]
790 Text feature [part] present in test data point [True]
796 Text feature [tumorigenesis] present in test data point [True]
797 Text feature [size] present in test data point [True]
800 Text feature [early] present in test data point [True]
801 Text feature [contained] present in test data point [True]
807 Text feature [normalized] present in test data point [True]
811 Text feature [able] present in test data point [True]
812 Text feature [resulting] present in test data point [True]
```

```
815 Text feature [biological] present in test data point [True]
817 Text feature [analyses] present in test data point [True]
821 Text feature [ng] present in test data point [True]
826 Text feature [alone] present in test data point [True]
832 Text feature [activities] present in test data point [True]
835 Text feature [1997] present in test data point [True]
837 Text feature [smad4] present in test data point [True]
844 Text feature [strongly] present in test data point [True]
846 Text feature [200] present in test data point [True]
848 Text feature [overexpression] present in test data point [True]
853 Text feature [carcinoma] present in test data point [True]
855 Text feature [stage] present in test data point [True]
859 Text feature [led] present in test data point [True]
860 Text feature [nature] present in test data point [True]
861 Text feature [2002] present in test data point [True]
862 Text feature [red] present in test data point [True]
863 Text feature [author] present in test data point [True]
865 Text feature [ca] present in test data point [True]
866 Text feature [decreased] present in test data point [True]
868 Text feature [contact] present in test data point [True]
869 Text feature [particular] present in test data point [True]
871 Text feature [caused] present in test data point [True]
878 Text feature [repeats] present in test data point [True]
879 Text feature [box] present in test data point [True]
883 Text feature [myc] present in test data point [True]
888 Text feature [48] present in test data point [True]
890 Text feature [manner] present in test data point [True]
891 Text feature [image] present in test data point [True]
893 Text feature [lanes] present in test data point [True]
895 Text feature [chain] present in test data point [True]
907 Text feature [smad2] present in test data point [True]
908 Text feature [driven] present in test data point [True]
911 Text feature [reaction] present in test data point [True]
913 Text feature [secondary] present in test data point [True]
916 Text feature [subjected] present in test data point [True]
922 Text feature [2003] present in test data point [True]
925 Text feature [top] present in test data point [True]
927 Text feature [luciferase] present in test data point [True]
933 Text feature [colorectal] present in test data point [True]
934 Text feature [finding] present in test data point [True]
935 Text feature [colon] present in test data point [True]
941 Text feature [lane] present in test data point [True]
943 Text feature [1996] present in test data point [True]
945 Text feature [1999] present in test data point [True]
954 Text feature [cyclin] present in test data point [True]
977 Text feature [smad3] present in test data point [True]
982 Text feature [bone] present in test data point [True]
Out of the top  1000  features  393 are present in query point
```

4.5.3. Hyper paramter tuning (With Response Coding)

In [263]:
```
# ------------------------------
# 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_decrea
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None, verbose=0, warm_start=Fa
# 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-cor
# ------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/generated/sklearn.calibra
# ----------------------------
# 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 = [10,50,100,200,500,1000]
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, train_y)
        sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
        sig_clf.fit(train_x_responseCoding, train_y)
        sig_clf_probs = sig_clf.predict_proba(cv_x_responseCoding)
        cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
        print("Log Loss :",log_loss(cv_y, 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(bes
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_responseCoding)
print('For values of best alpha = ', alpha[int(best_alpha/4)], "The train log loss is:",log_loss(y_train, predict
predict_y = sig_clf.predict_proba(cv_x_responseCoding)
print('For values of best alpha = ', alpha[int(best_alpha/4)], "The cross validation log loss is:",log_loss(y_cv,
predict_y = sig_clf.predict_proba(test_x_responseCoding)
print('For values of best alpha = ', alpha[int(best_alpha/4)], "The test log loss is:",log_loss(y_test, predict_y
```

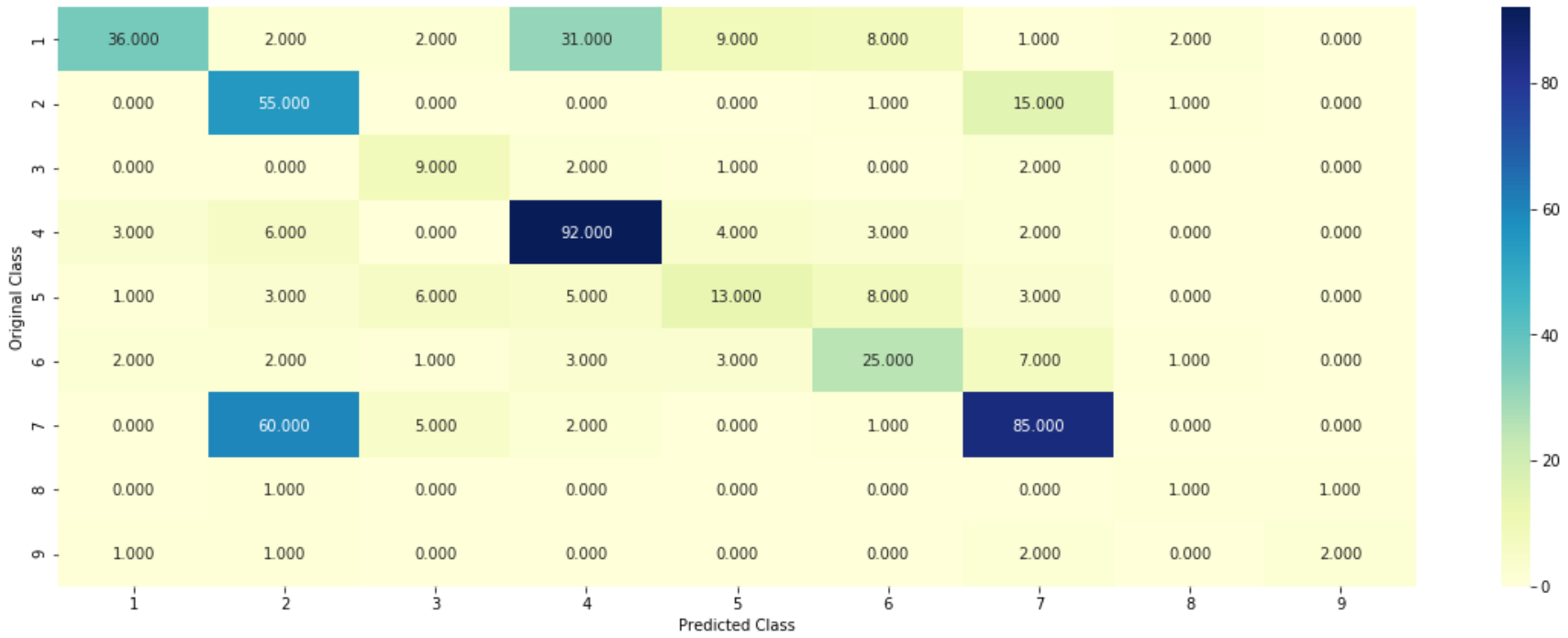
```
for n_estimators = 10 and max depth =  2
Log Loss : 2.1895959133756686
for n_estimators = 10 and max depth =  3
Log Loss : 1.6084889886524703
for n_estimators = 10 and max depth =  5
Log Loss : 1.5036769157053271
for n_estimators = 10 and max depth =  10
Log Loss : 1.6298985394357486
for n_estimators = 50 and max depth =  2
Log Loss : 1.8390658025579905
for n_estimators = 50 and max depth =  3
Log Loss : 1.520585121777144
for n_estimators = 50 and max depth =  5
Log Loss : 1.309920480885714
for n_estimators = 50 and max depth =  10
```

```
Log Loss : 1.614453006246426
for n_estimators = 100 and max depth =  2
Log Loss : 1.6258274226320277
for n_estimators = 100 and max depth =  3
Log Loss : 1.4940067268674646
for n_estimators = 100 and max depth =  5
Log Loss : 1.2100233173816997
for n_estimators = 100 and max depth =  10
Log Loss : 1.6740525658717225
for n_estimators = 200 and max depth =  2
Log Loss : 1.7004462847291089
for n_estimators = 200 and max depth =  3
Log Loss : 1.5059626102466777
for n_estimators = 200 and max depth =  5
Log Loss : 1.2963970694991371
for n_estimators = 200 and max depth =  10
Log Loss : 1.8009418246251776
for n_estimators = 500 and max depth =  2
Log Loss : 1.716566883123128
for n_estimators = 500 and max depth =  3
Log Loss : 1.5115769310355278
for n_estimators = 500 and max depth =  5
Log Loss : 1.2939991131246258
for n_estimators = 500 and max depth =  10
Log Loss : 1.8162757211717033
for n_estimators = 1000 and max depth =  2
Log Loss : 1.6930443402514093
for n_estimators = 1000 and max depth =  3
Log Loss : 1.5019327289926552
for n_estimators = 1000 and max depth =  5
Log Loss : 1.2867928989742734
for n_estimators = 1000 and max depth =  10
Log Loss : 1.7570579460013904
For values of best alpha =  100 The train log loss is: 0.05039010491056535
For values of best alpha =  100 The cross validation log loss is: 1.2100233173816997
For values of best alpha =  100 The test log loss is: 1.1779371370010046
```

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

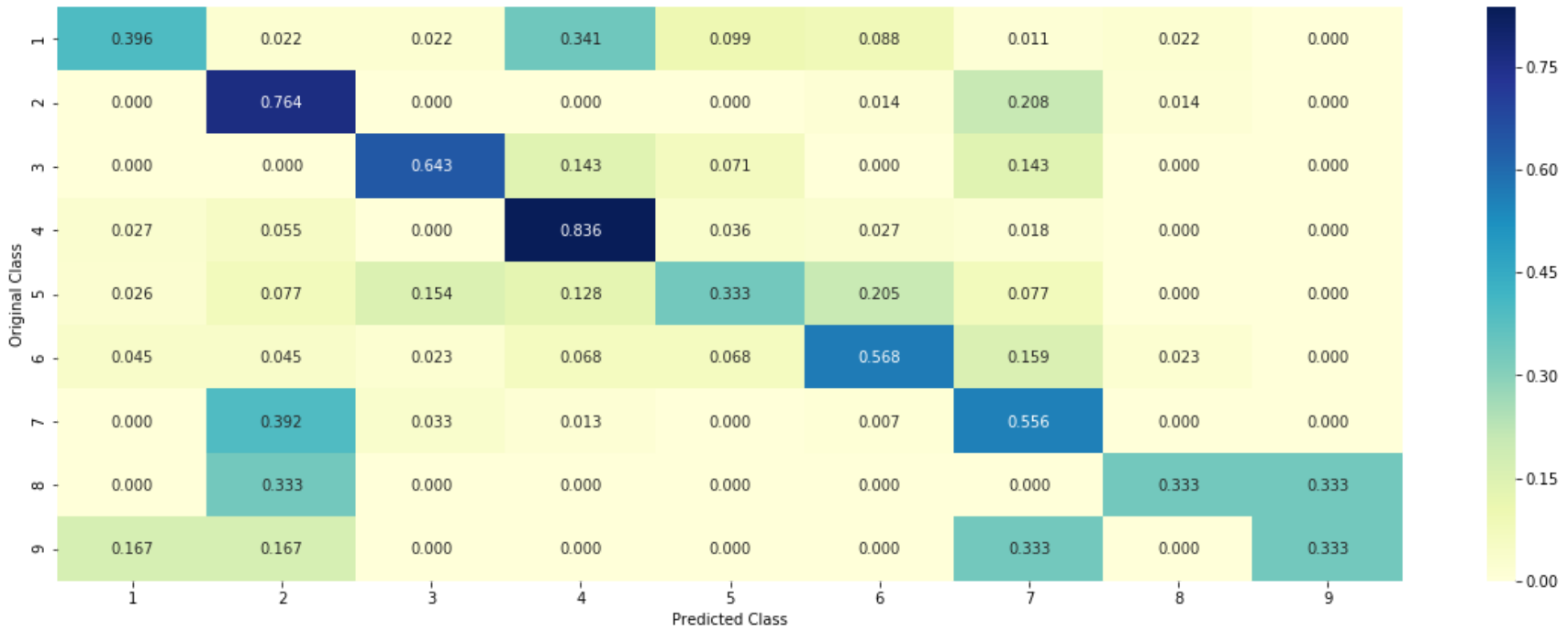
In [264]:
```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_samples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrea
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None, verbose=0, warm_start=Fa
# 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-cor
# --------------------------------

clf = RandomForestClassifier(max_depth=max_depth[int(best_alpha%4)], n_estimators=alpha[int(best_alpha/4)], crite
predict_and_plot_confusion_matrix(train_x_responseCoding, train_y,cv_x_responseCoding,cv_y, clf)
```

```
Log loss : 1.2100233173816997
Number of mis-classified points : 0.40225563909774437
-------------------- Confusion matrix --------------------
```




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


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



4.5.5. Feature Importance

4.5.5.1. Correctly Classified point

In [267]:
```
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/4)], criterion='gini', max_depth=max_depth[int(bes
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)


test_point_index = 1
no_feature = 1000
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].re
print("Actual Class :", test_y[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 : 4
Predicted Class Probabilities: [[0.0467 0.0249 0.3378 0.4993 0.0086 0.0208 0.0095 0.0237 0.0288]]
Actual Class : 4
--------------------------------------------------
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
Gene is important feature
Text is important feature
Gene is important feature
Variation is important feature
Text is important feature
Gene is important feature
Variation is important feature
Gene is important feature
Text is important feature
Text is important feature
Variation is important feature
Gene is important feature
Text is important feature
Gene is important feature
Gene is important feature
```

4.5.5.2. Incorrectly Classified point

In [268]:
```
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].re
print("Actual Class :", test_y[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 : 4
Predicted Class Probabilities: [[0.2103 0.0167 0.1255 0.4805 0.0286 0.0999 0.0083 0.0144 0.0159]]
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
Gene is important feature
Variation is important feature
Text is important feature
Gene is important feature
Variation is important feature
Gene is important feature
Text is important feature
Text is important feature
Variation is important feature
Gene is important feature
Text is important feature
Gene is important feature
Gene is important feature
```

4.7 Stack the models

4.7.1 testing with hyper parameter tuning

In [269]:
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDCl
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=N
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0
# 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/generat
# ------------------------------
# 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-co
# -------------------------------


# read more about support vector machines with linear kernals here http://scikit-learn.org/stable/modules/generat
# ------------------------------
# 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_decrea
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None, verbose=0, warm_start=Fa
# 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-cor
# -------------------------------


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

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


clf3 = MultinomialNB(alpha=0.001)
clf3.fit(train_x_onehotCoding, train_y)
sig_clf3 = CalibratedClassifierCV(clf3, method="sigmoid")

sig_clf1.fit(train_x_onehotCoding, train_y)
print("Logistic Regression :  Log Loss: %0.2f" % (log_loss(cv_y, sig_clf1.predict_proba(cv_x_onehotCoding))))
sig_clf2.fit(train_x_onehotCoding, train_y)
print("Support vector machines : Log Loss: %0.2f" % (log_loss(cv_y, sig_clf2.predict_proba(cv_x_onehotCoding))))
sig_clf3.fit(train_x_onehotCoding, train_y)
print("Naive Bayes : Log Loss: %0.2f" % (log_loss(cv_y, sig_clf3.predict_proba(cv_x_onehotCoding))))
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_probas=True)
    sclf.fit(train_x_onehotCoding, train_y)
    print("Stacking Classifer : for the value of alpha: %f Log Loss: %0.3f" % (i, log_loss(cv_y, sclf.predict_pro
    log_error =log_loss(cv_y, sclf.predict_proba(cv_x_onehotCoding))
    if best_alpha > log_error:
        best_alpha = log_error
```

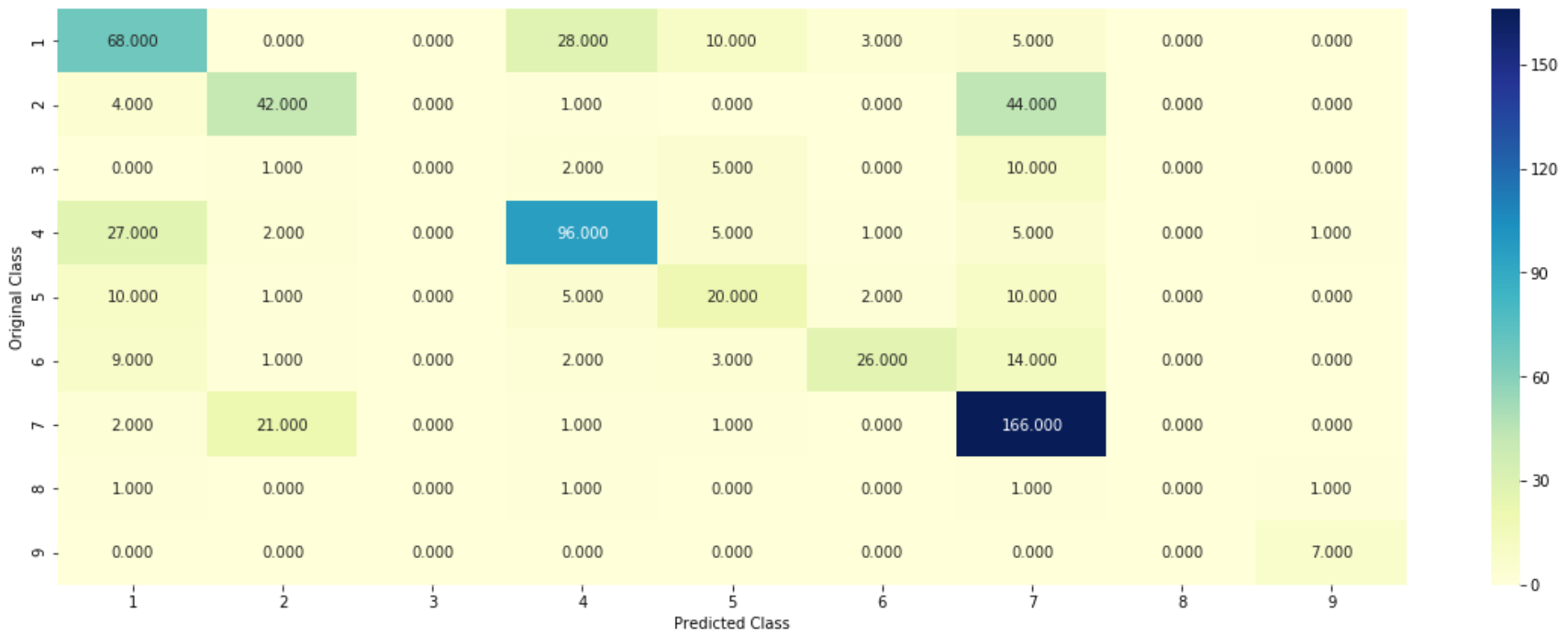
```
Logistic Regression :  Log Loss: 1.08
Support vector machines : Log Loss: 1.84
Naive Bayes : Log Loss: 1.21
--------------------------------------------------
Stacking Classifer : for the value of alpha: 0.000100 Log Loss: 2.178
Stacking Classifer : for the value of alpha: 0.001000 Log Loss: 2.031
```

```
Stacking Classifer : for the value of alpha: 0.010000 Log Loss: 1.500
Stacking Classifer : for the value of alpha: 0.100000 Log Loss: 1.166
Stacking Classifer : for the value of alpha: 1.000000 Log Loss: 1.378
Stacking Classifer : for the value of alpha: 10.000000 Log Loss: 1.800
```

4.7.2 testing the model with the best hyper parameters

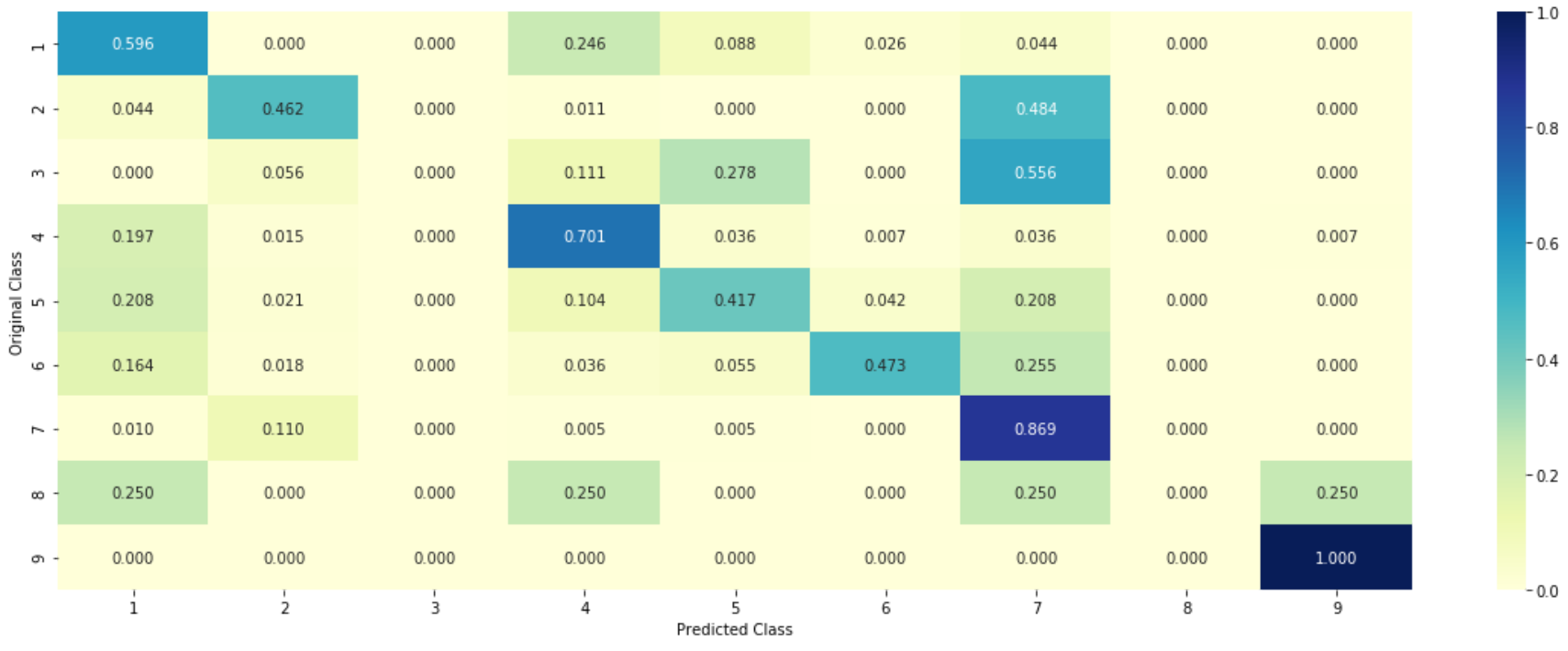
In [270]:
```
lr = LogisticRegression(C=0.1)
sclf = StackingClassifier(classifiers=[sig_clf1, sig_clf2, sig_clf3], meta_classifier=lr, use_probas=True)
sclf.fit(train_x_onehotCoding, train_y)

log_error = log_loss(train_y, sclf.predict_proba(train_x_onehotCoding))
print("Log loss (train) on the stacking classifier :",log_error)

log_error = log_loss(cv_y, sclf.predict_proba(cv_x_onehotCoding))
print("Log loss (CV) on the stacking classifier :",log_error)

log_error = log_loss(test_y, sclf.predict_proba(test_x_onehotCoding))
print("Log loss (test) on the stacking classifier :",log_error)

print("Number of missclassified point :", np.count_nonzero((sclf.predict(test_x_onehotCoding)- test_y))/test_y.sh
plot_confusion_matrix(test_y=test_y, predict_y=sclf.predict(test_x_onehotCoding))
```

```
Log loss (train) on the stacking classifier : 0.5444778395577321
Log loss (CV) on the stacking classifier : 1.1664760905643303
Log loss (test) on the stacking classifier : 1.1143784903828993
Number of missclassified point : 0.3609022556390977
-------------------- Confusion matrix --------------------
```




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

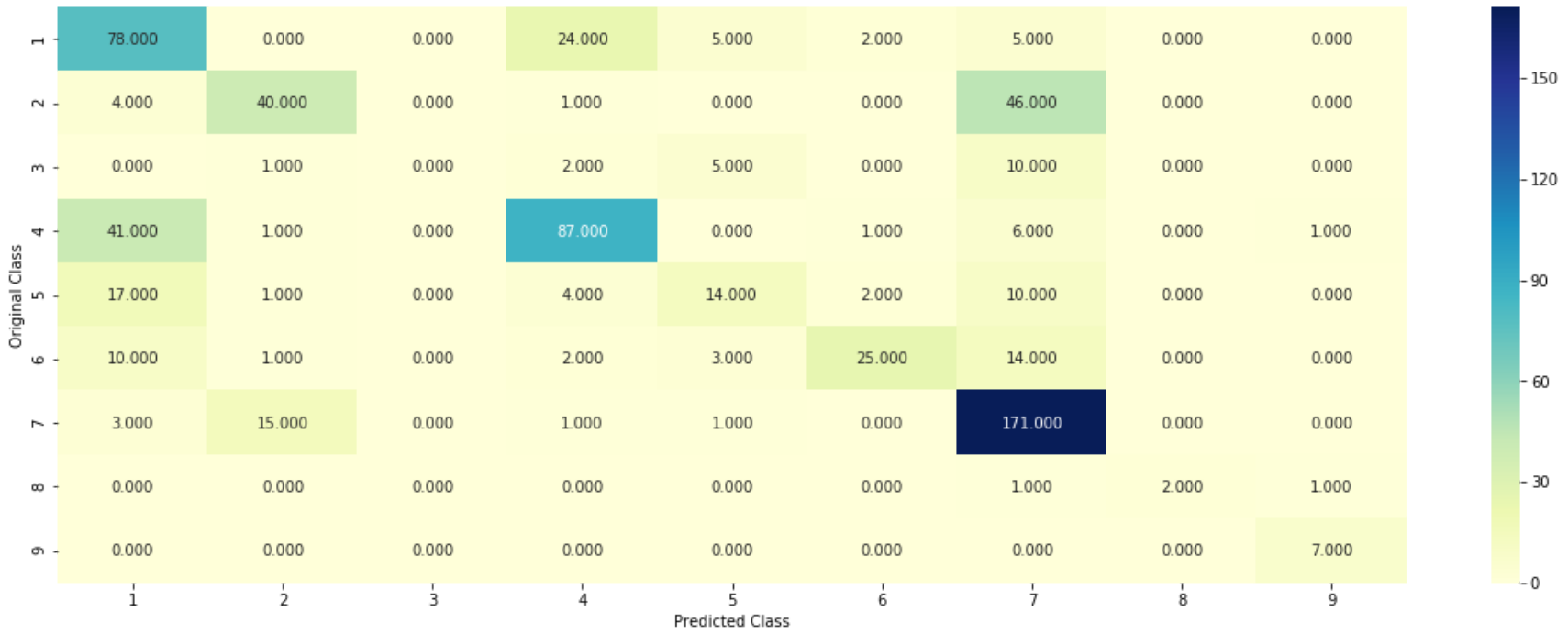

```
-------------------- Recall matrix (Row sum=1) --------------------
```



4.7.3 Maximum Voting classifier

In [271]:
```
#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_onehotCoding, train_y)
print("Log loss (train) on the VotingClassifier :", log_loss(train_y, vclf.predict_proba(train_x_onehotCoding)))
print("Log loss (CV) on the VotingClassifier :", log_loss(cv_y, vclf.predict_proba(cv_x_onehotCoding)))
print("Log loss (test) on the VotingClassifier :", log_loss(test_y, vclf.predict_proba(test_x_onehotCoding)))
print("Number of missclassified point :", np.count_nonzero((vclf.predict(test_x_onehotCoding)- test_y))/test_y.sh
plot_confusion_matrix(test_y=test_y, predict_y=vclf.predict(test_x_onehotCoding))
```

```
Log loss (train) on the VotingClassifier : 0.83492291034833
Log loss (CV) on the VotingClassifier : 1.213276637836091
Log loss (test) on the VotingClassifier : 1.1509826641959164
Number of missclassified point : 0.362406015037594
-------------------- Confusion matrix --------------------
```




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


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


Logistic Regression(CountVectorizer with Unigram and Bigram)

In [304]:
```
train_variation=train_df['Variation'].values
test_variation=test_df['Variation'].values
cv_variation=cv_df['Variation'].values
train_gene=train_df['Gene'].values
test_gene=test_df['Gene'].values
cv_gene=cv_df['Gene'].values
train_text=train_df['TEXT'].values
test_text=test_df['TEXT'].values
cv_text=cv_df['TEXT'].values


text_vectorizer=CountVectorizer(ngram_range=(1, 2))
train_variation=text_vectorizer.fit_transform(train_variation)
test_variation=text_vectorizer.transform(test_variation)
cv_variation=text_vectorizer.transform(cv_variation)

train_gene=text_vectorizer.fit_transform(train_gene)
test_gene=text_vectorizer.transform(test_gene)
cv_gene=text_vectorizer.transform(cv_gene)

text_vectorizer=CountVectorizer(min_df=10,ngram_range=(1,2))
train_variation=normalize(train_variation,axis=0)
test_variation=normalize(test_variation,axis=0)
cv_variation=normalize(cv_variation,axis=0)
train_gene=normalize(train_gene,axis=0)
test_gene=normalize(test_gene,axis=0)
cv_gene=normalize(cv_gene,axis=0)
print(train_gene.shape)
print(train_gene[1,:])
print(train_variation.shape)
print(train_variation[100,:])


train_text=text_vectorizer.fit_transform(train_text)
test_text=text_vectorizer.transform(test_text)
cv_text=text_vectorizer.transform(cv_text)
train_text=normalize(train_text,axis=0)
test_text=normalize(test_text,axis=0)
cv_text=normalize(cv_text,axis=0)
print(train_text.shape)
```

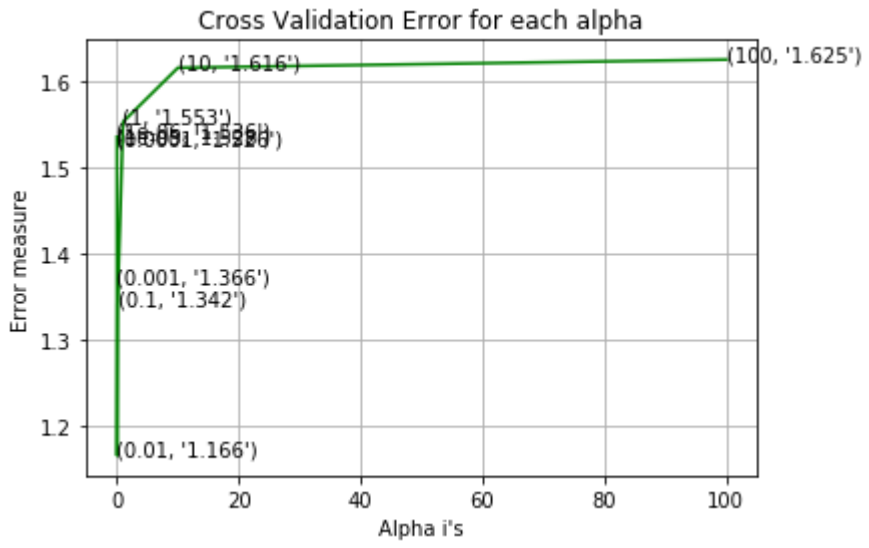
```
(2124, 235)
  (0, 194)      0.2581988897471611
(2124, 2065)
  (0, 104)      0.1414213562373095
(2124, 225052)
```

In [312]:
```
# hstacking the train,test, and CV Data

train_data=hstack([train_variation,train_gene,train_text]).tocsr()
cv_data=hstack([cv_variation,cv_gene,cv_text]).tocsr()
test_data=hstack([test_variation,test_gene,test_text]).tocsr()
```

In [314]:

```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDCl
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=N
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0
# 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/
#-------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/generated/sklearn.calibra
# ------------------------------
# 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_data, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_data, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_data)
    cv_log_error_array.append(log_loss(cv_y, 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(cv_y, 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', random_state=42)
clf.fit(train_data, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_data, train_y)

predict_y = sig_clf.predict_proba(train_data)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, predict_y, lab
predict_y = sig_clf.predict_proba(cv_data)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_cv, predic
predict_y = sig_clf.predict_proba(test_data)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predict_y, label
```

```
for alpha = 1e-06
Log Loss : 1.5356606542957567
for alpha = 1e-05
Log Loss : 1.5281435501820149
for alpha = 0.0001
Log Loss : 1.525671117661709
for alpha = 0.001
Log Loss : 1.3664457624226665
for alpha = 0.01
Log Loss : 1.1664031876485257
for alpha = 0.1
Log Loss : 1.3417879354578617
for alpha = 1
Log Loss : 1.5529809872270768
for alpha = 10
Log Loss : 1.6156199655139774
for alpha = 100
Log Loss : 1.625070451862977
```

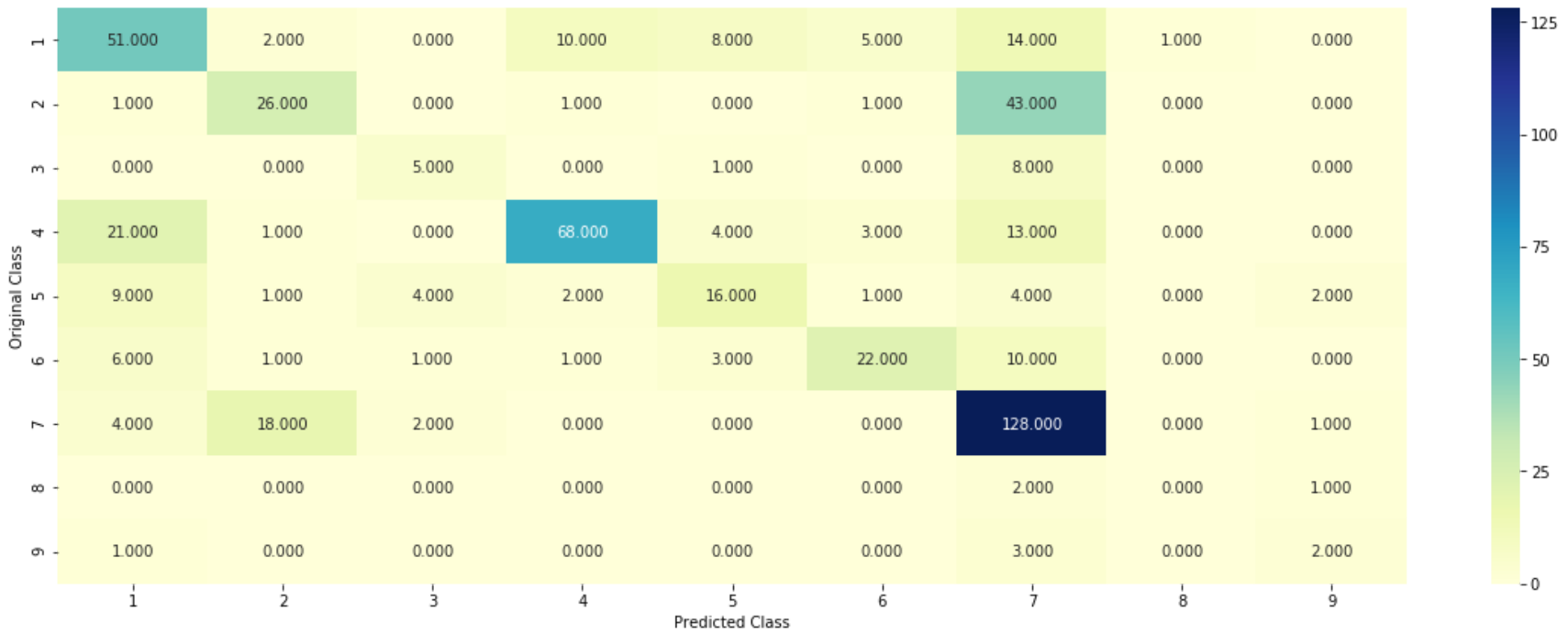



```
For values of best alpha =  0.01 The train log loss is: 0.8648003915897353
For values of best alpha =  0.01 The cross validation log loss is: 1.1664031876485257
For values of best alpha =  0.01 The test log loss is: 1.065964635372382
```

4.3.1.2. Testing the model with best hyper paramters

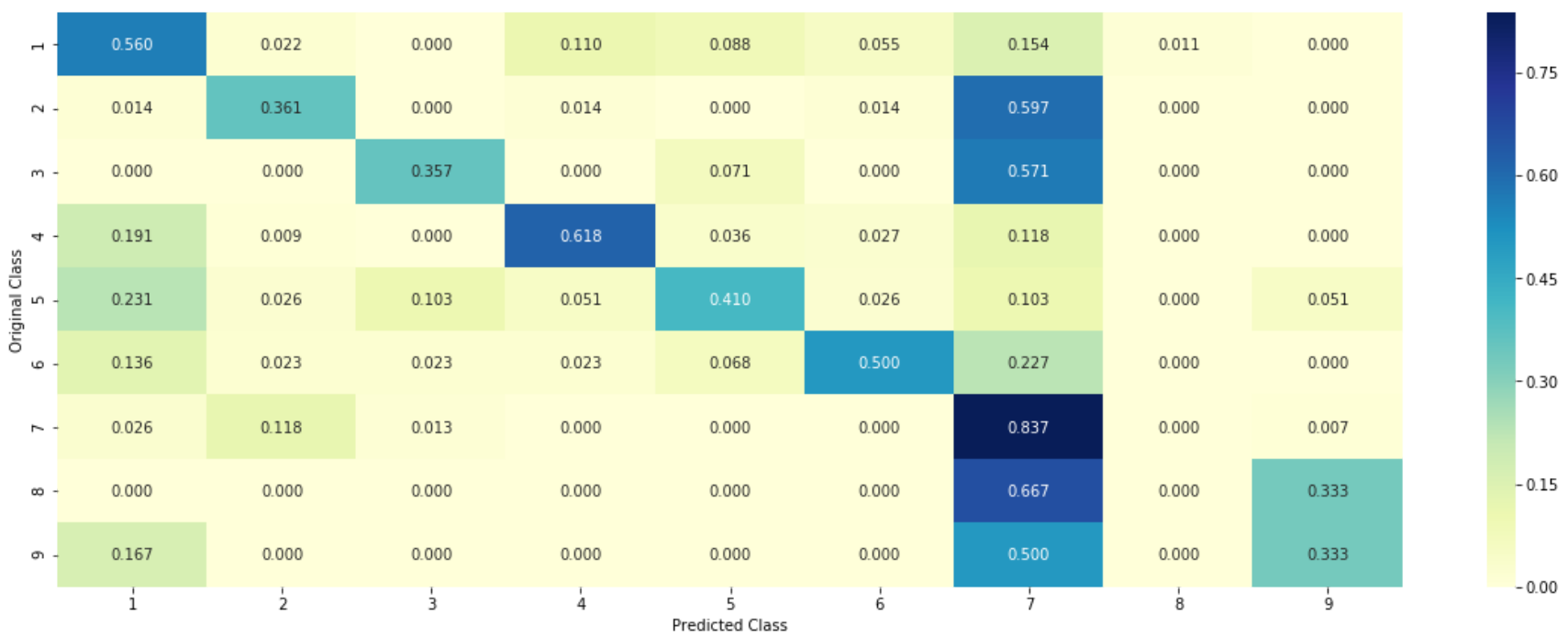
In [318]:
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDCl
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=N
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0
# 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/
#-------------------------------
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_data, train_y, cv_data, cv_y, clf)
```

```
Log loss : 1.1664031876485257
Number of mis-classified points : 0.40225563909774437
-------------------- Confusion matrix --------------------
```




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


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



Feature Engineering

In [371]:
```
#Code from Above Cells

result = pd.merge(data, data_text,on='ID', how='left')
result.loc[result['TEXT'].isnull(),'TEXT'] = result['Gene'] +' '+result['Variation']
y_true = result['Class'].values
result.Gene = result.Gene.str.replace('\s+', '_')
result.Variation = result.Variation.str.replace('\s+', '_')


x_train, x_test, y_train, y_test = train_test_split(result, y_true, stratify=y_true, test_size=0.2)
x_train, x_cv, y_train, y_cv = train_test_split(x_train, y_train, stratify=y_train, test_size=0.2)

#####################################################################

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))

#####################################################################
#FOR GENE FEATURE


gene_vectorizer = TfidfVectorizer()
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 [372]:
```
#FOR VARIATION FEATURE
#Code FRom Above cells

#####################################################################
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))

#####################################################################

vectorizer = TfidfVectorizer()
train_variation_feature_onehotCoding = vectorizer.fit_transform(x_train['Variation'])
test_variation_feature_onehotCoding = vectorizer.transform(x_test['Variation'])
cv_variation_feature_onehotCoding = vectorizer.transform(x_cv['Variation'])
```

In [373]:
```
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


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 [374]:
```
# For Text Feature


text_vectorizer = TfidfVectorizer()
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 features) vector
train_text_fea_counts = train_text_feature_onehotCoding.sum(axis=0).A1

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

test_text_feature_onehotCoding = text_vectorizer.transform(x_test['TEXT'])
cv_text_feature_onehotCoding = text_vectorizer.transform(x_cv['TEXT'])
########################################################################

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)

# 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(ax
test_text_feature_responseCoding = (test_text_feature_responseCoding.T/test_text_feature_responseCoding.sum(axis=
cv_text_feature_responseCoding = (cv_text_feature_responseCoding.T/cv_text_feature_responseCoding.sum(axis=1)).T
```

```
Total number of unique words in train data : 123837
```

In [375]:
```
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 [376]:
```
gene_variation = []

for gene in data['Gene'].values:
    gene_variation.append(gene)

for variation in data['Variation'].values:
    gene_variation.append(variation)



tfidfVectorizer = TfidfVectorizer(max_features=1000)
text2 = tfidfVectorizer.fit_transform(gene_variation)
gene_variation_features = tfidfVectorizer.get_feature_names()

train_text = tfidfVectorizer.transform(x_train['TEXT'])
test_text = tfidfVectorizer.transform(x_test['TEXT'])
cv_text = tfidfVectorizer.transform(x_cv['TEXT'])
```

In [377]:
```
train_x_onehotCoding = hstack((train_gene_feature_onehotCoding,train_variation_feature_onehotCoding, train_text,t
train_y = np.array(list(x_train['Class']))

test_x_onehotCoding = hstack((test_gene_feature_onehotCoding,test_variation_feature_onehotCoding, test_text,test_
test_y = np.array(list(x_test['Class']))

cv_x_onehotCoding = hstack((cv_gene_feature_onehotCoding,cv_variation_feature_onehotCoding, cv_text,cv_text_featu
cv_y = np.array(list(x_cv['Class']))
```

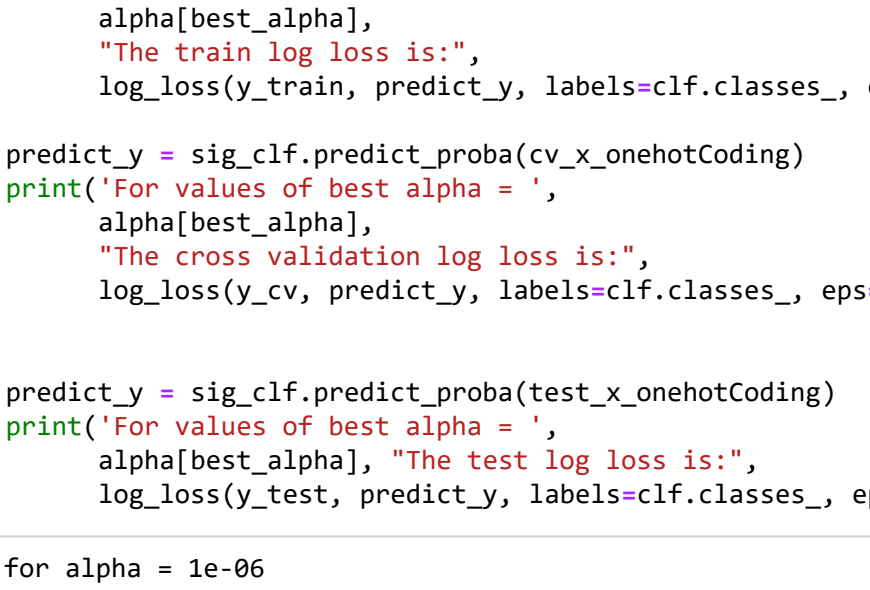
In [378]:
```
train_x_responseCoding = np.hstack((train_gene_feature_responseCoding,train_variation_feature_responseCoding,tra
test_x_responseCoding = np.hstack((test_gene_var_responseCoding, test_text_feature_responseCoding,test_gene_featu
cv_x_responseCoding = np.hstack((cv_gene_var_responseCoding, cv_text_feature_responseCoding,cv_gene_feature_respo
```

In [379]:
```
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)

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)
```

```
One hot encoding features :
(number of data points * number of features) in train data =  (2124, 127018)
(number of data points * number of features) in test data =  (665, 127018)
(number of data points * number of features) in cross validation data = (532, 127018)
 Response encoding features :
(number of data points * number of features) in train data =  (2124, 45)
(number of data points * number of features) in test data =  (665, 45)
(number of data points * number of features) in cross validation data = (532, 45)
```

In [380]:
```
alpha = [10 ** x for x in range(-6, 2)]
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, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, 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(cv_y, 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',)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_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_x_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_x_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))
```

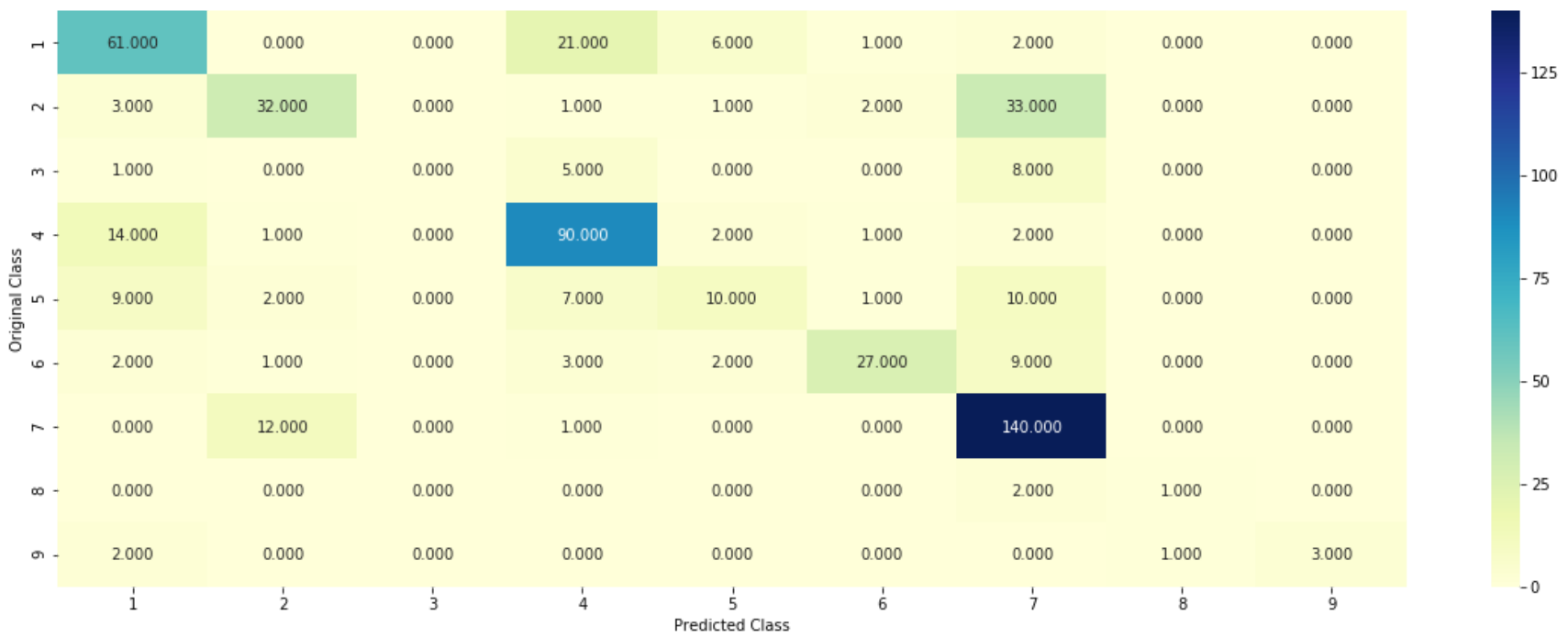
```
for alpha = 1e-06
Log Loss : 1.0869811936135794
for alpha = 1e-05
Log Loss : 1.0576362421467183
for alpha = 0.0001
Log Loss : 0.9685179670267624
for alpha = 0.001
Log Loss : 1.0369882810140187
for alpha = 0.01
Log Loss : 1.257802448842145
for alpha = 0.1
Log Loss : 1.6169754533282263
for alpha = 1
Log Loss : 1.7166560080012416
for alpha = 10
Log Loss : 1.726842098165162
```



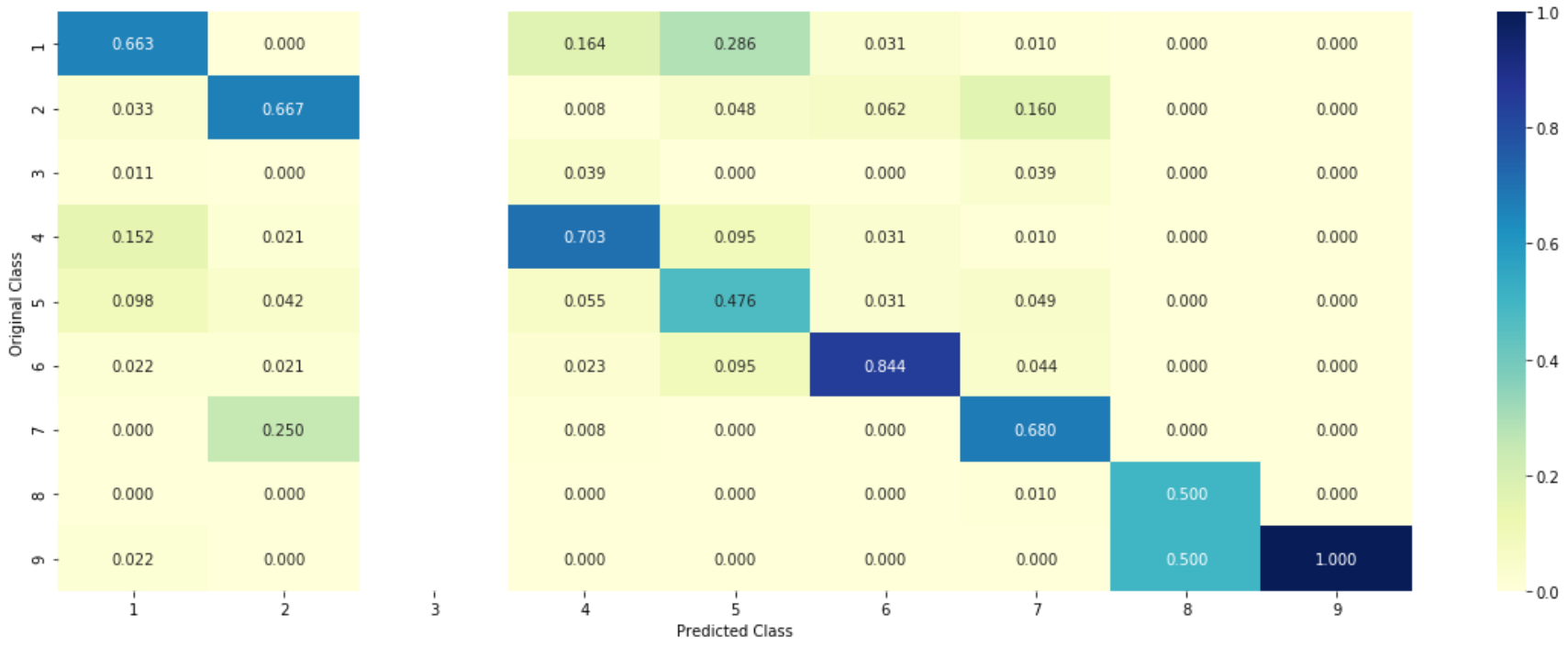

```
For values of best alpha =  0.0001 The train log loss is: 0.4479771216080555
For values of best alpha =  0.0001 The cross validation log loss is: 0.9882759278855332
For values of best alpha =  0.0001 The test log loss is: 1.0031279664365167
```

In [381]:
```
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log',)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_onehotCoding, cv_y, clf)
```

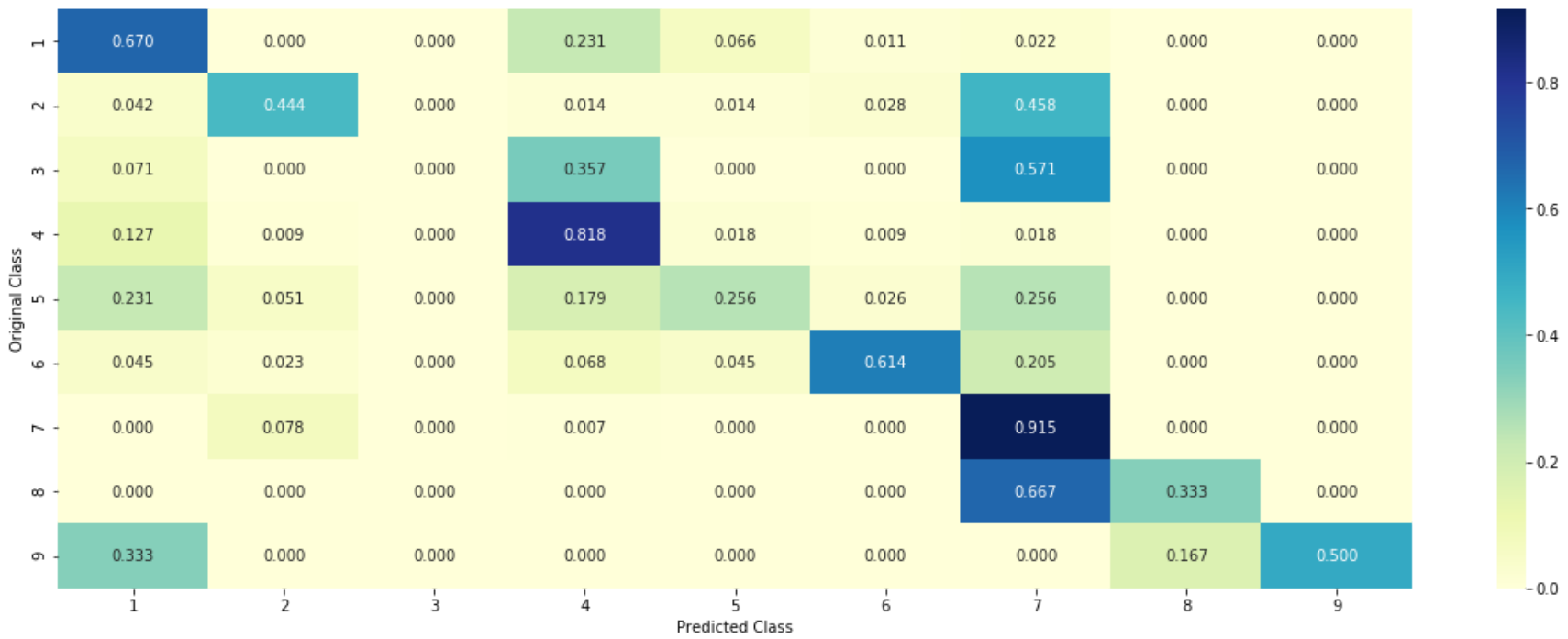
```
Log loss : 0.9881741288777217
Number of mis-classified points : 0.3157894736842105
-------------------- Confusion matrix --------------------
```




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




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



CONCLUSION

- We have taken top 1k features for both bow and tfidf representation.
- We did hyperparameter tuning to get best parameter and reduce the log loss.
- when we used logistic regression(After Feature Engineering) on tfidf we got log-loss less than 1 which is what we wanted.

After Feature Engineering ,we managed to get logloss<1 .

PROCEDURE OF SOLVING THE CASE STUDY:

- FIRSTLY,WE STARTED WITH LOADING DATA USING PANDAS LIBRARY.
- DID CLEANING,PREPROCESSING OF DATA USING VARIOUS NLP TECHNIQUES.
- DID EDA FOR UNDERSTANDING THE UNDERLYING PATTERN AND DISTRIBUTION OF DATA.
- CHECKED WHETHER EACH FEATURE IS IMPORTANT FOR FURTHER ANALYSIS USING UNIVARIATE ANALYSIS.
- CHECKED DISTRIBUTION OF CLASS LABEL AMONG TRAIN ,TEST, CV DATASET.
- BUILD A RANDOM MODEL TO FIND THE UPPERLIMIT OF THE LOGLOSS(== 2.5).
- GENERATED 9 CLASS PROBABILITIES RANDOMLY AND ENSURED THAT THEY SUM TO 1.
- USED ONE-HOT-ENCODING AND RESPONSE CODING FOR FEATURIZATION OF CATEGORICAL DATA.
- ON TEXT FEATURE WE USED BOW AND TFIDF AS FEATURIZATION TECHNIQUES AND TOOK ONLY TOP 1K FEATURES.
- BUILT MODEL ON EACH AND EVERY FEATURE TO ENSURE THAT THEY ARE USEFUL OR NOT.
- WE STARTED WITH NB AS OUR BASELINEMODEL AND THEN TRIED VARIOUS MODELS LIKE LOGISTIC REGRESSION,SVM,RF,STACKING,KNN ETC.
- SHOWED LOGLOSS FOR TRAIN,TEST,CV,MISCLASSIFICATION ERROR,CONFUSION MATRIX,PRECISION AND RECALL MATRIX FOR EACH MODEL .
- GAVE INTERPRETATION FOR EACH MODEL.
- DID FEATURE ENGINEERING BY COMBINING GIVEN FEATURE(GENE,VARIATION AND TEXT) AS A NEW FEATURE.
- FINALLY GOT A LOGLOSS OF < 1.

In [384]:
```
from prettytable import PrettyTable

# Names of models
model=['Naive Bayes  ','KNN','Logistic Regression With Class balancing  ','Logistic Regression Without Class bala


train =[0.52 ,0.65 ,0.44 ,0.44 ,0.58 ,0.87 ,0.05 ,0.54 ,0.83 ,0.86 ,0.44]
cv=    [1.21 ,1.03 ,1.06 ,1.10 ,1.09 ,1.23 ,1.21 ,1.16 ,1.21 ,1.16 ,0.98]
test = [1.14 ,0.98 ,0.92 ,0.95 ,1.00 ,1.16 ,1.17 ,1.11 ,1.15 ,1.06 ,1.00]
mp=    [44   ,36   ,36   ,36   ,34   ,44   ,40   ,36   ,36   ,40   ,31]
num=[1,2,3,4,5,6,7,8,9,10,11]
ptable = PrettyTable()

# Adding columns
ptable.add_column("S.NO.",num)
ptable.add_column("model",model)
ptable.add_column("train",train)
ptable.add_column("cv",cv)
ptable.add_column("test",test)
ptable.add_column("% Misclassified Points",mp)

# Printing the Table
print(ptable)
```

```
+-------+---------------------------------------------------------------+-------+------+------+----------------
--------+
| S.NO. |                             model                             | train |  cv  | test | % Misclassified
Points |
+-------+---------------------------------------------------------------+-------+------+------+----------------
--------+
|   1   |                         Naive Bayes                           |  0.52 | 1.21 | 1.14 |           44
|
|   2   |                             KNN                               |  0.65 | 1.03 | 0.98 |           36
|
|   3   |          Logistic Regression With Class balancing             |  0.44 | 1.06 | 0.92 |           36
|
|   4   |         Logistic Regression Without Class balancing           |  0.44 | 1.1  | 0.95 |           36
|
|   5   |                         Linear SVM                            |  0.58 | 1.09 | 1.0  |           34
|
|   6   |         Random Forest Classifier With One hot Encoding        |  0.87 | 1.23 | 1.16 |           44
|
|   7   |         Random Forest Classifier With Response Coding         |  0.05 | 1.21 | 1.17 |           40
|
|   8   |                   Stack Models:LR+NB+SVM                      |  0.54 | 1.16 | 1.11 |           36
|
|   9   |                  Maximum Voting classifier                    |  0.83 | 1.21 | 1.15 |           36
|
|   10  | CountVectorizer Features, including both unigrams and bigrams |  0.86 | 1.16 | 1.06 |           40
|
|   11  |                     Feature Engineering                       |  0.44 | 0.98 | 1.0  |           31
|
+-------+---------------------------------------------------------------+-------+------+------+----------------
--------+
```