Personalized cancer diagnosis

1. Business Problem

1.1. Description

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

Data: Memorial Sloan Kettering Cancer Center (MSKCC)

Download training_variants.zip and training_text.zip from Kaggle.

Context:

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

Problem statement :

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

1.2. Source/Useful Links

Some articles and reference blogs about the problem statement

1. https://www.forbes.com/sites/matthewherper/2017/06/03/a-new-cancer-drug-helped-almost-everyone-who-took-it-almost-heres-what-it-teaches-us/#2a44ee2f6b25

2. https://www.youtube.com/watch?v=UwbuW7oK8rk
3. https://www.youtube.com/watch?v=qxXRKVompI8

1.3. Real-world/Business objectives and constraints.

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

2. Machine Learning Problem Formulation

2.1. Data

2.1.1. Data Overview

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

2.1.2. Example Data Point

training_variants

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

training_text

ID,Text
0||Cyclin-dependent kinases (CDKs) regulate a variety of fundamental cellular processes. CDK10 stands out as one of the last orphan CDKs for which no activating cyclin has been identified and no kinase activity revealed. Previous work has shown that CDK10 silencing increases ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2)-driven activation of the MAPK pathway, which confers tamoxifen resistance to breast cancer cells. The precise mechanisms by which CDK10 modulates ETS2 activity, and more generally the functions of CDK10, remain elusive. Here we demonstrate that CDK10 is a cyclin-dependent kinase by identifying cyclin M as an activating cyclin. Cyclin M, an orphan cyclin, is the product of FAM58A, whose mutations cause STAR syndrome, a human developmental anomaly whose features include toe syndactyly, telecanthus, and anogenital and renal malformations. We show that STAR syndrome-associated cyclin M mutants are unable to interact with CDK10. Cyclin M silencing phenocopies CDK10 silencing in increasing c-Raf and in conferring tamoxifen resistance to breast cancer cells. CDK10/cyclin M phosphorylates ETS2 in vitro, and in cells it positively controls ETS2 degradation by the proteasome. ETS2 protein levels are increased in cells derived from a STAR patient, and this increase is attributable to decreased cyclin M levels. Altogether, our results reveal an additional regulatory mechanism for ETS2, which plays key roles in cancer and development. They also shed light on the molecular mechanisms underlying STAR syndrome.Cyclin-dependent kinases (CDKs) play a pivotal role in the control of a number of fundamental cellular processes (1). The human genome contains 21 genes encoding proteins

that can be considered as members of the CDK family owing to their sequence similarity with bona fide CDKs, those known to be activated by cyclins (2). Although discovered almost 20 y ago (3, 4), CDK10 remains one of the two CDKs without an identified cyclin partner. This knowledge gap has largely impeded the exploration of its biological functions. CDK10 can act as a positive cell cycle regulator in some cells (5, 6) or as a tumor suppressor in others (7, 8). CDK10 interacts with the ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2) transcription factor and inhibits its transcriptional activity through an unknown mechanism (9). CDK10 knockdown derepresses ETS2, which increases the expression of the c-Raf protein kinase, activates the MAPK pathway, and induces resistance of MCF7 cells to tamoxifen (6). ...

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

2.2.1. Type of Machine Learning Problem

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

2.2.2. Performance Metric

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

Metric(s):

- Multi class log-loss
- Confusion matrix

2.2.3. Machine Learing Objectives and Constraints

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

Constraints:

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

2.3. Train, CV and Test Datasets

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

3. Exploratory Data Analysis

```python
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
```

```python
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

```python
In [2]: data = pd.read_csv('training/training_variants')
print('Number of data points : ', data.shape[0])
print('Number of features : ', data.shape[1])
print('Features : ', data.columns.values)
data.head()
```

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

Out[2]:

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

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

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

3.1.2. Reading Text Data

In [3]:
```python
# note the seprator in this file
data_text =pd.read_csv("training/training_text",sep="\|\|",engine="pyth
on",names=["ID","TEXT"],skiprows=1)
print('Number of data points : ', data_text.shape[0])
print('Number of features : ', data_text.shape[1])
print('Features : ', data_text.columns.values)
data_text.head()
```

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

Out[3]:

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

3.1.3. Preprocessing of text

```python
In [4]:  # loading stop words from nltk library
         stop_words = set(stopwords.words('english'))


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

                 for word in total_text.split():
                 # if the word is a not a stop word then retain that word from t
         he data
                     if not word in stop_words:
                         string += word + " "

                 data_text[column][index] = string
```

```python
In [5]:  #text processing stage.
         start_time = time.clock()
         for index, row in data_text.iterrows():
```

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

```
there is no text description for id: 1109
there is no text description for id: 1277
there is no text description for id: 1407
there is no text description for id: 1639
there is no text description for id: 2755
Time took for preprocessing the text : 204.6077105 seconds
```

In [6]:
```
#merging both gene_variations and text data based on ID
result = pd.merge(data, data_text,on='ID', how='left')
result.head()
```

Out[6]:

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

In [8]:
```
result[result.isnull().any(axis=1)]
```

Out[8]:

	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

	ID	Gene	Variation	Class	TEXT
2755	2755	BRAF	G596C	7	NaN

In [7]:
```
result.loc[result['TEXT'].isnull(),'TEXT'] = result['Gene'] +' '+result
['Variation']
```

In [8]:
```
result[result['ID']==1109]
```

Out[8]:

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

3.1.4. Test, Train and Cross Validation Split

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

In [9]:
```
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 o
f output varaible 'y_true' [stratify=y_true]
X_train, test_df, y_train, y_test = train_test_split(result, y_true, st
ratify=y_true, test_size=0.2)
# split the train data into train and cross validation by maintaining s
ame distribution of output varaible 'y_train' [stratify=y_train]
train_df, cv_df, y_train, y_cv = train_test_split(X_train, y_train, str
atify=y_train, test_size=0.2)
```

In [10]:
```
from sklearn.feature_extraction.text import CountVectorizer
train_df.columns
train_gene=train_df['Gene'].values;test_gene=test_df['Gene'].values;cv_
gene=cv_df['Gene'].values
```

```python
train_variation=train_df['Variation'].values;test_variation=test_df['Va
riation'].values;cv_variation=cv_df['Variation'].values
train_text=train_df['TEXT'].values;test_text=test_df['TEXT'].values;cv_
text=cv_df['TEXT'].values

encd = CountVectorizer()
#For Gene Feature Encoding
train_gene=encd.fit_transform(train_gene)
test_gene=encd.transform(test_gene)
cv_gene=encd.transform(cv_gene)

#For Variation Feature Encoding
train_variation=encd.fit_transform(train_variation)
test_variation=encd.transform(test_variation)
cv_variation=encd.transform(cv_variation)

#For Text Feature
train_text=encd.fit_transform(train_text);test_text=encd.transform(test
_text);cv_text=encd.transform(cv_text)

print(train_gene.shape)
print(train_variation.shape)
print(train_text.shape)
```

```
(2124, 224)
(2124, 1971)
(2124, 124536)
```

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

```python
In [11]: 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 [12]: # it returns a dict, keys as class labels and values as the number of d
         ata points in that class
         import matplotlib.pyplot as plt
         import seaborn as sns
         train_class_distribution = train_df['Class'].value_counts().sort_index
         ()
         test_class_distribution = test_df['Class'].value_counts().sort_index()
         cv_class_distribution = cv_df['Class'].value_counts().sort_index()

         my_colors = ['r', 'g', 'b', 'k', 'y', 'm', 'c']
         train_class_distribution.plot(kind='bar', color=my_colors)
         plt.xlabel('Class')
         plt.ylabel('Data points per Class')
         plt.title('Distribution of yi in train data')
         plt.grid()
         plt.show()

         # ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/num
         py.argsort.html
         # -(train_class_distribution.values): the minus sign will give us in de
         creasing 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_distri
         bution.values[i], '(', np.round((train_class_distribution.values[i]/tra
         in_df.shape[0]*100), 3), '%)')


         print('-'*80)
         my_colors = ['r', 'g', 'b', 'k', 'y', 'm', 'c']
         test_class_distribution.plot(kind='bar', color=my_colors)
         plt.xlabel('Class')
         plt.ylabel('Data points per Class')
         plt.title('Distribution of yi in test data')
         plt.grid()
```

```python
plt.show()

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/num
py.argsort.html
# -(train_class_distribution.values): the minus sign will give us in de
creasing 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_distrib
ution.values[i], '(', np.round((test_class_distribution.values[i]/test_
df.shape[0]*100), 3), '%)')

print('-'*80)
my_colors = ['r', 'g', 'b', 'k', 'y', 'm', 'c']
cv_class_distribution.plot(kind='bar', color=my_colors)
plt.xlabel('Class')
plt.ylabel('Data points per Class')
plt.title('Distribution of yi in cross validation data')
plt.grid()
plt.show()

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/num
py.argsort.html
# -(train_class_distribution.values): the minus sign will give us in de
creasing order
sorted_yi = np.argsort(-cv_class_distribution.values)
for i in sorted_yi:
    print('Number of data points in class', i+1, ':',cv_class_distribut
ion.values[i], '(', np.round((cv_class_distribution.values[i]/cv_df.sha
pe[0]*100), 3), '%)')
```
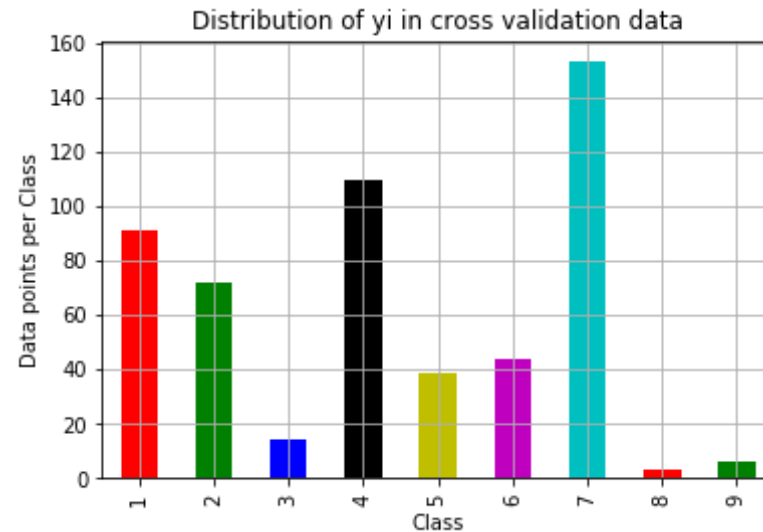

Distribution of yi in train data
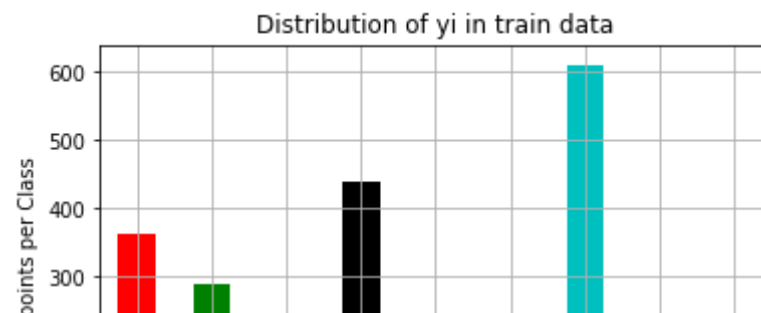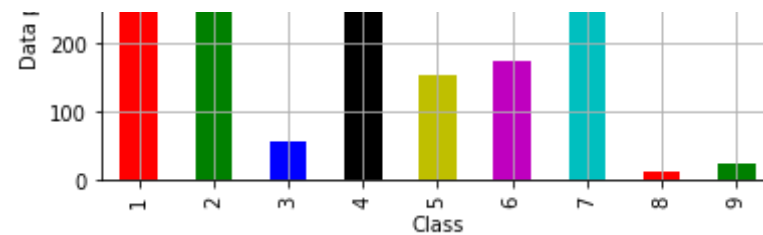
```
Number of data points in class 7 : 609 ( 28.672 %)
Number of data points in class 4 : 439 ( 20.669 %)

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 %)
----------------------------------------------------------------------
---------
```


Distribution of yi in test data
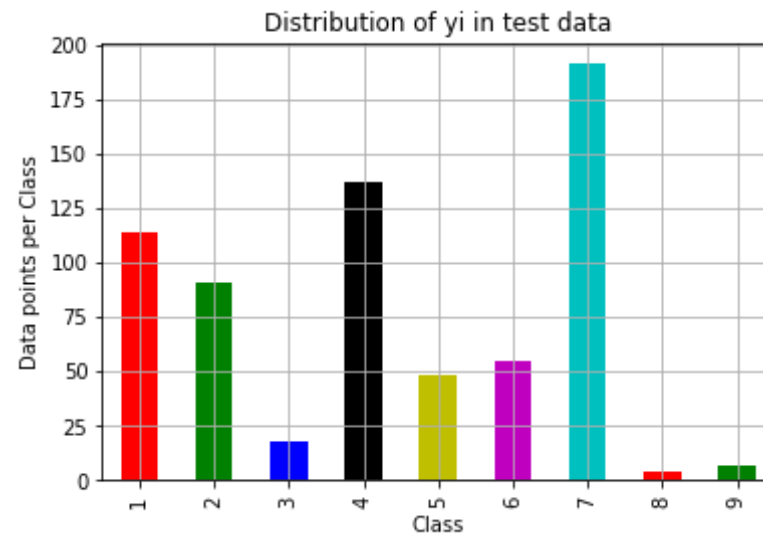
```
Number of data points in class 7 : 191 ( 28.722 %)
Number of data points in class 4 : 137 ( 20.602 %)
Number of data points in class 1 : 114 ( 17.143 %)
```

```
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 %)
-----------------------------------------------------------------------
---------
```


Distribution of yi in cross validation data

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

3.2 Prediction using a 'Random' Model

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

In [13]:
```python
# 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 correspo
nds 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 correspo
nds 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]]
```

```python
        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=la
bels, 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=la
bels, 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=la
bels, yticklabels=labels)
        plt.xlabel('Predicted Class')
        plt.ylabel('Original Class')
        plt.show()
```

In [16]:
```python
# 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])
```

```python
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_p
redicted_y, eps=1e-15))

predicted_y =np.argmax(test_predicted_y, axis=1)
plot_confusion_matrix(y_test, predicted_y+1)
```
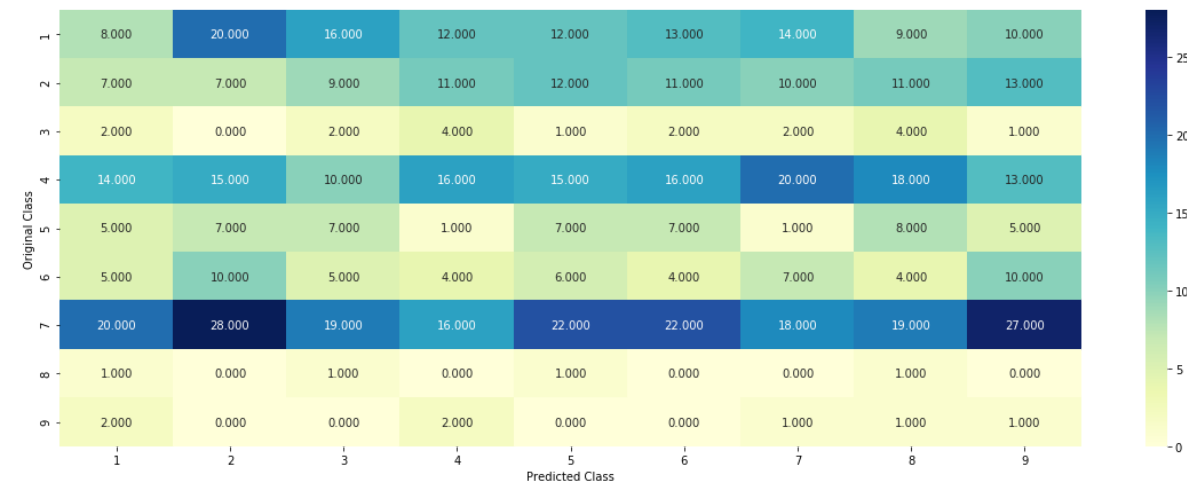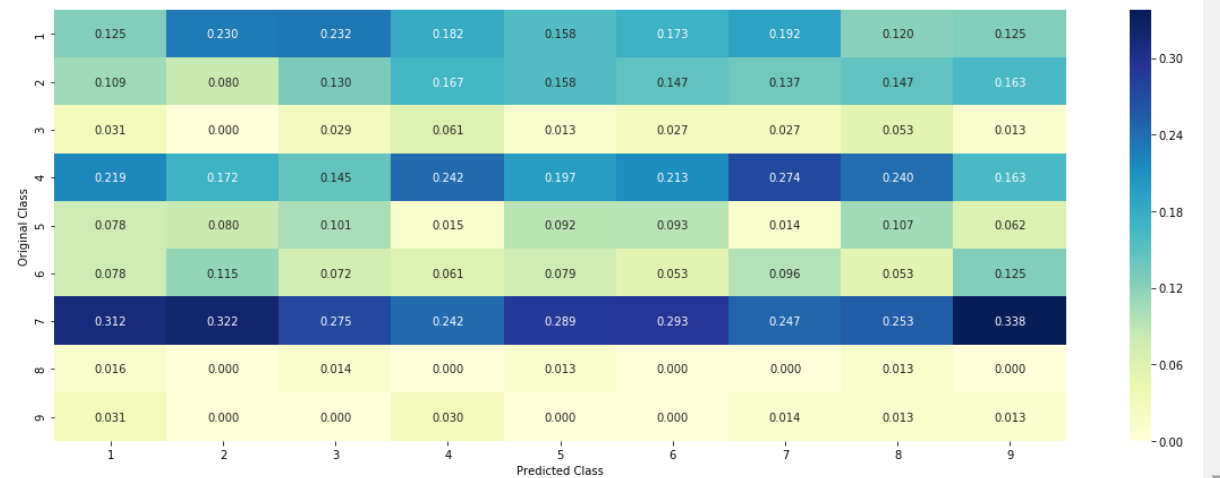
```
Log loss on Cross Validation Data using Random Model 2.45605457070145
94
Log loss on Test Data using Random Model 2.4552398240762274
------------------- Confusion matrix -------------------
```



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

Original Class \ Predicted Class	1	2	3	4	5	6	7	8	9
1	0.125	0.230	0.232	0.182	0.158	0.173	0.192	0.120	0.125
2	0.109	0.080	0.130	0.167	0.158	0.147	0.137	0.147	0.163
3	0.031	0.000	0.029	0.061	0.013	0.027	0.027	0.053	0.013
4	0.219	0.172	0.145	0.242	0.197	0.213	0.274	0.240	0.163
5	0.078	0.080	0.101	0.015	0.092	0.093	0.014	0.107	0.062
6	0.078	0.115	0.072	0.061	0.079	0.053	0.096	0.053	0.125
7	0.312	0.322	0.275	0.242	0.289	0.293	0.247	0.253	0.338
8	0.016	0.000	0.014	0.000	0.013	0.000	0.000	0.013	0.000
9	0.031	0.000	0.000	0.030	0.000	0.000	0.014	0.013	0.013

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

Original Class \ Predicted Class	1	2	3	4	5	6	7	8	9
1	0.070	0.175	0.140	0.105	0.105	0.114	0.123	0.079	0.088
2	0.077	0.077	0.099	0.121	0.132	0.121	0.110	0.121	0.143
3	0.111	0.000	0.111	0.222	0.056	0.111	0.111	0.222	0.056
4	0.102	0.109	0.073	0.117	0.109	0.117	0.146	0.131	0.095
5	0.104	0.146	0.146	0.021	0.146	0.146	0.021	0.167	0.104
6	0.091	0.182	0.091	0.073	0.109	0.073	0.127	0.073	0.182
7	0.105	0.147	0.099	0.084	0.115	0.115	0.094	0.099	0.141
8	0.250	0.000	0.250	0.000	0.250	0.000	0.000	0.250	0.000
9	0.286	0.000	0.000	0.286	0.000	0.000	0.143	0.143	0.143

3.3 Univariate Analysis

In [14]:
```python
# code for response coding with Laplace smoothing.
# alpha : used for laplace smoothing
```

```python
# feature: ['gene', 'variation']
# df: ['train_df', 'test_df', 'cv_df']
# algorithm
# ----------
# Consider all unique values and the number of occurances of given feat
ure in train data dataframe
# build a vector (1*9) , the first element = (number of times it occure
d in class1 + 10*alpha / number of time it occurred in total data+90*al
pha)
# gv_dict is like a look up table, for every gene it store a (1*9) repr
esentation 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_f
ea'
# 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
```

```python
    # 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 arr
ay for each gene/variation
    gv_dict = dict()

    # denominator will contain the number of time that particular featu
re occured in whole data
    for i, denominator in value_count.items():
        # vec will contain (p(yi==1/Gi) probability of gene/variation b
elongs 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['Ge
ne']=='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[f
eature]==i)]
```

```python
            # cls_cnt.shape[0](numerator) will contain the number of ti
me that particular feature occured in whole data
            vec.append((cls_cnt.shape[0] + alpha*10)/ (denominator + 90
*alpha))

        # we are adding the gene/variation to the dict as key and vec a
s 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, 0.068181
818181818177, 0.13636363636363635, 0.25, 0.19318181818181818, 0.0378787
8787878788, 0.03787878787878788, 0.03787878787878788],
    #       'TP53': [0.32142857142857145, 0.061224489795918366, 0.061224
489795918366, 0.27040816326530615, 0.061224489795918366, 0.066326530612
244902, 0.051020408163265307, 0.051020408163265307, 0.05612244897959183
7],
    #       'EGFR': [0.056818181818181816, 0.21590909090909091, 0.0625,
 0.068181818181818177, 0.068181818181818177, 0.0625, 0.3465909090909091
2, 0.0625, 0.056818181818181816],
    #       'BRCA2': [0.13333333333333333, 0.060606060606060608, 0.06060
6060606060608, 0.078787878787878782, 0.1393939393939394, 0.345454545454
54546, 0.060606060606060608, 0.060606060606060608, 0.06060606060606060
8],
    #       'PTEN': [0.069182389937106917, 0.062893081761006289, 0.06918
2389937106917, 0.46540880503144655, 0.075471698113207544, 0.06289308176
1006289, 0.069182389937106917, 0.062893081761006289, 0.0628930817610062
89],
    #       'KIT': [0.066225165562913912, 0.25165562913907286, 0.0728476
82119205295, 0.072847682119205295, 0.066225165562913912, 0.066225165562
913912, 0.2715231788079702, 0.066225165562913912, 0.06622516556291391
2],
    #       'BRAF': [0.06666666666666666, 0.17999999999999999, 0.073333
333333333334, 0.073333333333333334, 0.09333333333333338, 0.08000000000
0000002, 0.29999999999999999, 0.06666666666666666, 0.06666666666666666
6],
```

```
    #       ...
    #     }
    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 e
ach feature value in the data
    gv_fea = []
    # for every feature values in the given data frame we will check if
 it is there in the train data then we will add the feature to gv_fea
    # if not we will add [1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9] to gv_fe
a
    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 [15]: unique_genes = train_df['Gene'].value_counts()
         print('Number of Unique Genes :', unique_genes.shape[0])
```

```python
# the top 10 genes that occured most
print(unique_genes.head(10))
```

```
Number of Unique Genes : 224
BRCA1      183
TP53       103
EGFR        91
BRCA2       86
PTEN        76
KIT         65
BRAF        62
ERBB2       50
ALK         45
PIK3CA      37
Name: Gene, dtype: int64
```

In [16]:
```python
print("Ans: There are", unique_genes.shape[0] ,"different categories of
 genes in the train data, and they are distibuted as follows",)
```

```
Ans: There are 224 different categories of genes in the train data, and
they are distibuted as follows
```

In [17]:
```python
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()
```

Histrogram of Genes
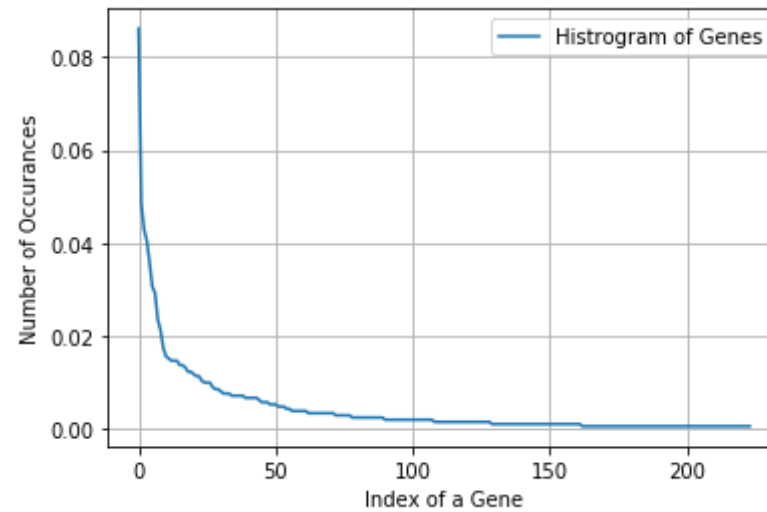
```
In [18]:  c = np.cumsum(h)
          plt.plot(c,label='Cumulative distribution of Genes')
          plt.grid()
          plt.legend()
          plt.show()
```


Cumulative distribution of Genes
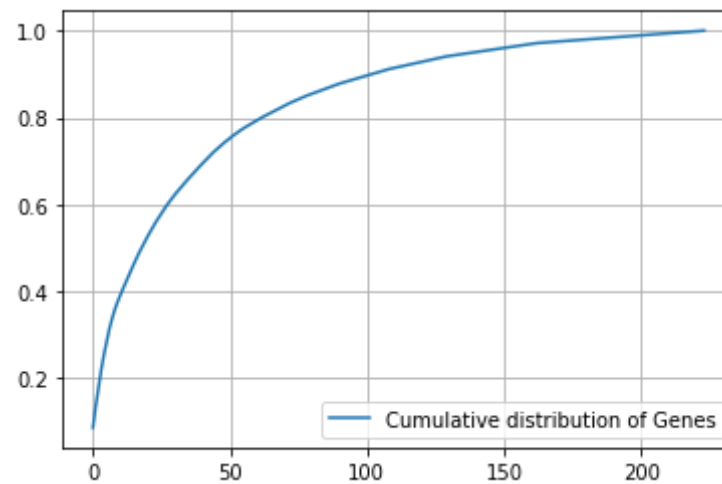
Q3. How to featurize this Gene feature ?

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

1. One hot Encoding
2. Response coding

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

```
In [19]: #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, "Gen
         e", train_df))
         # test gene feature
         test_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gen
         e", test_df))
         # cross validation gene feature
         cv_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene",
          cv_df))
```

```
In [20]: print("train_gene_feature_responseCoding is converted feature using res
         pone coding method. The shape of gene feature:", train_gene_feature_res
         ponseCoding.shape)
```

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

```
In [21]: # one-hot encoding of Gene feature.
         gene_vectorizer = TfidfVectorizer()
         train_gene_feature_onehotCoding = gene_vectorizer.fit_transform(train_d
         f['Gene'])
         test_gene_feature_onehotCoding = gene_vectorizer.transform(test_df['Gen
```

```
      e'])
      cv_gene_feature_onehotCoding = gene_vectorizer.transform(cv_df['Gene'])
```

In [22]: `train_df['Gene'].head()`

Out[22]:
```
1986      CTNNB1
1636      MAP2K4
2767        BRAF
2657       BRCA1
314         ELF3
Name: Gene, dtype: object
```

In [23]: `gene_vectorizer.get_feature_names()`

Out[23]:
```
['abl1',
 'acvr1',
 'ago2',
 'akt1',
 'akt2',
 'akt3',
 'alk',
 'apc',
 'ar',
 'araf',
 'arid1a',
 'arid1b',
 'arid2',
 'arid5b',
 'asxl1',
 'atm',
 'atr',
 'atrx',
 'aurka',
 'aurkb',
 'axin1',
 'b2m',
 'bap1',
 'bcl10',
 'bcl2',
```

```
'bcl2l11',
'bcor',
'braf',
'brca1',
'brca2',
'brd4',
'brip1',
'btk',
'card11',
'carm1',
'casp8',
'cbl',
'ccnd1',
'ccnd2',
'ccnd3',
'cdh1',
'cdk12',
'cdk4',
'cdk6',
'cdk8',
'cdkn1a',
'cdkn1b',
'cdkn2a',
'cdkn2b',
'cdkn2c',
'chek2',
'cic',
'crebbp',
'ctcf',
'ctnnb1',
'ddr2',
'dicer1',
'dnmt3a',
'dnmt3b',
'egfr',
'elf3',
'ep300',
'epas1',
'epcam',
```

```
'erbb2',
'erbb3',
'erbb4',
'ercc2',
'ercc3',
'ercc4',
'erg',
'esr1',
'etv1',
'etv6',
'ewsr1',
'ezh2',
'fanca',
'fat1',
'fbxw7',
'fgf19',
'fgfr1',
'fgfr2',
'fgfr3',
'fgfr4',
'flt3',
'foxa1',
'foxl2',
'foxp1',
'fubp1',
'gata3',
'gnaq',
'gnas',
'h3f3a',
'hist1h1c',
'hla',
'hnf1a',
'hras',
'idh1',
'idh2',
'inpp4b',
'jak1',
'jak2',
'jun',
```

```
                        'kdm5a',
                        'kdm6a',
                        'kdr',
                        'keap1',
                        'kit',
                        'klf4',
                        'kmt2a',
                        'kmt2b',
                        'kmt2c',
                        'knstrn',
                        'kras',
                        'lats2',
                        'map2k1',
                        'map2k2',
                        'map2k4',
                        'map3k1',
                        'mapk1',
                        'med12',
                        'mef2b',
                        'men1',
                        'met',
                        'mga',
                        'mlh1',
                        'mpl',
                        'msh2',
                        'msh6',
                        'mtor',
                        'myc',
                        'mycn',
                        'nf1',
                        'nf2',
                        'nfe2l2',
                        'nfkbia',
                        'nkx2',
                        'notch1',
                        'notch2',
                        'npm1',
                        'nras',
                        'nsd1',
```

```
'ntrk1',
'ntrk3',
'nup93',
'pak1',
'pax8',
'pbrm1',
'pdgfra',
'pdgfrb',
'pik3ca',
'pik3cb',
'pik3cd',
'pik3r1',
'pik3r2',
'pik3r3',
'pim1',
'pms2',
'pole',
'ppm1d',
'ppp2r1a',
'ppp6c',
'prdm1',
'ptch1',
'pten',
'ptpn11',
'ptprd',
'ptprt',
'rab35',
'rac1',
'rad21',
'rad50',
'rad51b',
'rad51c',
'rad51d',
'rad54l',
'raf1',
'rasa1',
'rb1',
'rbm10',
'ret',
```

```
                              'rheb',
                              'rhoa',
                              'rictor',
                              'rit1',
                              'rnf43',
                              'ros1',
                              'rras2',
                              'runx1',
                              'rxra',
                              'rybp',
                              'sdhc',
                              'setd2',
                              'sf3b1',
                              'shq1',
                              'smad2',
                              'smad3',
                              'smad4',
                              'smarca4',
                              'smarcb1',
                              'smo',
                              'sos1',
                              'sox9',
                              'spop',
                              'src',
                              'srsf2',
                              'stag2',
                              'stat3',
                              'stk11',
                              'tert',
                              'tet2',
                              'tgfbr1',
                              'tgfbr2',
                              'tmprss2',
                              'tp53',
                              'tp53bp1',
                              'tsc1',
                              'tsc2',
                              'u2af1',
                              'vegfa',
```

```
            'vhl',
            'xpo1',
            'xrcc2',
            'yap1']
```

In [24]:
```python
print("train_gene_feature_onehotCoding is converted feature using one-h
ot encoding method. The shape of gene feature:", train_gene_feature_one
hotCoding.shape)
```

```
train_gene_feature_onehotCoding is converted feature using one-hot enco
ding method. The shape of gene feature: (2124, 224)
```

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

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

In [25]:
```python
alpha = [10 ** x for x in range(-5, 1)] # hyperparam for SGD classifie
r.

# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# ----------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_rate='optimal', eta0=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

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

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

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

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


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

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

```
))
predict_y = sig_clf.predict_proba(cv_gene_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross vali
dation log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_, eps
=1e-15))
predict_y = sig_clf.predict_proba(test_gene_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log l
oss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.1590184964124919
For values of alpha =  0.0001 The log loss is: 1.1284964366531633
For values of alpha =  0.001 The log loss is: 1.173934730626959
For values of alpha =  0.01 The log loss is: 1.3089558839857254
For values of alpha =  0.1 The log loss is: 1.4161616357713804
For values of alpha =  1 The log loss is: 1.4588133497400888
```
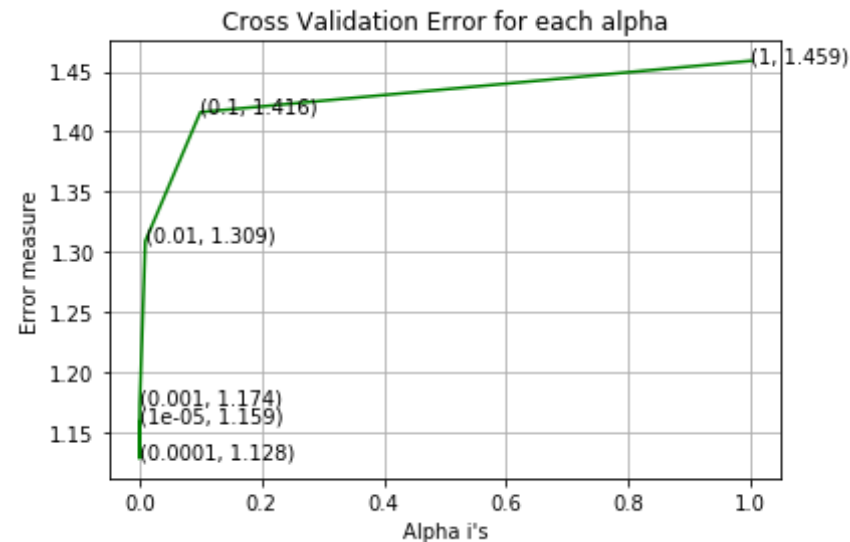


```
For values of best alpha =  0.0001 The train log loss is: 1.01540115367
2015
For values of best alpha =  0.0001 The cross validation log loss is: 1.
1284964366531633
For values of best alpha =  0.0001 The test log loss is: 1.218627734943
0026
```

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

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

In [26]:
```python
print("Q6. How many data points in Test and CV datasets are covered by
 the ", unique_genes.shape[0], " genes in train dataset?")

test_coverage=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'])))].shap
e[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  22
4  genes in train dataset?
Ans
1. In test data 636 out of 665 : 95.6390977443609
2. In cross validation data 507 out of  532 : 95.30075187969925
```

3.2.2 Univariate Analysis on Variation Feature

Q7. Variation, What type of feature is it ?

Ans. Variation is a categorical variable

Q8. How many categories are there?

In [27]:
```python
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 : 1939
Truncating_Mutations    56
Deletion                44
Amplification           41
Fusions                 21
Overexpression           4
Q61L                     3
G12V                     3
Q22K                     2
Y42C                     2
T286A                    2
Name: Variation, dtype: int64
```

In [28]:
```
print("Ans: There are", unique_variations.shape[0] ,"different categori
es of variations in the train data, and they are distibuted as follows"
,)
```

Ans: There are 1939 different categories of variations in the train dat
a, and they are distibuted as follows

In [29]:
```
s = sum(unique_variations.values);
h = unique_variations.values/s;
plt.plot(h, label="Histrogram of Variations")
plt.xlabel('Index of a Variation')
plt.ylabel('Number of Occurances')
plt.legend()
plt.grid()
plt.show()
```
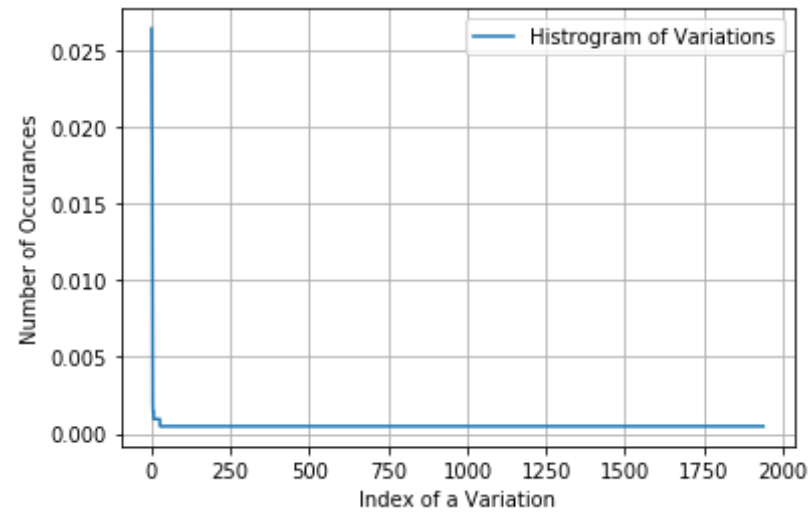
```
In [30]: c = np.cumsum(h)
         print(c)
         plt.plot(c,label='Cumulative distribution of Variations')
         plt.grid()
         plt.legend()
         plt.show()
```

[0.02636535 0.04708098 0.06638418 ... 0.99905838 0.99952919 1.   ]
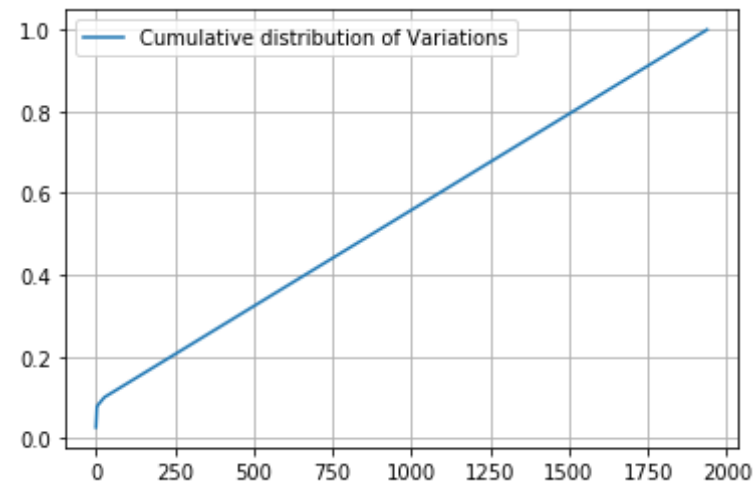
Q9. How to featurize this Variation feature ?

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

1. One hot Encoding
2. Response coding

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

In [31]:
```python
# 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, "V
ariation", cv_df))
```

In [32]:
```python
print("train_variation_feature_responseCoding is a converted feature us
ing the response coding method. The shape of Variation feature:", train
_variation_feature_responseCoding.shape)
```

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

In [33]:
```python
# one-hot encoding of variation feature.
variation_vectorizer = TfidfVectorizer()
train_variation_feature_onehotCoding = variation_vectorizer.fit_transfo
rm(train_df['Variation'])
test_variation_feature_onehotCoding = variation_vectorizer.transform(te
st_df['Variation'])
```

```
cv_variation_feature_onehotCoding = variation_vectorizer.transform(cv_d
f['Variation'])
```

In [34]:
```
print("train_variation_feature_onehotEncoded is converted feature using
 the onne-hot encoding method. The shape of Variation feature:", train_
variation_feature_onehotCoding.shape)
```

train_variation_feature_onehotEncoded is converted feature using the on
ne-hot encoding method. The shape of Variation feature: (2124, 1971)

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

Let's build a model just like the earlier!

In [35]:
```
alpha = [10 ** x for x in range(-5, 1)]

# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_rate='optimal', eta0=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

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

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


cv_log_error_array=[]
```

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

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


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

predict_y = sig_clf.predict_proba(train_variation_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log
 loss is:",log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15
))
predict_y = sig_clf.predict_proba(cv_variation_feature_onehotCoding)
```

```
print('For values of best alpha = ', alpha[best_alpha], "The cross vali
dation log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_, eps
=1e-15))
predict_y = sig_clf.predict_proba(test_variation_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log l
oss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

For values of alpha = 1e-05 The log loss is: 1.6895004561097002

For values of alpha = 0.0001 The log loss is: 1.683489213094863
For values of alpha = 0.001 The log loss is: 1.6911384771032505
For values of alpha = 0.01 The log loss is: 1.6970520700607605
For values of alpha = 0.1 The log loss is: 1.710108994133302
For values of alpha = 1 The log loss is: 1.7117350010700811
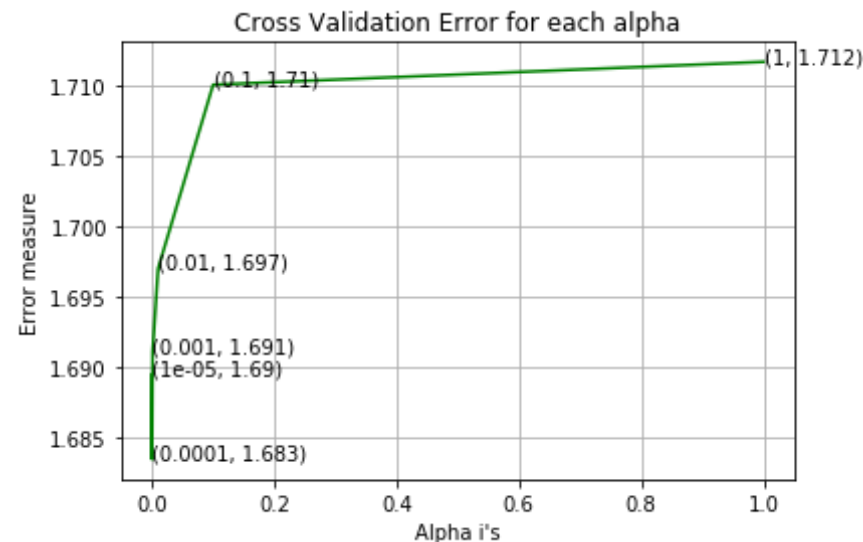
For values of best alpha = 0.0001 The train log loss is: 0.74626785315
63928
For values of best alpha = 0.0001 The cross validation log loss is: 1.
683489213094863
For values of best alpha = 0.0001 The test log loss is: 1.703803067322
2617

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 [36]:
```python
print("Q12. How many data points are covered by total ", unique_variati
ons.shape[0], " genes in test and cross validation data sets?")
test_coverage=test_df[test_df['Variation'].isin(list(set(train_df['Vari
ation'])))].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  1939  genes in test and
cross validation data sets?
Ans
1. In test data 76 out of 665 : 11.428571428571429
2. In cross validation data 61 out of  532 : 11.466165413533833
```

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 [37]:
```python
# cls_text is a data frame
# for every row in data fram consider the 'TEXT'
# split the words by space
# make a dict with those words
# increment its count whenever we see that word
```

```python
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 [38]:
```python
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 )/(t
otal_dict.get(word,0)+90)))
            text_feature_responseCoding[row_index][i] = math.exp(sum_pr
ob/len(row['TEXT'].split()))
            row_index += 1
    return text_feature_responseCoding
```

In [41]:
```python
def top_tfidf_feats(row, features, top_n=25):
    ''' Get top n tfidf values in row and return them with their corres
ponding feature names.'''
    topn_ids = np.argsort(row)[::-1][:top_n]
    top_feats = [(features[i], row[i]) for i in topn_ids]
    df = pd.DataFrame(top_feats)
    df.columns = ['feature', 'tfidf']
    return df

def top_mean_feats(Xtr, features, min_tfidf=0.1, grp_ids=None, top_n=25
):
    ''' Return the top n features that on average are most important am
ongst documents in rows
        indentified by indices in grp_ids. '''
```

```
            if grp_ids:
                D = Xtr[grp_ids].toarray()
            else:
                D = Xtr.toarray()

            D[D < min_tfidf] = 0
            tfidf_means = np.mean(D, axis=0)
            return top_tfidf_feats(tfidf_means, features, top_n)
```

In [39]:
```
# building a TfidfVectorizer with all the words that occured minimum 3
 times in train data
text_vectorizer = TfidfVectorizer(min_df=3,max_features=1000)
train_text_feature_onehotCoding = text_vectorizer.fit_transform(train_d
f['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 num
ber of times it occured
text_fea_dict = dict(zip(list(train_text_features),train_text_fea_count
s))


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

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

In [40]:
```
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
```

```python
# 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 [41]:
```python
#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 [42]:
```python
# https://stackoverflow.com/a/16202486
# we convert each row values such that they sum to 1
train_text_feature_responseCoding = (train_text_feature_responseCoding.
T/train_text_feature_responseCoding.sum(axis=1)).T
test_text_feature_responseCoding = (test_text_feature_responseCoding.T/
test_text_feature_responseCoding.sum(axis=1)).T
cv_text_feature_responseCoding = (cv_text_feature_responseCoding.T/cv_t
ext_feature_responseCoding.sum(axis=1)).T
```

In [43]:
```python
# don't forget to normalize every feature
train_text_feature_onehotCoding = normalize(train_text_feature_onehotCo
ding, axis=0)

# we use the same vectorizer that was trained on train data
test_text_feature_onehotCoding = text_vectorizer.transform(test_df['TEX
T'])
# don't forget to normalize every feature
test_text_feature_onehotCoding = normalize(test_text_feature_onehotCodi
```

```
ng, 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 [44]:
```
#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 [45]:
```
# Number of words for a given frequency.
print(Counter(sorted_text_occur))
```

```
Counter({250.03474732415123: 1, 175.96818638881825: 1, 143.722562467840
45: 1, 130.39550874780656: 1, 124.9578217075219: 1, 117.18264881724652:
1, 116.3826276578983: 1, 113.71616552517335: 1, 113.06872979018696: 1,
110.77463665502441: 1, 104.3500437905872: 1, 87.69152110392537: 1, 87.1
9390650483537: 1, 86.34082136723569: 1, 80.66954591682727: 1, 80.448769
83976017: 1, 79.31487609098448: 1, 77.73395416881046: 1, 77.60256865270
341: 1, 77.4269016201933: 1, 76.36267008280785: 1, 74.76809167262614:
1, 71.3970319026008: 1, 68.57000073572546: 1, 67.05064676533851: 1, 66.
88617010132002: 1, 66.3857145835632: 1, 65.25775713195524: 1, 65.052931
62314706: 1, 63.2131725392625: 1, 63.13512403214793: 1, 62.99464559903
6: 1, 62.82250261994918: 1, 60.77364465158595: 1, 58.50331076289059: 1,
55.746323449705926: 1, 55.2297254941024: 1, 53.05961613732826: 1, 52.61
143243266872: 1, 51.339873924872: 1, 51.219815740686144: 1, 50.72807977
082189: 1, 49.588702345197966: 1, 49.026506602094685: 1, 47.66821306470
7075: 1, 47.15580842143257: 1, 45.70545734464458: 1, 45.42233277906975
5: 1, 45.05097364439181: 1, 44.70870929661169: 1, 44.188167429034316:
1, 43.08313645683813: 1, 43.068331775997066: 1, 43.04322366579178: 1, 4
2.80096063295079: 1, 42.698030511701134: 1, 42.607664811007155: 1, 42.4
8082839053412: 1, 42.21226097258875: 1, 42.1245457270988: 1, 41.2457735
0009634: 1, 41.05803583815108: 1, 40.96473567810854: 1, 40.525295862488
68: 1, 39.89778393774366: 1, 39.69630932506647: 1, 39.684991649627015:
1, 38.29196984046447: 1, 38.25675149708367: 1, 38.25357593629001: 1, 3
8.14499185480327: 1, 38.10051013967206: 1, 37.84185584070669: 1, 37.521
```

04574152389: 1, 37.35188676352496: 1, 36.952389544009314: 1, 36.3005438
38073025: 1, 35.34193715478422: 1, 35.267535783748855: 1, 34.9607836562
97764: 1, 34.9175867503184: 1, 34.90205567841248: 1, 34.8765194772418:
1, 34.710913334107694: 1, 34.52153248021165: 1, 34.39141791440859: 1, 3
3.79569743049251: 1, 33.57749414222369: 1, 33.26316434804345: 1, 33.182
61632690109: 1, 32.644661197842495: 1, 32.5883702132252: 1, 32.39129505
8448684: 1, 32.3823468846539: 1, 32.24542426166109: 1, 32.1116865579334
8: 1, 32.08064196256643: 1, 31.995911444966893: 1, 31.99470646097628:
1, 31.989531559124316: 1, 31.980403956873847: 1, 31.944754074070637: 1,
31.767549345338004: 1, 31.692549562996458: 1, 31.433421583571466: 1, 3
1.195177413646768: 1, 31.088306203843523: 1, 30.968846061889668: 1, 30.
841125196447326: 1, 30.761973881845805: 1, 30.6277626139082: 1, 30.5783
27184802273: 1, 30.546413785997146: 1, 30.380924264443664: 1, 30.364148
66177054: 1, 30.04931444620254: 1, 30.034263877089902: 1, 29.9791064761
7556: 1, 29.819315526117556: 1, 29.729422521384805: 1, 29.4310509181258
87: 1, 29.413957430761602: 1, 29.385531247824144: 1, 29.0394155714509:
1, 28.86079974425628: 1, 28.732703309872214: 1, 28.41817556144125: 1, 2
8.40907299152425: 1, 28.39455420223587: 1, 28.233808389166196: 1, 28.19
5383466632958: 1, 28.13285125637795: 1, 27.79442764399916: 1, 27.612718
30410803: 1, 27.570712387723056: 1, 27.484013674941764: 1, 27.398644064
892387: 1, 27.21141008593759: 1, 27.14466046152745: 1, 26.9698773560610
87: 1, 26.95921255409716: 1, 26.85979147689292: 1, 26.643714897861177:
1, 26.253167536063: 1, 26.157890838954323: 1, 25.95340746477933: 1, 25.
93266312682881: 1, 25.876808293501522: 1, 25.83974959008824: 1, 25.7166
05246009703: 1, 25.694575379040298: 1, 25.682945586052092: 1, 25.614314
889160006: 1, 25.43868084934155: 1, 25.345251927465686: 1, 25.340648221
4361: 1, 25.308986757210512: 1, 25.28276892189394: 1, 25.1905112802507
9: 1, 25.151593583755417: 1, 25.15151737565609: 1, 25.134828598852604:
1, 25.07533481923498: 1, 24.983819976273086: 1, 24.86546281808818: 1, 2
4.638472012489483: 1, 24.5627747791463: 1, 24.32173799785921: 1, 24.308
67079330682: 1, 24.231926138923278: 1, 24.113554124506216: 1, 24.045373
13587733: 1, 23.99600061540534: 1, 23.919593717391145: 1, 23.9071678112
10318: 1, 23.8866253494004: 1, 23.733088197338922: 1, 23.70875944989663
6: 1, 23.615295496331093: 1, 23.444537556773483: 1, 23.439152509879527:
1, 23.32368358517786: 1, 23.28327182156675: 1, 23.105352862879915: 1, 2
3.093484185917575: 1, 23.020212025986265: 1, 22.997457450356773: 1, 22.
98584935227899: 1, 22.78822695154163: 1, 22.659362801415803: 1, 22.5719
1599100145: 1, 22.566967989216057: 1, 22.47882595270522: 1, 22.47384693
6559195: 1, 22.466771360298054: 1, 22.389424264510094: 1, 22.3565499864

40644: 1, 22.32476854320074: 1, 22.30507111877814: 1, 22.126808148482873: 1, 22.001581775642563: 1, 21.987307356016878: 1, 21.90904345085246: 1, 21.895509883489844: 1, 21.87061097224795: 1, 21.825122090120257: 1, 21.747830169923507: 1, 21.74144537219869: 1, 21.673436003849186: 1, 21.617602965908045: 1, 21.580708239814477: 1, 21.517375577853826: 1, 21.485796744054063: 1, 21.46044653354556: 1, 21.41579334672493: 1, 21.413267302708554: 1, 21.405990644501575: 1, 21.316899517512233: 1, 21.304552550516693: 1, 21.29564417299079: 1, 21.258161955266566: 1, 21.243199218636768: 1, 21.2079752629721: 1, 21.20202478724957: 1, 21.198947167149118: 1, 21.181372255577326: 1, 21.163652575183608: 1, 21.087173581219417: 1, 20.99437876479766: 1, 20.944401603868542: 1, 20.93292348606574: 1, 20.902276795370014: 1, 20.892162474088572: 1, 20.78644739761582: 1, 20.72993608739457: 1, 20.675027752844283: 1, 20.592170876469165: 1, 20.586928336401435: 1, 20.341261155589013: 1, 20.32852912305122: 1, 20.30463506235834: 1, 20.237552505308162: 1, 20.19276124563212: 1, 20.17757739780441: 1, 20.100769588460135: 1, 20.071202621296656: 1, 19.969551797202012: 1, 19.961866421862887: 1, 19.956013045199615: 1, 19.95521168222404: 1, 19.893453629645467: 1, 19.893349857053046: 1, 19.8859678984311: 1, 19.883168937871904: 1, 19.809088937854774: 1, 19.773493935718626: 1, 19.731742215305882: 1, 19.720989449698767: 1, 19.67474191254145: 1, 19.65211589434607: 1, 19.6507428028312: 1, 19.627021715356086: 1, 19.568487074355183: 1, 19.514441068521986: 1, 19.48128416608206: 1, 19.443369461302904: 1, 19.326406954971027: 1, 19.26013042792094: 1, 19.24911915740106: 1, 19.210724179019095: 1, 19.153289666189163: 1, 19.141765929727416: 1, 19.120864358833185: 1, 19.049186130892277: 1, 18.95050512720911: 1, 18.902035086251303: 1, 18.806864753319907: 1, 18.793105889460907: 1, 18.765234354750746: 1, 18.68562949556452: 1, 18.617165799603107: 1, 18.604515085846586: 1, 18.559747312312986: 1, 18.552863166113973: 1, 18.543493621110457: 1, 18.534799624016177: 1, 18.42479808818444: 1, 18.424013315433335: 1, 18.405553191416683: 1, 18.36724508178581: 1, 18.363963001358446: 1, 18.348965622964457: 1, 18.32414542435796: 1, 18.21255391733154: 1, 18.17614302530952: 1, 18.16047411083081: 1, 18.134257536970914: 1, 18.126945238600946: 1, 18.110702406949684: 1, 18.096953500218195: 1, 18.03000838592906: 1, 18.008446521373425: 1, 18.00656219190012: 1, 17.955976784319013: 1, 17.950801811978508: 1, 17.88692280050977: 1, 17.87316385729534: 1, 17.786223519204523: 1, 17.774377690029944: 1, 17.750253691563252: 1, 17.743482225689668: 1, 17.743292410733105: 1, 17.716990217451738: 1, 17.680795519760117: 1, 17.662574277742692: 1, 17.6322284779241: 1, 17.610193631117472: 1, 17.552309061634222: 1, 17.537553475681523: 1,

17.485915138666694: 1, 17.408378099484942: 1, 17.368394421097285: 1, 1
7.335333868572278: 1, 17.313440442698866: 1, 17.23788559860935: 1, 17.1
79665716749273: 1, 17.1316902700721: 1, 17.107878427829995: 1, 17.08273
6501905533: 1, 17.063546654147583: 1, 17.047448054180517: 1, 17.0177938
6541951: 1, 17.01180293151438: 1, 17.003692405580953: 1, 16.99343642614
5887: 1, 16.968461095933353: 1, 16.934874246461177: 1, 16.8685354491585
1: 1, 16.841624556685257: 1, 16.83921406459264: 1, 16.806358626725434:
1, 16.784930794882307: 1, 16.76177801995078: 1, 16.70084121386191: 1, 1
6.63668576012497: 1, 16.62264806035414: 1, 16.60268388003539: 1, 16.575
728653188218: 1, 16.57553862468636: 1, 16.57323767211677: 1, 16.5678462
04175996: 1, 16.52999908501935: 1, 16.515025445831153: 1, 16.4753453593
85177: 1, 16.464012484392708: 1, 16.44426194381821: 1, 16.4301790083116
54: 1, 16.39407346500856: 1, 16.393708639585007: 1, 16.323479312091262:
1, 16.297427501753294: 1, 16.26881405163095: 1, 16.212176177416087: 1,
16.089105206304144: 1, 16.07592538354891: 1, 16.05674792762714: 1, 16.0
38473273089004: 1, 15.98551397954115: 1, 15.975296000152392: 1, 15.9571
71023992679: 1, 15.944012311528848: 1, 15.88381336599075: 1, 15.8619588
98107714: 1, 15.84964337357056: 1, 15.848337854846497: 1, 15.7437588829
6365: 1, 15.743025187212636: 1, 15.693010102485045: 1, 15.6532404725540
07: 1, 15.65090932171562: 1, 15.65079410171174: 1, 15.648166033762228:
1, 15.632188636837897: 1, 15.581119860540973: 1, 15.57434251384246: 1,
15.547024052573141: 1, 15.521487309395559: 1, 15.492467684602806: 1, 1
5.478036221034593: 1, 15.46411759934719: 1, 15.360560530690675: 1, 15.2
89590681599513: 1, 15.28000730784837: 1, 15.269907328521928: 1, 15.2177
14780797497: 1, 15.215358524994564: 1, 15.17957817465293: 1, 15.1719012
64867897: 1, 15.115932879639807: 1, 15.105957514428392: 1, 15.096230445
835594: 1, 15.093587987196111: 1, 15.078970493585109: 1, 15.07254247009
6396: 1, 15.067658897630547: 1, 15.028206791553252: 1, 14.9413300360572
52: 1, 14.933911447760307: 1, 14.918163997702203: 1, 14.91312467883821
4: 1, 14.897718596455407: 1, 14.886458738179064: 1, 14.884120610107335:
1, 14.882933922548387: 1, 14.87598811967212: 1, 14.873236450289184: 1,
14.86546107029632: 1, 14.812230032468833: 1, 14.793722314265446: 1, 14.
771318753205302: 1, 14.755151179607797: 1, 14.744331657501492: 1, 14.73
8417663228741: 1, 14.723020648212223: 1, 14.71355248703073: 1, 14.68112
3991182265: 1, 14.675406601078977: 1, 14.652021917795027: 1, 14.6450781
35571085: 1, 14.640990780760383: 1, 14.623754459117844: 1, 14.614626758
317034: 1, 14.598194655450751: 1, 14.593625846552639: 1, 14.58669305866
4716: 1, 14.574308733758398: 1, 14.561937819682178: 1, 14.5509124121259
87: 1, 14.527150733436017: 1, 14.50875188053441: 1, 14.505531448761447:

1, 14.504218507882278: 1, 14.498705547725425: 1, 14.473741978223204: 1, 14.418583091547157: 1, 14.398933455692733: 1, 14.364917596892191: 1, 14.347282648381055: 1, 14.345305064890535: 1, 14.33735037169306: 1, 14.309785294854574: 1, 14.231813761316424: 1, 14.217642748745009: 1, 14.154999372738494: 1, 14.143841887944932: 1, 14.141623121384892: 1, 14.129229213817796: 1, 14.128191761432763: 1, 14.097075666301345: 1, 14.073362523960913: 1, 14.07328781890436: 1, 14.057431573717128: 1, 14.024031271870763: 1, 14.017741002926257: 1, 13.976619156342778: 1, 13.954811254229016: 1, 13.93174589802636: 1, 13.916710444643034: 1, 13.900541021674842: 1, 13.879618301326795: 1, 13.852562928199067: 1, 13.8495271658734: 1, 13.808861739835836: 1, 13.801727823648793: 1, 13.799955652438285: 1, 13.79798355934972: 1, 13.78207788134253: 1, 13.781887318904358: 1, 13.727378036507291: 1, 13.689371588032547: 1, 13.599133367304667: 1, 13.555154182653789: 1, 13.540042696929406: 1, 13.534589078790384: 1, 13.515489633944492: 1, 13.499358384027472: 1, 13.486496765952703: 1, 13.473678212335615: 1, 13.473069233008722: 1, 13.466438950245502: 1, 13.437003348195633: 1, 13.378465343503734: 1, 13.374658856088931: 1, 13.357723751456215: 1, 13.350546169371956: 1, 13.324066544171975: 1, 13.31857411242175: 1, 13.312842840786107: 1, 13.3060229202557: 1, 13.294399388016696: 1, 13.292245972231788: 1, 13.259899220839186: 1, 13.247741370809562: 1, 13.19151001538211: 1, 13.167123411199313: 1, 13.165226142570287: 1, 13.135330126907453: 1, 13.105837325392875: 1, 13.102074861062508: 1, 13.08184605855811: 1, 13.080334686822649: 1, 13.076349435703982: 1, 12.982060135205224: 1, 12.979451281256248: 1, 12.945701561576186: 1, 12.867073518461948: 1, 12.844130086697925: 1, 12.80467274073049: 1, 12.804354801521484: 1, 12.750518952430857: 1, 12.74805436609893: 1, 12.691586980732717: 1, 12.688711223238187: 1, 12.6814410501988: 1, 12.667227057707317: 1, 12.652370335739578: 1, 12.65217882336745: 1, 12.647728742586299: 1, 12.612308905896194: 1, 12.59233825785606: 1, 12.579322301536065: 1, 12.571089490256856: 1, 12.528010912047776: 1, 12.516288521454266: 1, 12.482967206063806: 1, 12.474233294539566: 1, 12.467037707575471: 1, 12.442516575725325: 1, 12.431415417869475: 1, 12.428978593603604: 1, 12.417677042800833: 1, 12.390195260484587: 1, 12.389410758353959: 1, 12.365469960605058: 1, 12.342302902251392: 1, 12.320987137842783: 1, 12.310633549254591: 1, 12.29051220733526: 1, 12.287259932062232: 1, 12.26937210780466: 1, 12.26800703272652: 1, 12.254884798023886: 1, 12.225714808805247: 1, 12.196776080261909: 1, 12.184767079382155: 1, 12.177358528318056: 1, 12.144617178724827: 1, 12.142311904740325: 1, 12.124701671227662: 1, 12.12081866392057: 1, 12.115455370475939: 1, 12.061839535796423: 1, 11.9

96814977674834: 1, 11.990410945643667: 1, 11.980265917348929: 1, 11.973
522532641786: 1, 11.969187298430441: 1, 11.967432946091032: 1, 11.96275
5751720614: 1, 11.950814606318318: 1, 11.933567441260097: 1, 11.9302402
68735604: 1, 11.911674371042047: 1, 11.8784871058399: 1, 11.85985138463
527: 1, 11.813354431405877: 1, 11.760833829267014: 1, 11.73809336342106
1: 1, 11.70425737591388: 1, 11.703190607964128: 1, 11.674400169256218:
1, 11.67384488974617: 1, 11.660366294007224: 1, 11.62065046586628: 1, 1
1.618634440691455: 1, 11.605486111668748: 1, 11.591821738919606: 1, 11.
578639730209511: 1, 11.545900994034634: 1, 11.524107006236624: 1, 11.49
3374265991415: 1, 11.488518494237182: 1, 11.487568464806476: 1, 11.4830
10505038102: 1, 11.478667428484364: 1, 11.45968270091798: 1, 11.4425939
9511333: 1, 11.34790819757489: 1, 11.343877780336479: 1, 11.34019171551
2692: 1, 11.314839317868964: 1, 11.297537107200382: 1, 11.2905605514721
4: 1, 11.26868554481359: 1, 11.268527506680437: 1, 11.263327810575744:
1, 11.261729956336525: 1, 11.252266949113212: 1, 11.247627678293922: 1,
11.24621574391395: 1, 11.211123839601678: 1, 11.206201151432701: 1, 11.
201267905866027: 1, 11.196703296460544: 1, 11.17301493124604: 1, 11.164
498315539054: 1, 11.142439009675423: 1, 11.1009838036381: 1, 11.0945063
99679277: 1, 11.082384303078205: 1, 11.081139738569759: 1, 11.076124823
546673: 1, 11.075682199307554: 1, 11.072252776355722: 1, 11.05244417748
57: 1, 11.036981711955283: 1, 11.031628212811047: 1, 11.02090940573841
3: 1, 11.012936812584117: 1, 10.986890514650943: 1, 10.93684369456662:
1, 10.929206369616743: 1, 10.928695402318795: 1, 10.915193766473047: 1,
10.87890019815634: 1, 10.878555720028439: 1, 10.872122428585424: 1, 10.
864651032490604: 1, 10.835304272908958: 1, 10.835193563984872: 1, 10.83
4427186606032: 1, 10.806577213149478: 1, 10.803635992652007: 1, 10.7825
070507625: 1, 10.777918779619759: 1, 10.776959842278549: 1, 10.77200744
9724775: 1, 10.771263202190092: 1, 10.765112138502937: 1, 10.7428921587
15645: 1, 10.730299296426187: 1, 10.719677015283308: 1, 10.713713185519
58: 1, 10.700412141712784: 1, 10.693009685434204: 1, 10.65958767931870
8: 1, 10.652447907710483: 1, 10.649467982379631: 1, 10.629522435781617:
1, 10.607373385825934: 1, 10.606246677210015: 1, 10.602990370744855: 1,
10.560038987126992: 1, 10.552740644609864: 1, 10.549486921173285: 1, 1
0.53797240289367: 1, 10.537535385918403: 1, 10.53077949110144: 1, 10.52
1210499718494: 1, 10.52059877598766: 1, 10.516503442313994: 1, 10.50113
7974401406: 1, 10.498721771920925: 1, 10.490421077789076: 1, 10.4700951
1244476: 1, 10.445519064321493: 1, 10.444111506684287: 1, 10.4360181593
8943: 1, 10.43468830473354: 1, 10.428941016943197: 1, 10.39666771510527
9: 1, 10.355546686072646: 1, 10.341611948498416: 1, 10.33611322314327:

1, 10.316132284234989: 1, 10.313739126457346: 1, 10.308958188845915: 1, 10.300083114252022: 1, 10.29862015502261: 1, 10.278794936933375: 1, 10.26993185350685: 1, 10.266373523012765: 1, 10.256855002202363: 1, 10.254166090646699: 1, 10.244456397374849: 1, 10.238699074277408: 1, 10.23274822136888: 1, 10.232470909109955: 1, 10.221200717684615: 1, 10.207916843644563: 1, 10.15566930531784: 1, 10.149576175721768: 1, 10.145014517792424: 1, 10.129855873948916: 1, 10.129229187853674: 1, 10.115967860179701: 1, 10.11101988995488: 1, 10.110572202034286: 1, 10.096586358062215: 1, 10.094405852677005: 1, 10.08428482327594: 1, 10.066995786995104: 1, 10.063918138307395: 1, 10.044268066205976: 1, 10.032592284855852: 1, 9.99106023311477: 1, 9.973180814306778: 1, 9.969480206316453: 1, 9.9463066389329: 1, 9.945257319092587: 1, 9.917150756818387: 1, 9.9135829367364: 1, 9.893430558370136: 1, 9.892018842926891: 1, 9.87925021752224: 1, 9.845261543479275: 1, 9.839093022686024: 1, 9.829727681469057: 1, 9.82267682160674: 1, 9.767831895000452: 1, 9.763778258456211: 1, 9.762553159393489: 1, 9.754076315585477: 1, 9.70962396535891: 1, 9.67816931428281: 1, 9.677446899225151: 1, 9.672915802826788: 1, 9.665840956843724: 1, 9.658345898255247: 1, 9.656149625564716: 1, 9.651895685109148: 1, 9.633991493369189: 1, 9.616293440127375: 1, 9.607089851034791: 1, 9.604574206791497: 1, 9.602752791646413: 1, 9.581058403917613: 1, 9.577825043553396: 1, 9.57140494106809: 1, 9.570413550820854: 1, 9.56166108312655: 1, 9.538833770346315: 1, 9.532161721427784: 1, 9.527138856676235: 1, 9.51775360856059: 1, 9.514689250161176: 1, 9.508777115359994: 1, 9.494320482979285: 1, 9.493681612305151: 1, 9.476589568836683: 1, 9.467828864717688: 1, 9.463298518230348: 1, 9.45889700084438: 1, 9.446930838700732: 1, 9.438121910686087: 1, 9.42127185471366: 1, 9.410517519212904: 1, 9.406747473480818: 1, 9.389155911529024: 1, 9.384792300907932: 1, 9.38084523933283: 1, 9.375073272531482: 1, 9.370001778063227: 1, 9.369931962438015: 1, 9.356755549260573: 1, 9.326854340805708: 1, 9.315411048125906: 1, 9.296760034231909: 1, 9.28805251739 3858: 1, 9.283959867307152: 1, 9.253310607996873: 1, 9.23953026611918: 1, 9.238035078637322: 1, 9.20455706652068: 1, 9.20191057523233: 1, 9.185266618282977: 1, 9.181059245394925: 1, 9.16993421351546: 1, 9.145989086099782: 1, 9.143645387781783: 1, 9.132745145546306: 1, 9.127463117730594: 1, 9.11878290706506: 1, 9.117265858601618: 1, 9.112494404329498: 1, 9.099196884673162: 1, 9.096930898932811: 1, 9.094010568155065: 1, 9.090862700452995: 1, 9.085950376935367: 1, 9.064176472541437: 1, 9.05949887042003: 1, 9.05935862875613: 1, 9.056451683296363: 1, 9.042691771039266: 1, 9.04097788722348: 1, 9.024385689534775: 1, 9.004931588299488: 1, 9.004443648040183: 1, 9.0024231666808

1: 1, 8.991200539733367: 1, 8.985907742821283: 1, 8.982007895364191: 1,
8.971709170446019: 1, 8.967615317417625: 1, 8.967494530711958: 1, 8.957
564093340553: 1, 8.948825804670022: 1, 8.916334830410266: 1, 8.91147226
0219812: 1, 8.90972725369847: 1, 8.90641509973929: 1, 8.90365402098080
8: 1, 8.897125231508255: 1, 8.882592141376827: 1, 8.878563846233689: 1,
8.873752146242182: 1, 8.867575877030683: 1, 8.86546161703496: 1, 8.8653
52212849368: 1, 8.8623643002727: 1, 8.852300199202245: 1, 8.84673394420
149: 1, 8.839274680753832: 1, 8.795359106391075: 1, 8.795312614333263:
1, 8.78998824352581: 1, 8.785559233461939: 1, 8.77879096962508: 1, 8.77
7096816322917: 1, 8.771739214603695: 1, 8.764889941739325: 1, 8.7591134
42717847: 1, 8.748554312265236: 1, 8.735227592040868: 1, 8.727387662729
3: 1, 8.715036189775068: 1, 8.712521076480073: 1, 8.69313580632389: 1,
8.681555453497394: 1, 8.656677678471237: 1, 8.652043721712563: 1, 8.580
354920020026: 1, 8.571417638565872: 1, 8.541341149750837: 1, 8.53531681
3309912: 1, 8.531835767472238: 1, 8.527392799902284: 1, 8.5186169774083
09: 1, 8.511227059728784: 1, 8.501578883318098: 1, 8.490633472993302:
1, 8.483150510448613: 1, 8.480395216855575: 1, 8.477158157113095: 1, 8.
469836239147861: 1, 8.437248225478363: 1, 8.437115841751288: 1, 8.43618
5414752575: 1, 8.424012731643842: 1, 8.421916058073236: 1, 8.3756683720
44663: 1, 8.369820825089178: 1, 8.36885491554583: 1, 8.348755600448545:
1, 8.331391806008366: 1, 8.319447035772294: 1, 8.31239266464547: 1, 8.3
10635261267597: 1, 8.308128599476245: 1, 8.288399461940221: 1, 8.276888
764855151: 1, 8.27369111780222: 1, 8.271079300038034: 1, 8.257289538167
289: 1, 8.23655739318425: 1, 8.22558913018967: 1, 8.220816084872446: 1,
8.213813219079773: 1, 8.211905060737989: 1, 8.1877685134449: 1, 8.18385
8138845265: 1, 8.17362772116125: 1, 8.171107490364976: 1, 8.15742035174
3855: 1, 8.144145624639224: 1, 8.134216123391035: 1, 8.124336779870053:
1, 8.114916250386214: 1, 8.106270205443911: 1, 8.090977030888297: 1, 8.
085604877328882: 1, 8.078280550894934: 1, 8.070461955475905: 1, 8.04716
9957550846: 1, 8.04111861163527: 1, 8.039286603000482: 1, 8.02869145318
9749: 1, 8.02496192998756: 1, 8.0160546370718: 1, 7.935693519815968: 1,
7.911262097153478: 1, 7.908291698461302: 1, 7.907960712129177: 1, 7.905
49428918233: 1, 7.904911998165901: 1, 7.894613908779852: 1, 7.884341074
363364: 1, 7.871433924875626: 1, 7.855048097639604: 1, 7.84555219228736
2: 1, 7.811244340182322: 1, 7.809670029757706: 1, 7.80662707928: 1, 7.8
03088795796641: 1, 7.796723450054868: 1, 7.794293246704388: 1, 7.792391
941585682: 1, 7.787810740128156: 1, 7.773280893977649: 1, 7.77187406153
9344: 1, 7.758860776975461: 1, 7.7432696702121335: 1, 7.73323523993343
5: 1, 7.729466175750535: 1, 7.725862291821469: 1, 7.7247040240030085:

```
1, 7.720501554214147: 1, 7.7184855415274: 1, 7.7181204113770585: 1, 7.7
12341142076051: 1, 7.689122978578691: 1, 7.67743486079945: 1, 7.6756374
9284111: 1, 7.657640605140514: 1, 7.6465314449611235: 1, 7.635030750172
473: 1, 7.628781054418943: 1, 7.616070798341814: 1, 7.613510603009687:
1, 7.604135342101209: 1, 7.564141304329447: 1, 7.563943127657511: 1, 7.
5538892865743446: 1, 7.540766494610485: 1, 7.530703055263175: 1, 7.5231
22654817884: 1, 7.477817527201547: 1, 7.47172956693088: 1, 7.4629978781
557265: 1, 7.446189976482725: 1, 7.3995437873710905: 1, 7.3836873353889
35: 1, 7.383562049200109: 1, 7.382894517780061: 1, 7.379717782698696:
1, 7.364037433569303: 1, 7.322594730938315: 1, 7.267146759552682: 1, 7.
2670524309416: 1, 7.262636452510385: 1, 7.22620412626707: 1, 7.20668596
7586005: 1, 7.165015892193955: 1, 7.159543482230967: 1, 7.1504115884784
32: 1, 7.141721942452791: 1, 7.141512651160529: 1, 7.107891444374135:
1, 7.095797927639719: 1, 7.072307456529319: 1, 7.071759652709824: 1, 7.
0604291584132675: 1, 7.0367389558090165: 1, 6.941678861062921: 1, 6.911
735210164857: 1, 6.896942841762888: 1, 6.893203621579399: 1, 6.88984446
4399743: 1, 6.874574850052018: 1, 6.868497531774119: 1, 6.8221355937576
496: 1, 6.7897364162613: 1, 6.685751275175943: 1, 6.679933057361155: 1,
6.629365555513137: 1, 6.581331409097498: 1, 6.564530424105985: 1, 6.327
806979807444: 1, 6.137549292875256: 1})
```

In [46]:
```python
# Train a Logistic regression+Calibration model using text features whi
cha re on-hot encoded
alpha = [10 ** x for x in range(-5, 1)]

# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_rate='optimal', eta0=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, …])     Fit linear model with S
tochastic Gradient Descent.
# predict(X)     Predict class labels for samples in X.
```

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

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


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

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

```
For values of alpha =  1e-05 The log loss is: 1.1009894909411904
For values of alpha =  0.0001 The log loss is: 1.125213696223533
For values of alpha =  0.001 The log loss is: 1.4563593054036552
For values of alpha =  0.01 The log loss is: 1.9288630428489024
For values of alpha =  0.1 The log loss is: 2.006050422524432
For values of alpha =  1 The log loss is: 1.994114243558515
```


Cross Validation Error for each alpha
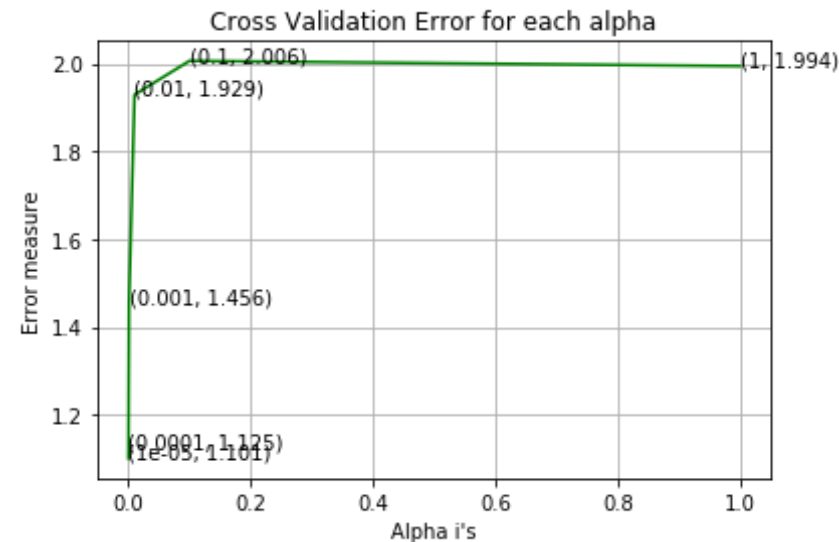
```
For values of best alpha =  1e-05 The train log loss is: 0.723844565696
9569
For values of best alpha =  1e-05 The cross validation log loss is: 1.1
009894909411904
For values of best alpha =  1e-05 The test log loss is: 1.206780454???
```

For values of best alpha = 1e-05 The test log loss is: 1.206789454382777
777

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

Ans. Yes, it seems like!

```python
In [47]: def get_intersec_text(df):
             df_text_vec = TfidfVectorizer(max_features=1000)
             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
```

```python
In [48]: 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")
```

```
93.4 % of word of test data appeared in train data
94.1 % of word of Cross Validation appeared in train data
```

4. Machine Learning Models

```python
In [49]: #Data preparation for ML models.

         #Misc. functionns for ML models
```

```python
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 probabilit
ies 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 [50]:
```python
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 [51]:
```python
# 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(min_df=3, max_features=1000)

    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 p
oint [{}]".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, "ar
e present in query point")
```

Stacking the three types of features

In [52]:
```
# 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,t
rain_variation_feature_onehotCoding))
test_gene_var_onehotCoding = hstack((test_gene_feature_onehotCoding,tes
t_variation_feature_onehotCoding))
cv_gene_var_onehotCoding = hstack((cv_gene_feature_onehotCoding,cv_vari
ation_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_fea
ture_onehotCoding)).tocsr()
test_y = np.array(list(test_df['Class']))

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


train_gene_var_responseCoding = np.hstack((train_gene_feature_responseC
oding,train_variation_feature_responseCoding))
test_gene_var_responseCoding = np.hstack((test_gene_feature_responseCod
ing,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, trai
n_text_feature_responseCoding))
test_x_responseCoding = np.hstack((test_gene_var_responseCoding, test_t
ext_feature_responseCoding))
cv_x_responseCoding = np.hstack((cv_gene_var_responseCoding, cv_text_fe
ature_responseCoding))
```

```
In [53]:  print("One hot encoding features :")
```

```python
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 = ", t
est_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, 31
95)
(number of data points * number of features) in test data =  (665, 319
5)
(number of data points * number of features) in cross validation data =
(532, 3195)
```

In [54]:
```python
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 = ", t
est_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, 2
7)
(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 [55]:  # find more about Multinomial Naive base function here http://scikit-le
          arn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html
          # -------------------------
          # default paramters
          # sklearn.naive_bayes.MultinomialNB(alpha=1.0, fit_prior=True, class_pr
          ior=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 v
          ector 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.or
          g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
          tml
          # ----------------------------
          # default paramters
          # sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
          d='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]
```

```python
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_a
rray[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, labels=clf.classes_, eps=1e-15
))
predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
```

```python
print('For values of best alpha = ', alpha[best_alpha], "The cross vali
dation 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 l
oss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-05
Log Loss : 1.1701347769511954
for alpha = 0.0001
Log Loss : 1.169420794167966
for alpha = 0.001
Log Loss : 1.16905461991396
for alpha = 0.1
Log Loss : 1.2161298208324283
for alpha = 1
Log Loss : 1.2791821747509506
for alpha = 10
Log Loss : 1.4891739311004868
for alpha = 100
Log Loss : 1.463042405008789
for alpha = 1000
Log Loss : 1.45269248735883
```


Cross Validation Error for each alpha
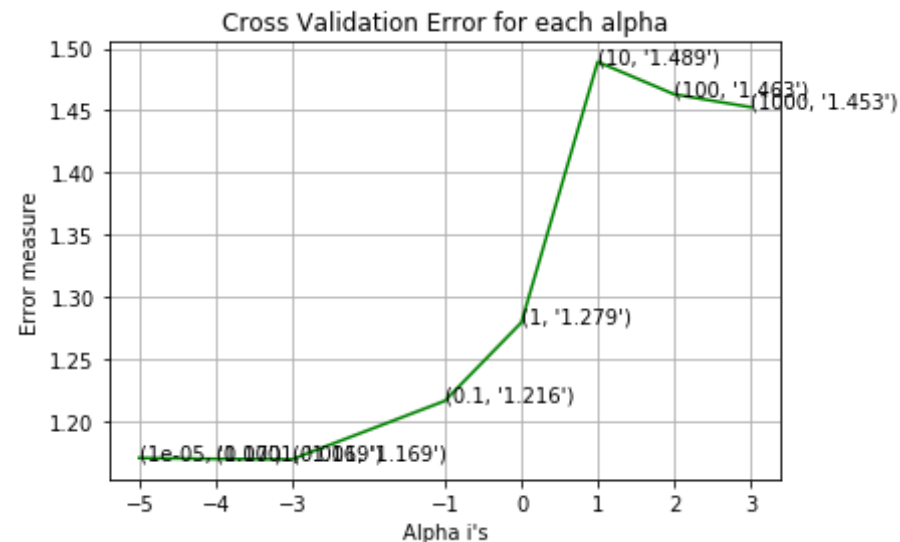
```
For values of best alpha =  0.001 The train log loss is: 0.505189404611
9969
For values of best alpha =  0.001 The cross validation log loss is: 1.1
6905461991396
For values of best alpha =  0.001 The test log loss is: 1.2256012062877
36
```

4.1.1.2. Testing the model with best hyper paramters

In [56]:
```python
# find more about Multinomial Naive base function here http://scikit-le
arn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html
# -------------------------
# default paramters
# sklearn.naive_bayes.MultinomialNB(alpha=1.0, fit_prior=True, class_pr
ior=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 v
ector 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.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
tml
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
```

```python
# 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-pro
bability estimates
print("Log Loss :",log_loss(cv_y, sig_clf_probs))
print("Number of missclassified point :", np.count_nonzero((sig_clf.pre
dict(cv_x_onehotCoding)- cv_y))/cv_y.shape[0])
plot_confusion_matrix(cv_y, sig_clf.predict(cv_x_onehotCoding.toarray
()))
```

```
Log Loss : 1.16905461991396
Number of missclassified point : 0.37406015037593987
-------------------- Confusion matrix --------------------
```



```
-------------------- Precision matrix (Columm Sum=1) -----------------
--
```
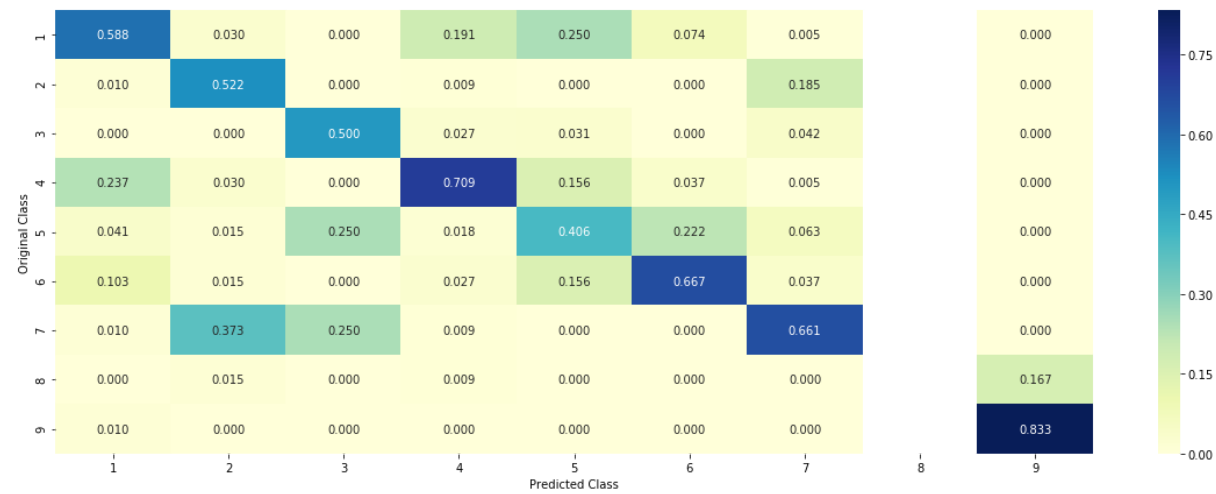
-------------------- Recall matrix (Row sum=1) --------------------
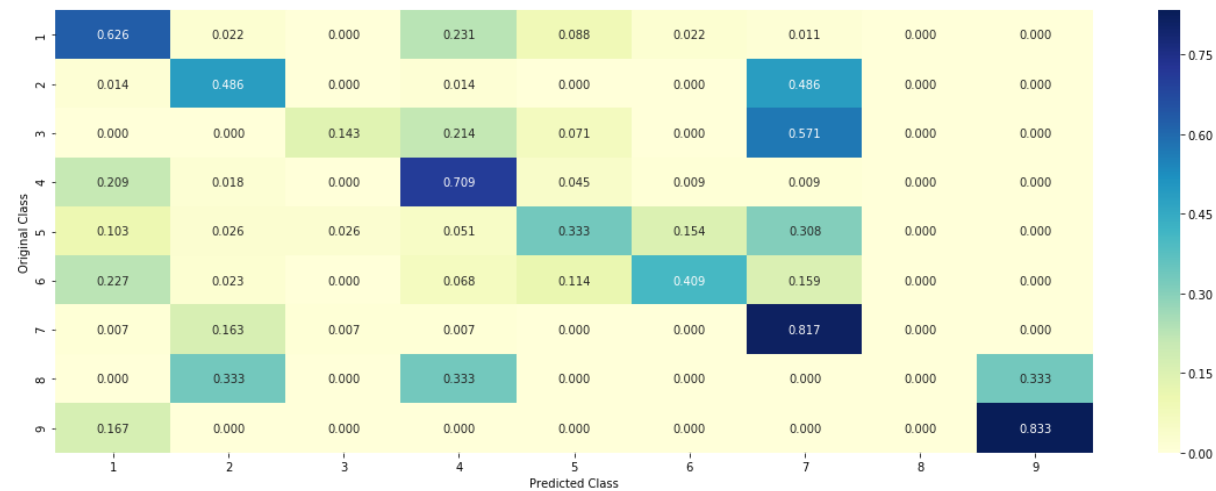
4.1.1.3. Feature Importance, Correctly classified point

```
In [57]: 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],test_df['Variation'].iloc[test
         _point_index], no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[0.5923 0.0512 0.0104 0.1872 0.0339 0.0
311 0.0876 0.0039 0.0025]]
Actual Class : 1
--------------------------------------------------
10 Text feature [one] present in test data point [True]
14 Text feature [results] present in test data point [True]
15 Text feature [protein] present in test data point [True]
16 Text feature [type] present in test data point [True]
18 Text feature [loss] present in test data point [True]
19 Text feature [therefore] present in test data point [True]
20 Text feature [two] present in test data point [True]
21 Text feature [function] present in test data point [True]
22 Text feature [also] present in test data point [True]
23 Text feature [result] present in test data point [True]
24 Text feature [possible] present in test data point [True]
25 Text feature [wild] present in test data point [True]
26 Text feature [however] present in test data point [True]
27 Text feature [dna] present in test data point [True]
```

```
28 Text feature [using] present in test data point [True]
29 Text feature [containing] present in test data point [True]
30 Text feature [role] present in test data point [True]
32 Text feature [discussion] present in test data point [True]
33 Text feature [control] present in test data point [True]
34 Text feature [present] present in test data point [True]
35 Text feature [least] present in test data point [True]
36 Text feature [may] present in test data point [True]
37 Text feature [determined] present in test data point [True]
38 Text feature [region] present in test data point [True]
39 Text feature [used] present in test data point [True]
40 Text feature [25] present in test data point [True]
41 Text feature [specific] present in test data point [True]
42 Text feature [large] present in test data point [True]
43 Text feature [deletion] present in test data point [True]
44 Text feature [effect] present in test data point [True]
45 Text feature [binding] present in test data point [True]
46 Text feature [important] present in test data point [True]
47 Text feature [analysis] present in test data point [True]
48 Text feature [three] present in test data point [True]
49 Text feature [data] present in test data point [True]
50 Text feature [likely] present in test data point [True]
51 Text feature [four] present in test data point [True]
52 Text feature [human] present in test data point [True]
53 Text feature [shown] present in test data point [True]
54 Text feature [based] present in test data point [True]
55 Text feature [cancer] present in test data point [True]
56 Text feature [including] present in test data point [True]
57 Text feature [whether] present in test data point [True]
58 Text feature [remaining] present in test data point [True]
59 Text feature [example] present in test data point [True]
60 Text feature [expression] present in test data point [True]
61 Text feature [additional] present in test data point [True]
63 Text feature [performed] present in test data point [True]
64 Text feature [previously] present in test data point [True]
67 Text feature [gene] present in test data point [True]
68 Text feature [within] present in test data point [True]
70 Text feature [another] present in test data point [True]
71 Text feature [majority] present in test data point [True]
```

```
72 Text feature [mutations] present in test data point [True]
73 Text feature [genes] present in test data point [True]
75 Text feature [different] present in test data point [True]
76 Text feature [similar] present in test data point [True]
77 Text feature [observed] present in test data point [True]
78 Text feature [addition] present in test data point [True]
79 Text feature [mutation] present in test data point [True]
80 Text feature [associated] present in test data point [True]
81 Text feature [30] present in test data point [True]
82 Text feature [methods] present in test data point [True]
83 Text feature [although] present in test data point [True]
85 Text feature [identified] present in test data point [True]
86 Text feature [several] present in test data point [True]
87 Text feature [10] present in test data point [True]
88 Text feature [significant] present in test data point [True]
89 Text feature [highly] present in test data point [True]
90 Text feature [defined] present in test data point [True]
91 Text feature [corresponding] present in test data point [True]
92 Text feature [sequence] present in test data point [True]
93 Text feature [indicate] present in test data point [True]
94 Text feature [single] present in test data point [True]
96 Text feature [significantly] present in test data point [True]
97 Text feature [found] present in test data point [True]
98 Text feature [studies] present in test data point [True]
99 Text feature [involved] present in test data point [True]
101 Text feature [proteins] present in test data point [True]
102 Text feature [domains] present in test data point [True]
103 Text feature [genetic] present in test data point [True]
104 Text feature [yet] present in test data point [True]
106 Text feature [cell] present in test data point [True]
107 Text feature [following] present in test data point [True]
108 Text feature [15] present in test data point [True]
109 Text feature [first] present in test data point [True]
110 Text feature [determine] present in test data point [True]
111 Text feature [32] present in test data point [True]
114 Text feature [described] present in test data point [True]
115 Text feature [together] present in test data point [True]
116 Text feature [many] present in test data point [True]
117 Text feature [essential] present in test data point [True]
```

```
118 Text feature [critical] present in test data point [True]
119 Text feature [respectively] present in test data point [True]
120 Text feature [identify] present in test data point [True]
121 Text feature [whereas] present in test data point [True]
122 Text feature [could] present in test data point [True]
123 Text feature [indicating] present in test data point [True]
124 Text feature [specificity] present in test data point [True]
126 Text feature [directly] present in test data point [True]
128 Text feature [required] present in test data point [True]
129 Text feature [16] present in test data point [True]
130 Text feature [fig] present in test data point [True]
131 Text feature [consistent] present in test data point [True]
132 Text feature [functional] present in test data point [True]
134 Text feature [furthermore] present in test data point [True]
135 Text feature [effects] present in test data point [True]
136 Text feature [reported] present in test data point [True]
137 Text feature [structure] present in test data point [True]
138 Text feature [37] present in test data point [True]
139 Text feature [domain] present in test data point [True]
140 Text feature [transcriptional] present in test data point [True]
141 Text feature [suggests] present in test data point [True]
145 Text feature [assays] present in test data point [True]
146 Text feature [according] present in test data point [True]
148 Text feature [high] present in test data point [True]
149 Text feature [five] present in test data point [True]
150 Text feature [study] present in test data point [True]
151 Text feature [ability] present in test data point [True]
152 Text feature [mutant] present in test data point [True]
153 Text feature [would] present in test data point [True]
154 Text feature [figure] present in test data point [True]
155 Text feature [complex] present in test data point [True]
156 Text feature [system] present in test data point [True]
157 Text feature [provide] present in test data point [True]
158 Text feature [established] present in test data point [True]
159 Text feature [due] present in test data point [True]
160 Text feature [conserved] present in test data point [True]
161 Text feature [contains] present in test data point [True]
162 Text feature [full] present in test data point [True]
164 Text feature [tumor] present in test data point [True]
```

```
166 Text feature [sufficient] present in test data point [True]
168 Text feature [major] present in test data point [True]
169 Text feature [obtained] present in test data point [True]
170 Text feature [cancers] present in test data point [True]
171 Text feature [analyzed] present in test data point [True]
172 Text feature [missense] present in test data point [True]
173 Text feature [calculated] present in test data point [True]
175 Text feature [strong] present in test data point [True]
177 Text feature [site] present in test data point [True]
178 Text feature [size] present in test data point [True]
179 Text feature [show] present in test data point [True]
180 Text feature [surface] present in test data point [True]
181 Text feature [expected] present in test data point [True]
182 Text feature [negative] present in test data point [True]
183 Text feature [expressed] present in test data point [True]
184 Text feature [cells] present in test data point [True]
186 Text feature [deletions] present in test data point [True]
187 Text feature [dependent] present in test data point [True]
188 Text feature [finding] present in test data point [True]
189 Text feature [even] present in test data point [True]
190 Text feature [recently] present in test data point [True]
191 Text feature [range] present in test data point [True]
192 Text feature [levels] present in test data point [True]
193 Text feature [shows] present in test data point [True]
194 Text feature [35] present in test data point [True]
195 Text feature [specifically] present in test data point [True]
196 Text feature [small] present in test data point [True]
197 Text feature [derived] present in test data point [True]
198 Text feature [p53] present in test data point [True]
199 Text feature [number] present in test data point [True]
200 Text feature [sequences] present in test data point [True]
202 Text feature [form] present in test data point [True]
203 Text feature [frequency] present in test data point [True]
204 Text feature [amino] present in test data point [True]
206 Text feature [remains] present in test data point [True]
208 Text feature [showed] present in test data point [True]
209 Text feature [29] present in test data point [True]
210 Text feature [taken] present in test data point [True]
211 Text feature [terminal] present in test data point [True]
```

```
213 Text feature [indicates] present in test data point [True]
214 Text feature [finally] present in test data point [True]
215 Text feature [vitro] present in test data point [True]
216 Text feature [strongly] present in test data point [True]
218 Text feature [early] present in test data point [True]
219 Text feature [rather] present in test data point [True]
220 Text feature [position] present in test data point [True]
221 Text feature [low] present in test data point [True]
222 Text feature [materials] present in test data point [True]
224 Text feature [28] present in test data point [True]
225 Text feature [100] present in test data point [True]
226 Text feature [confirmed] present in test data point [True]
227 Text feature [increased] present in test data point [True]
228 Text feature [changes] present in test data point [True]
229 Text feature [sequenced] present in test data point [True]
232 Text feature [20] present in test data point [True]
233 Text feature [known] present in test data point [True]
234 Text feature [bind] present in test data point [True]
235 Text feature [lower] present in test data point [True]
236 Text feature [observation] present in test data point [True]
237 Text feature [structural] present in test data point [True]
238 Text feature [mean] present in test data point [True]
239 Text feature [development] present in test data point [True]
240 Text feature [less] present in test data point [True]
241 Text feature [27] present in test data point [True]
243 Text feature [cellular] present in test data point [True]
246 Text feature [13] present in test data point [True]
247 Text feature [complete] present in test data point [True]
248 Text feature [sequencing] present in test data point [True]
249 Text feature [include] present in test data point [True]
252 Text feature [might] present in test data point [True]
255 Text feature [lack] present in test data point [True]
256 Text feature [detected] present in test data point [True]
257 Text feature [via] present in test data point [True]
259 Text feature [assay] present in test data point [True]
261 Text feature [50] present in test data point [True]
262 Text feature [common] present in test data point [True]
263 Text feature [length] present in test data point [True]
265 Text feature [change] present in test data point [True]
```

267 Text feature [constructs] present in test data point [True]
269 Text feature [characterized] present in test data point [True]
270 Text feature [regions] present in test data point [True]
271 Text feature [normal] present in test data point [True]
272 Text feature [higher] present in test data point [True]
275 Text feature [types] present in test data point [True]
276 Text feature [carried] present in test data point [True]
280 Text feature [target] present in test data point [True]
286 Text feature [pcr] present in test data point [True]
287 Text feature [acid] present in test data point [True]
288 Text feature [acids] present in test data point [True]
289 Text feature [lead] present in test data point [True]
290 Text feature [cannot] present in test data point [True]
291 Text feature [deficient] present in test data point [True]
292 Text feature [presence] present in test data point [True]
293 Text feature [displayed] present in test data point [True]
294 Text feature [without] present in test data point [True]
295 Text feature [sites] present in test data point [True]
296 Text feature [represent] present in test data point [True]
297 Text feature [individual] present in test data point [True]
298 Text feature [predicted] present in test data point [True]
299 Text feature [transcription] present in test data point [True]
301 Text feature [fold] present in test data point [True]
304 Text feature [activities] present in test data point [True]
305 Text feature [key] present in test data point [True]
306 Text feature [stability] present in test data point [True]
307 Text feature [part] present in test data point [True]
308 Text feature [see] present in test data point [True]
312 Text feature [demonstrated] present in test data point [True]
313 Text feature [cases] present in test data point [True]
314 Text feature [26] present in test data point [True]
316 Text feature [manufacturer] present in test data point [True]
317 Text feature [overall] present in test data point [True]
319 Text feature [repair] present in test data point [True]
320 Text feature [purified] present in test data point [True]
321 Text feature [examined] present in test data point [True]
322 Text feature [approximately] present in test data point [True]
323 Text feature [variants] present in test data point [True]
325 Text feature [often] present in test data point [True]

```
326 Text feature [17] present in test data point [True]
327 Text feature [potential] present in test data point [True]
328 Text feature [findings] present in test data point [True]
329 Text feature [activity] present in test data point [True]
330 Text feature [positions] present in test data point [True]
331 Text feature [11] present in test data point [True]
332 Text feature [frequently] present in test data point [True]
336 Text feature [suggested] present in test data point [True]
337 Text feature [residues] present in test data point [True]
338 Text feature [gel] present in test data point [True]
339 Text feature [samples] present in test data point [True]
340 Text feature [interact] present in test data point [True]
342 Text feature [selected] present in test data point [True]
343 Text feature [fact] present in test data point [True]
344 Text feature [somatic] present in test data point [True]
345 Text feature [suppressor] present in test data point [True]
346 Text feature [48] present in test data point [True]
347 Text feature [recent] present in test data point [True]
349 Text feature [ref] present in test data point [True]
350 Text feature [et] present in test data point [True]
351 Text feature [occur] present in test data point [True]
352 Text feature [experiments] present in test data point [True]
353 Text feature [phenotype] present in test data point [True]
354 Text feature [42] present in test data point [True]
356 Text feature [larger] present in test data point [True]
357 Text feature [al] present in test data point [True]
358 Text feature [appears] present in test data point [True]
360 Text feature [indeed] present in test data point [True]
362 Text feature [18] present in test data point [True]
364 Text feature [evaluated] present in test data point [True]
365 Text feature [series] present in test data point [True]
366 Text feature [primers] present in test data point [True]
367 Text feature [contrast] present in test data point [True]
371 Text feature [relatively] present in test data point [True]
372 Text feature [similarly] present in test data point [True]
373 Text feature [vector] present in test data point [True]
374 Text feature [3b] present in test data point [True]
377 Text feature [absence] present in test data point [True]
379 Text feature [comparison] present in test data point [True]
```

```
380 Text feature [http] present in test data point [True]
381 Text feature [evidence] present in test data point [True]
382 Text feature [generated] present in test data point [True]
383 Text feature [43] present in test data point [True]
384 Text feature [12] present in test data point [True]
385 Text feature [44] present in test data point [True]
386 Text feature [database] present in test data point [True]
387 Text feature [predict] present in test data point [True]
389 Text feature [able] present in test data point [True]
394 Text feature [independent] present in test data point [True]
398 Text feature [interactions] present in test data point [True]
399 Text feature [genomic] present in test data point [True]
401 Text feature [19] present in test data point [True]
404 Text feature [cause] present in test data point [True]
405 Text feature [observations] present in test data point [True]
407 Text feature [second] present in test data point [True]
409 Text feature [biological] present in test data point [True]
410 Text feature [reduction] present in test data point [True]
411 Text feature [red] present in test data point [True]
413 Text feature [differences] present in test data point [True]
415 Text feature [model] present in test data point [True]
416 Text feature [view] present in test data point [True]
417 Text feature [complexes] present in test data point [True]
418 Text feature [co] present in test data point [True]
419 Text feature [six] present in test data point [True]
420 Text feature [side] present in test data point [True]
421 Text feature [functionally] present in test data point [True]
422 Text feature [measured] present in test data point [True]
423 Text feature [products] present in test data point [True]
425 Text feature [plasmid] present in test data point [True]
426 Text feature [chain] present in test data point [True]
427 Text feature [41] present in test data point [True]
428 Text feature [conditions] present in test data point [True]
429 Text feature [residue] present in test data point [True]
430 Text feature [allele] present in test data point [True]
431 Text feature [alterations] present in test data point [True]
433 Text feature [binds] present in test data point [True]
434 Text feature [core] present in test data point [True]
438 Text feature [non] present in test data point [True]
```

```
441 Text feature [hypothesis] present in test data point [True]
443 Text feature [controls] present in test data point [True]
444 Text feature [process] present in test data point [True]
445 Text feature [approach] present in test data point [True]
446 Text feature [repeats] present in test data point [True]
447 Text feature [isolated] present in test data point [True]
450 Text feature [regulated] present in test data point [True]
451 Text feature [correlation] present in test data point [True]
453 Text feature [distribution] present in test data point [True]
456 Text feature [half] present in test data point [True]
457 Text feature [groups] present in test data point [True]
462 Text feature [particular] present in test data point [True]
466 Text feature [line] present in test data point [True]
467 Text feature [various] present in test data point [True]
468 Text feature [increase] present in test data point [True]
469 Text feature [showing] present in test data point [True]
470 Text feature [34] present in test data point [True]
471 Text feature [sample] present in test data point [True]
473 Text feature [affected] present in test data point [True]
474 Text feature [peptide] present in test data point [True]
475 Text feature [status] present in test data point [True]
478 Text feature [induced] present in test data point [True]
481 Text feature [14] present in test data point [True]
482 Text feature [manner] present in test data point [True]
484 Text feature [mechanism] present in test data point [True]
485 Text feature [mutated] present in test data point [True]
488 Text feature [stable] present in test data point [True]
490 Text feature [unique] present in test data point [True]
493 Text feature [tumors] present in test data point [True]
494 Text feature [direct] present in test data point [True]
495 Text feature [software] present in test data point [True]
496 Text feature [difference] present in test data point [True]
497 Text feature [21] present in test data point [True]
498 Text feature [cycle] present in test data point [True]
501 Text feature [added] present in test data point [True]
502 Text feature [molecular] present in test data point [True]
504 Text feature [nucleus] present in test data point [True]
505 Text feature [incubated] present in test data point [True]
506 Text feature [mechanisms] present in test data point [True]
```

```
508 Text feature [new] present in test data point [True]
509 Text feature [interaction] present in test data point [True]
510 Text feature [defective] present in test data point [True]
511 Text feature [45] present in test data point [True]
512 Text feature [antibody] present in test data point [True]
513 Text feature [testing] present in test data point [True]
515 Text feature [reverse] present in test data point [True]
517 Text feature [upon] present in test data point [True]
519 Text feature [signal] present in test data point [True]
521 Text feature [molecules] present in test data point [True]
522 Text feature [defects] present in test data point [True]
523 Text feature [vivo] present in test data point [True]
525 Text feature [developed] present in test data point [True]
526 Text feature [mutants] present in test data point [True]
527 Text feature [nuclear] present in test data point [True]
529 Text feature [tissues] present in test data point [True]
530 Text feature [s2] present in test data point [True]
534 Text feature [seen] present in test data point [True]
535 Text feature [exons] present in test data point [True]
536 Text feature [www] present in test data point [True]
537 Text feature [terminus] present in test data point [True]
548 Text feature [suppression] present in test data point [True]
552 Text feature [substitution] present in test data point [True]
553 Text feature [antibodies] present in test data point [True]
554 Text feature [primer] present in test data point [True]
555 Text feature [subset] present in test data point [True]
556 Text feature [use] present in test data point [True]
557 Text feature [induce] present in test data point [True]
558 Text feature [screening] present in test data point [True]
559 Text feature [activation] present in test data point [True]
560 Text feature [point] present in test data point [True]
561 Text feature [image] present in test data point [True]
562 Text feature [western] present in test data point [True]
563 Text feature [46] present in test data point [True]
564 Text feature [crystal] present in test data point [True]
565 Text feature [clinical] present in test data point [True]
570 Text feature [apoptosis] present in test data point [True]
571 Text feature [nucleotide] present in test data point [True]
572 Text feature [caused] present in test data point [True]
```

574 Text feature [transfected] present in test data point [True]
576 Text feature [targets] present in test data point [True]
577 Text feature [contact] present in test data point [True]
578 Text feature [green] present in test data point [True]
579 Text feature [tissue] present in test data point [True]
582 Text feature [like] present in test data point [True]
585 Text feature [numbers] present in test data point [True]
590 Text feature [growth] present in test data point [True]
598 Text feature [39] present in test data point [True]
599 Text feature [nature] present in test data point [True]
600 Text feature [experimental] present in test data point [True]
601 Text feature [200] present in test data point [True]
603 Text feature [factor] present in test data point [True]
606 Text feature [tumorigenesis] present in test data point [True]
610 Text feature [positive] present in test data point [True]
612 Text feature [motif] present in test data point [True]
613 Text feature [transactivation] present in test data point [True]
616 Text feature [90] present in test data point [True]
618 Text feature [patients] present in test data point [True]
620 Text feature [75] present in test data point [True]
623 Text feature [time] present in test data point [True]
627 Text feature [52] present in test data point [True]
628 Text feature [factors] present in test data point [True]
629 Text feature [ii] present in test data point [True]
632 Text feature [bound] present in test data point [True]
633 Text feature [2001] present in test data point [True]
634 Text feature [diagnosis] present in test data point [True]
635 Text feature [damage] present in test data point [True]
639 Text feature [rates] present in test data point [True]
640 Text feature [hydrophobic] present in test data point [True]
644 Text feature [regulatory] present in test data point [True]
646 Text feature [structures] present in test data point [True]
648 Text feature [inactivation] present in test data point [True]
650 Text feature [properties] present in test data point [True]
653 Text feature [carcinoma] present in test data point [True]
654 Text feature [blue] present in test data point [True]
655 Text feature [pattern] present in test data point [True]
656 Text feature [dominant] present in test data point [True]
657 Text feature [sensitivity] present in test data point [True]

```
658 Text feature [group] present in test data point [True]
659 Text feature [figures] present in test data point [True]
662 Text feature [anti] present in test data point [True]
665 Text feature [buffer] present in test data point [True]
666 Text feature [promoter] present in test data point [True]
667 Text feature [tumour] present in test data point [True]
668 Text feature [targeted] present in test data point [True]
670 Text feature [strand] present in test data point [True]
675 Text feature [case] present in test data point [True]
677 Text feature [min] present in test data point [True]
679 Text feature [codon] present in test data point [True]
680 Text feature [activate] present in test data point [True]
684 Text feature [pathways] present in test data point [True]
685 Text feature [affinity] present in test data point [True]
692 Text feature [involving] present in test data point [True]
694 Text feature [rare] present in test data point [True]
695 Text feature [years] present in test data point [True]
697 Text feature [yeast] present in test data point [True]
699 Text feature [patient] present in test data point [True]
702 Text feature [helix] present in test data point [True]
703 Text feature [research] present in test data point [True]
704 Text feature [wt] present in test data point [True]
706 Text feature [combination] present in test data point [True]
709 Text feature [null] present in test data point [True]
711 Text feature [invitrogen] present in test data point [True]
712 Text feature [right] present in test data point [True]
715 Text feature [lysates] present in test data point [True]
716 Text feature [initial] present in test data point [True]
717 Text feature [impaired] present in test data point [True]
718 Text feature [partial] present in test data point [True]
719 Text feature [class] present in test data point [True]
721 Text feature [blot] present in test data point [True]
730 Text feature [tables] present in test data point [True]
731 Text feature [medium] present in test data point [True]
732 Text feature [pathway] present in test data point [True]
734 Text feature [overexpression] present in test data point [True]
735 Text feature [phase] present in test data point [True]
737 Text feature [tp53] present in test data point [True]
740 Text feature [colon] present in test data point [True]
```

```
741 Text feature [inhibit] present in test data point [True]
742 Text feature [box] present in test data point [True]
744 Text feature [events] present in test data point [True]
748 Text feature [unable] present in test data point [True]
750 Text feature [expressing] present in test data point [True]
757 Text feature [molecule] present in test data point [True]
758 Text feature [express] present in test data point [True]
759 Text feature [insertion] present in test data point [True]
762 Text feature [models] present in test data point [True]
764 Text feature [concentration] present in test data point [True]
766 Text feature [hotspots] present in test data point [True]
771 Text feature [tumours] present in test data point [True]
773 Text feature [reaction] present in test data point [True]
774 Text feature [require] present in test data point [True]
775 Text feature [colonies] present in test data point [True]
777 Text feature [substitutions] present in test data point [True]
783 Text feature [death] present in test data point [True]
784 Text feature [cluster] present in test data point [True]
785 Text feature [mice] present in test data point [True]
786 Text feature [amplified] present in test data point [True]
787 Text feature [treatment] present in test data point [True]
789 Text feature [interface] present in test data point [True]
790 Text feature [transformed] present in test data point [True]
795 Text feature [free] present in test data point [True]
796 Text feature [000] present in test data point [True]
807 Text feature [versus] present in test data point [True]
808 Text feature [95] present in test data point [True]
810 Text feature [lane] present in test data point [True]
812 Text feature [hotspot] present in test data point [True]
813 Text feature [state] present in test data point [True]
818 Text feature [chemotherapy] present in test data point [True]
819 Text feature [response] present in test data point [True]
821 Text feature [syndrome] present in test data point [True]
823 Text feature [degradation] present in test data point [True]
826 Text feature [activated] present in test data point [True]
827 Text feature [therapeutic] present in test data point [True]
829 Text feature [concentrations] present in test data point [True]
830 Text feature [lanes] present in test data point [True]
833 Text feature [phosphorylation] present in test data point [True]
```

```
836 Text feature [promote] present in test data point [True]
837 Text feature [alternative] present in test data point [True]
839 Text feature [ng] present in test data point [True]
843 Text feature [conformation] present in test data point [True]
844 Text feature [screen] present in test data point [True]
845 Text feature [mm] present in test data point [True]
846 Text feature [sensitive] present in test data point [True]
847 Text feature [insertions] present in test data point [True]
849 Text feature [inhibitory] present in test data point [True]
854 Text feature [subunit] present in test data point [True]
858 Text feature [lesions] present in test data point [True]
860 Text feature [active] present in test data point [True]
865 Text feature [days] present in test data point [True]
900 Text feature [56] present in test data point [True]
903 Text feature [atp] present in test data point [True]
906 Text feature [effective] present in test data point [True]
910 Text feature [gain] present in test data point [True]
911 Text feature [alleles] present in test data point [True]
912 Text feature [aberrant] present in test data point [True]
913 Text feature [inhibitor] present in test data point [True]
915 Text feature [inhibition] present in test data point [True]
916 Text feature [inhibited] present in test data point [True]
918 Text feature [kinases] present in test data point [True]
920 Text feature [loop] present in test data point [True]
926 Text feature [amplification] present in test data point [True]
932 Text feature [kinase] present in test data point [True]
933 Text feature [melanoma] present in test data point [True]
935 Text feature [tyrosine] present in test data point [True]
936 Text feature [therapy] present in test data point [True]
937 Text feature [oncogenic] present in test data point [True]
939 Text feature [lung] present in test data point [True]
940 Text feature [drug] present in test data point [True]
944 Text feature [trials] present in test data point [True]
950 Text feature [inactive] present in test data point [True]
952 Text feature [adenocarcinoma] present in test data point [True]
953 Text feature [activating] present in test data point [True]
954 Text feature [cyclin] present in test data point [True]
957 Text feature [ligand] present in test data point [True]
968 Text feature [mg] present in test data point [True]
```

```
975 Text feature [resistant] present in test data point [True]
978 Text feature [ph] present in test data point [True]
Out of the top  1000  features  562 are present in query point
```

4.1.1.4. Feature Importance, Incorrectly classified point

In [59]:
```python
test_point_index = 55
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],test_df['Variation'].iloc[test
_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.1238 0.1181 0.0111 0.1181 0.0355 0.0
325 0.554  0.0042 0.0028]]
Actual Class : 4
--------------------------------------------------
14 Text feature [activation] present in test data point [True]
17 Text feature [activated] present in test data point [True]
18 Text feature [kinase] present in test data point [True]
19 Text feature [cells] present in test data point [True]
22 Text feature [downstream] present in test data point [True]
23 Text feature [expressing] present in test data point [True]
24 Text feature [inhibitor] present in test data point [True]
25 Text feature [also] present in test data point [True]
26 Text feature [signaling] present in test data point [True]
27 Text feature [contrast] present in test data point [True]
28 Text feature [growth] present in test data point [True]
29 Text feature [independent] present in test data point [True]
30 Text feature [however] present in test data point [True]
33 Text feature [treatment] present in test data point [True]
```

33 Text feature [treatment] present in test data point [True]
34 Text feature [shown] present in test data point [True]
35 Text feature [10] present in test data point [True]
36 Text feature [factor] present in test data point [True]
37 Text feature [addition] present in test data point [True]
38 Text feature [phosphorylation] present in test data point [True]
39 Text feature [similar] present in test data point [True]
40 Text feature [compared] present in test data point [True]
41 Text feature [well] present in test data point [True]
43 Text feature [sensitive] present in test data point [True]
44 Text feature [cell] present in test data point [True]
45 Text feature [constitutive] present in test data point [True]
46 Text feature [higher] present in test data point [True]
47 Text feature [treated] present in test data point [True]
48 Text feature [previously] present in test data point [True]
49 Text feature [potential] present in test data point [True]
50 Text feature [mutations] present in test data point [True]
51 Text feature [may] present in test data point [True]
52 Text feature [presence] present in test data point [True]
53 Text feature [recently] present in test data point [True]
54 Text feature [found] present in test data point [True]
55 Text feature [showed] present in test data point [True]
56 Text feature [inhibitors] present in test data point [True]
57 Text feature [3b] present in test data point [True]
58 Text feature [activate] present in test data point [True]
59 Text feature [suggest] present in test data point [True]
60 Text feature [increased] present in test data point [True]
61 Text feature [tyrosine] present in test data point [True]
62 Text feature [inhibition] present in test data point [True]
63 Text feature [consistent] present in test data point [True]
64 Text feature [oncogenic] present in test data point [True]
68 Text feature [receptor] present in test data point [True]
69 Text feature [3a] present in test data point [True]
70 Text feature [although] present in test data point [True]
71 Text feature [pathways] present in test data point [True]
72 Text feature [described] present in test data point [True]
73 Text feature [mutant] present in test data point [True]
74 Text feature [mutation] present in test data point [True]
75 Text feature [mechanism] present in test data point [True]
76 Text feature [constitutively] present in test data point [True]

```
78 Text feature [total] present in test data point [True]
79 Text feature [interestingly] present in test data point [True]
80 Text feature [without] present in test data point [True]
81 Text feature [identified] present in test data point [True]
82 Text feature [fig] present in test data point [True]
83 Text feature [results] present in test data point [True]
84 Text feature [12] present in test data point [True]
85 Text feature [absence] present in test data point [True]
86 Text feature [proliferation] present in test data point [True]
89 Text feature [different] present in test data point [True]
90 Text feature [therapeutic] present in test data point [True]
91 Text feature [observed] present in test data point [True]
92 Text feature [various] present in test data point [True]
93 Text feature [two] present in test data point [True]
94 Text feature [increase] present in test data point [True]
95 Text feature [respectively] present in test data point [True]
96 Text feature [including] present in test data point [True]
97 Text feature [could] present in test data point [True]
98 Text feature [discussion] present in test data point [True]
99 Text feature [using] present in test data point [True]
100 Text feature [recent] present in test data point [True]
101 Text feature [either] present in test data point [True]
102 Text feature [enhanced] present in test data point [True]
104 Text feature [survival] present in test data point [True]
105 Text feature [demonstrated] present in test data point [True]
107 Text feature [trials] present in test data point [True]
108 Text feature [pathway] present in test data point [True]
109 Text feature [whereas] present in test data point [True]
110 Text feature [figure] present in test data point [True]
111 Text feature [13] present in test data point [True]
112 Text feature [antibodies] present in test data point [True]
113 Text feature [mechanisms] present in test data point [True]
114 Text feature [studies] present in test data point [True]
115 Text feature [20] present in test data point [True]
116 Text feature [lines] present in test data point [True]
117 Text feature [expressed] present in test data point [True]
118 Text feature [previous] present in test data point [True]
119 Text feature [phospho] present in test data point [True]
120 Text feature [induced] present in test data point [True]
```

```
121 Text feature [point] present in test data point [True]
123 Text feature [occur] present in test data point [True]
125 Text feature [small] present in test data point [True]
126 Text feature [clinical] present in test data point [True]
127 Text feature [thus] present in test data point [True]
128 Text feature [suggesting] present in test data point [True]
130 Text feature [due] present in test data point [True]
131 Text feature [several] present in test data point [True]
133 Text feature [three] present in test data point [True]
134 Text feature [inhibited] present in test data point [True]
135 Text feature [express] present in test data point [True]
136 Text feature [expression] present in test data point [True]
139 Text feature [whether] present in test data point [True]
140 Text feature [domain] present in test data point [True]
141 Text feature [24] present in test data point [True]
143 Text feature [present] present in test data point [True]
144 Text feature [vitro] present in test data point [True]
147 Text feature [high] present in test data point [True]
148 Text feature [4a] present in test data point [True]
149 Text feature [tumor] present in test data point [True]
150 Text feature [1a] present in test data point [True]
151 Text feature [activity] present in test data point [True]
152 Text feature [mutants] present in test data point [True]
154 Text feature [next] present in test data point [True]
155 Text feature [suggests] present in test data point [True]
157 Text feature [analysis] present in test data point [True]
158 Text feature [detected] present in test data point [True]
159 Text feature [effective] present in test data point [True]
160 Text feature [might] present in test data point [True]
161 Text feature [progression] present in test data point [True]
162 Text feature [anti] present in test data point [True]
164 Text feature [2b] present in test data point [True]
165 Text feature [patients] present in test data point [True]
166 Text feature [common] present in test data point [True]
167 Text feature [study] present in test data point [True]
168 Text feature [target] present in test data point [True]
169 Text feature [confirmed] present in test data point [True]
170 Text feature [lead] present in test data point [True]
171 Text feature [fold] present in test data point [True]
```

```
172 Text feature [first] present in test data point [True]
173 Text feature [additional] present in test data point [True]
174 Text feature [together] present in test data point [True]
175 Text feature [2a] present in test data point [True]
177 Text feature [levels] present in test data point [True]
180 Text feature [within] present in test data point [True]
181 Text feature [50] present in test data point [True]
182 Text feature [level] present in test data point [True]
183 Text feature [single] present in test data point [True]
184 Text feature [similarly] present in test data point [True]
185 Text feature [role] present in test data point [True]
186 Text feature [drug] present in test data point [True]
187 Text feature [one] present in test data point [True]
188 Text feature [experiments] present in test data point [True]
189 Text feature [suggested] present in test data point [True]
191 Text feature [19] present in test data point [True]
192 Text feature [25] present in test data point [True]
193 Text feature [active] present in test data point [True]
195 Text feature [human] present in test data point [True]
198 Text feature [performed] present in test data point [True]
199 Text feature [less] present in test data point [True]
200 Text feature [1b] present in test data point [True]
201 Text feature [resulting] present in test data point [True]
204 Text feature [primary] present in test data point [True]
205 Text feature [mediated] present in test data point [True]
206 Text feature [result] present in test data point [True]
207 Text feature [specific] present in test data point [True]
208 Text feature [another] present in test data point [True]
209 Text feature [promote] present in test data point [True]
210 Text feature [transformation] present in test data point [True]
213 Text feature [directly] present in test data point [True]
215 Text feature [findings] present in test data point [True]
217 Text feature [disease] present in test data point [True]
218 Text feature [associated] present in test data point [True]
219 Text feature [genomic] present in test data point [True]
220 Text feature [table] present in test data point [True]
223 Text feature [data] present in test data point [True]
224 Text feature [line] present in test data point [True]
225 Text feature [dependent] present in test data point [True]
```

```
226 Text feature [four] present in test data point [True]
228 Text feature [differences] present in test data point [True]
229 Text feature [major] present in test data point [True]
230 Text feature [revealed] present in test data point [True]
231 Text feature [used] present in test data point [True]
233 Text feature [harboring] present in test data point [True]
235 Text feature [novel] present in test data point [True]
236 Text feature [overexpression] present in test data point [True]
238 Text feature [possible] present in test data point [True]
243 Text feature [form] present in test data point [True]
244 Text feature [introduction] present in test data point [True]
245 Text feature [development] present in test data point [True]
246 Text feature [derived] present in test data point [True]
247 Text feature [highly] present in test data point [True]
248 Text feature [evaluated] present in test data point [True]
249 Text feature [14] present in test data point [True]
250 Text feature [tumors] present in test data point [True]
251 Text feature [despite] present in test data point [True]
252 Text feature [significantly] present in test data point [True]
253 Text feature [part] present in test data point [True]
254 Text feature [100] present in test data point [True]
255 Text feature [led] present in test data point [True]
256 Text feature [cancer] present in test data point [True]
258 Text feature [therapy] present in test data point [True]
259 Text feature [strongly] present in test data point [True]
261 Text feature [30] present in test data point [True]
264 Text feature [analyzed] present in test data point [True]
266 Text feature [go] present in test data point [True]
267 Text feature [targeted] present in test data point [True]
269 Text feature [new] present in test data point [True]
270 Text feature [lower] present in test data point [True]
271 Text feature [fact] present in test data point [True]
272 Text feature [cellular] present in test data point [True]
273 Text feature [positive] present in test data point [True]
274 Text feature [generation] present in test data point [True]
275 Text feature [observations] present in test data point [True]
276 Text feature [leading] present in test data point [True]
278 Text feature [developed] present in test data point [True]
281 Text feature [contribute] present in test data point [True]
```

```
282 Text feature [wt] present in test data point [True]
283 Text feature [stable] present in test data point [True]
286 Text feature [established] present in test data point [True]
288 Text feature [provide] present in test data point [True]
289 Text feature [distinct] present in test data point [True]
292 Text feature [17] present in test data point [True]
294 Text feature [cancers] present in test data point [True]
295 Text feature [indicating] present in test data point [True]
296 Text feature [indicate] present in test data point [True]
298 Text feature [identification] present in test data point [True]
300 Text feature [mutated] present in test data point [True]
301 Text feature [number] present in test data point [True]
303 Text feature [possibility] present in test data point [True]
305 Text feature [gene] present in test data point [True]
306 Text feature [32] present in test data point [True]
309 Text feature [would] present in test data point [True]
312 Text feature [even] present in test data point [True]
313 Text feature [response] present in test data point [True]
314 Text feature [transformed] present in test data point [True]
316 Text feature [following] present in test data point [True]
318 Text feature [time] present in test data point [True]
319 Text feature [important] present in test data point [True]
320 Text feature [targets] present in test data point [True]
321 Text feature [11] present in test data point [True]
323 Text feature [days] present in test data point [True]
324 Text feature [type] present in test data point [True]
326 Text feature [patient] present in test data point [True]
327 Text feature [selected] present in test data point [True]
328 Text feature [driven] present in test data point [True]
329 Text feature [16] present in test data point [True]
330 Text feature [able] present in test data point [True]
331 Text feature [28] present in test data point [True]
332 Text feature [remains] present in test data point [True]
333 Text feature [endogenous] present in test data point [True]
334 Text feature [indeed] present in test data point [True]
338 Text feature [moreover] present in test data point [True]
339 Text feature [samples] present in test data point [True]
341 Text feature [initial] present in test data point [True]
342 Text feature [kinases] present in test data point [True]
```

```
343 Text feature [combination] present in test data point [True]
344 Text feature [normal] present in test data point [True]
347 Text feature [effect] present in test data point [True]
348 Text feature [forms] present in test data point [True]
349 Text feature [sequencing] present in test data point [True]
351 Text feature [observation] present in test data point [True]
355 Text feature [control] present in test data point [True]
356 Text feature [corresponding] present in test data point [True]
357 Text feature [overall] present in test data point [True]
358 Text feature [finding] present in test data point [True]
359 Text feature [negative] present in test data point [True]
360 Text feature [indicated] present in test data point [True]
361 Text feature [region] present in test data point [True]
362 Text feature [frequently] present in test data point [True]
363 Text feature [subsequent] present in test data point [True]
366 Text feature [amplification] present in test data point [True]
367 Text feature [protein] present in test data point [True]
368 Text feature [experimental] present in test data point [True]
369 Text feature [formation] present in test data point [True]
373 Text feature [finally] present in test data point [True]
374 Text feature [subset] present in test data point [True]
375 Text feature [representative] present in test data point [True]
382 Text feature [sample] present in test data point [True]
384 Text feature [complete] present in test data point [True]
386 Text feature [known] present in test data point [True]
387 Text feature [acid] present in test data point [True]
388 Text feature [alone] present in test data point [True]
390 Text feature [non] present in test data point [True]
391 Text feature [required] present in test data point [True]
393 Text feature [targeting] present in test data point [True]
394 Text feature [include] present in test data point [True]
397 Text feature [22] present in test data point [True]
398 Text feature [37] present in test data point [True]
401 Text feature [pcr] present in test data point [True]
402 Text feature [direct] present in test data point [True]
403 Text feature [least] present in test data point [True]
408 Text feature [upon] present in test data point [True]
409 Text feature [malignant] present in test data point [True]
410 Text feature [note] present in test data point [True]
```

```
411 Text feature [somatic] present in test data point [True]
413 Text feature [et] present in test data point [True]
415 Text feature [key] present in test data point [True]
417 Text feature [among] present in test data point [True]
418 Text feature [caused] present in test data point [True]
420 Text feature [al] present in test data point [True]
422 Text feature [tested] present in test data point [True]
423 Text feature [oncogene] present in test data point [True]
424 Text feature [biological] present in test data point [True]
425 Text feature [often] present in test data point [True]
428 Text feature [rate] present in test data point [True]
430 Text feature [factors] present in test data point [True]
431 Text feature [green] present in test data point [True]
432 Text feature [wild] present in test data point [True]
433 Text feature [expected] present in test data point [True]
435 Text feature [majority] present in test data point [True]
436 Text feature [40] present in test data point [True]
438 Text feature [multiple] present in test data point [True]
439 Text feature [many] present in test data point [True]
440 Text feature [transforming] present in test data point [True]
442 Text feature [top] present in test data point [True]
443 Text feature [29] present in test data point [True]
444 Text feature [since] present in test data point [True]
446 Text feature [comparison] present in test data point [True]
450 Text feature [regulated] present in test data point [True]
451 Text feature [screen] present in test data point [True]
452 Text feature [use] present in test data point [True]
453 Text feature [reports] present in test data point [True]
455 Text feature [limited] present in test data point [True]
457 Text feature [seen] present in test data point [True]
459 Text feature [akt] present in test data point [True]
461 Text feature [long] present in test data point [True]
466 Text feature [reduced] present in test data point [True]
467 Text feature [inhibitory] present in test data point [True]
469 Text feature [proteins] present in test data point [True]
471 Text feature [current] present in test data point [True]
472 Text feature [reverse] present in test data point [True]
475 Text feature [stage] present in test data point [True]
476 Text feature [context] present in test data point [True]
```

```
477 Text feature [numbers] present in test data point [True]
479 Text feature [sites] present in test data point [True]
480 Text feature [particular] present in test data point [True]
484 Text feature [test] present in test data point [True]
485 Text feature [inhibit] present in test data point [True]
486 Text feature [critical] present in test data point [True]
487 Text feature [regulation] present in test data point [True]
488 Text feature [cause] present in test data point [True]
489 Text feature [41] present in test data point [True]
490 Text feature [strong] present in test data point [True]
491 Text feature [frequent] present in test data point [True]
492 Text feature [status] present in test data point [True]
495 Text feature [yet] present in test data point [True]
497 Text feature [mouse] present in test data point [True]
499 Text feature [induce] present in test data point [True]
501 Text feature [given] present in test data point [True]
503 Text feature [early] present in test data point [True]
505 Text feature [model] present in test data point [True]
507 Text feature [methods] present in test data point [True]
509 Text feature [amino] present in test data point [True]
514 Text feature [second] present in test data point [True]
515 Text feature [whole] present in test data point [True]
516 Text feature [decreased] present in test data point [True]
518 Text feature [26] present in test data point [True]
519 Text feature [involving] present in test data point [True]
523 Text feature [genes] present in test data point [True]
524 Text feature [sufficient] present in test data point [True]
525 Text feature [likely] present in test data point [True]
528 Text feature [coding] present in test data point [True]
529 Text feature [35] present in test data point [True]
530 Text feature [gain] present in test data point [True]
531 Text feature [day] present in test data point [True]
532 Text feature [frequency] present in test data point [True]
535 Text feature [events] present in test data point [True]
536 Text feature [hypothesis] present in test data point [True]
539 Text feature [cases] present in test data point [True]
541 Text feature [essential] present in test data point [True]
542 Text feature [phase] present in test data point [True]
545 Text feature [relatively] present in test data point [True]
```

```
546 Text feature [included] present in test data point [True]
550 Text feature [site] present in test data point [True]
554 Text feature [via] present in test data point [True]
555 Text feature [70] present in test data point [True]
556 Text feature [nucleotide] present in test data point [True]
558 Text feature [based] present in test data point [True]
560 Text feature [involved] present in test data point [True]
561 Text feature [molecule] present in test data point [True]
562 Text feature [31] present in test data point [True]
565 Text feature [selection] present in test data point [True]
566 Text feature [position] present in test data point [True]
568 Text feature [like] present in test data point [True]
570 Text feature [recurrent] present in test data point [True]
571 Text feature [genetic] present in test data point [True]
572 Text feature [encoding] present in test data point [True]
574 Text feature [specifically] present in test data point [True]
576 Text feature [panel] present in test data point [True]
577 Text feature [induction] present in test data point [True]
579 Text feature [phenotype] present in test data point [True]
581 Text feature [appears] present in test data point [True]
584 Text feature [available] present in test data point [True]
585 Text feature [published] present in test data point [True]
590 Text feature [ability] present in test data point [True]
592 Text feature [products] present in test data point [True]
593 Text feature [association] present in test data point [True]
596 Text feature [regions] present in test data point [True]
600 Text feature [alterations] present in test data point [True]
605 Text feature [mutational] present in test data point [True]
607 Text feature [assess] present in test data point [True]
611 Text feature [process] present in test data point [True]
612 Text feature [larger] present in test data point [True]
614 Text feature [displayed] present in test data point [True]
615 Text feature [frame] present in test data point [True]
618 Text feature [see] present in test data point [True]
620 Text feature [changes] present in test data point [True]
623 Text feature [loop] present in test data point [True]
624 Text feature [bound] present in test data point [True]
625 Text feature [binding] present in test data point [True]
629 Text feature [affinity] present in test data point [True]
```

```
635 Text feature [relevant] present in test data point [True]
638 Text feature [assay] present in test data point [True]
644 Text feature [dose] present in test data point [True]
645 Text feature [sequence] present in test data point [True]
646 Text feature [bone] present in test data point [True]
649 Text feature [members] present in test data point [True]
650 Text feature [nature] present in test data point [True]
652 Text feature [positions] present in test data point [True]
655 Text feature [screening] present in test data point [True]
657 Text feature [affect] present in test data point [True]
660 Text feature [large] present in test data point [True]
661 Text feature [2005] present in test data point [True]
663 Text feature [chain] present in test data point [True]
665 Text feature [residues] present in test data point [True]
668 Text feature [state] present in test data point [True]
672 Text feature [case] present in test data point [True]
673 Text feature [partial] present in test data point [True]
685 Text feature [fusion] present in test data point [True]
687 Text feature [lack] present in test data point [True]
689 Text feature [interactions] present in test data point [True]
691 Text feature [clear] present in test data point [True]
692 Text feature [figures] present in test data point [True]
693 Text feature [showing] present in test data point [True]
695 Text feature [genome] present in test data point [True]
696 Text feature [average] present in test data point [True]
699 Text feature [pi3k] present in test data point [True]
700 Text feature [2004] present in test data point [True]
701 Text feature [blood] present in test data point [True]
702 Text feature [rna] present in test data point [True]
708 Text feature [reduction] present in test data point [True]
709 Text feature [cannot] present in test data point [True]
710 Text feature [acids] present in test data point [True]
711 Text feature [binds] present in test data point [True]
717 Text feature [substitutions] present in test data point [True]
721 Text feature [weeks] present in test data point [True]
722 Text feature [secondary] present in test data point [True]
724 Text feature [years] present in test data point [True]
725 Text feature [evidence] present in test data point [True]
726 Text feature [function] present in test data point [True]
```

```
727 Text feature [mrna] present in test data point [True]
730 Text feature [pattern] present in test data point [True]
731 Text feature [chemotherapy] present in test data point [True]
732 Text feature [residue] present in test data point [True]
733 Text feature [reaction] present in test data point [True]
735 Text feature [altered] present in test data point [True]
738 Text feature [regulatory] present in test data point [True]
741 Text feature [domains] present in test data point [True]
742 Text feature [mice] present in test data point [True]
743 Text feature [lesions] present in test data point [True]
746 Text feature [deletion] present in test data point [True]
747 Text feature [complex] present in test data point [True]
750 Text feature [rare] present in test data point [True]
752 Text feature [2008] present in test data point [True]
753 Text feature [helix] present in test data point [True]
756 Text feature [2001] present in test data point [True]
757 Text feature [value] present in test data point [True]
760 Text feature [across] present in test data point [True]
763 Text feature [impaired] present in test data point [True]
764 Text feature [38] present in test data point [True]
766 Text feature [motif] present in test data point [True]
768 Text feature [tagged] present in test data point [True]
769 Text feature [unique] present in test data point [True]
771 Text feature [2006] present in test data point [True]
776 Text feature [related] present in test data point [True]
779 Text feature [defined] present in test data point [True]
780 Text feature [supplemental] present in test data point [True]
781 Text feature [half] present in test data point [True]
782 Text feature [contact] present in test data point [True]
783 Text feature [models] present in test data point [True]
785 Text feature [bind] present in test data point [True]
786 Text feature [combined] present in test data point [True]
787 Text feature [functions] present in test data point [True]
789 Text feature [ref] present in test data point [True]
790 Text feature [transcription] present in test data point [True]
791 Text feature [cycle] present in test data point [True]
792 Text feature [remaining] present in test data point [True]
794 Text feature [basis] present in test data point [True]
796 Text feature [2009] present in test data point [True]
```

```
802 Text feature [dna] present in test data point [True]
803 Text feature [mtor] present in test data point [True]
812 Text feature [structural] present in test data point [True]
814 Text feature [construct] present in test data point [True]
816 Text feature [locus] present in test data point [True]
817 Text feature [distribution] present in test data point [True]
818 Text feature [interact] present in test data point [True]
820 Text feature [differentiation] present in test data point [True]
821 Text feature [set] present in test data point [True]
822 Text feature [alternative] present in test data point [True]
823 Text feature [52] present in test data point [True]
825 Text feature [cluster] present in test data point [True]
826 Text feature [2010] present in test data point [True]
827 Text feature [linked] present in test data point [True]
828 Text feature [population] present in test data point [True]
833 Text feature [chromosome] present in test data point [True]
838 Text feature [conserved] present in test data point [True]
839 Text feature [deletions] present in test data point [True]
840 Text feature [2012] present in test data point [True]
843 Text feature [allele] present in test data point [True]
846 Text feature [2007] present in test data point [True]
847 Text feature [structures] present in test data point [True]
848 Text feature [gfp] present in test data point [True]
849 Text feature [interaction] present in test data point [True]
851 Text feature [structure] present in test data point [True]
852 Text feature [controls] present in test data point [True]
853 Text feature [sporadic] present in test data point [True]
855 Text feature [promoter] present in test data point [True]
857 Text feature [terminus] present in test data point [True]
859 Text feature [functional] present in test data point [True]
860 Text feature [dominant] present in test data point [True]
863 Text feature [akt1] present in test data point [True]
865 Text feature [affected] present in test data point [True]
866 Text feature [fraction] present in test data point [True]
867 Text feature [heterozygous] present in test data point [True]
868 Text feature [1997] present in test data point [True]
872 Text feature [family] present in test data point [True]
878 Text feature [loss] present in test data point [True]
880 Text feature [co] present in test data point [True]
```

```
883 Text feature [surface] present in test data point [True]
885 Text feature [95] present in test data point [True]
887 Text feature [2013] present in test data point [True]
888 Text feature [knockdown] present in test data point [True]
889 Text feature [src] present in test data point [True]
890 Text feature [rt] present in test data point [True]
895 Text feature [2011] present in test data point [True]
897 Text feature [rearrangements] present in test data point [True]
904 Text feature [cyclin] present in test data point [True]
905 Text feature [catalytic] present in test data point [True]
906 Text feature [d1] present in test data point [True]
907 Text feature [pik3ca] present in test data point [True]
919 Text feature [stem] present in test data point [True]
920 Text feature [59] present in test data point [True]
921 Text feature [enzyme] present in test data point [True]
927 Text feature [degradation] present in test data point [True]
928 Text feature [tables] present in test data point [True]
929 Text feature [myc] present in test data point [True]
931 Text feature [1996] present in test data point [True]
934 Text feature [germline] present in test data point [True]
935 Text feature [apoptosis] present in test data point [True]
936 Text feature [56] present in test data point [True]
941 Text feature [aberrant] present in test data point [True]
947 Text feature [substrate] present in test data point [True]
948 Text feature [suppression] present in test data point [True]
949 Text feature [lymphoma] present in test data point [True]
954 Text feature [specificity] present in test data point [True]
961 Text feature [individuals] present in test data point [True]
962 Text feature [alleles] present in test data point [True]
965 Text feature [phosphatase] present in test data point [True]
967 Text feature [transcriptional] present in test data point [True]
972 Text feature [missense] present in test data point [True]
973 Text feature [bcr] present in test data point [True]
979 Text feature [cdk4] present in test data point [True]
Out of the top  1000  features  555 are present in query point
```

4.2. K Nearest Neighbour Classification

4.2.1. Hyper parameter tuning

In [60]:
```python
# find more about KNeighborsClassifier() here http://scikit-learn.org/s
table/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
# -------------------------
# default parameter
# KNeighborsClassifier(n_neighbors=5, weights='uniform', algorithm='aut
o', 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 va
lues
# 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-geometric-intuition-with-a-toy-examp
le-1/
#------------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
tml
# ---------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='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:
```

```python
#---------------------------------------

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, labels=clf.classes_, eps=1e-15
))
```

```python
predict_y = sig_clf.predict_proba(cv_x_responseCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross vali
dation log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_, eps
=1e-15))
predict_y = sig_clf.predict_proba(test_x_responseCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log l
oss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 5
Log Loss : 1.009289796186825
for alpha = 11
Log Loss : 1.0098711830977605
for alpha = 15
Log Loss : 1.0088184029103286
for alpha = 21
Log Loss : 1.0157605335451443
for alpha = 31
Log Loss : 1.006811358196161
for alpha = 41
Log Loss : 1.0077544329599115
for alpha = 51
Log Loss : 1.0157040623257463
for alpha = 99
Log Loss : 1.036441186404738
```

Cross Validation Error for each alpha
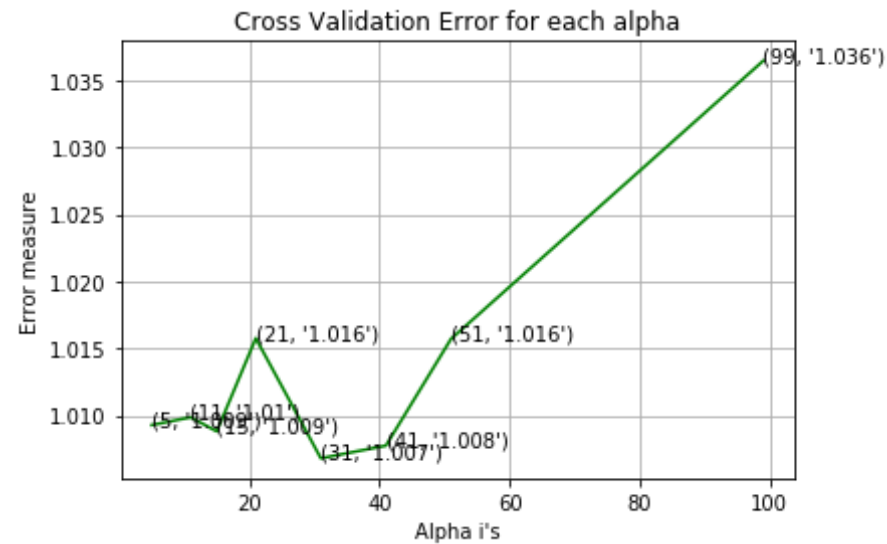
For values of best alpha = 31 The train log loss is: 0.8159966136934
336
For values of best alpha = 31 The cross validation log loss is: 1.00
6811358196161
For values of best alpha = 31 The test log loss is: 1.11492567096578

4.2.2. Testing the model with best hyper paramters

```
In [61]: # find more about KNeighborsClassifier() here http://scikit-learn.org/s
         table/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
         # -------------------------
         # default parameter
         # KNeighborsClassifier(n_neighbors=5, weights='uniform', algorithm='aut
         o', 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 va
         lues
         # 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-geometric-intuition-with-a-toy-examp
le-1/
#-------------------------------------
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
predict_and_plot_confusion_matrix(train_x_responseCoding, train_y, cv_x
_responseCoding, cv_y, clf)
```
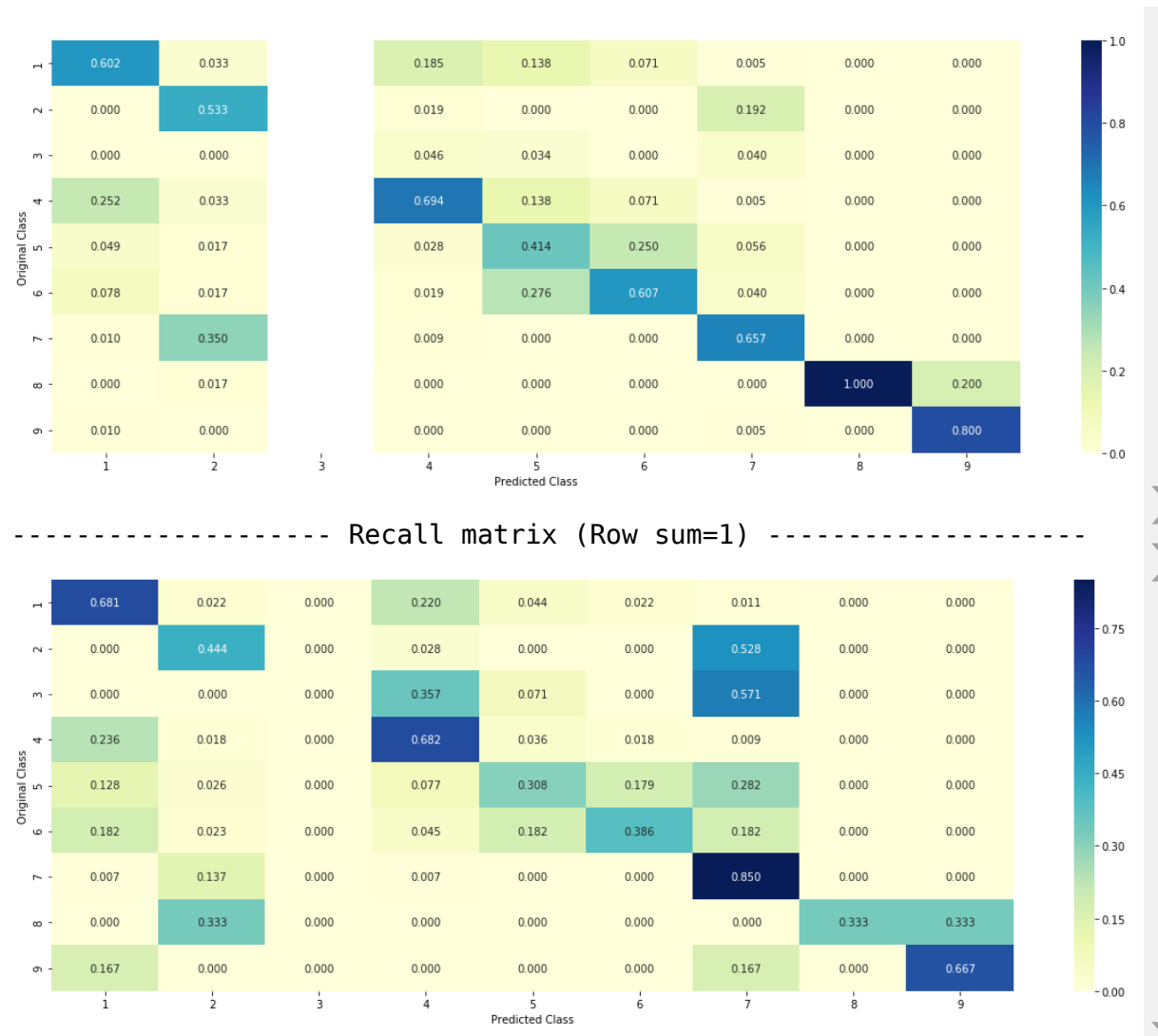
Log loss : 1.006811358196161
Number of mis-classified points : 0.37406015037593987
-------------------- Confusion matrix --------------------

-------------------- Precision matrix (Columm Sum=1) ----------------
----
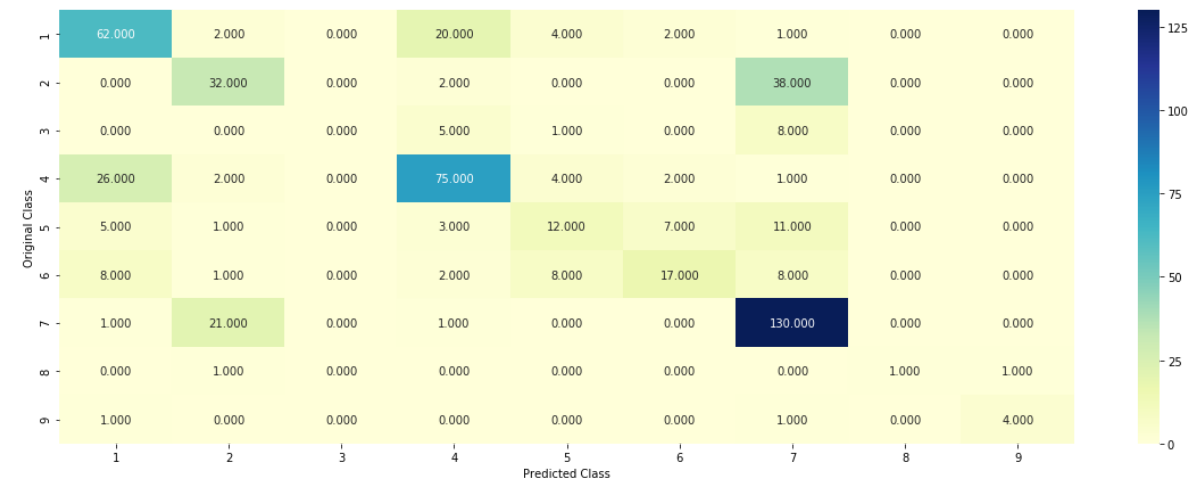
------------------- Recall matrix (Row sum=1) -------------------

4.2.3.Sample Query point -1

```
In [62]: 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].resh
ape(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 : 7
Actual Class : 1
The  31  nearest neighbours of the test points belongs to classes [3 1
1 4 4 4 4 4 1 1 4 4 4 4 5 1 4 1 1 1 1 4 4 4 4 1 4 1 4 1 1]
Fequency of nearest points : Counter({4: 16, 1: 13, 3: 1, 5: 1})
```

4.2.4. Sample Query Point-2

In [63]:
```
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 = 100

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].resh
ape(1, -1), alpha[best_alpha])
print("the k value for knn is",alpha[best_alpha],"and the nearest neigh
bours 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 : 7
Actual Class : 7
```

Actual class : 7

```
the k value for knn is 31 and the nearest neighbours of the test points
belongs to classes [7 7 7 7 7 5 7 7 2 7 2 7 7 7 1 2 7 2 2 7 3 7 3 3 7 3
3 7 3 7 7]
Fequency of nearest points : Counter({7: 18, 3: 6, 2: 5, 5: 1, 1: 1})
```

4.3. Logistic Regression

4.3.1. With Class balancing

4.3.1.1. Hyper paramter tuning

In [64]:
```python
# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_rate='optimal', eta0=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, …])     Fit linear model with S
tochastic 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.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
```

```
tml
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='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], p
enalty='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, 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 vali
dation 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 l
oss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.1558184685360064
for alpha = 1e-05
Log Loss : 1.0747091593756521
for alpha = 0.0001
Log Loss : 1.013845869676537
for alpha = 0.001
Log Loss : 1.011553133045781
for alpha = 0.01
Log Loss : 1.1703330443509026
for alpha = 0.1
Log Loss : 1.6090238049423566
for alpha = 1
Log Loss : 1.7468688641379477
for alpha = 10
Log Loss : 1.7623967186227167
for alpha = 100
Log Loss : 1.7642824325118118
```

Cross Validation Error for each alpha
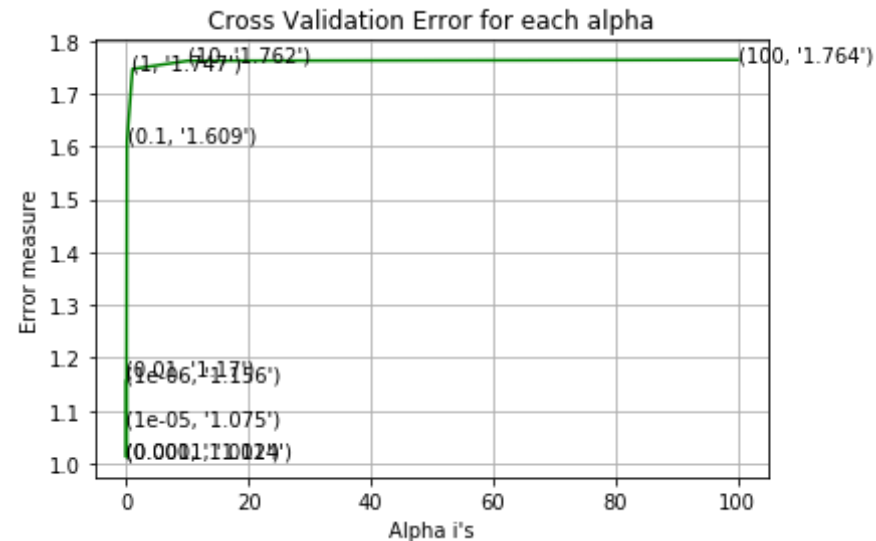
For values of best alpha = 0.001 The train log loss is: 0.7204015243541387
For values of best alpha = 0.001 The cross validation log loss is: 1.011553133045781
For values of best alpha = 0.001 The test log loss is: 1.1077803423049677

4.3.1.2. Testing the model with best hyper paramters

```
In [65]: # read more about SGDClassifier() at http://scikit-learn.org/stable/mod
         ules/generated/sklearn.linear_model.SGDClassifier.html
         # ----------------------------
         # default parameters
         # SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
         5, fit_intercept=True, max_iter=None, tol=None,
         # shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
         arning_rate='optimal', eta0=0.0, power_t=0.5,
         # class_weight=None, warm_start=False, average=False, n_iter=None)
```

```
# some of methods
# fit(X, y[, coef_init, intercept_init, …])     Fit linear model with S
tochastic 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], p
enalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_o
nehotCoding, cv_y, clf)
```

```
Log loss : 1.011553133045781
Number of mis-classified points : 0.34962406015037595
------------------- Confusion matrix --------------------
```



```
------------------- Precision matrix (Columm Sum=1) ---------------
----
```
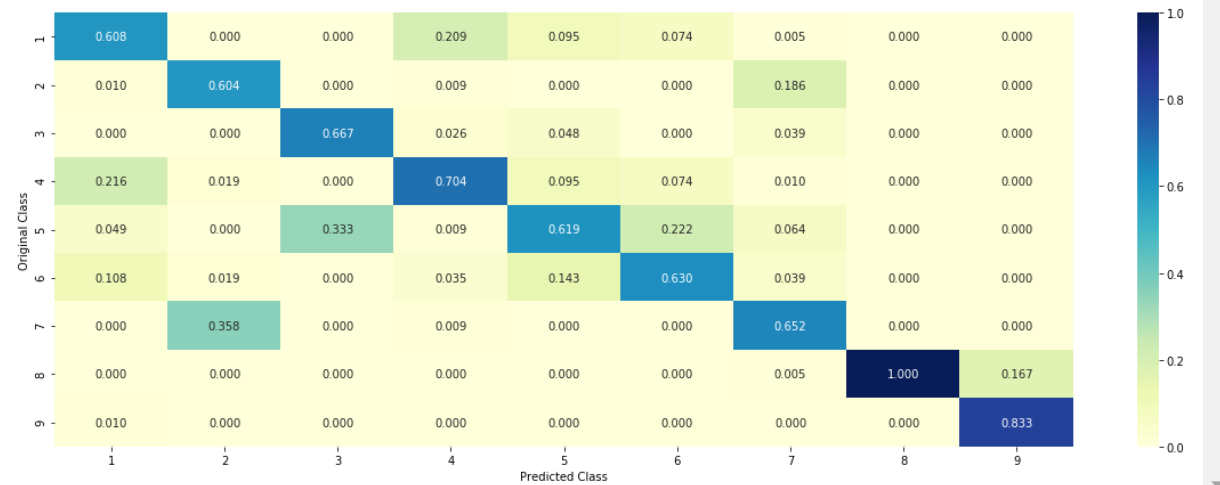
-------------------- Recall matrix (Row sum=1) --------------------
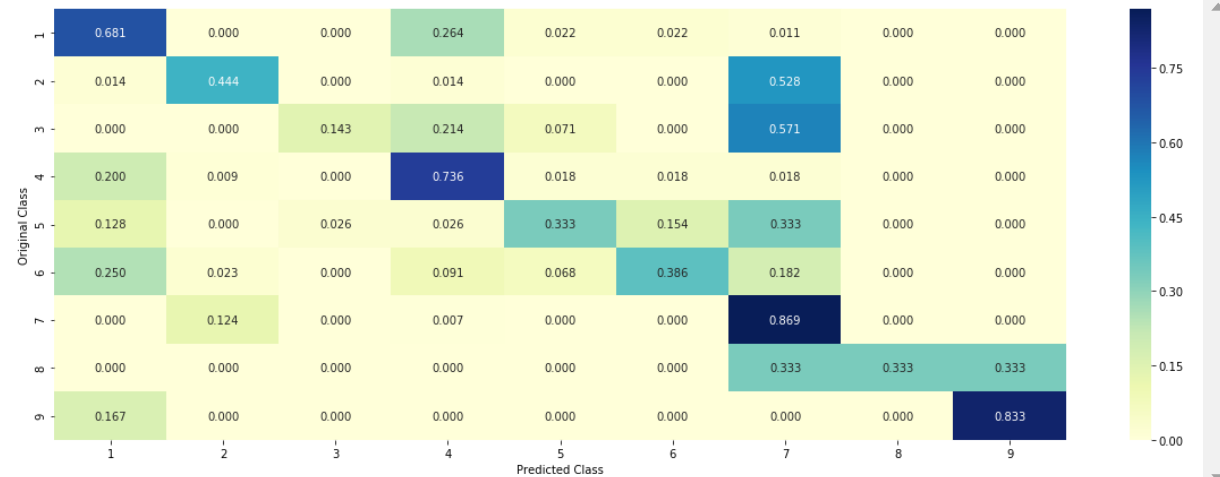
4.3.1.3. Feature Importance

```
In [66]: def get_imp_feature_names(text, indices, removed_ind = []):
             word_present = 0
             tabulte_list = []
```

```python
        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 que
ry point")
    print("-"*50)
    print("The features that are most importent of the ",predicted_cls[
0]," class:")
    print (tabulate(tabulte_list, headers=["Index",'Feature name', 'Pre
sent or Not']))
```

4.3.1.3.1. Correctly Classified point

```python
In [67]:   # from tabulate import tabulate
           clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], p
           enalty='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],test_df['Variation'].iloc[test
_point_index], no_feature)
```

Predicted Class : 1
Predicted Class Probabilities: [[6.712e-01 5.100e-03 9.000e-04 3.048e-0
1 5.400e-03 8.600e-03 1.600e-03
  2.400e-03 1.000e-04]]
Actual Class : 1
----
14 Text feature [repeats] present in test data point [True]
15 Text feature [surface] present in test data point [True]
16 Text feature [hotspots] present in test data point [True]
20 Text feature [specificity] present in test data point [True]
22 Text feature [deletion] present in test data point [True]
25 Text feature [deficient] present in test data point [True]
26 Text feature [transcriptional] present in test data point [True]
27 Text feature [structure] present in test data point [True]
29 Text feature [defined] present in test data point [True]
32 Text feature [nucleus] present in test data point [True]
35 Text feature [loss] present in test data point [True]
68 Text feature [deletions] present in test data point [True]
133 Text feature [peptide] present in test data point [True]
140 Text feature [remaining] present in test data point [True]
204 Text feature [sequenced] present in test data point [True]
209 Text feature [hydrophobic] present in test data point [True]
223 Text feature [repair] present in test data point [True]
225 Text feature [position] present in test data point [True]
231 Text feature [region] present in test data point [True]
237 Text feature [dna] present in test data point [True]
241 Text feature [domains] present in test data point [True]
245 Text feature [primers] present in test data point [True]
283 Text feature [transcription] present in test data point [True]
298 Text feature [box] present in test data point [True]
316 Text feature [early] present in test data point [True]
333 Text feature [somatic] present in test data point [True]
336 Text feature [corresponding] present in test data point [True]
342 Text feature [essential] present in test data point [True]
345 Text feature [structural] present in test data point [True]
346 Text feature [binding] present in test data point [True]
```

347 Text feature [length] present in test data point [True]

350 Text feature [rather] present in test data point [True]
353 Text feature [hotspot] present in test data point [True]
355 Text feature [calculated] present in test data point [True]
358 Text feature [genes] present in test data point [True]
360 Text feature [molecules] present in test data point [True]
364 Text feature [size] present in test data point [True]
374 Text feature [effect] present in test data point [True]
375 Text feature [complexes] present in test data point [True]
376 Text feature [stability] present in test data point [True]
378 Text feature [sequencing] present in test data point [True]
381 Text feature [crystal] present in test data point [True]
384 Text feature [identify] present in test data point [True]
387 Text feature [structures] present in test data point [True]
389 Text feature [function] present in test data point [True]
393 Text feature [defects] present in test data point [True]
396 Text feature [therefore] present in test data point [True]
401 Text feature [diagnosis] present in test data point [True]
410 Text feature [p53] present in test data point [True]
412 Text feature [insertion] present in test data point [True]
413 Text feature [side] present in test data point [True]
415 Text feature [strong] present in test data point [True]
416 Text feature [conserved] present in test data point [True]
418 Text feature [individual] present in test data point [True]
420 Text feature [large] present in test data point [True]
422 Text feature [interact] present in test data point [True]
425 Text feature [regions] present in test data point [True]
426 Text feature [types] present in test data point [True]
427 Text feature [gel] present in test data point [True]
428 Text feature [positions] present in test data point [True]
430 Text feature [mean] present in test data point [True]
431 Text feature [nuclear] present in test data point [True]
432 Text feature [contains] present in test data point [True]
433 Text feature [strand] present in test data point [True]
435 Text feature [subunit] present in test data point [True]
438 Text feature [reduction] present in test data point [True]
439 Text feature [one] present in test data point [True]
446 Text feature [insertions] present in test data point [True]
448 Text feature [example] present in test data point [True]
449 Text feature [role] present in test data point [True]

450 Text feature [assays] present in test data point [True]
457 Text feature [present] present in test data point [True]
459 Text feature [regulatory] present in test data point [True]
462 Text feature [tumour] present in test data point [True]
463 Text feature [allele] present in test data point [True]
464 Text feature [finding] present in test data point [True]
465 Text feature [core] present in test data point [True]
466 Text feature [ref] present in test data point [True]
470 Text feature [control] present in test data point [True]
472 Text feature [nucleotide] present in test data point [True]
474 Text feature [via] present in test data point [True]
477 Text feature [significantly] present in test data point [True]
482 Text feature [complex] present in test data point [True]
483 Text feature [containing] present in test data point [True]
485 Text feature [activities] present in test data point [True]
486 Text feature [derived] present in test data point [True]
488 Text feature [chain] present in test data point [True]
491 Text feature [screening] present in test data point [True]
492 Text feature [residue] present in test data point [True]
494 Text feature [established] present in test data point [True]
495 Text feature [co] present in test data point [True]
496 Text feature [protein] present in test data point [True]
497 Text feature [change] present in test data point [True]
499 Text feature [reverse] present in test data point [True]
504 Text feature [colon] present in test data point [True]
505 Text feature [additional] present in test data point [True]
506 Text feature [displayed] present in test data point [True]
507 Text feature [involved] present in test data point [True]
513 Text feature [least] present in test data point [True]
515 Text feature [specifically] present in test data point [True]
518 Text feature [following] present in test data point [True]
519 Text feature [shows] present in test data point [True]
520 Text feature [yet] present in test data point [True]
522 Text feature [variants] present in test data point [True]
524 Text feature [constructs] present in test data point [True]
525 Text feature [pcr] present in test data point [True]
527 Text feature [based] present in test data point [True]
529 Text feature [human] present in test data point [True]
532 Text feature [wild] present in test data point [True]

532 Text feature [mild] present in test data point [True]
535 Text feature [frequency] present in test data point [True]
537 Text feature [http] present in test data point [True]
539 Text feature [interactions] present in test data point [True]
540 Text feature [terminal] present in test data point [True]
541 Text feature [alterations] present in test data point [True]
544 Text feature [sample] present in test data point [True]
546 Text feature [years] present in test data point [True]
548 Text feature [full] present in test data point [True]
552 Text feature [16] present in test data point [True]
553 Text feature [cancers] present in test data point [True]
555 Text feature [products] present in test data point [True]
556 Text feature [21] present in test data point [True]
557 Text feature [suppression] present in test data point [True]
559 Text feature [appears] present in test data point [True]
560 Text feature [manufacturer] present in test data point [True]
562 Text feature [lack] present in test data point [True]
563 Text feature [possible] present in test data point [True]
567 Text feature [result] present in test data point [True]
568 Text feature [tumours] present in test data point [True]
571 Text feature [type] present in test data point [True]
572 Text feature [exons] present in test data point [True]
574 Text feature [larger] present in test data point [True]
575 Text feature [genetic] present in test data point [True]
579 Text feature [critical] present in test data point [True]
580 Text feature [samples] present in test data point [True]
581 Text feature [fold] present in test data point [True]
583 Text feature [often] present in test data point [True]
586 Text feature [proteins] present in test data point [True]
587 Text feature [controls] present in test data point [True]
588 Text feature [primer] present in test data point [True]
589 Text feature [predict] present in test data point [True]
592 Text feature [ability] present in test data point [True]
595 Text feature [residues] present in test data point [True]
597 Text feature [defective] present in test data point [True]
598 Text feature [cancer] present in test data point [True]
599 Text feature [series] present in test data point [True]
600 Text feature [cases] present in test data point [True]
602 Text feature [damage] present in test data point [True]
603 Text feature [bind] present in test data point [True]

603 Text feature [bind] present in test data point [True]
604 Text feature [methods] present in test data point [True]
605 Text feature [binds] present in test data point [True]
606 Text feature [purified] present in test data point [True]
607 Text feature [performed] present in test data point [True]
609 Text feature [sequences] present in test data point [True]
613 Text feature [materials] present in test data point [True]
616 Text feature [small] present in test data point [True]
617 Text feature [within] present in test data point [True]
619 Text feature [tp53] present in test data point [True]
621 Text feature [39] present in test data point [True]
622 Text feature [strongly] present in test data point [True]
623 Text feature [biological] present in test data point [True]
624 Text feature [testing] present in test data point [True]
625 Text feature [missense] present in test data point [True]
626 Text feature [molecule] present in test data point [True]
627 Text feature [32] present in test data point [True]
631 Text feature [identified] present in test data point [True]
633 Text feature [sufficient] present in test data point [True]
634 Text feature [25] present in test data point [True]
635 Text feature [colonies] present in test data point [True]
641 Text feature [major] present in test data point [True]
645 Text feature [cell] present in test data point [True]
646 Text feature [chemotherapy] present in test data point [True]
648 Text feature [confirmed] present in test data point [True]
649 Text feature [five] present in test data point [True]
652 Text feature [functional] present in test data point [True]
654 Text feature [46] present in test data point [True]
656 Text feature [sequence] present in test data point [True]
657 Text feature [form] present in test data point [True]
659 Text feature [majority] present in test data point [True]
660 Text feature [death] present in test data point [True]
662 Text feature [non] present in test data point [True]
663 Text feature [highly] present in test data point [True]
664 Text feature [key] present in test data point [True]
667 Text feature [selected] present in test data point [True]
668 Text feature [likely] present in test data point [True]
670 Text feature [cluster] present in test data point [True]
675 Text feature [obtained] present in test data point [True]
682 Text feature [negative] present in test data point [True]

```
683 Text feature [class] present in test data point [True]
691 Text feature [vivo] present in test data point [True]
693 Text feature [furthermore] present in test data point [True]
694 Text feature [domain] present in test data point [True]
696 Text feature [like] present in test data point [True]
699 Text feature [development] present in test data point [True]
708 Text feature [2001] present in test data point [True]
714 Text feature [figures] present in test data point [True]
733 Text feature [s2] present in test data point [True]
748 Text feature [degradation] present in test data point [True]
756 Text feature [another] present in test data point [True]
757 Text feature [remains] present in test data point [True]
789 Text feature [gene] present in test data point [True]
800 Text feature [sites] present in test data point [True]
816 Text feature [specific] present in test data point [True]
824 Text feature [contact] present in test data point [True]
849 Text feature [alleles] present in test data point [True]
904 Text feature [see] present in test data point [True]
917 Text feature [red] present in test data point [True]
918 Text feature [carried] present in test data point [True]
944 Text feature [together] present in test data point [True]
982 Text feature [showing] present in test data point [True]
Out of the top  1000  features  209 are present in query point
```

4.3.1.3.2. Incorrectly Classified point

In [68]:
```python
test_point_index = 55
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],test_df['Variation'].iloc[test
_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.085  0.1627 0.0042 0.168  0.0184 0.0
116 0.5367 0.0099 0.0034]]
Actual Class : 4
-----------------------------------------------------
15 Text feature [constitutive] present in test data point [True]
16 Text feature [activation] present in test data point [True]
18 Text feature [activated] present in test data point [True]
22 Text feature [transformed] present in test data point [True]
24 Text feature [oncogene] present in test data point [True]
26 Text feature [downstream] present in test data point [True]
37 Text feature [transforming] present in test data point [True]
40 Text feature [activate] present in test data point [True]
44 Text feature [oncogenic] present in test data point [True]
48 Text feature [transformation] present in test data point [True]
49 Text feature [constitutively] present in test data point [True]
50 Text feature [overexpression] present in test data point [True]
52 Text feature [tyrosine] present in test data point [True]
53 Text feature [phospho] present in test data point [True]
55 Text feature [expressing] present in test data point [True]
58 Text feature [signaling] present in test data point [True]
63 Text feature [phosphorylation] present in test data point [True]
65 Text feature [pathways] present in test data point [True]
76 Text feature [enhanced] present in test data point [True]
77 Text feature [inhibited] present in test data point [True]
78 Text feature [factor] present in test data point [True]
83 Text feature [akt] present in test data point [True]
86 Text feature [inhibitor] present in test data point [True]
88 Text feature [malignant] present in test data point [True]
89 Text feature [2a] present in test data point [True]
94 Text feature [promote] present in test data point [True]
95 Text feature [bone] present in test data point [True]
111 Text feature [express] present in test data point [True]
113 Text feature [fold] present in test data point [True]
118 Text feature [amplification] present in test data point [True]
120 Text feature [kinase] present in test data point [True]
122 Text feature [proliferation] present in test data point [True]
```

```
132 Text feature [sensitive] present in test data point [True]
134 Text feature [3b] present in test data point [True]
138 Text feature [growth] present in test data point [True]
146 Text feature [survival] present in test data point [True]
149 Text feature [cells] present in test data point [True]
153 Text feature [independent] present in test data point [True]
167 Text feature [factors] present in test data point [True]
168 Text feature [receptor] present in test data point [True]
171 Text feature [days] present in test data point [True]
182 Text feature [inhibition] present in test data point [True]
184 Text feature [treated] present in test data point [True]
200 Text feature [pathway] present in test data point [True]
206 Text feature [supplemental] present in test data point [True]
207 Text feature [akt1] present in test data point [True]
211 Text feature [d1] present in test data point [True]
260 Text feature [effective] present in test data point [True]
284 Text feature [wt] present in test data point [True]
287 Text feature [harboring] present in test data point [True]
290 Text feature [driven] present in test data point [True]
326 Text feature [leading] present in test data point [True]
345 Text feature [interestingly] present in test data point [True]
412 Text feature [induction] present in test data point [True]
415 Text feature [various] present in test data point [True]
417 Text feature [absence] present in test data point [True]
440 Text feature [fusion] present in test data point [True]
452 Text feature [2b] present in test data point [True]
468 Text feature [increased] present in test data point [True]
531 Text feature [demonstrated] present in test data point [True]
559 Text feature [patient] present in test data point [True]
597 Text feature [derived] present in test data point [True]
611 Text feature [genomic] present in test data point [True]
620 Text feature [mtor] present in test data point [True]
636 Text feature [mechanism] present in test data point [True]
639 Text feature [positive] present in test data point [True]
649 Text feature [reverse] present in test data point [True]
665 Text feature [progression] present in test data point [True]
685 Text feature [combination] present in test data point [True]
687 Text feature [kinases] present in test data point [True]
707 Text feature [tumors] present in test data point [True]
708 Text feature [trials] present in test data point [True]
```

```
716 Text feature [regulated] present in test data point [True]
721 Text feature [observations] present in test data point [True]
731 Text feature [mutants] present in test data point [True]
732 Text feature [src] present in test data point [True]
764 Text feature [potential] present in test data point [True]
782 Text feature [occur] present in test data point [True]
785 Text feature [mechanisms] present in test data point [True]
792 Text feature [stage] present in test data point [True]
801 Text feature [lines] present in test data point [True]
802 Text feature [presence] present in test data point [True]
805 Text feature [years] present in test data point [True]
808 Text feature [primary] present in test data point [True]
810 Text feature [screen] present in test data point [True]
811 Text feature [contrast] present in test data point [True]
815 Text feature [long] present in test data point [True]
820 Text feature [treatment] present in test data point [True]
827 Text feature [inhibitors] present in test data point [True]
829 Text feature [without] present in test data point [True]
830 Text feature [like] present in test data point [True]
832 Text feature [pi3k] present in test data point [True]
834 Text feature [increase] present in test data point [True]
840 Text feature [3a] present in test data point [True]
843 Text feature [lead] present in test data point [True]
844 Text feature [transcription] present in test data point [True]
845 Text feature [anti] present in test data point [True]
848 Text feature [addition] present in test data point [True]
853 Text feature [upon] present in test data point [True]
854 Text feature [fig] present in test data point [True]
855 Text feature [form] present in test data point [True]
856 Text feature [blood] present in test data point [True]
857 Text feature [nucleotide] present in test data point [True]
859 Text feature [endogenous] present in test data point [True]
860 Text feature [induced] present in test data point [True]
863 Text feature [cyclin] present in test data point [True]
864 Text feature [total] present in test data point [True]
866 Text feature [inhibitory] present in test data point [True]
867 Text feature [14] present in test data point [True]
869 Text feature [high] present in test data point [True]
870 Text feature [regulatory] present in test data point [True]
```

```
873 Text feature [somatic] present in test data point [True]
874 Text feature [myc] present in test data point [True]
875 Text feature [showed] present in test data point [True]
876 Text feature [target] present in test data point [True]
879 Text feature [led] present in test data point [True]
884 Text feature [gain] present in test data point [True]
886 Text feature [differentiation] present in test data point [True]
887 Text feature [point] present in test data point [True]
888 Text feature [well] present in test data point [True]
890 Text feature [early] present in test data point [True]
892 Text feature [mice] present in test data point [True]
894 Text feature [therapeutic] present in test data point [True]
899 Text feature [strongly] present in test data point [True]
900 Text feature [targeting] present in test data point [True]
901 Text feature [stem] present in test data point [True]
904 Text feature [lesions] present in test data point [True]
906 Text feature [fact] present in test data point [True]
908 Text feature [mouse] present in test data point [True]
909 Text feature [tumor] present in test data point [True]
910 Text feature [consistent] present in test data point [True]
911 Text feature [generation] present in test data point [True]
912 Text feature [frequent] present in test data point [True]
913 Text feature [position] present in test data point [True]
914 Text feature [often] present in test data point [True]
915 Text feature [mutated] present in test data point [True]
921 Text feature [cdk4] present in test data point [True]
922 Text feature [59] present in test data point [True]
924 Text feature [coding] present in test data point [True]
925 Text feature [higher] present in test data point [True]
926 Text feature [forms] present in test data point [True]
928 Text feature [sites] present in test data point [True]
932 Text feature [therapy] present in test data point [True]
934 Text feature [involving] present in test data point [True]
935 Text feature [common] present in test data point [True]
936 Text feature [mutant] present in test data point [True]
937 Text feature [cell] present in test data point [True]
938 Text feature [gfp] present in test data point [True]
940 Text feature [22] present in test data point [True]
941 Text feature [drug] present in test data point [True]
```

```
943 Text feature [mediated] present in test data point [True]
944 Text feature [24] present in test data point [True]
945 Text feature [events] present in test data point [True]
946 Text feature [since] present in test data point [True]
949 Text feature [cluster] present in test data point [True]
950 Text feature [differences] present in test data point [True]
953 Text feature [sample] present in test data point [True]
954 Text feature [corresponding] present in test data point [True]
960 Text feature [12] present in test data point [True]
963 Text feature [targets] present in test data point [True]
966 Text feature [molecule] present in test data point [True]
967 Text feature [experimental] present in test data point [True]
968 Text feature [might] present in test data point [True]
970 Text feature [products] present in test data point [True]
971 Text feature [promoter] present in test data point [True]
976 Text feature [strong] present in test data point [True]
978 Text feature [suggest] present in test data point [True]
979 Text feature [1a] present in test data point [True]
984 Text feature [found] present in test data point [True]
Out of the top  1000  features  169 are present in query point
```

4.3.2. Without Class balancing

4.3.2.1. Hyper paramter tuning

In [69]:
```python
# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# -----------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_rate='optimal', eta0=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
```

```python
# fit(X, y[, coef_init, intercept_init, …])       Fit linear model with S
tochastic 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.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
tml
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='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.
```

```python
        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, 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 vali
dation 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 l
oss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.1684087417355444
for alpha = 1e-05
Log Loss : 1.108817263154273
for alpha = 0.0001
Log Loss : 1.044262722473795
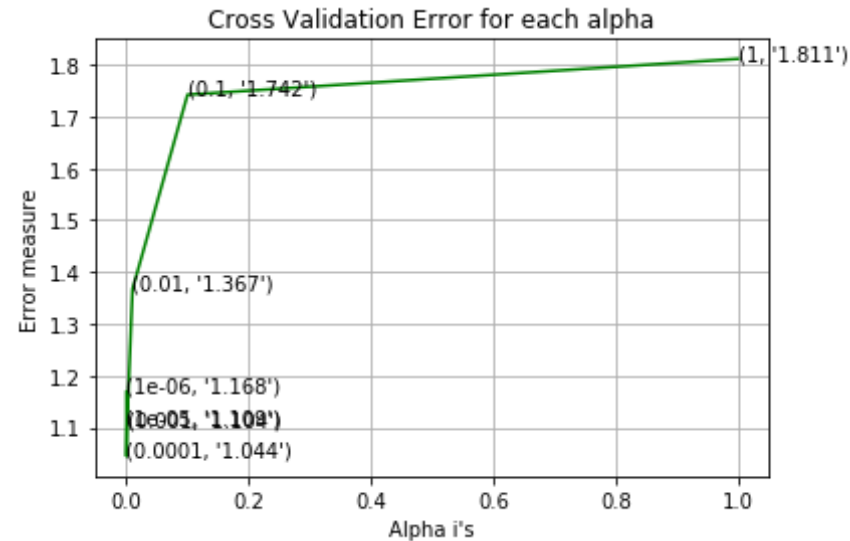for alpha = 0.001
```

```
Log Loss : 1.1039238023343
for alpha = 0.01
Log Loss : 1.3669263154131837
for alpha = 0.1
Log Loss : 1.7421605653561714

for alpha = 1
Log Loss : 1.811332701949407
```


Cross Validation Error for each alpha

```
For values of best alpha =  0.0001 The train log loss is: 0.43238101586
44072
For values of best alpha =  0.0001 The cross validation log loss is: 1.
044262722473795
For values of best alpha =  0.0001 The test log loss is: 1.117625914227
8473
```

4.3.2.2. Testing model with best hyper parameters

```
In [70]:  # read more about SGDClassifier() at http://scikit-learn.org/stable/mod
          ules/generated/sklearn.linear_model.SGDClassifier.html
          # ------------------------------
          # default parameters
```

```
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_rate='optimal', eta0=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

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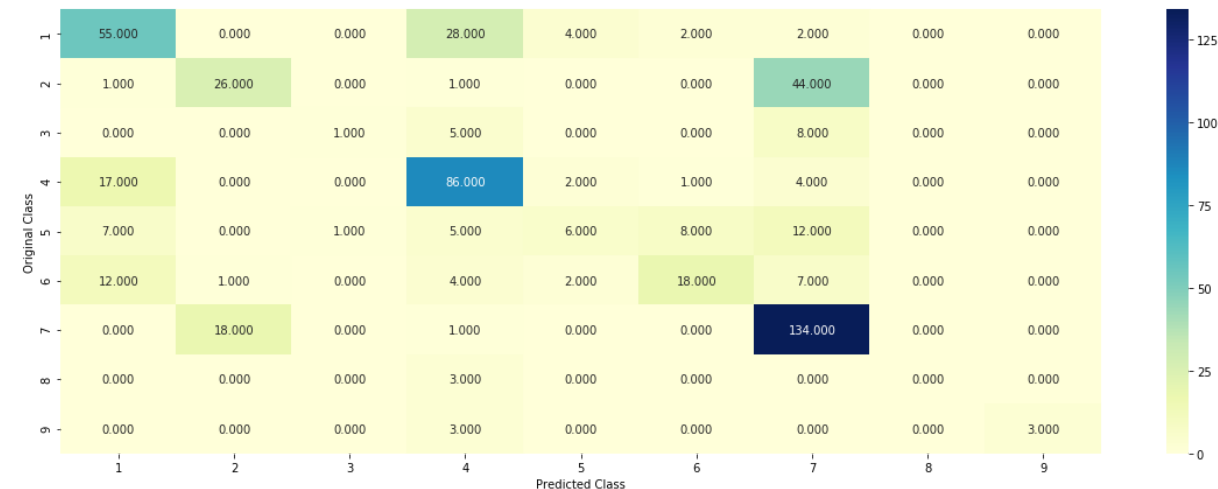# predict(X)    Predict class labels for samples in X.

#-------------------------------
# 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_o
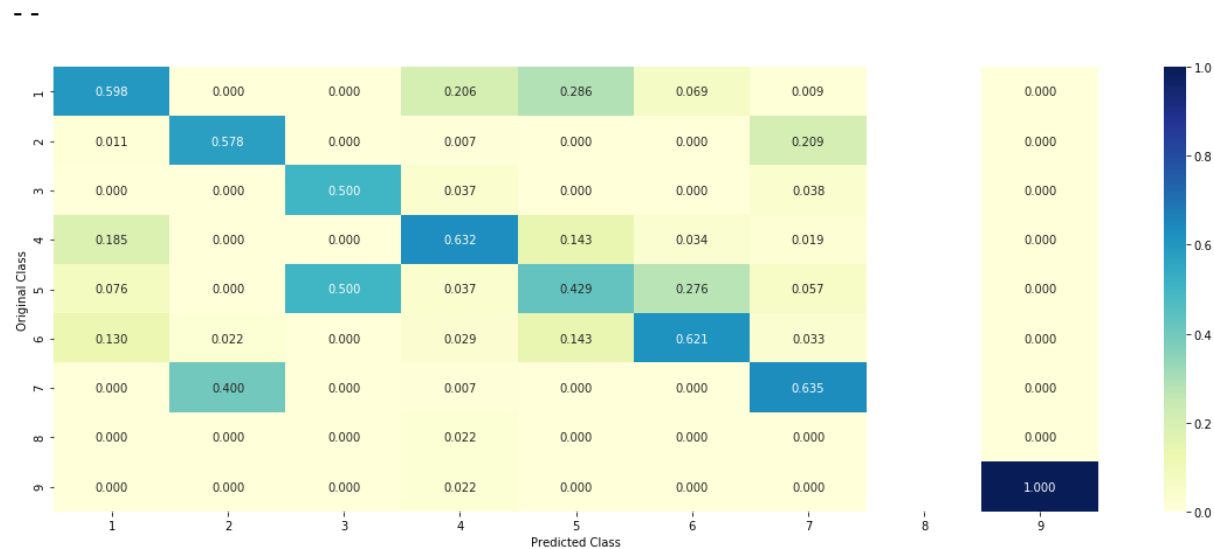nehotCoding, cv_y, clf)
```

```
Log loss : 1.044262722473795
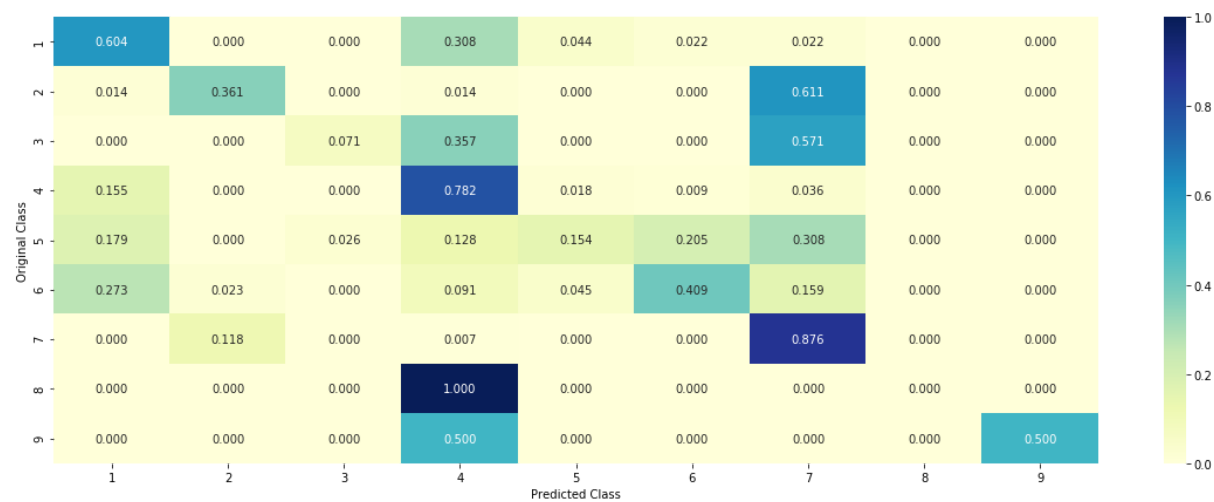Number of mis-classified points : 0.3815789473684211
------------------- Confusion matrix -------------------
```



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

--

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

4.3.2.3. Feature Importance, Correctly Classified point

```python
In [71]: clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log',
         random_state=42)
         clf.fit(train_x_onehotCoding,train_y)
         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],test_df['Variation'].iloc[test
         _point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0201 0.0548 0.0623 0.0368 0.0373 0.0
034 0.7805 0.004  0.0008]]
Actual Class : 7
--------------------------------------------------
20 Text feature [activation] present in test data point [True]
21 Text feature [transformed] present in test data point [True]
24 Text feature [activated] present in test data point [True]
25 Text feature [constitutive] present in test data point [True]
30 Text feature [downstream] present in test data point [True]
40 Text feature [stably] present in test data point [True]
42 Text feature [bone] present in test data point [True]
57 Text feature [oncogene] present in test data point [True]
58 Text feature [membrane] present in test data point [True]
71 Text feature [codon] present in test data point [True]
```

```
80 Text feature [phosphorylated] present in test data point [True]

85 Text feature [fold] present in test data point [True]
86 Text feature [activate] present in test data point [True]
95 Text feature [2a] present in test data point [True]
108 Text feature [overexpression] present in test data point [True]
160 Text feature [pathways] present in test data point [True]
164 Text feature [factor] present in test data point [True]
178 Text feature [oncogenic] present in test data point [True]
179 Text feature [colony] present in test data point [True]
184 Text feature [phospho] present in test data point [True]
201 Text feature [transforming] present in test data point [True]
204 Text feature [3b] present in test data point [True]
209 Text feature [examined] present in test data point [True]
214 Text feature [activating] present in test data point [True]
215 Text feature [akt] present in test data point [True]
231 Text feature [phosphorylation] present in test data point [True]
244 Text feature [express] present in test data point [True]
251 Text feature [factors] present in test data point [True]
254 Text feature [observations] present in test data point [True]
265 Text feature [expressing] present in test data point [True]
271 Text feature [elevated] present in test data point [True]
273 Text feature [positive] present in test data point [True]
282 Text feature [inhibitor] present in test data point [True]
283 Text feature [added] present in test data point [True]
293 Text feature [high] present in test data point [True]
297 Text feature [serum] present in test data point [True]
309 Text feature [000] present in test data point [True]
316 Text feature [ras] present in test data point [True]
346 Text feature [signaling] present in test data point [True]
348 Text feature [transformation] present in test data point [True]
351 Text feature [genomic] present in test data point [True]
354 Text feature [sensitive] present in test data point [True]
371 Text feature [consistent] present in test data point [True]
377 Text feature [demonstrated] present in test data point [True]
402 Text feature [increased] present in test data point [True]
411 Text feature [transcription] present in test data point [True]
417 Text feature [various] present in test data point [True]
423 Text feature [cdna] present in test data point [True]
424 Text feature [tyrosine] present in test data point [True]
```

```
434 Text feature [regions] present in test data point [True]

442 Text feature [wt] present in test data point [True]
447 Text feature [per] present in test data point [True]
474 Text feature [days] present in test data point [True]
475 Text feature [properties] present in test data point [True]
476 Text feature [conditions] present in test data point [True]
477 Text feature [reverse] present in test data point [True]
482 Text feature [supplemental] present in test data point [True]
491 Text feature [medium] present in test data point [True]
494 Text feature [enhanced] present in test data point [True]
500 Text feature [approach] present in test data point [True]
502 Text feature [position] present in test data point [True]
506 Text feature [ph] present in test data point [True]
513 Text feature [2b] present in test data point [True]
528 Text feature [download] present in test data point [True]
529 Text feature [years] present in test data point [True]
547 Text feature [3a] present in test data point [True]
552 Text feature [2000] present in test data point [True]
557 Text feature [constitutively] present in test data point [True]
560 Text feature [primer] present in test data point [True]
565 Text feature [common] present in test data point [True]
570 Text feature [cells] present in test data point [True]
572 Text feature [long] present in test data point [True]
573 Text feature [somatic] present in test data point [True]
589 Text feature [well] present in test data point [True]
592 Text feature [basal] present in test data point [True]
603 Text feature [promote] present in test data point [True]
613 Text feature [proliferation] present in test data point [True]
617 Text feature [colonies] present in test data point [True]
631 Text feature [signals] present in test data point [True]
632 Text feature [upon] present in test data point [True]
634 Text feature [interestingly] present in test data point [True]
635 Text feature [leads] present in test data point [True]
643 Text feature [like] present in test data point [True]
644 Text feature [mechanism] present in test data point [True]
645 Text feature [major] present in test data point [True]
661 Text feature [example] present in test data point [True]
668 Text feature [might] present in test data point [True]
671 Text feature [fusion] present in test data point [True]
```

```
672 Text feature [total] present in test data point [True]

675 Text feature [pathway] present in test data point [True]
682 Text feature [assessed] present in test data point [True]
690 Text feature [fig] present in test data point [True]
691 Text feature [kinases] present in test data point [True]
693 Text feature [five] present in test data point [True]
696 Text feature [endogenous] present in test data point [True]
699 Text feature [addition] present in test data point [True]
713 Text feature [free] present in test data point [True]
716 Text feature [strong] present in test data point [True]
719 Text feature [concentrations] present in test data point [True]
725 Text feature [75] present in test data point [True]
728 Text feature [driven] present in test data point [True]
735 Text feature [mediated] present in test data point [True]
739 Text feature [strongly] present in test data point [True]
744 Text feature [vector] present in test data point [True]
745 Text feature [could] present in test data point [True]
749 Text feature [24] present in test data point [True]
750 Text feature [tumors] present in test data point [True]
754 Text feature [independent] present in test data point [True]
757 Text feature [mutant] present in test data point [True]
765 Text feature [amplification] present in test data point [True]
772 Text feature [region] present in test data point [True]
777 Text feature [include] present in test data point [True]
781 Text feature [whereas] present in test data point [True]
786 Text feature [specific] present in test data point [True]
788 Text feature [generation] present in test data point [True]
793 Text feature [suggest] present in test data point [True]
794 Text feature [contrast] present in test data point [True]
800 Text feature [cancers] present in test data point [True]
804 Text feature [kinase] present in test data point [True]
805 Text feature [products] present in test data point [True]
806 Text feature [resistant] present in test data point [True]
808 Text feature [comparison] present in test data point [True]
811 Text feature [myc] present in test data point [True]
813 Text feature [mutants] present in test data point [True]
817 Text feature [22] present in test data point [True]
819 Text feature [14] present in test data point [True]
821 Text feature [presence] present in test data point [True]
```

831 Text feature [incubated] present in test data point [True]

833 Text feature [regulatory] present in test data point [True]
836 Text feature [previously] present in test data point [True]
838 Text feature [determined] present in test data point [True]
841 Text feature [tumor] present in test data point [True]
846 Text feature [clinically] present in test data point [True]
849 Text feature [biochemical] present in test data point [True]
851 Text feature [directly] present in test data point [True]
852 Text feature [showed] present in test data point [True]
853 Text feature [sites] present in test data point [True]
855 Text feature [associated] present in test data point [True]
856 Text feature [domain] present in test data point [True]
857 Text feature [use] present in test data point [True]
859 Text feature [point] present in test data point [True]
863 Text feature [pi3k] present in test data point [True]
864 Text feature [growth] present in test data point [True]
866 Text feature [nucleotide] present in test data point [True]
868 Text feature [primers] present in test data point [True]
869 Text feature [higher] present in test data point [True]
870 Text feature [within] present in test data point [True]
873 Text feature [sensitivity] present in test data point [True]
877 Text feature [report] present in test data point [True]
879 Text feature [13] present in test data point [True]
883 Text feature [increase] present in test data point [True]
886 Text feature [developed] present in test data point [True]
887 Text feature [levels] present in test data point [True]
888 Text feature [cellular] present in test data point [True]
890 Text feature [tumorigenesis] present in test data point [True]
891 Text feature [taken] present in test data point [True]
892 Text feature [corresponding] present in test data point [True]
900 Text feature [obtained] present in test data point [True]
901 Text feature [available] present in test data point [True]
904 Text feature [29] present in test data point [True]
906 Text feature [found] present in test data point [True]
907 Text feature [also] present in test data point [True]
912 Text feature [terminus] present in test data point [True]
915 Text feature [mm] present in test data point [True]
920 Text feature [12] present in test data point [True]
924 Text feature [biological] present in test data point [True]

```
926 Text feature [nm] present in test data point [True]

930 Text feature [cultured] present in test data point [True]
931 Text feature [kit] present in test data point [True]
932 Text feature [standard] present in test data point [True]
935 Text feature [clones] present in test data point [True]
936 Text feature [level] present in test data point [True]
937 Text feature [met] present in test data point [True]
946 Text feature [recently] present in test data point [True]
948 Text feature [direct] present in test data point [True]
956 Text feature [70] present in test data point [True]
958 Text feature [experimental] present in test data point [True]
960 Text feature [sample] present in test data point [True]
970 Text feature [two] present in test data point [True]
979 Text feature [similar] present in test data point [True]
983 Text feature [gain] present in test data point [True]
984 Text feature [culture] present in test data point [True]
986 Text feature [possible] present in test data point [True]
991 Text feature [transfected] present in test data point [True]
992 Text feature [mtor] present in test data point [True]
993 Text feature [screen] present in test data point [True]
994 Text feature [generated] present in test data point [True]
Out of the top  1000  features  187 are present in query point
```

4.3.2.4. Feature Importance, Inorrectly Classified point

```python
In [73]: test_point_index = 2
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],test_df['Variation'].iloc[test
_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[3.000e-04 1.067e-01 0.000e+00 1.000e-0
3 8.000e-04 2.000e-04 8.899e-01
  1.100e-03 0.000e+00]]
Actual Class : 6
----------------------------------------------------
20 Text feature [activation] present in test data point [True]
24 Text feature [activated] present in test data point [True]
25 Text feature [constitutive] present in test data point [True]
30 Text feature [downstream] present in test data point [True]
36 Text feature [insertion] present in test data point [True]
40 Text feature [stably] present in test data point [True]
42 Text feature [bone] present in test data point [True]
57 Text feature [oncogene] present in test data point [True]
58 Text feature [membrane] present in test data point [True]
71 Text feature [codon] present in test data point [True]
80 Text feature [phosphorylated] present in test data point [True]
85 Text feature [fold] present in test data point [True]
86 Text feature [activate] present in test data point [True]
95 Text feature [2a] present in test data point [True]
99 Text feature [derived] present in test data point [True]
108 Text feature [overexpression] present in test data point [True]
117 Text feature [inhibited] present in test data point [True]
143 Text feature [specimens] present in test data point [True]
160 Text feature [pathways] present in test data point [True]
164 Text feature [factor] present in test data point [True]
178 Text feature [oncogenic] present in test data point [True]
179 Text feature [colony] present in test data point [True]
190 Text feature [receptors] present in test data point [True]
195 Text feature [approximately] present in test data point [True]
196 Text feature [her2] present in test data point [True]
198 Text feature [carcinomas] present in test data point [True]
201 Text feature [transforming] present in test data point [True]
206 Text feature [epithelial] present in test data point [True]
209 Text feature [examined] present in test data point [True]
214 Text feature [activating] present in test data point [True]
215 Text feature [akt] present in test data point [True]
230 Text feature [provided] present in test data point [True]
233 Text feature [versus] present in test data point [True]
```

```
244 Text feature [express] present in test data point [True]
248 Text feature [patient] present in test data point [True]
251 Text feature [factors] present in test data point [True]
258 Text feature [lung] present in test data point [True]
265 Text feature [expressing] present in test data point [True]
269 Text feature [leading] present in test data point [True]
271 Text feature [elevated] present in test data point [True]
273 Text feature [positive] present in test data point [True]
282 Text feature [inhibitor] present in test data point [True]
283 Text feature [added] present in test data point [True]
293 Text feature [high] present in test data point [True]
297 Text feature [serum] present in test data point [True]
309 Text feature [000] present in test data point [True]
312 Text feature [ligand] present in test data point [True]
334 Text feature [3t3] present in test data point [True]
346 Text feature [signaling] present in test data point [True]
348 Text feature [transformation] present in test data point [True]
351 Text feature [genomic] present in test data point [True]
354 Text feature [sensitive] present in test data point [True]
371 Text feature [consistent] present in test data point [True]
372 Text feature [form] present in test data point [True]
377 Text feature [demonstrated] present in test data point [True]
402 Text feature [increased] present in test data point [True]
406 Text feature [liver] present in test data point [True]
410 Text feature [hours] present in test data point [True]
411 Text feature [transcription] present in test data point [True]
417 Text feature [various] present in test data point [True]
419 Text feature [raf] present in test data point [True]
423 Text feature [cdna] present in test data point [True]
424 Text feature [tyrosine] present in test data point [True]
434 Text feature [regions] present in test data point [True]
447 Text feature [per] present in test data point [True]
465 Text feature [leukemia] present in test data point [True]
466 Text feature [occur] present in test data point [True]
474 Text feature [days] present in test data point [True]
475 Text feature [properties] present in test data point [True]
476 Text feature [conditions] present in test data point [True]
477 Text feature [reverse] present in test data point [True]
479 Text feature [coding] present in test data point [True]
```

```
486 Text feature [insertions] present in test data point [True]
491 Text feature [medium] present in test data point [True]
494 Text feature [enhanced] present in test data point [True]
496 Text feature [erlotinib] present in test data point [True]
498 Text feature [metastatic] present in test data point [True]
502 Text feature [position] present in test data point [True]
512 Text feature [regulated] present in test data point [True]
513 Text feature [2b] present in test data point [True]
515 Text feature [tissues] present in test data point [True]
520 Text feature [harboring] present in test data point [True]
524 Text feature [51] present in test data point [True]
527 Text feature [egfr] present in test data point [True]
529 Text feature [years] present in test data point [True]
532 Text feature [malignant] present in test data point [True]
534 Text feature [survival] present in test data point [True]
550 Text feature [blood] present in test data point [True]
552 Text feature [2000] present in test data point [True]
553 Text feature [bp] present in test data point [True]
557 Text feature [constitutively] present in test data point [True]
560 Text feature [primer] present in test data point [True]
565 Text feature [common] present in test data point [True]
570 Text feature [cells] present in test data point [True]
572 Text feature [long] present in test data point [True]
573 Text feature [somatic] present in test data point [True]
589 Text feature [well] present in test data point [True]
591 Text feature [stimulated] present in test data point [True]
592 Text feature [basal] present in test data point [True]
600 Text feature [lead] present in test data point [True]
603 Text feature [promote] present in test data point [True]
606 Text feature [72] present in test data point [True]
611 Text feature [locus] present in test data point [True]
613 Text feature [proliferation] present in test data point [True]
621 Text feature [mechanisms] present in test data point [True]
623 Text feature [mapk] present in test data point [True]
631 Text feature [signals] present in test data point [True]
632 Text feature [upon] present in test data point [True]
633 Text feature [disease] present in test data point [True]
634 Text feature [interestingly] present in test data point [True]
635 Text feature [leads] present in test data point [True]
```

```
643 Text feature [like] present in test data point [True]
644 Text feature [mechanism] present in test data point [True]
645 Text feature [major] present in test data point [True]
648 Text feature [pdgfr] present in test data point [True]
659 Text feature [melanoma] present in test data point [True]
660 Text feature [stage] present in test data point [True]
661 Text feature [example] present in test data point [True]
663 Text feature [treated] present in test data point [True]
668 Text feature [might] present in test data point [True]
669 Text feature [greater] present in test data point [True]
671 Text feature [fusion] present in test data point [True]
672 Text feature [total] present in test data point [True]
675 Text feature [pathway] present in test data point [True]
681 Text feature [university] present in test data point [True]
682 Text feature [assessed] present in test data point [True]
683 Text feature [additional] present in test data point [True]
687 Text feature [fact] present in test data point [True]
690 Text feature [fig] present in test data point [True]
691 Text feature [kinases] present in test data point [True]
693 Text feature [five] present in test data point [True]
696 Text feature [endogenous] present in test data point [True]
699 Text feature [addition] present in test data point [True]
704 Text feature [carcinoma] present in test data point [True]
705 Text feature [series] present in test data point [True]
713 Text feature [free] present in test data point [True]
716 Text feature [strong] present in test data point [True]
719 Text feature [concentrations] present in test data point [True]
720 Text feature [er] present in test data point [True]
721 Text feature [prostate] present in test data point [True]
723 Text feature [59] present in test data point [True]
725 Text feature [75] present in test data point [True]
726 Text feature [nsclc] present in test data point [True]
728 Text feature [driven] present in test data point [True]
731 Text feature [often] present in test data point [True]
735 Text feature [mediated] present in test data point [True]
739 Text feature [strongly] present in test data point [True]
744 Text feature [vector] present in test data point [True]
745 Text feature [could] present in test data point [True]
749 Text feature [24] present in test data point [True]
```

```
750 Text feature [tumors] present in test data point [True]
751 Text feature [efficacy] present in test data point [True]
754 Text feature [independent] present in test data point [True]
755 Text feature [extracellular] present in test data point [True]
757 Text feature [mutant] present in test data point [True]
762 Text feature [process] present in test data point [True]
764 Text feature [progression] present in test data point [True]
765 Text feature [amplification] present in test data point [True]
767 Text feature [open] present in test data point [True]
769 Text feature [fusions] present in test data point [True]
772 Text feature [region] present in test data point [True]
777 Text feature [include] present in test data point [True]
780 Text feature [staining] present in test data point [True]
781 Text feature [whereas] present in test data point [True]
782 Text feature [tissue] present in test data point [True]
786 Text feature [specific] present in test data point [True]
787 Text feature [stimulation] present in test data point [True]
788 Text feature [generation] present in test data point [True]
789 Text feature [effective] present in test data point [True]
793 Text feature [suggest] present in test data point [True]
794 Text feature [contrast] present in test data point [True]
800 Text feature [cancers] present in test data point [True]
801 Text feature [responses] present in test data point [True]
804 Text feature [kinase] present in test data point [True]
805 Text feature [products] present in test data point [True]
806 Text feature [resistant] present in test data point [True]
808 Text feature [comparison] present in test data point [True]
812 Text feature [negative] present in test data point [True]
813 Text feature [mutants] present in test data point [True]
817 Text feature [22] present in test data point [True]
818 Text feature [ret] present in test data point [True]
819 Text feature [14] present in test data point [True]
820 Text feature [frequent] present in test data point [True]
821 Text feature [presence] present in test data point [True]
824 Text feature [brain] present in test data point [True]
828 Text feature [tkis] present in test data point [True]
830 Text feature [since] present in test data point [True]
831 Text feature [incubated] present in test data point [True]
832 Text feature [fgfr2] present in test data point [True]
```

833 Text feature [regulatory] present in test data point [True]
835 Text feature [without] present in test data point [True]
836 Text feature [previously] present in test data point [True]
838 Text feature [determined] present in test data point [True]
839 Text feature [green] present in test data point [True]
840 Text feature [exon] present in test data point [True]
841 Text feature [tumor] present in test data point [True]
844 Text feature [induction] present in test data point [True]
845 Text feature [distinct] present in test data point [True]
846 Text feature [clinically] present in test data point [True]
847 Text feature [49] present in test data point [True]
851 Text feature [directly] present in test data point [True]
852 Text feature [showed] present in test data point [True]
853 Text feature [sites] present in test data point [True]
854 Text feature [f3] present in test data point [True]
855 Text feature [associated] present in test data point [True]
856 Text feature [domain] present in test data point [True]
857 Text feature [use] present in test data point [True]
859 Text feature [point] present in test data point [True]
863 Text feature [pi3k] present in test data point [True]
864 Text feature [growth] present in test data point [True]
865 Text feature [contribute] present in test data point [True]
867 Text feature [context] present in test data point [True]
868 Text feature [primers] present in test data point [True]
869 Text feature [higher] present in test data point [True]
870 Text feature [within] present in test data point [True]
871 Text feature [inhibitory] present in test data point [True]
873 Text feature [sensitivity] present in test data point [True]
877 Text feature [report] present in test data point [True]
879 Text feature [13] present in test data point [True]
882 Text feature [52] present in test data point [True]
883 Text feature [increase] present in test data point [True]
884 Text feature [flt3] present in test data point [True]
886 Text feature [developed] present in test data point [True]
887 Text feature [levels] present in test data point [True]
888 Text feature [cellular] present in test data point [True]
890 Text feature [tumorigenesis] present in test data point [True]
892 Text feature [corresponding] present in test data point [True]
894 Text feature [early] present in test data point [True]

```
895 Text feature [led] present in test data point [True]
898 Text feature [double] present in test data point [True]
900 Text feature [obtained] present in test data point [True]
901 Text feature [available] present in test data point [True]
902 Text feature [box] present in test data point [True]
904 Text feature [29] present in test data point [True]
906 Text feature [found] present in test data point [True]
907 Text feature [also] present in test data point [True]
915 Text feature [mm] present in test data point [True]
916 Text feature [lesions] present in test data point [True]
917 Text feature [materials] present in test data point [True]
920 Text feature [12] present in test data point [True]
924 Text feature [biological] present in test data point [True]
925 Text feature [published] present in test data point [True]
926 Text feature [nm] present in test data point [True]
930 Text feature [cultured] present in test data point [True]
931 Text feature [kit] present in test data point [True]
932 Text feature [standard] present in test data point [True]
934 Text feature [44] present in test data point [True]
935 Text feature [clones] present in test data point [True]
936 Text feature [level] present in test data point [True]
937 Text feature [met] present in test data point [True]
943 Text feature [braf] present in test data point [True]
944 Text feature [current] present in test data point [True]
946 Text feature [recently] present in test data point [True]
948 Text feature [direct] present in test data point [True]
949 Text feature [differences] present in test data point [True]
950 Text feature [blot] present in test data point [True]
956 Text feature [70] present in test data point [True]
957 Text feature [differentiation] present in test data point [True]
958 Text feature [experimental] present in test data point [True]
959 Text feature [carried] present in test data point [True]
960 Text feature [sample] present in test data point [True]
961 Text feature [features] present in test data point [True]
965 Text feature [events] present in test data point [True]
969 Text feature [18] present in test data point [True]
970 Text feature [two] present in test data point [True]
971 Text feature [subset] present in test data point [True]
979 Text feature [similar] present in test data point [True]
```

```
982 Text feature [combination] present in test data point [True]
983 Text feature [gain] present in test data point [True]
984 Text feature [culture] present in test data point [True]
986 Text feature [possible] present in test data point [True]
991 Text feature [transfected] present in test data point [True]
992 Text feature [mtor] present in test data point [True]
993 Text feature [screen] present in test data point [True]
994 Text feature [generated] present in test data point [True]
Out of the top  1000  features  275 are present in query point
```

4.4. Linear Support Vector Machines

4.4.1. Hyper paramter tuning

In [74]:
```python
# read more about support vector machines with linear kernals here htt
p://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

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

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

```python
# find more about CalibratedClassifierCV here at http://scikit-learn.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
tml
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='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='bal
anced')
    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")
```

```python
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='balance
d')
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], p
enalty='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, 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 vali
dation 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 l
oss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for C = 1e-05
Log Loss :  1.1282957501676352
for C = 0.0001
Log Loss :  1.0407439082223864
for C = 0.001
Log Loss :  1.0581738455183807
for C = 0.01
Log Loss :  1.273270173714094
for C = 0.1
Log Loss :  1.594711756785334
for C = 1
Log Loss :  1.7647998959809297
for C = 10
Log Loss :  1.7648052394114275
```

```
for C = 100
Log Loss : 1.7648005346630697
```


Cross Validation Error for each alpha
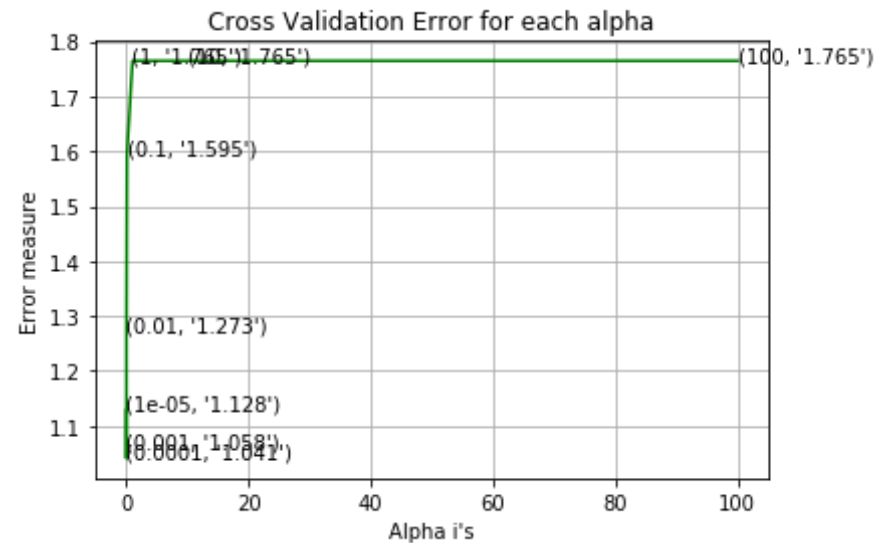
```
For values of best alpha =  0.0001 The train log loss is: 0.38421654903
625363
For values of best alpha =  0.0001 The cross validation log loss is: 1.
0407439082223864
For values of best alpha =  0.0001 The test log loss is: 1.109469372590
8253
```

4.4.2. Testing model with best hyper parameters

```
In [75]:  # read more about support vector machines with linear kernals here htt
          p://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

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

```python
# 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, decisi
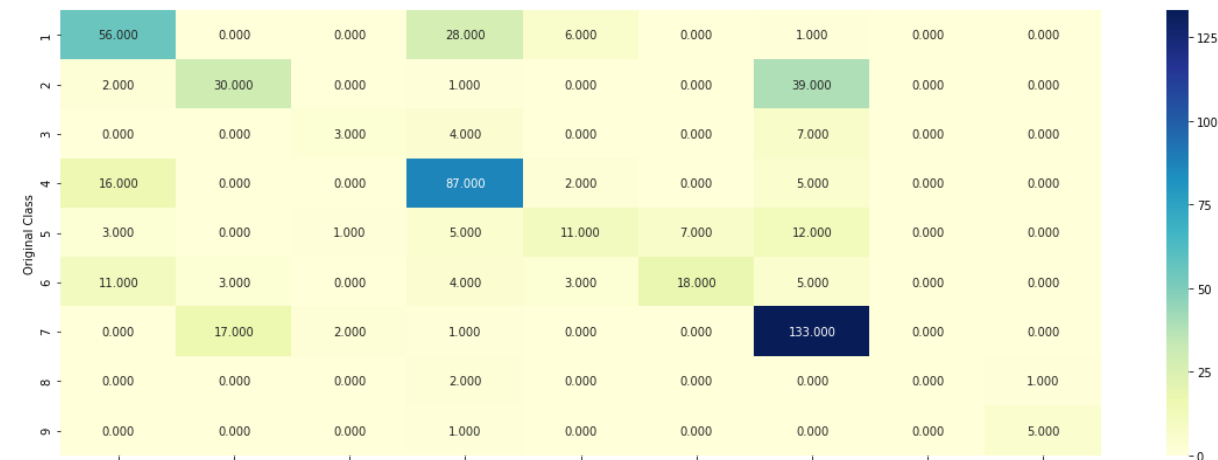on_function_shape='ovr', random_state=None)

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


# clf = SVC(C=alpha[best_alpha],kernel='linear',probability=True, class
_weight='balanced')
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='hinge'
, random_state=42,class_weight='balanced')
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y,cv_x_on
ehotCoding,cv_y, clf)
```

```
Log loss : 1.0407439082223864
Number of mis-classified points : 0.35526315789473684
------------------- Confusion matrix -------------------
```

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

------------------ Recall matrix (Row sum=1) -------------------
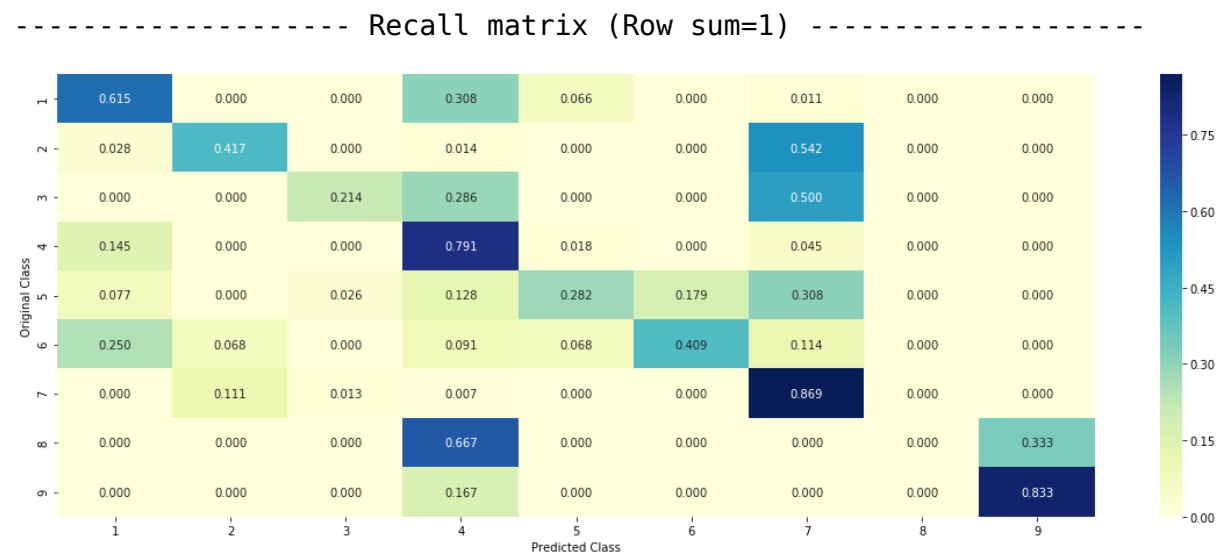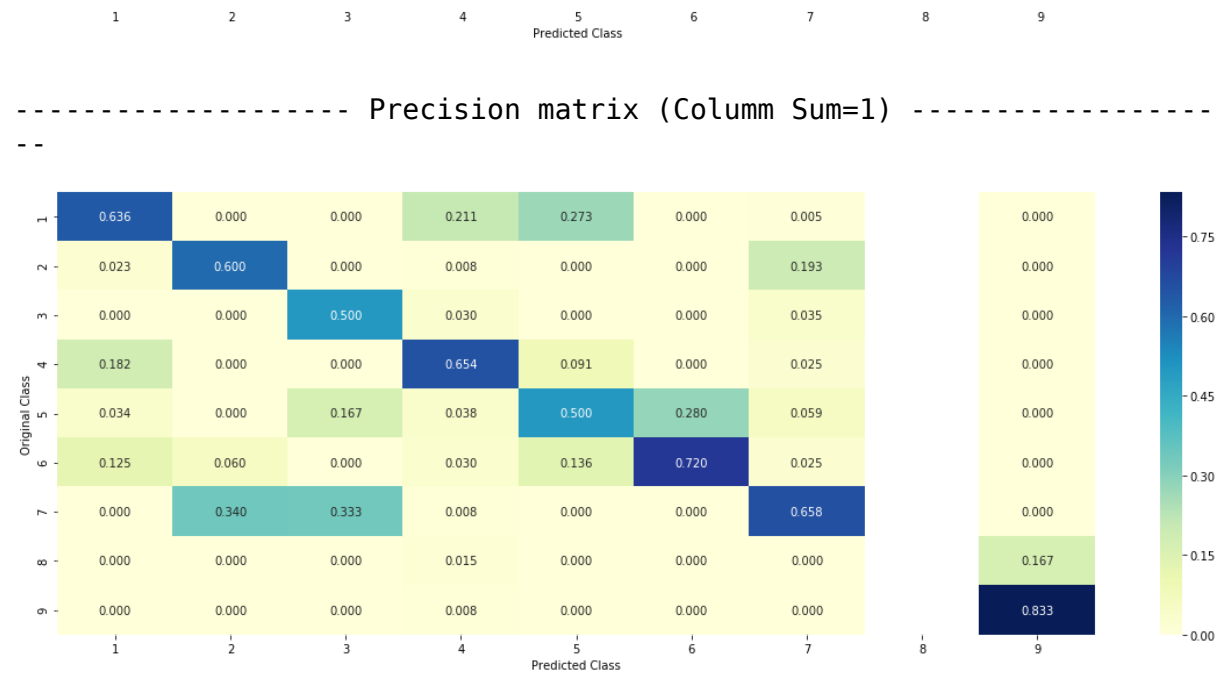
4.3.3. Feature Importance

4.3.3.1. For Correctly classified point

In [78]:
```python
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='hinge'
, random_state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 100
# test_point_index = 1000
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],test_df['Variation'].iloc[test
_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0315 0.045  0.0231 0.0783 0.0558 0.0
111 0.7496 0.0037 0.0019]]
Actual Class : 7
--------------------------------------------------
31 Text feature [transformed] present in test data point [True]
72 Text feature [membrane] present in test data point [True]
74 Text feature [downstream] present in test data point [True]
75 Text feature [bone] present in test data point [True]
```

```
79 Text feature [codon] present in test data point [True]
81 Text feature [phosphorylated] present in test data point [True]
138 Text feature [fold] present in test data point [True]
143 Text feature [constitutive] present in test data point [True]
144 Text feature [activation] present in test data point [True]
150 Text feature [activate] present in test data point [True]
153 Text feature [consistent] present in test data point [True]
241 Text feature [2a] present in test data point [True]
246 Text feature [activated] present in test data point [True]
250 Text feature [positive] present in test data point [True]
251 Text feature [examined] present in test data point [True]
254 Text feature [stably] present in test data point [True]
259 Text feature [oncogene] present in test data point [True]
260 Text feature [high] present in test data point [True]
262 Text feature [regions] present in test data point [True]
263 Text feature [elevated] present in test data point [True]
270 Text feature [genomic] present in test data point [True]
272 Text feature [observations] present in test data point [True]
274 Text feature [000] present in test data point [True]
276 Text feature [factor] present in test data point [True]
280 Text feature [colony] present in test data point [True]
282 Text feature [express] present in test data point [True]
395 Text feature [free] present in test data point [True]
397 Text feature [phospho] present in test data point [True]
403 Text feature [akt] present in test data point [True]
408 Text feature [added] present in test data point [True]
409 Text feature [approach] present in test data point [True]
410 Text feature [factors] present in test data point [True]
412 Text feature [common] present in test data point [True]
413 Text feature [conditions] present in test data point [True]
415 Text feature [3b] present in test data point [True]
416 Text feature [properties] present in test data point [True]
417 Text feature [pathways] present in test data point [True]
418 Text feature [demonstrated] present in test data point [True]
425 Text feature [transcription] present in test data point [True]
426 Text feature [2000] present in test data point [True]
428 Text feature [various] present in test data point [True]
429 Text feature [position] present in test data point [True]
431 Text feature [major] present in test data point [True]
434 Text feature [determined] present in test data point [True]
```

435 Text feature [transforming] present in test data point [True]

436 Text feature [overexpression] present in test data point [True]
437 Text feature [ph] present in test data point [True]
439 Text feature [ras] present in test data point [True]
440 Text feature [long] present in test data point [True]
442 Text feature [years] present in test data point [True]
446 Text feature [expressing] present in test data point [True]
447 Text feature [sensitive] present in test data point [True]
449 Text feature [might] present in test data point [True]
455 Text feature [endogenous] present in test data point [True]
458 Text feature [region] present in test data point [True]
578 Text feature [activating] present in test data point [True]
582 Text feature [five] present in test data point [True]
583 Text feature [generation] present in test data point [True]
584 Text feature [well] present in test data point [True]
585 Text feature [phosphorylation] present in test data point [True]
586 Text feature [proliferation] present in test data point [True]
587 Text feature [example] present in test data point [True]
588 Text feature [specific] present in test data point [True]
590 Text feature [increased] present in test data point [True]
591 Text feature [directly] present in test data point [True]
593 Text feature [13] present in test data point [True]
594 Text feature [use] present in test data point [True]
597 Text feature [supplemental] present in test data point [True]
599 Text feature [cdna] present in test data point [True]
603 Text feature [mediated] present in test data point [True]
604 Text feature [report] present in test data point [True]
605 Text feature [associated] present in test data point [True]
606 Text feature [primer] present in test data point [True]
607 Text feature [22] present in test data point [True]
608 Text feature [upon] present in test data point [True]
609 Text feature [download] present in test data point [True]
611 Text feature [oncogenic] present in test data point [True]
615 Text feature [incubated] present in test data point [True]
617 Text feature [reverse] present in test data point [True]
618 Text feature [total] present in test data point [True]
620 Text feature [signaling] present in test data point [True]
622 Text feature [could] present in test data point [True]
623 Text feature [strong] present in test data point [True]

624 Text feature [assessed] present in test data point [True]

625 Text feature [24] present in test data point [True]
633 Text feature [3a] present in test data point [True]
634 Text feature [clinically] present in test data point [True]
635 Text feature [available] present in test data point [True]
636 Text feature [mutant] present in test data point [True]
637 Text feature [sensitivity] present in test data point [True]
638 Text feature [per] present in test data point [True]
640 Text feature [products] present in test data point [True]
642 Text feature [within] present in test data point [True]
644 Text feature [taken] present in test data point [True]
646 Text feature [cells] present in test data point [True]
653 Text feature [deletion] present in test data point [True]
654 Text feature [wt] present in test data point [True]
655 Text feature [serum] present in test data point [True]
660 Text feature [previously] present in test data point [True]
665 Text feature [vector] present in test data point [True]
670 Text feature [colonies] present in test data point [True]
671 Text feature [mm] present in test data point [True]
672 Text feature [terminus] present in test data point [True]
676 Text feature [2b] present in test data point [True]
679 Text feature [point] present in test data point [True]
680 Text feature [signals] present in test data point [True]
685 Text feature [addition] present in test data point [True]
686 Text feature [somatic] present in test data point [True]
687 Text feature [75] present in test data point [True]
690 Text feature [enhanced] present in test data point [True]
694 Text feature [recently] present in test data point [True]
695 Text feature [driven] present in test data point [True]
772 Text feature [members] present in test data point [True]
774 Text feature [like] present in test data point [True]
775 Text feature [pathway] present in test data point [True]
776 Text feature [inhibitor] present in test data point [True]
779 Text feature [fusion] present in test data point [True]
780 Text feature [contrast] present in test data point [True]
782 Text feature [domain] present in test data point [True]
783 Text feature [medium] present in test data point [True]
788 Text feature [whereas] present in test data point [True]
796 Text feature [mutants] present in test data point [True]

```
802 Text feature [days] present in test data point [True]

804 Text feature [direct] present in test data point [True]
812 Text feature [tyrosine] present in test data point [True]
813 Text feature [fig] present in test data point [True]
814 Text feature [level] present in test data point [True]
815 Text feature [phosphatase] present in test data point [True]
816 Text feature [promote] present in test data point [True]
818 Text feature [suggest] present in test data point [True]
820 Text feature [transfected] present in test data point [True]
822 Text feature [concentrations] present in test data point [True]
824 Text feature [basal] present in test data point [True]
826 Text feature [cancers] present in test data point [True]
827 Text feature [independent] present in test data point [True]
829 Text feature [regulatory] present in test data point [True]
834 Text feature [70] present in test data point [True]
835 Text feature [mean] present in test data point [True]
838 Text feature [sites] present in test data point [True]
840 Text feature [levels] present in test data point [True]
843 Text feature [include] present in test data point [True]
845 Text feature [amplification] present in test data point [True]
848 Text feature [transformation] present in test data point [True]
849 Text feature [comparison] present in test data point [True]
851 Text feature [myc] present in test data point [True]
852 Text feature [key] present in test data point [True]
856 Text feature [two] present in test data point [True]
858 Text feature [kinases] present in test data point [True]
859 Text feature [showed] present in test data point [True]
863 Text feature [14] present in test data point [True]
864 Text feature [strongly] present in test data point [True]
865 Text feature [25] present in test data point [True]
866 Text feature [nm] present in test data point [True]
867 Text feature [ca] present in test data point [True]
868 Text feature [found] present in test data point [True]
870 Text feature [targeting] present in test data point [True]
871 Text feature [biochemical] present in test data point [True]
873 Text feature [data] present in test data point [True]
874 Text feature [leads] present in test data point [True]
876 Text feature [tumorigenesis] present in test data point [True]
878 Text feature [selected] present in test data point [True]
```

879 Text feature [developed] present in test data point [True]

881 Text feature [obtained] present in test data point [True]
884 Text feature [introduction] present in test data point [True]
885 Text feature [less] present in test data point [True]
886 Text feature [reaction] present in test data point [True]
887 Text feature [measured] present in test data point [True]
888 Text feature [cellular] present in test data point [True]
889 Text feature [pi3k] present in test data point [True]
891 Text feature [corresponding] present in test data point [True]
892 Text feature [presence] present in test data point [True]
893 Text feature [directed] present in test data point [True]
895 Text feature [29] present in test data point [True]
897 Text feature [experimental] present in test data point [True]
898 Text feature [interestingly] present in test data point [True]
899 Text feature [mechanism] present in test data point [True]
903 Text feature [constitutively] present in test data point [True]
905 Text feature [concentration] present in test data point [True]
907 Text feature [48] present in test data point [True]
908 Text feature [moreover] present in test data point [True]
910 Text feature [cycle] present in test data point [True]
912 Text feature [12] present in test data point [True]
913 Text feature [tumor] present in test data point [True]
914 Text feature [characterized] present in test data point [True]
916 Text feature [see] present in test data point [True]
917 Text feature [possible] present in test data point [True]
918 Text feature [regulation] present in test data point [True]
983 Text feature [resistant] present in test data point [True]
984 Text feature [standard] present in test data point [True]
985 Text feature [also] present in test data point [True]
986 Text feature [tumors] present in test data point [True]
987 Text feature [higher] present in test data point [True]
988 Text feature [sequenced] present in test data point [True]
989 Text feature [positions] present in test data point [True]
990 Text feature [either] present in test data point [True]
991 Text feature [experiments] present in test data point [True]
997 Text feature [page] present in test data point [True]
998 Text feature [primers] present in test data point [True]
999 Text feature [american] present in test data point [True]
Out of the top 1000 features 199 are present in query point

4.3.3.2. For Incorrectly classified point

```
In [79]:  test_point_index = 31
          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],test_df['Variation'].iloc[test
          _point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0285 0.0691 0.0071 0.0459 0.016  0.0
727 0.7551 0.0033 0.0022]]
Actual Class : 2
--------------------------------------------------
72 Text feature [membrane] present in test data point [True]
73 Text feature [insertion] present in test data point [True]
74 Text feature [downstream] present in test data point [True]
79 Text feature [codon] present in test data point [True]
138 Text feature [fold] present in test data point [True]
143 Text feature [constitutive] present in test data point [True]
144 Text feature [activation] present in test data point [True]
150 Text feature [activate] present in test data point [True]
152 Text feature [inhibited] present in test data point [True]
246 Text feature [activated] present in test data point [True]
259 Text feature [oncogene] present in test data point [True]
260 Text feature [high] present in test data point [True]
262 Text feature [regions] present in test data point [True]
265 Text feature [nf1] present in test data point [True]
271 Text feature [form] present in test data point [True]
273 Text feature [motif] present in test data point [True]
274 Text feature [000] present in test data point [True]
276 Text feature [factor] present in test data point [True]
```

276 Text feature [factor] present in test data point [True]
281 Text feature [carcinomas] present in test data point [True]
395 Text feature [free] present in test data point [True]
398 Text feature [patient] present in test data point [True]
399 Text feature [raf] present in test data point [True]
400 Text feature [versus] present in test data point [True]
405 Text feature [leading] present in test data point [True]
409 Text feature [approach] present in test data point [True]
410 Text feature [factors] present in test data point [True]
412 Text feature [common] present in test data point [True]
413 Text feature [conditions] present in test data point [True]
416 Text feature [properties] present in test data point [True]
418 Text feature [demonstrated] present in test data point [True]
425 Text feature [transcription] present in test data point [True]
426 Text feature [2000] present in test data point [True]
429 Text feature [position] present in test data point [True]
430 Text feature [lead] present in test data point [True]
432 Text feature [negative] present in test data point [True]
434 Text feature [determined] present in test data point [True]
435 Text feature [transforming] present in test data point [True]
439 Text feature [ras] present in test data point [True]
443 Text feature [additional] present in test data point [True]
448 Text feature [lung] present in test data point [True]
449 Text feature [might] present in test data point [True]
458 Text feature [region] present in test data point [True]
460 Text feature [erk] present in test data point [True]
584 Text feature [well] present in test data point [True]
585 Text feature [phosphorylation] present in test data point [True]
588 Text feature [specific] present in test data point [True]
590 Text feature [increased] present in test data point [True]
591 Text feature [directly] present in test data point [True]
593 Text feature [13] present in test data point [True]
594 Text feature [use] present in test data point [True]
603 Text feature [mediated] present in test data point [True]
604 Text feature [report] present in test data point [True]
608 Text feature [upon] present in test data point [True]
610 Text feature [gtp] present in test data point [True]
611 Text feature [oncogenic] present in test data point [True]
620 Text feature [signaling] present in test data point [True]
622 Text feature [could] present in test data point [True]

```
623 Text feature [strong] present in test data point [True]
625 Text feature [24] present in test data point [True]
630 Text feature [gap] present in test data point [True]
635 Text feature [available] present in test data point [True]
636 Text feature [mutant] present in test data point [True]
642 Text feature [within] present in test data point [True]
645 Text feature [liver] present in test data point [True]
646 Text feature [cells] present in test data point [True]
650 Text feature [inhibitory] present in test data point [True]
653 Text feature [deletion] present in test data point [True]
654 Text feature [wt] present in test data point [True]
660 Text feature [previously] present in test data point [True]
662 Text feature [association] present in test data point [True]
666 Text feature [kd] present in test data point [True]
679 Text feature [point] present in test data point [True]
680 Text feature [signals] present in test data point [True]
684 Text feature [stimulated] present in test data point [True]
685 Text feature [addition] present in test data point [True]
688 Text feature [59] present in test data point [True]
689 Text feature [screening] present in test data point [True]
690 Text feature [enhanced] present in test data point [True]
692 Text feature [insertions] present in test data point [True]
696 Text feature [leukemia] present in test data point [True]
772 Text feature [members] present in test data point [True]
774 Text feature [like] present in test data point [True]
775 Text feature [pathway] present in test data point [True]
780 Text feature [contrast] present in test data point [True]
782 Text feature [domain] present in test data point [True]
784 Text feature [often] present in test data point [True]
785 Text feature [exon] present in test data point [True]
786 Text feature [fact] present in test data point [True]
788 Text feature [whereas] present in test data point [True]
789 Text feature [61] present in test data point [True]
790 Text feature [mechanisms] present in test data point [True]
796 Text feature [mutants] present in test data point [True]
799 Text feature [substitution] present in test data point [True]
820 Text feature [transfected] present in test data point [True]
822 Text feature [concentrations] present in test data point [True]
824 Text feature [basal] present in test data point [True]
```

```
825 Text feature [since] present in test data point [True]
827 Text feature [independent] present in test data point [True]
829 Text feature [regulatory] present in test data point [True]
836 Text feature [would] present in test data point [True]
843 Text feature [include] present in test data point [True]
845 Text feature [amplification] present in test data point [True]
846 Text feature [44] present in test data point [True]
847 Text feature [fraction] present in test data point [True]
849 Text feature [comparison] present in test data point [True]
855 Text feature [affinity] present in test data point [True]
856 Text feature [two] present in test data point [True]
864 Text feature [strongly] present in test data point [True]
865 Text feature [25] present in test data point [True]
866 Text feature [nm] present in test data point [True]
868 Text feature [found] present in test data point [True]
871 Text feature [biochemical] present in test data point [True]
873 Text feature [data] present in test data point [True]
874 Text feature [leads] present in test data point [True]
881 Text feature [obtained] present in test data point [True]
882 Text feature [18] present in test data point [True]
884 Text feature [introduction] present in test data point [True]
885 Text feature [less] present in test data point [True]
886 Text feature [reaction] present in test data point [True]
887 Text feature [measured] present in test data point [True]
888 Text feature [cellular] present in test data point [True]
892 Text feature [presence] present in test data point [True]
893 Text feature [directed] present in test data point [True]
896 Text feature [led] present in test data point [True]
903 Text feature [constitutively] present in test data point [True]
905 Text feature [concentration] present in test data point [True]
907 Text feature [48] present in test data point [True]
910 Text feature [cycle] present in test data point [True]
912 Text feature [12] present in test data point [True]
914 Text feature [characterized] present in test data point [True]
985 Text feature [also] present in test data point [True]
986 Text feature [tumors] present in test data point [True]
987 Text feature [higher] present in test data point [True]
990 Text feature [either] present in test data point [True]
```

```
991 Text feature [experiments] present in test data point [True]
Out of the top  1000  features  135 are present in query point
```

4.5 Random Forest Classifier

4.5.1. Hyper paramter tuning (With One hot Encoding)

In [80]:
```python
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='g
ini', max_depth=None, min_samples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='aut
o', max_leaf_nodes=None, min_impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, r
andom_state=None, verbose=0, warm_start=False,
# class_weight=None)

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

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

# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/random-forest-and-their-construction-2/
# --------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
```

```
tml
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='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]).ra
vel()
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)), (featur
es[i],cv_log_error_array[i]))
plt.grid()
```

```python
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)], cri
terion='gini', max_depth=max_depth[int(best_alpha%2)], 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)

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, predict_y, labels=clf.classes_,
eps=1e-15))
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_cv, predict_y, labels=clf.cl
asses_, eps=1e-15))
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, predict_y, labels=clf.classes_, ep
s=1e-15))
```

```
for n_estimators = 100 and max depth =  5
Log Loss : 1.2087611921071726
for n_estimators = 100 and max depth =  10
Log Loss : 1.228147564913086
for n_estimators = 200 and max depth =  5
Log Loss : 1.1990613240400025
for n_estimators = 200 and max depth =  10
Log Loss : 1.2073289926450084
for n_estimators = 500 and max depth =  5
Log Loss : 1.1841376952635174
for n_estimators = 500 and max depth =  10
Log Loss : 1.1989322662176276
for n_estimators = 1000 and max depth =  5
```

```
Log Loss : 1.18649541331408
for n_estimators = 1000 and max depth =  10
Log Loss : 1.199498792184227
for n_estimators = 2000 and max depth =  5
Log Loss : 1.1894233566015975
for n_estimators = 2000 and max depth =  10
Log Loss : 1.2007953796641582
For values of best estimator =  500 The train log loss is: 0.8414275254
235917

For values of best estimator =  500 The cross validation log loss is:
1.1841376952635174
For values of best estimator =  500 The test log loss is: 1.22034628111
1821
```

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

In [81]:
```python
# -------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='g
ini', max_depth=None, min_samples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='aut
o', max_leaf_nodes=None, min_impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, r
andom_state=None, verbose=0, warm_start=False,
# class_weight=None)

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

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

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

```
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/random-forest-and-their-construction-2/
# --------------------------------

clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], cri
terion='gini', max_depth=max_depth[int(best_alpha%2)], random_state=42,
 n_jobs=-1)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y,cv_x_on
ehotCoding,cv_y, clf)
```

```
Log loss : 1.1841376952635176
Number of mis-classified points : 0.4041353383458647
-------------------- Confusion matrix --------------------
```



```
-------------------- Precision matrix (Columm Sum=1) -----------------
--
```
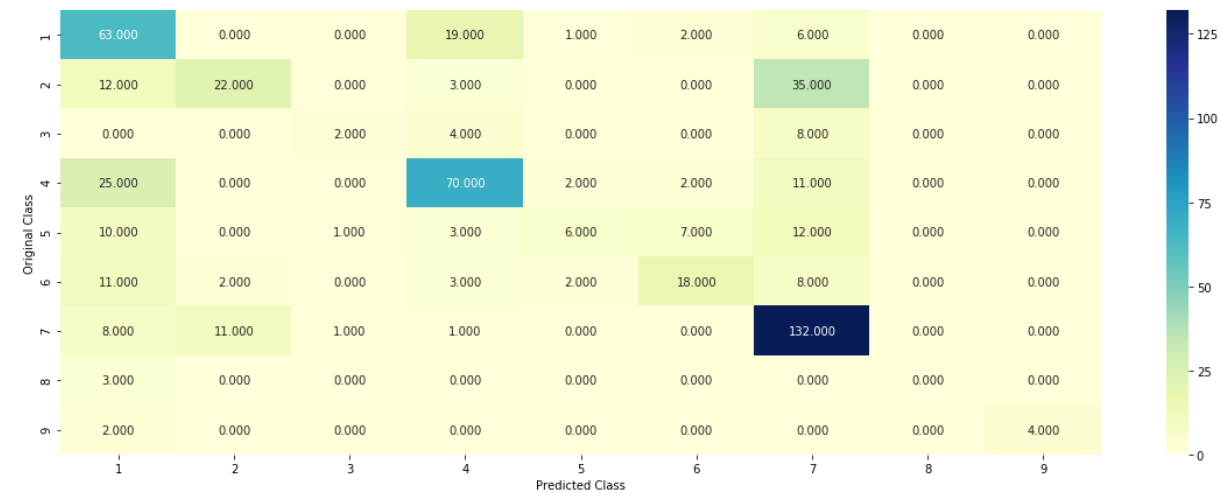
------------------- Recall matrix (Row sum=1) -------------------

4.5.3. Feature Importance

4.5.3.1. Correctly Classified point

In [86]:
```python
# test_point_index = 10
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], cri
terion='gini', max_depth=max_depth[int(best_alpha%2)], 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)

test_point_index = 100
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
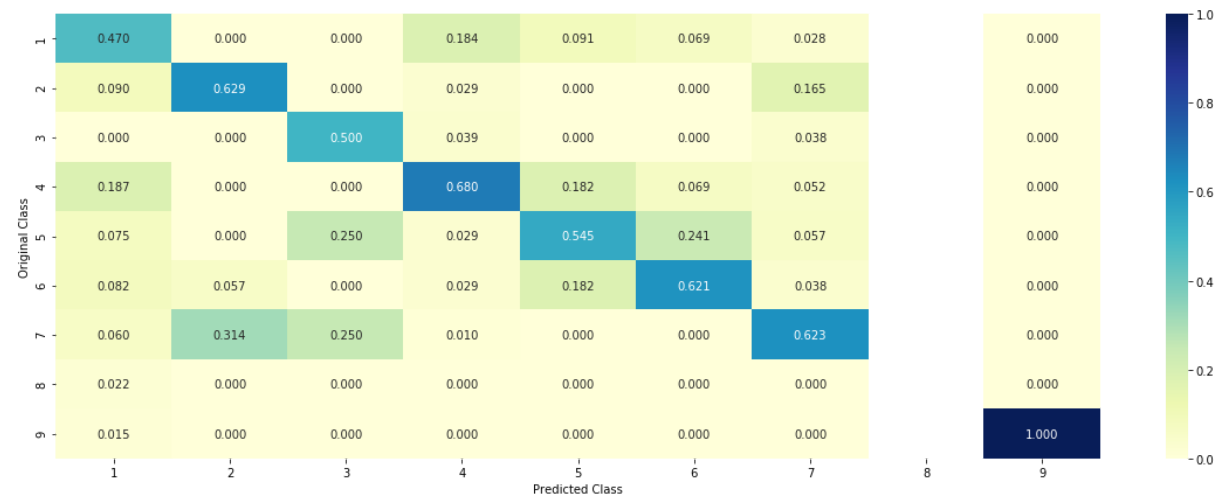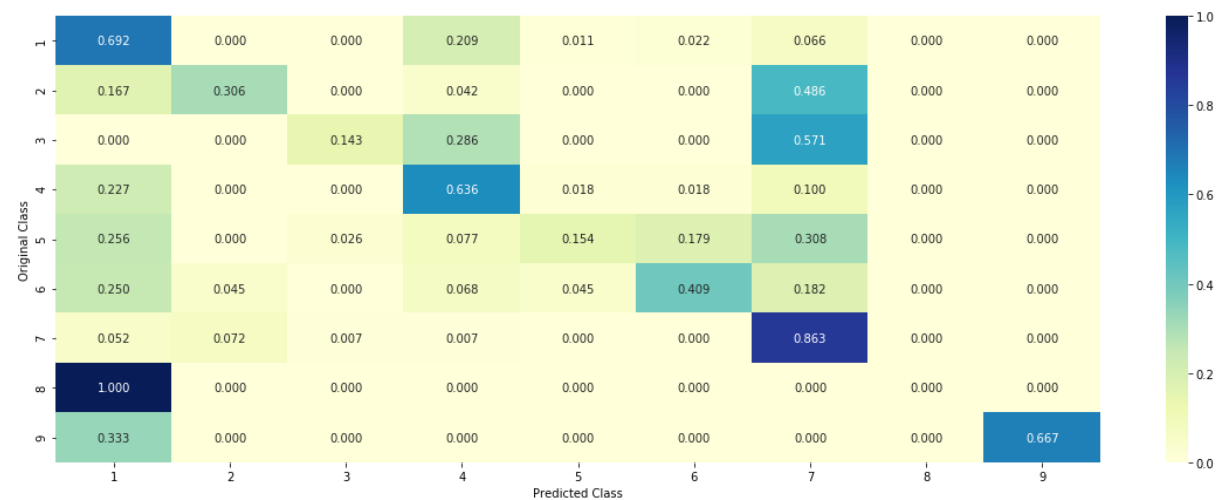print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(
test_x_onehotCoding[test_point_index]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.feature_importances_)
print("-"*50)
get_impfeature_names(indices[:no_feature], test_df['TEXT'].iloc[test_po
int_index],test_df['Gene'].iloc[test_point_index],test_df['Variation'].
iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0553 0.0863 0.0251 0.0918 0.0541 0.0
348 0.6447 0.0052 0.0027]]
Actual Class : 7
--------------------------------------------------
0 Text feature [kinase] present in test data point [True]
1 Text feature [activating] present in test data point [True]
2 Text feature [phosphorylation] present in test data point [True]
3 Text feature [inhibitors] present in test data point [True]
4 Text feature [inhibitor] present in test data point [True]
5 Text feature [activated] present in test data point [True]
6 Text feature [constitutive] present in test data point [True]
7 Text feature [activation] present in test data point [True]
8 Text feature [function] present in test data point [True]
```

8 Text feature [function] present in test data point [True]

9 Text feature [tyrosine] present in test data point [True]
11 Text feature [loss] present in test data point [True]
14 Text feature [constitutively] present in test data point [True]
15 Text feature [receptor] present in test data point [True]
17 Text feature [signaling] present in test data point [True]
18 Text feature [protein] present in test data point [True]
21 Text feature [pten] present in test data point [True]
25 Text feature [functional] present in test data point [True]
28 Text feature [oncogenic] present in test data point [True]
31 Text feature [kinases] present in test data point [True]
32 Text feature [cell] present in test data point [True]
34 Text feature [expression] present in test data point [True]
38 Text feature [transforming] present in test data point [True]
42 Text feature [akt] present in test data point [True]
45 Text feature [phosphatase] present in test data point [True]
49 Text feature [cells] present in test data point [True]
50 Text feature [expected] present in test data point [True]
53 Text feature [response] present in test data point [True]
57 Text feature [oncogene] present in test data point [True]
62 Text feature [activate] present in test data point [True]
63 Text feature [predicted] present in test data point [True]
68 Text feature [inhibition] present in test data point [True]
69 Text feature [functions] present in test data point [True]
70 Text feature [growth] present in test data point [True]
71 Text feature [proliferation] present in test data point [True]
72 Text feature [yeast] present in test data point [True]
73 Text feature [clinical] present in test data point [True]
74 Text feature [proteins] present in test data point [True]
75 Text feature [therapeutic] present in test data point [True]
76 Text feature [serum] present in test data point [True]
77 Text feature [sensitivity] present in test data point [True]
78 Text feature [phospho] present in test data point [True]
79 Text feature [expressing] present in test data point [True]
80 Text feature [downstream] present in test data point [True]
82 Text feature [dna] present in test data point [True]
85 Text feature [assays] present in test data point [True]
86 Text feature [combined] present in test data point [True]
91 Text feature [mammalian] present in test data point [True]
92 Text feature [assay] present in test data point [True]

92 Text feature [assay] present in test data point [True]

93 Text feature [activity] present in test data point [True]
97 Text feature [tagged] present in test data point [True]
101 Text feature [identified] present in test data point [True]
103 Text feature [likely] present in test data point [True]
105 Text feature [transformation] present in test data point [True]
106 Text feature [ability] present in test data point [True]
107 Text feature [basis] present in test data point [True]
108 Text feature [affected] present in test data point [True]
111 Text feature [binding] present in test data point [True]
112 Text feature [genes] present in test data point [True]
113 Text feature [amplification] present in test data point [True]
114 Text feature [catalytic] present in test data point [True]
115 Text feature [tumors] present in test data point [True]
116 Text feature [based] present in test data point [True]
118 Text feature [weeks] present in test data point [True]
120 Text feature [phosphorylated] present in test data point [True]
121 Text feature [potential] present in test data point [True]
125 Text feature [interaction] present in test data point [True]
126 Text feature [indicated] present in test data point [True]
130 Text feature [resistant] present in test data point [True]
131 Text feature [evidence] present in test data point [True]
133 Text feature [sequencing] present in test data point [True]
136 Text feature [type] present in test data point [True]
137 Text feature [one] present in test data point [True]
139 Text feature [results] present in test data point [True]
140 Text feature [substrate] present in test data point [True]
141 Text feature [pathway] present in test data point [True]
142 Text feature [site] present in test data point [True]
144 Text feature [affect] present in test data point [True]
150 Text feature [sensitive] present in test data point [True]
151 Text feature [experiments] present in test data point [True]
153 Text feature [32] present in test data point [True]
154 Text feature [41] present in test data point [True]
155 Text feature [50] present in test data point [True]
156 Text feature [conserved] present in test data point [True]
158 Text feature [p53] present in test data point [True]
159 Text feature [database] present in test data point [True]
160 Text feature [mutants] present in test data point [True]
162 Text feature [given] present in test data point [True]

162 Text feature [given] present in test data point [True]
164 Text feature [majority] present in test data point [True]
166 Text feature [none] present in test data point [True]
167 Text feature [transcription] present in test data point [True]
168 Text feature [21] present in test data point [True]
170 Text feature [recently] present in test data point [True]
171 Text feature [human] present in test data point [True]
172 Text feature [use] present in test data point [True]
173 Text feature [lysates] present in test data point [True]
174 Text feature [approach] present in test data point [True]
177 Text feature [35] present in test data point [True]
178 Text feature [atp] present in test data point [True]
179 Text feature [suggesting] present in test data point [True]
180 Text feature [many] present in test data point [True]
181 Text feature [34] present in test data point [True]
182 Text feature [deletion] present in test data point [True]
183 Text feature [transfected] present in test data point [True]
186 Text feature [ras] present in test data point [True]
187 Text feature [available] present in test data point [True]
188 Text feature [vector] present in test data point [True]
189 Text feature [clinically] present in test data point [True]
191 Text feature [previously] present in test data point [True]
192 Text feature [first] present in test data point [True]
193 Text feature [pi3k] present in test data point [True]
195 Text feature [wild] present in test data point [True]
196 Text feature [containing] present in test data point [True]
197 Text feature [large] present in test data point [True]
199 Text feature [plasmid] present in test data point [True]
200 Text feature [active] present in test data point [True]
201 Text feature [two] present in test data point [True]
202 Text feature [present] present in test data point [True]
207 Text feature [days] present in test data point [True]
208 Text feature [confer] present in test data point [True]
209 Text feature [2a] present in test data point [True]
210 Text feature [sequence] present in test data point [True]
211 Text feature [presence] present in test data point [True]
213 Text feature [somatic] present in test data point [True]
214 Text feature [consistent] present in test data point [True]
217 Text feature [3a] present in test data point [True]
218 Text feature [basal] present in test data point [True]

218 Text feature [basal] present in test data point [True]

220 Text feature [discussion] present in test data point [True]
222 Text feature [introduction] present in test data point [True]
223 Text feature [possible] present in test data point [True]
225 Text feature [26] present in test data point [True]
226 Text feature [mutagenesis] present in test data point [True]
227 Text feature [therefore] present in test data point [True]
228 Text feature [known] present in test data point [True]
229 Text feature [family] present in test data point [True]
230 Text feature [31] present in test data point [True]
231 Text feature [70] present in test data point [True]
232 Text feature [39] present in test data point [True]
237 Text feature [12] present in test data point [True]
238 Text feature [role] present in test data point [True]
240 Text feature [mediated] present in test data point [True]
241 Text feature [criteria] present in test data point [True]
242 Text feature [expressed] present in test data point [True]
245 Text feature [terminal] present in test data point [True]
247 Text feature [levels] present in test data point [True]
248 Text feature [significantly] present in test data point [True]
251 Text feature [detected] present in test data point [True]
253 Text feature [length] present in test data point [True]
254 Text feature [45] present in test data point [True]
255 Text feature [cancers] present in test data point [True]
256 Text feature [induced] present in test data point [True]
258 Text feature [single] present in test data point [True]
260 Text feature [whereas] present in test data point [True]
261 Text feature [suggest] present in test data point [True]
264 Text feature [independent] present in test data point [True]
265 Text feature [gene] present in test data point [True]
267 Text feature [american] present in test data point [True]
268 Text feature [complex] present in test data point [True]
269 Text feature [11] present in test data point [True]
273 Text feature [mechanism] present in test data point [True]
274 Text feature [36] present in test data point [True]
275 Text feature [demonstrated] present in test data point [True]
276 Text feature [observed] present in test data point [True]
279 Text feature [figure] present in test data point [True]
281 Text feature [data] present in test data point [True]
282 Text feature [kit] present in test data point [True]

282 Text feature [kit] present in test data point [True]

283 Text feature [primary] present in test data point [True]
286 Text feature [mm] present in test data point [True]
288 Text feature [associated] present in test data point [True]
290 Text feature [described] present in test data point [True]
291 Text feature [specific] present in test data point [True]
292 Text feature [developed] present in test data point [True]
293 Text feature [genetic] present in test data point [True]
295 Text feature [increased] present in test data point [True]
297 Text feature [transfection] present in test data point [True]
299 Text feature [mutant] present in test data point [True]
300 Text feature [25] present in test data point [True]
301 Text feature [factor] present in test data point [True]
303 Text feature [directly] present in test data point [True]
304 Text feature [consequences] present in test data point [True]
305 Text feature [effect] present in test data point [True]
306 Text feature [structure] present in test data point [True]
307 Text feature [using] present in test data point [True]
308 Text feature [14] present in test data point [True]
309 Text feature [1a] present in test data point [True]
310 Text feature [confirmed] present in test data point [True]
312 Text feature [least] present in test data point [True]
315 Text feature [acids] present in test data point [True]
316 Text feature [next] present in test data point [True]
317 Text feature [whether] present in test data point [True]
319 Text feature [membrane] present in test data point [True]
320 Text feature [wt] present in test data point [True]
323 Text feature [generation] present in test data point [True]
324 Text feature [19] present in test data point [True]
326 Text feature [anti] present in test data point [True]
328 Text feature [control] present in test data point [True]
329 Text feature [formation] present in test data point [True]
330 Text feature [mouse] present in test data point [True]
331 Text feature [pcr] present in test data point [True]
333 Text feature [even] present in test data point [True]
334 Text feature [plates] present in test data point [True]
337 Text feature [molecule] present in test data point [True]
338 Text feature [half] present in test data point [True]
339 Text feature [total] present in test data point [True]
341 Text feature [related] present in test data point [True]

341 Text feature [related] present in test data point [True]
342 Text feature [mutation] present in test data point [True]
344 Text feature [level] present in test data point [True]
345 Text feature [domain] present in test data point [True]
346 Text feature [lack] present in test data point [True]
347 Text feature [testing] present in test data point [True]
349 Text feature [tested] present in test data point [True]
350 Text feature [signals] present in test data point [True]
352 Text feature [study] present in test data point [True]
354 Text feature [showed] present in test data point [True]
355 Text feature [discovery] present in test data point [True]
356 Text feature [shown] present in test data point [True]
357 Text feature [addition] present in test data point [True]
359 Text feature [mice] present in test data point [True]
360 Text feature [another] present in test data point [True]
361 Text feature [mutated] present in test data point [True]
362 Text feature [antibody] present in test data point [True]
366 Text feature [co] present in test data point [True]
367 Text feature [multiple] present in test data point [True]
368 Text feature [22] present in test data point [True]
371 Text feature [33] present in test data point [True]
373 Text feature [point] present in test data point [True]
375 Text feature [low] present in test data point [True]
376 Text feature [medium] present in test data point [True]
377 Text feature [37] present in test data point [True]
378 Text feature [3b] present in test data point [True]
379 Text feature [46] present in test data point [True]
380 Text feature [domains] present in test data point [True]
381 Text feature [29] present in test data point [True]
382 Text feature [region] present in test data point [True]
384 Text feature [change] present in test data point [True]
385 Text feature [17] present in test data point [True]
386 Text feature [lower] present in test data point [True]
388 Text feature [analyzed] present in test data point [True]
389 Text feature [also] present in test data point [True]
390 Text feature [page] present in test data point [True]
398 Text feature [used] present in test data point [True]
401 Text feature [five] present in test data point [True]
402 Text feature [controls] present in test data point [True]
403 Text feature [determine] present in test data point [True]

405 Text feature [determine] present in test data point [True]
405 Text feature [27] present in test data point [True]
406 Text feature [displayed] present in test data point [True]
409 Text feature [full] present in test data point [True]
411 Text feature [although] present in test data point [True]
412 Text feature [ml] present in test data point [True]
413 Text feature [including] present in test data point [True]
414 Text feature [resulting] present in test data point [True]
415 Text feature [significant] present in test data point [True]
416 Text feature [loop] present in test data point [True]
417 Text feature [molecular] present in test data point [True]
418 Text feature [direct] present in test data point [True]
419 Text feature [defined] present in test data point [True]
421 Text feature [number] present in test data point [True]
423 Text feature [contrast] present in test data point [True]
426 Text feature [within] present in test data point [True]
429 Text feature [representative] present in test data point [True]
432 Text feature [cellular] present in test data point [True]
433 Text feature [culture] present in test data point [True]
435 Text feature [increase] present in test data point [True]
436 Text feature [times] present in test data point [True]
437 Text feature [acid] present in test data point [True]
439 Text feature [moreover] present in test data point [True]
441 Text feature [respectively] present in test data point [True]
442 Text feature [mutational] present in test data point [True]
443 Text feature [elevated] present in test data point [True]
444 Text feature [carrying] present in test data point [True]
445 Text feature [well] present in test data point [True]
447 Text feature [added] present in test data point [True]
449 Text feature [showing] present in test data point [True]
455 Text feature [report] present in test data point [True]
456 Text feature [similar] present in test data point [True]
458 Text feature [measured] present in test data point [True]
459 Text feature [changes] present in test data point [True]
463 Text feature [analyses] present in test data point [True]
464 Text feature [higher] present in test data point [True]
467 Text feature [however] present in test data point [True]
468 Text feature [30] present in test data point [True]
469 Text feature [15] present in test data point [True]
471 Text feature [found] present in test data point [True]

471 Text feature [found] present in test data point [True]
472 Text feature [directed] present in test data point [True]
474 Text feature [altered] present in test data point [True]
475 Text feature [performed] present in test data point [True]
476 Text feature [sites] present in test data point [True]
477 Text feature [stably] present in test data point [True]
479 Text feature [screen] present in test data point [True]
480 Text feature [nature] present in test data point [True]
481 Text feature [gel] present in test data point [True]
482 Text feature [types] present in test data point [True]
483 Text feature [amplified] present in test data point [True]
484 Text feature [concentrations] present in test data point [True]
485 Text feature [driven] present in test data point [True]
486 Text feature [antibodies] present in test data point [True]
487 Text feature [90] present in test data point [True]
489 Text feature [determined] present in test data point [True]
491 Text feature [identification] present in test data point [True]
493 Text feature [important] present in test data point [True]
494 Text feature [molecules] present in test data point [True]
495 Text feature [development] present in test data point [True]
498 Text feature [sequenced] present in test data point [True]
499 Text feature [tumorigenesis] present in test data point [True]
501 Text feature [high] present in test data point [True]
504 Text feature [size] present in test data point [True]
506 Text feature [purified] present in test data point [True]
507 Text feature [transformed] present in test data point [True]
508 Text feature [group] present in test data point [True]
510 Text feature [cancer] present in test data point [True]
511 Text feature [common] present in test data point [True]
513 Text feature [activities] present in test data point [True]
516 Text feature [2b] present in test data point [True]
517 Text feature [gain] present in test data point [True]
520 Text feature [alone] present in test data point [True]
523 Text feature [presented] present in test data point [True]
524 Text feature [1b] present in test data point [True]
525 Text feature [cause] present in test data point [True]
526 Text feature [target] present in test data point [True]
528 Text feature [75] present in test data point [True]
529 Text feature [right] present in test data point [True]
532 Text feature [recent] present in test data point [True]

532 Text feature [recent] present in test data point [True]
535 Text feature [fig] present in test data point [True]
536 Text feature [located] present in test data point [True]
540 Text feature [fusion] present in test data point [True]
541 Text feature [whole] present in test data point [True]
544 Text feature [generated] present in test data point [True]
546 Text feature [small] present in test data point [True]
547 Text feature [analysis] present in test data point [True]
548 Text feature [tumor] present in test data point [True]
549 Text feature [13] present in test data point [True]
550 Text feature [20] present in test data point [True]
552 Text feature [selection] present in test data point [True]
553 Text feature [resulted] present in test data point [True]
556 Text feature [genomic] present in test data point [True]
558 Text feature [nucleotide] present in test data point [True]
559 Text feature [table] present in test data point [True]
561 Text feature [isolated] present in test data point [True]
562 Text feature [terminus] present in test data point [True]
563 Text feature [42] present in test data point [True]
565 Text feature [previous] present in test data point [True]
567 Text feature [top] present in test data point [True]
569 Text feature [vivo] present in test data point [True]
570 Text feature [endogenous] present in test data point [True]
571 Text feature [position] present in test data point [True]
572 Text feature [result] present in test data point [True]
573 Text feature [revealed] present in test data point [True]
574 Text feature [sample] present in test data point [True]
576 Text feature [23] present in test data point [True]
577 Text feature [thus] present in test data point [True]
579 Text feature [experimental] present in test data point [True]
580 Text feature [examined] present in test data point [True]
582 Text feature [24] present in test data point [True]
583 Text feature [positive] present in test data point [True]
584 Text feature [constructs] present in test data point [True]
585 Text feature [16] present in test data point [True]
587 Text feature [interestingly] present in test data point [True]
588 Text feature [established] present in test data point [True]
590 Text feature [western] present in test data point [True]
591 Text feature [cycle] present in test data point [True]
593 Text feature [among] present in test data point [True]

595 Text feature [among] present in test data point [True]
597 Text feature [contains] present in test data point [True]
599 Text feature [complete] present in test data point [True]
600 Text feature [non] present in test data point [True]
602 Text feature [formed] present in test data point [True]
607 Text feature [provide] present in test data point [True]
609 Text feature [fold] present in test data point [True]
610 Text feature [three] present in test data point [True]
611 Text feature [regulatory] present in test data point [True]
612 Text feature [cultured] present in test data point [True]
615 Text feature [60] present in test data point [True]
616 Text feature [characterized] present in test data point [True]
619 Text feature [compared] present in test data point [True]
620 Text feature [alterations] present in test data point [True]
621 Text feature [overexpression] present in test data point [True]
628 Text feature [suggested] present in test data point [True]
629 Text feature [four] present in test data point [True]
630 Text feature [comparison] present in test data point [True]
631 Text feature [include] present in test data point [True]
634 Text feature [either] present in test data point [True]
638 Text feature [studies] present in test data point [True]
641 Text feature [residues] present in test data point [True]
644 Text feature [enhanced] present in test data point [True]
645 Text feature [findings] present in test data point [True]
648 Text feature [show] present in test data point [True]
650 Text feature [buffer] present in test data point [True]
651 Text feature [standard] present in test data point [True]
652 Text feature [relative] present in test data point [True]
653 Text feature [nm] present in test data point [True]
656 Text feature [clones] present in test data point [True]
657 Text feature [regions] present in test data point [True]
658 Text feature [complexes] present in test data point [True]
659 Text feature [represent] present in test data point [True]
661 Text feature [vitro] present in test data point [True]
662 Text feature [primers] present in test data point [True]
664 Text feature [manufacturer] present in test data point [True]
667 Text feature [strong] present in test data point [True]
668 Text feature [long] present in test data point [True]
671 Text feature [several] present in test data point [True]
672 Text feature [case] present in test data point [True]

672 Text feature [case] present in test data point [True]
673 Text feature [sequences] present in test data point [True]
674 Text feature [may] present in test data point [True]
675 Text feature [seen] present in test data point [True]
676 Text feature [amino] present in test data point [True]
678 Text feature [day] present in test data point [True]
680 Text feature [pathways] present in test data point [True]
681 Text feature [involved] present in test data point [True]
682 Text feature [key] present in test data point [True]
687 Text feature [exhibited] present in test data point [True]
689 Text feature [gst] present in test data point [True]
691 Text feature [2005] present in test data point [True]
692 Text feature [blue] present in test data point [True]
694 Text feature [normalized] present in test data point [True]
696 Text feature [caused] present in test data point [True]
699 Text feature [observations] present in test data point [True]
703 Text feature [see] present in test data point [True]
704 Text feature [4a] present in test data point [True]
706 Text feature [helix] present in test data point [True]
713 Text feature [targeting] present in test data point [True]
714 Text feature [biological] present in test data point [True]
717 Text feature [could] present in test data point [True]
718 Text feature [assessed] present in test data point [True]
719 Text feature [detection] present in test data point [True]
722 Text feature [indicate] present in test data point [True]
723 Text feature [washed] present in test data point [True]
726 Text feature [express] present in test data point [True]
727 Text feature [might] present in test data point [True]
728 Text feature [genome] present in test data point [True]
729 Text feature [example] present in test data point [True]
730 Text feature [biochemical] present in test data point [True]
731 Text feature [000] present in test data point [True]
732 Text feature [support] present in test data point [True]
734 Text feature [mutations] present in test data point [True]
735 Text feature [new] present in test data point [True]
738 Text feature [100] present in test data point [True]
742 Text feature [line] present in test data point [True]
743 Text feature [mean] present in test data point [True]
744 Text feature [regulation] present in test data point [True]
747 Text feature [shows] present in test data point [True]

747 Text feature [shows] present in test data point [True]
748 Text feature [obtained] present in test data point [True]
749 Text feature [signal] present in test data point [True]
752 Text feature [factors] present in test data point [True]
754 Text feature [end] present in test data point [True]
763 Text feature [38] present in test data point [True]
764 Text feature [cdna] present in test data point [True]
766 Text feature [secondary] present in test data point [True]
770 Text feature [ii] present in test data point [True]
771 Text feature [different] present in test data point [True]
772 Text feature [48] present in test data point [True]
775 Text feature [flag] present in test data point [True]
777 Text feature [corresponding] present in test data point [True]
780 Text feature [contact] present in test data point [True]
781 Text feature [iii] present in test data point [True]
782 Text feature [members] present in test data point [True]
783 Text feature [myc] present in test data point [True]
785 Text feature [demonstrate] present in test data point [True]
786 Text feature [cases] present in test data point [True]
787 Text feature [crystal] present in test data point [True]
788 Text feature [positions] present in test data point [True]
789 Text feature [time] present in test data point [True]
796 Text feature [state] present in test data point [True]
797 Text feature [stable] present in test data point [True]
798 Text feature [primer] present in test data point [True]
799 Text feature [numbers] present in test data point [True]
800 Text feature [foci] present in test data point [True]
804 Text feature [10] present in test data point [True]
805 Text feature [able] present in test data point [True]
806 Text feature [test] present in test data point [True]
809 Text feature [products] present in test data point [True]
812 Text feature [selected] present in test data point [True]
814 Text feature [highly] present in test data point [True]
815 Text feature [subjected] present in test data point [True]
816 Text feature [view] present in test data point [True]
819 Text feature [per] present in test data point [True]
820 Text feature [mtor] present in test data point [True]
821 Text feature [codon] present in test data point [True]
822 Text feature [bone] present in test data point [True]
826 Text feature [conditions] present in test data point [True]

```
826 Text feature [conditions] present in test data point [True]
827 Text feature [followed] present in test data point [True]
828 Text feature [2000] present in test data point [True]
831 Text feature [various] present in test data point [True]
832 Text feature [unique] present in test data point [True]
835 Text feature [aberrant] present in test data point [True]
841 Text feature [part] present in test data point [True]
842 Text feature [ph] present in test data point [True]
846 Text feature [strongly] present in test data point [True]
848 Text feature [impact] present in test data point [True]
850 Text feature [major] present in test data point [True]
852 Text feature [difference] present in test data point [True]
853 Text feature [taken] present in test data point [True]
855 Text feature [finding] present in test data point [True]
858 Text feature [observation] present in test data point [True]
861 Text feature [groups] present in test data point [True]
865 Text feature [less] present in test data point [True]
873 Text feature [concentration] present in test data point [True]
874 Text feature [repeats] present in test data point [True]
876 Text feature [critical] present in test data point [True]
877 Text feature [ca] present in test data point [True]
878 Text feature [applied] present in test data point [True]
879 Text feature [sds] present in test data point [True]
882 Text feature [assess] present in test data point [True]
883 Text feature [years] present in test data point [True]
887 Text feature [alternative] present in test data point [True]
888 Text feature [subunit] present in test data point [True]
889 Text feature [free] present in test data point [True]
890 Text feature [indeed] present in test data point [True]
891 Text feature [copy] present in test data point [True]
894 Text feature [40] present in test data point [True]
895 Text feature [according] present in test data point [True]
899 Text feature [min] present in test data point [True]
905 Text feature [promote] present in test data point [True]
911 Text feature [eight] present in test data point [True]
915 Text feature [download] present in test data point [True]
921 Text feature [larger] present in test data point [True]
923 Text feature [stained] present in test data point [True]
925 Text feature [properties] present in test data point [True]
927 Text feature [side] present in test data point [True]
```

927 Text feature [side] present in test data point [True]
928 Text feature [incubated] present in test data point [True]
933 Text feature [considered] present in test data point [True]
934 Text feature [colonies] present in test data point [True]
936 Text feature [characteristics] present in test data point [True]
939 Text feature [colony] present in test data point [True]
945 Text feature [like] present in test data point [True]
947 Text feature [upon] present in test data point [True]
961 Text feature [met] present in test data point [True]
963 Text feature [reaction] present in test data point [True]
966 Text feature [2004] present in test data point [True]
974 Text feature [pik3ca] present in test data point [True]
978 Text feature [supplemental] present in test data point [True]
984 Text feature [reverse] present in test data point [True]
992 Text feature [leads] present in test data point [True]
Out of the top 1000 features 530 are present in query point

4.5.3.2. Inorrectly Classified point

In [89]:
```python
test_point_index = 55
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_po
int_index],test_df['Gene'].iloc[test_point_index],test_df['Variation'].
iloc[test_point_index], no_feature)
```

Predicted Class : 7
Predicted Class Probabilities: [[0.1268 0.2221 0.0185 0.1523 0.0574 0.0
435 0.3618 0.0072 0.0103]]
Actuall Class : 4
----
0 Text feature [kinase] present in test data point [True]

```
2 Text feature [phosphorylation] present in test data point [True]
3 Text feature [inhibitors] present in test data point [True]
4 Text feature [inhibitor] present in test data point [True]
5 Text feature [activated] present in test data point [True]
6 Text feature [constitutive] present in test data point [True]
7 Text feature [activation] present in test data point [True]
8 Text feature [function] present in test data point [True]
9 Text feature [tyrosine] present in test data point [True]
10 Text feature [suppressor] present in test data point [True]
11 Text feature [loss] present in test data point [True]
12 Text feature [missense] present in test data point [True]
14 Text feature [constitutively] present in test data point [True]
15 Text feature [receptor] present in test data point [True]
16 Text feature [treatment] present in test data point [True]
17 Text feature [signaling] present in test data point [True]
18 Text feature [protein] present in test data point [True]
21 Text feature [pten] present in test data point [True]
22 Text feature [variants] present in test data point [True]
24 Text feature [treated] present in test data point [True]
25 Text feature [functional] present in test data point [True]
28 Text feature [oncogenic] present in test data point [True]
29 Text feature [stability] present in test data point [True]
31 Text feature [kinases] present in test data point [True]
32 Text feature [cell] present in test data point [True]
33 Text feature [trials] present in test data point [True]
34 Text feature [expression] present in test data point [True]
38 Text feature [transforming] present in test data point [True]
41 Text feature [drug] present in test data point [True]
42 Text feature [akt] present in test data point [True]
45 Text feature [phosphatase] present in test data point [True]
46 Text feature [inhibited] present in test data point [True]
47 Text feature [therapy] present in test data point [True]
49 Text feature [cells] present in test data point [True]
50 Text feature [expected] present in test data point [True]
53 Text feature [response] present in test data point [True]
57 Text feature [oncogene] present in test data point [True]
62 Text feature [activate] present in test data point [True]
64 Text feature [patients] present in test data point [True]
65 Text feature [survival] present in test data point [True]
68 Text feature [inhibition] present in test data point [True]
```

```
69 Text feature [functions] present in test data point [True]
70 Text feature [growth] present in test data point [True]
71 Text feature [proliferation] present in test data point [True]
73 Text feature [clinical] present in test data point [True]
74 Text feature [proteins] present in test data point [True]
75 Text feature [therapeutic] present in test data point [True]
78 Text feature [phospho] present in test data point [True]
79 Text feature [expressing] present in test data point [True]
80 Text feature [downstream] present in test data point [True]
82 Text feature [dna] present in test data point [True]
83 Text feature [inactivation] present in test data point [True]
86 Text feature [combined] present in test data point [True]
87 Text feature [splice] present in test data point [True]
92 Text feature [assay] present in test data point [True]
93 Text feature [activity] present in test data point [True]
94 Text feature [substitutions] present in test data point [True]
95 Text feature [model] present in test data point [True]
96 Text feature [classification] present in test data point [True]
97 Text feature [tagged] present in test data point [True]
100 Text feature [lines] present in test data point [True]
101 Text feature [identified] present in test data point [True]
103 Text feature [likely] present in test data point [True]
104 Text feature [partial] present in test data point [True]
105 Text feature [transformation] present in test data point [True]
106 Text feature [ability] present in test data point [True]
107 Text feature [basis] present in test data point [True]
108 Text feature [affected] present in test data point [True]
111 Text feature [binding] present in test data point [True]
112 Text feature [genes] present in test data point [True]
113 Text feature [amplification] present in test data point [True]
114 Text feature [catalytic] present in test data point [True]
115 Text feature [tumors] present in test data point [True]
116 Text feature [based] present in test data point [True]
118 Text feature [weeks] present in test data point [True]
121 Text feature [potential] present in test data point [True]
123 Text feature [effective] present in test data point [True]
125 Text feature [interaction] present in test data point [True]
126 Text feature [indicated] present in test data point [True]
129 Text feature [patient] present in test data point [True]
```

```
131 Text feature [evidence] present in test data point [True]
133 Text feature [sequencing] present in test data point [True]
136 Text feature [type] present in test data point [True]
137 Text feature [one] present in test data point [True]
139 Text feature [results] present in test data point [True]
140 Text feature [substrate] present in test data point [True]
141 Text feature [pathway] present in test data point [True]
142 Text feature [site] present in test data point [True]
144 Text feature [affect] present in test data point [True]
145 Text feature [dose] present in test data point [True]
146 Text feature [bind] present in test data point [True]
148 Text feature [structural] present in test data point [True]
149 Text feature [variant] present in test data point [True]
150 Text feature [sensitive] present in test data point [True]
151 Text feature [experiments] present in test data point [True]
153 Text feature [32] present in test data point [True]
154 Text feature [41] present in test data point [True]
155 Text feature [50] present in test data point [True]
156 Text feature [conserved] present in test data point [True]
158 Text feature [p53] present in test data point [True]
160 Text feature [mutants] present in test data point [True]
162 Text feature [given] present in test data point [True]
164 Text feature [majority] present in test data point [True]
167 Text feature [transcription] present in test data point [True]
170 Text feature [recently] present in test data point [True]
171 Text feature [human] present in test data point [True]
172 Text feature [use] present in test data point [True]
177 Text feature [35] present in test data point [True]
179 Text feature [suggesting] present in test data point [True]
180 Text feature [many] present in test data point [True]
182 Text feature [deletion] present in test data point [True]
187 Text feature [available] present in test data point [True]
190 Text feature [inhibitory] present in test data point [True]
191 Text feature [previously] present in test data point [True]
192 Text feature [first] present in test data point [True]
193 Text feature [pi3k] present in test data point [True]
195 Text feature [wild] present in test data point [True]
197 Text feature [large] present in test data point [True]
198 Text feature [progression] present in test data point [True]
```

```
200 Text feature [active] present in test data point [True]
201 Text feature [two] present in test data point [True]
202 Text feature [present] present in test data point [True]
204 Text feature [malignant] present in test data point [True]
205 Text feature [harboring] present in test data point [True]
207 Text feature [days] present in test data point [True]
209 Text feature [2a] present in test data point [True]
210 Text feature [sequence] present in test data point [True]
211 Text feature [presence] present in test data point [True]
212 Text feature [inhibit] present in test data point [True]
213 Text feature [somatic] present in test data point [True]
214 Text feature [consistent] present in test data point [True]
215 Text feature [stem] present in test data point [True]
217 Text feature [3a] present in test data point [True]
219 Text feature [allele] present in test data point [True]
220 Text feature [discussion] present in test data point [True]
221 Text feature [status] present in test data point [True]
222 Text feature [introduction] present in test data point [True]
223 Text feature [possible] present in test data point [True]
224 Text feature [cannot] present in test data point [True]
225 Text feature [26] present in test data point [True]
228 Text feature [known] present in test data point [True]
229 Text feature [family] present in test data point [True]
230 Text feature [31] present in test data point [True]
231 Text feature [70] present in test data point [True]
233 Text feature [reduced] present in test data point [True]
236 Text feature [individuals] present in test data point [True]
237 Text feature [12] present in test data point [True]
238 Text feature [role] present in test data point [True]
239 Text feature [enzyme] present in test data point [True]
240 Text feature [mediated] present in test data point [True]
242 Text feature [expressed] present in test data point [True]
244 Text feature [binds] present in test data point [True]
247 Text feature [levels] present in test data point [True]
248 Text feature [significantly] present in test data point [True]
251 Text feature [detected] present in test data point [True]
255 Text feature [cancers] present in test data point [True]
256 Text feature [induced] present in test data point [True]
258 Text feature [single] present in test data point [True]
```

```
260 Text feature [whereas] present in test data point [True]
261 Text feature [suggest] present in test data point [True]
263 Text feature [56] present in test data point [True]
264 Text feature [independent] present in test data point [True]
265 Text feature [gene] present in test data point [True]
266 Text feature [dependent] present in test data point [True]
268 Text feature [complex] present in test data point [True]
269 Text feature [11] present in test data point [True]
272 Text feature [absence] present in test data point [True]
273 Text feature [mechanism] present in test data point [True]
275 Text feature [demonstrated] present in test data point [True]
276 Text feature [observed] present in test data point [True]
278 Text feature [green] present in test data point [True]
279 Text feature [figure] present in test data point [True]
281 Text feature [data] present in test data point [True]
283 Text feature [primary] present in test data point [True]
288 Text feature [associated] present in test data point [True]
289 Text feature [published] present in test data point [True]
290 Text feature [described] present in test data point [True]
291 Text feature [specific] present in test data point [True]
292 Text feature [developed] present in test data point [True]
293 Text feature [genetic] present in test data point [True]
295 Text feature [increased] present in test data point [True]
299 Text feature [mutant] present in test data point [True]
300 Text feature [25] present in test data point [True]
301 Text feature [factor] present in test data point [True]
302 Text feature [suppression] present in test data point [True]
303 Text feature [directly] present in test data point [True]
305 Text feature [effect] present in test data point [True]
306 Text feature [structure] present in test data point [True]
307 Text feature [using] present in test data point [True]
308 Text feature [14] present in test data point [True]
309 Text feature [1a] present in test data point [True]
310 Text feature [confirmed] present in test data point [True]
311 Text feature [recognition] present in test data point [True]
312 Text feature [least] present in test data point [True]
315 Text feature [acids] present in test data point [True]
316 Text feature [next] present in test data point [True]
317 Text feature [whether] present in test data point [True]
```

```
318 Text feature [models] present in test data point [True]
320 Text feature [wt] present in test data point [True]
323 Text feature [generation] present in test data point [True]
324 Text feature [19] present in test data point [True]
326 Text feature [anti] present in test data point [True]
327 Text feature [degradation] present in test data point [True]
328 Text feature [control] present in test data point [True]
329 Text feature [formation] present in test data point [True]
330 Text feature [mouse] present in test data point [True]
331 Text feature [pcr] present in test data point [True]
333 Text feature [even] present in test data point [True]
335 Text feature [52] present in test data point [True]
337 Text feature [molecule] present in test data point [True]
338 Text feature [half] present in test data point [True]
339 Text feature [total] present in test data point [True]
341 Text feature [related] present in test data point [True]
342 Text feature [mutation] present in test data point [True]
343 Text feature [deficient] present in test data point [True]
344 Text feature [level] present in test data point [True]
345 Text feature [domain] present in test data point [True]
346 Text feature [lack] present in test data point [True]
349 Text feature [tested] present in test data point [True]
352 Text feature [study] present in test data point [True]
354 Text feature [showed] present in test data point [True]
356 Text feature [shown] present in test data point [True]
357 Text feature [addition] present in test data point [True]
358 Text feature [overall] present in test data point [True]
359 Text feature [mice] present in test data point [True]
360 Text feature [another] present in test data point [True]
361 Text feature [mutated] present in test data point [True]
364 Text feature [population] present in test data point [True]
365 Text feature [value] present in test data point [True]
366 Text feature [co] present in test data point [True]
367 Text feature [multiple] present in test data point [True]
368 Text feature [22] present in test data point [True]
370 Text feature [interactions] present in test data point [True]
372 Text feature [deletions] present in test data point [True]
373 Text feature [point] present in test data point [True]
374 Text feature [ref] present in test data point [True]
```

```
377 Text feature [37] present in test data point [True]
378 Text feature [3b] present in test data point [True]
380 Text feature [domains] present in test data point [True]
381 Text feature [29] present in test data point [True]
382 Text feature [region] present in test data point [True]
383 Text feature [occur] present in test data point [True]
385 Text feature [17] present in test data point [True]
386 Text feature [lower] present in test data point [True]
388 Text feature [analyzed] present in test data point [True]
389 Text feature [also] present in test data point [True]
393 Text feature [heterozygous] present in test data point [True]
394 Text feature [rate] present in test data point [True]
396 Text feature [hotspot] present in test data point [True]
398 Text feature [used] present in test data point [True]
399 Text feature [59] present in test data point [True]
400 Text feature [hypothesis] present in test data point [True]
402 Text feature [controls] present in test data point [True]
404 Text feature [specificity] present in test data point [True]
406 Text feature [displayed] present in test data point [True]
408 Text feature [clear] present in test data point [True]
410 Text feature [frequently] present in test data point [True]
411 Text feature [although] present in test data point [True]
413 Text feature [including] present in test data point [True]
414 Text feature [resulting] present in test data point [True]
416 Text feature [loop] present in test data point [True]
418 Text feature [direct] present in test data point [True]
419 Text feature [defined] present in test data point [True]
420 Text feature [leading] present in test data point [True]
421 Text feature [number] present in test data point [True]
423 Text feature [contrast] present in test data point [True]
426 Text feature [within] present in test data point [True]
427 Text feature [via] present in test data point [True]
429 Text feature [representative] present in test data point [True]
430 Text feature [impaired] present in test data point [True]
431 Text feature [alleles] present in test data point [True]
432 Text feature [cellular] present in test data point [True]
435 Text feature [increase] present in test data point [True]
437 Text feature [acid] present in test data point [True]
439 Text feature [moreover] present in test data point [True]
```

```
440 Text feature [remaining] present in test data point [True]
441 Text feature [respectively] present in test data point [True]
442 Text feature [mutational] present in test data point [True]
445 Text feature [well] present in test data point [True]
446 Text feature [tables] present in test data point [True]
448 Text feature [apoptosis] present in test data point [True]
449 Text feature [showing] present in test data point [True]
451 Text feature [interact] present in test data point [True]
454 Text feature [essential] present in test data point [True]
456 Text feature [similar] present in test data point [True]
459 Text feature [changes] present in test data point [True]
460 Text feature [second] present in test data point [True]
464 Text feature [higher] present in test data point [True]
465 Text feature [would] present in test data point [True]
467 Text feature [however] present in test data point [True]
468 Text feature [30] present in test data point [True]
470 Text feature [forms] present in test data point [True]
471 Text feature [found] present in test data point [True]
473 Text feature [included] present in test data point [True]
474 Text feature [altered] present in test data point [True]
475 Text feature [performed] present in test data point [True]
476 Text feature [sites] present in test data point [True]
478 Text feature [src] present in test data point [True]
479 Text feature [screen] present in test data point [True]
480 Text feature [nature] present in test data point [True]
485 Text feature [driven] present in test data point [True]
486 Text feature [antibodies] present in test data point [True]
488 Text feature [additional] present in test data point [True]
490 Text feature [set] present in test data point [True]
491 Text feature [identification] present in test data point [True]
492 Text feature [form] present in test data point [True]
493 Text feature [important] present in test data point [True]
495 Text feature [development] present in test data point [True]
496 Text feature [negative] present in test data point [True]
500 Text feature [encoding] present in test data point [True]
501 Text feature [high] present in test data point [True]
505 Text feature [required] present in test data point [True]
507 Text feature [transformed] present in test data point [True]
509 Text feature [bcr] present in test data point [True]
```

```
510 Text feature [cancer] present in test data point [True]
511 Text feature [common] present in test data point [True]
512 Text feature [due] present in test data point [True]
514 Text feature [2013] present in test data point [True]
515 Text feature [suggests] present in test data point [True]
516 Text feature [2b] present in test data point [True]
517 Text feature [gain] present in test data point [True]
518 Text feature [lead] present in test data point [True]
520 Text feature [alone] present in test data point [True]
521 Text feature [al] present in test data point [True]
522 Text feature [disease] present in test data point [True]
524 Text feature [1b] present in test data point [True]
525 Text feature [cause] present in test data point [True]
526 Text feature [target] present in test data point [True]
530 Text feature [reports] present in test data point [True]
531 Text feature [et] present in test data point [True]
532 Text feature [recent] present in test data point [True]
533 Text feature [recurrent] present in test data point [True]
534 Text feature [normal] present in test data point [True]
535 Text feature [fig] present in test data point [True]
537 Text feature [95] present in test data point [True]
540 Text feature [fusion] present in test data point [True]
541 Text feature [whole] present in test data point [True]
542 Text feature [relatively] present in test data point [True]
545 Text feature [early] present in test data point [True]
546 Text feature [small] present in test data point [True]
547 Text feature [analysis] present in test data point [True]
548 Text feature [tumor] present in test data point [True]
549 Text feature [13] present in test data point [True]
550 Text feature [20] present in test data point [True]
552 Text feature [selection] present in test data point [True]
554 Text feature [frequency] present in test data point [True]
555 Text feature [limited] present in test data point [True]
556 Text feature [genomic] present in test data point [True]
558 Text feature [nucleotide] present in test data point [True]
559 Text feature [table] present in test data point [True]
560 Text feature [28] present in test data point [True]
562 Text feature [terminus] present in test data point [True]
565 Text feature [previous] present in test data point [True]
```

```
567 Text feature [top] present in test data point [True]
568 Text feature [together] present in test data point [True]
570 Text feature [endogenous] present in test data point [True]
571 Text feature [position] present in test data point [True]
572 Text feature [result] present in test data point [True]
573 Text feature [revealed] present in test data point [True]
574 Text feature [sample] present in test data point [True]
577 Text feature [thus] present in test data point [True]
578 Text feature [differentiation] present in test data point [True]
579 Text feature [experimental] present in test data point [True]
582 Text feature [24] present in test data point [True]
583 Text feature [positive] present in test data point [True]
585 Text feature [16] present in test data point [True]
586 Text feature [methods] present in test data point [True]
587 Text feature [interestingly] present in test data point [True]
588 Text feature [established] present in test data point [True]
589 Text feature [decreased] present in test data point [True]
591 Text feature [cycle] present in test data point [True]
593 Text feature [among] present in test data point [True]
596 Text feature [subset] present in test data point [True]
598 Text feature [surface] present in test data point [True]
599 Text feature [complete] present in test data point [True]
600 Text feature [non] present in test data point [True]
604 Text feature [subsequent] present in test data point [True]
606 Text feature [linked] present in test data point [True]
607 Text feature [provide] present in test data point [True]
609 Text feature [fold] present in test data point [True]
610 Text feature [three] present in test data point [True]
611 Text feature [regulatory] present in test data point [True]
613 Text feature [pattern] present in test data point [True]
614 Text feature [combination] present in test data point [True]
618 Text feature [fraction] present in test data point [True]
619 Text feature [compared] present in test data point [True]
620 Text feature [alterations] present in test data point [True]
621 Text feature [overexpression] present in test data point [True]
623 Text feature [novel] present in test data point [True]
624 Text feature [mrna] present in test data point [True]
625 Text feature [without] present in test data point [True]
626 Text feature [construct] present in test data point [True]
```

```
627 Text feature [induce] present in test data point [True]
628 Text feature [suggested] present in test data point [True]
629 Text feature [four] present in test data point [True]
630 Text feature [comparison] present in test data point [True]
631 Text feature [include] present in test data point [True]
632 Text feature [distribution] present in test data point [True]
633 Text feature [promoter] present in test data point [True]
634 Text feature [either] present in test data point [True]
636 Text feature [transactivation] present in test data point [True]
637 Text feature [current] present in test data point [True]
638 Text feature [studies] present in test data point [True]
639 Text feature [dominant] present in test data point [True]
641 Text feature [residues] present in test data point [True]
644 Text feature [enhanced] present in test data point [True]
645 Text feature [findings] present in test data point [True]
646 Text feature [involving] present in test data point [True]
647 Text feature [frame] present in test data point [True]
654 Text feature [mechanisms] present in test data point [True]
657 Text feature [regions] present in test data point [True]
660 Text feature [hotspots] present in test data point [True]
661 Text feature [vitro] present in test data point [True]
665 Text feature [locus] present in test data point [True]
666 Text feature [screening] present in test data point [True]
667 Text feature [strong] present in test data point [True]
668 Text feature [long] present in test data point [True]
671 Text feature [several] present in test data point [True]
672 Text feature [case] present in test data point [True]
674 Text feature [may] present in test data point [True]
675 Text feature [seen] present in test data point [True]
676 Text feature [amino] present in test data point [True]
678 Text feature [day] present in test data point [True]
679 Text feature [differences] present in test data point [True]
680 Text feature [pathways] present in test data point [True]
681 Text feature [involved] present in test data point [True]
682 Text feature [key] present in test data point [True]
683 Text feature [evaluated] present in test data point [True]
684 Text feature [initial] present in test data point [True]
685 Text feature [2011] present in test data point [True]
686 Text feature [remains] present in test data point [True]
```

```
690 Text feature [led] present in test data point [True]
691 Text feature [2005] present in test data point [True]
695 Text feature [across] present in test data point [True]
696 Text feature [caused] present in test data point [True]
697 Text feature [go] present in test data point [True]
699 Text feature [observations] present in test data point [True]
700 Text feature [chemotherapy] present in test data point [True]
702 Text feature [association] present in test data point [True]
703 Text feature [see] present in test data point [True]
704 Text feature [4a] present in test data point [True]
705 Text feature [2010] present in test data point [True]
706 Text feature [helix] present in test data point [True]
710 Text feature [chromosome] present in test data point [True]
711 Text feature [2006] present in test data point [True]
712 Text feature [events] present in test data point [True]
713 Text feature [targeting] present in test data point [True]
714 Text feature [biological] present in test data point [True]
715 Text feature [phase] present in test data point [True]
716 Text feature [particular] present in test data point [True]
717 Text feature [could] present in test data point [True]
720 Text feature [following] present in test data point [True]
721 Text feature [targeted] present in test data point [True]
722 Text feature [indicate] present in test data point [True]
724 Text feature [transcriptional] present in test data point [True]
726 Text feature [express] present in test data point [True]
727 Text feature [might] present in test data point [True]
728 Text feature [genome] present in test data point [True]
734 Text feature [mutations] present in test data point [True]
735 Text feature [new] present in test data point [True]
738 Text feature [100] present in test data point [True]
739 Text feature [affinity] present in test data point [True]
741 Text feature [structures] present in test data point [True]
742 Text feature [line] present in test data point [True]
744 Text feature [regulation] present in test data point [True]
745 Text feature [relevant] present in test data point [True]
746 Text feature [bound] present in test data point [True]
752 Text feature [factors] present in test data point [True]
756 Text feature [average] present in test data point [True]
757 Text feature [rna] present in test data point [True]
```

```
759 Text feature [rare] present in test data point [True]
763 Text feature [38] present in test data point [True]
765 Text feature [tp53] present in test data point [True]
766 Text feature [secondary] present in test data point [True]
767 Text feature [specifically] present in test data point [True]
768 Text feature [context] present in test data point [True]
769 Text feature [gfp] present in test data point [True]
771 Text feature [different] present in test data point [True]
774 Text feature [distinct] present in test data point [True]
777 Text feature [corresponding] present in test data point [True]
778 Text feature [residue] present in test data point [True]
780 Text feature [contact] present in test data point [True]
782 Text feature [members] present in test data point [True]
783 Text feature [myc] present in test data point [True]
786 Text feature [cases] present in test data point [True]
788 Text feature [positions] present in test data point [True]
789 Text feature [time] present in test data point [True]
792 Text feature [akt1] present in test data point [True]
793 Text feature [samples] present in test data point [True]
796 Text feature [state] present in test data point [True]
797 Text feature [stable] present in test data point [True]
799 Text feature [numbers] present in test data point [True]
804 Text feature [10] present in test data point [True]
805 Text feature [able] present in test data point [True]
806 Text feature [test] present in test data point [True]
807 Text feature [blood] present in test data point [True]
808 Text feature [class] present in test data point [True]
809 Text feature [products] present in test data point [True]
812 Text feature [selected] present in test data point [True]
813 Text feature [frequent] present in test data point [True]
814 Text feature [highly] present in test data point [True]
817 Text feature [process] present in test data point [True]
818 Text feature [possibility] present in test data point [True]
820 Text feature [mtor] present in test data point [True]
822 Text feature [bone] present in test data point [True]
823 Text feature [motif] present in test data point [True]
824 Text feature [induction] present in test data point [True]
829 Text feature [sufficient] present in test data point [True]
831 Text feature [various] present in test data point [True]
```

```
832 Text feature [unique] present in test data point [True]
834 Text feature [indicating] present in test data point [True]
835 Text feature [aberrant] present in test data point [True]
836 Text feature [recombination] present in test data point [True]
841 Text feature [part] present in test data point [True]
846 Text feature [strongly] present in test data point [True]
849 Text feature [phenotype] present in test data point [True]
850 Text feature [major] present in test data point [True]
855 Text feature [finding] present in test data point [True]
856 Text feature [often] present in test data point [True]
858 Text feature [observation] present in test data point [True]
862 Text feature [germline] present in test data point [True]
863 Text feature [fact] present in test data point [True]
864 Text feature [lymphoma] present in test data point [True]
865 Text feature [less] present in test data point [True]
866 Text feature [rearrangements] present in test data point [True]
868 Text feature [contribute] present in test data point [True]
869 Text feature [lesions] present in test data point [True]
876 Text feature [critical] present in test data point [True]
880 Text feature [targets] present in test data point [True]
881 Text feature [regulated] present in test data point [True]
882 Text feature [assess] present in test data point [True]
883 Text feature [years] present in test data point [True]
887 Text feature [alternative] present in test data point [True]
888 Text feature [subunit] present in test data point [True]
890 Text feature [indeed] present in test data point [True]
894 Text feature [40] present in test data point [True]
896 Text feature [sporadic] present in test data point [True]
900 Text feature [box] present in test data point [True]
904 Text feature [since] present in test data point [True]
905 Text feature [promote] present in test data point [True]
907 Text feature [note] present in test data point [True]
909 Text feature [chain] present in test data point [True]
912 Text feature [yet] present in test data point [True]
913 Text feature [reduction] present in test data point [True]
914 Text feature [cluster] present in test data point [True]
916 Text feature [derived] present in test data point [True]
917 Text feature [finally] present in test data point [True]
920 Text feature [knockdown] present in test data point [True]
```

```
921 Text feature [larger] present in test data point [True]
926 Text feature [figures] present in test data point [True]
932 Text feature [stage] present in test data point [True]
938 Text feature [coding] present in test data point [True]
942 Text feature [2001] present in test data point [True]
944 Text feature [similarly] present in test data point [True]
945 Text feature [like] present in test data point [True]
946 Text feature [panel] present in test data point [True]
947 Text feature [upon] present in test data point [True]
954 Text feature [1996] present in test data point [True]
962 Text feature [rt] present in test data point [True]
963 Text feature [reaction] present in test data point [True]
964 Text feature [1997] present in test data point [True]
965 Text feature [despite] present in test data point [True]
966 Text feature [2004] present in test data point [True]
972 Text feature [2009] present in test data point [True]
973 Text feature [appears] present in test data point [True]
974 Text feature [pik3ca] present in test data point [True]
977 Text feature [2008] present in test data point [True]
978 Text feature [supplemental] present in test data point [True]
984 Text feature [reverse] present in test data point [True]
991 Text feature [cyclin] present in test data point [True]
Out of the top  1000  features  570 are present in query point
```

4.5.3. Hyper paramter tuning (With Response Coding)

```
In [94]:  # -------------------------------
          # default parameters
          # sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='g
          ini', max_depth=None, min_samples_split=2,
          # min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='aut
          o', max_leaf_nodes=None, min_impurity_decrease=0.0,
          # min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, r
          andom_state=None, verbose=0, warm_start=False,
          # class_weight=None)

          # Some of methods of RandomForestClassifier()
```

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

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

# ----------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/random-forest-and-their-construction-2/
# ----------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight])    Fit the calibrated model
# get_params([deep])    Get parameters for this estimator.
# predict(X)    Predict the target of new samples.
# predict_proba(X)      Posterior probabilities of classification
#-----------------------------------
# video link:
#-----------------------------------

alpha = [10,50,100,200,500,1000]
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',
```

```python
                              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]).ra
vel()
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)), (featur
es[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)], cri
terion='gini', max_depth=max_depth[int(best_alpha%4)], 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)

predict_y = sig_clf.predict_proba(train_x_responseCoding)
print('For values of best alpha = ', alpha[int(best_alpha/4)], "The tra
in log loss is:",log_loss(y_train, predict_y, labels=clf.classes_, eps=
1e-15))
predict_y = sig_clf.predict_proba(cv_x_responseCoding)
print('For values of best alpha = ', alpha[int(best_alpha/4)], "The cro
ss validation log loss is:",log_loss(y_cv, predict_y, labels=clf.classe
s_, eps=1e-15))
```

```
predict_y = sig_clf.predict_proba(test_x_responseCoding)
print('For values of best alpha = ', alpha[int(best_alpha/4)], "The tes
t log loss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e
-15))
```

```
for n_estimators = 10 and max depth =  2
Log Loss : 2.205010902896129
for n_estimators = 10 and max depth =  3
Log Loss : 1.6036729102430995
for n_estimators = 10 and max depth =  5
Log Loss : 1.2720634690809804
for n_estimators = 10 and max depth =  10
Log Loss : 2.031969833322205
for n_estimators = 50 and max depth =  2
Log Loss : 1.6521070870598387
for n_estimators = 50 and max depth =  3
Log Loss : 1.4149813072776047
for n_estimators = 50 and max depth =  5
Log Loss : 1.2945296226152692
for n_estimators = 50 and max depth =  10
Log Loss : 1.6799539729161637
for n_estimators = 100 and max depth =  2
Log Loss : 1.4915951147277886
for n_estimators = 100 and max depth =  3
Log Loss : 1.4594948822256295
for n_estimators = 100 and max depth =  5
Log Loss : 1.3441231725903056
for n_estimators = 100 and max depth =  10
Log Loss : 1.7043432279007016
for n_estimators = 200 and max depth =  2
Log Loss : 1.63653034044923
for n_estimators = 200 and max depth =  3
Log Loss : 1.4681404499188615
for n_estimators = 200 and max depth =  5
Log Loss : 1.3603531172139833
for n_estimators = 200 and max depth =  10
Log Loss : 1.6924644656791237
for n_estimators = 500 and max depth =  2
Log Loss : 1.6730932860288907
for n_estimators = 500 and max depth =  3
```

```
Log Loss : 1.5096321745272052
for n_estimators = 500 and max depth =  5
Log Loss : 1.3669836108428683
for n_estimators = 500 and max depth =  10
Log Loss : 1.7577865498622776
for n_estimators = 1000 and max depth =  2
Log Loss : 1.6443986760966114
for n_estimators = 1000 and max depth =  3

Log Loss : 1.5046126836056442
for n_estimators = 1000 and max depth =  5
Log Loss : 1.3638168549665657
for n_estimators = 1000 and max depth =  10
Log Loss : 1.7775847077708469
For values of best alpha =  10 The train log loss is: 0.077717967649555
69
For values of best alpha =  10 The cross validation log loss is: 1.2720
634690809804
For values of best alpha =  10 The test log loss is: 1.329313933250481
```

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

In [95]:
```python
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='g
ini', max_depth=None, min_samples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='aut
o', max_leaf_nodes=None, min_impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, r
andom_state=None, verbose=0, warm_start=False,
# class_weight=None)

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

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

# ----------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/random-forest-and-their-construction-2/
# ----------------------------------

clf = RandomForestClassifier(max_depth=max_depth[int(best_alpha%4)], n_
estimators=alpha[int(best_alpha/4)], criterion='gini', max_features='au
to',random_state=42)
predict_and_plot_confusion_matrix(train_x_responseCoding, train_y,cv_x_
responseCoding,cv_y, clf)
```

```
Log loss : 1.2720634690809804
Number of mis-classified points : 0.4116541353383459
-------------------- Confusion matrix --------------------
```



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

Precision matrix (Predicted Class vs Original Class)

Original Class \ Predicted Class	1	2	3	4	5	6	7	8	9
1	0.507	0.026	0.000	0.102	0.214	0.071	0.006	0.200	0.000
2	0.007	0.513	0.167	0.000	0.024	0.000	0.162	0.200	0.000
3	0.015	0.000	0.000	0.017	0.071	0.000	0.048	0.000	0.000
4	0.353	0.013	0.000	0.864	0.119	0.071	0.006	0.100	0.167
5	0.029	0.013	0.000	0.000	0.405	0.214	0.060	0.100	0.000
6	0.074	0.064	0.000	0.000	0.119	0.643	0.030	0.100	0.000
7	0.007	0.359	0.833	0.000	0.048	0.000	0.683	0.300	0.000
8	0.000	0.013	0.000	0.017	0.000	0.000	0.000	0.000	0.167
9	0.007	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.667

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

Original Class \ Predicted Class	1	2	3	4	5	6	7	8	9
1	0.758	0.022	0.000	0.066	0.099	0.022	0.011	0.022	0.000
2	0.014	0.556	0.014	0.000	0.014	0.000	0.375	0.028	0.000
3	0.143	0.000	0.000	0.071	0.214	0.000	0.571	0.000	0.000
4	0.436	0.009	0.000	0.464	0.045	0.018	0.009	0.009	0.009
5	0.103	0.026	0.000	0.000	0.436	0.154	0.256	0.026	0.000
6	0.227	0.114	0.000	0.000	0.114	0.409	0.114	0.023	0.000
7	0.007	0.183	0.033	0.000	0.013	0.000	0.745	0.020	0.000
8	0.000	0.333	0.000	0.333	0.000	0.000	0.000	0.000	0.333
9	0.167	0.000	0.000	0.000	0.000	0.000	0.167	0.000	0.667

4.5.5. Feature Importance

4.5.5.1. Correctly Classified point

```
In [96]:  clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/4)], cri
          terion='gini', max_depth=max_depth[int(best_alpha%4)], 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)


          test_point_index = 1
          no_feature = 27
          predicted_cls = sig_clf.predict(test_x_responseCoding[test_point_index]
          .reshape(1,-1))
          print("Predicted Class :", predicted_cls[0])
          print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(
          test_x_responseCoding[test_point_index].reshape(1,-1)),4))
          print("Actual Class :", 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 : 1
Predicted Class Probabilities: [[0.5366 0.0188 0.1147 0.2471 0.0203 0.0
129 0.0074 0.0126 0.0295]]
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
Text 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
Gene is important feature
Variation is important feature
Text is important feature
Variation is important feature
Variation is important feature
Gene is important feature
Gene is important feature
Text is important feature
Text is important feature
Text 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 [98]:
```python
test_point_index = 31
predicted_cls = sig_clf.predict(test_x_responseCoding[test_point_index]
.reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(
test_x_responseCoding[test_point_index].reshape(1,-1)),4))
print("Actual Class :", 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 : 7
Predicted Class Probabilities: [[0.0252 0.2051 0.2283 0.0192 0.0487 0.0
505 0.3571 0.0289 0.0369]]
Actual Class : 2
----------------------------------------------------
Variation is important feature
Variation is important feature
Variation is important feature
Variation is important feature
Gene is important feature
Variation is important feature
Text 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
Gene is important feature
Variation is important feature
Text is important feature
Variation is important feature
Variation is important feature
Gene is important feature
Gene is important feature
Text is important feature
Text is important feature
Text 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 [99]:
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# -------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_rate='optimal', eta0=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, …])    Fit linear model with S
tochastic 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 htt
p://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
# -------------------------------
# default parameters
# SVC(C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking
=True, probability=False, tol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decisi
on_function_shape='ovr', random_state=None)

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

```python
# read more about support vector machines with linear kernals here htt
p://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomFo
restClassifier.html
# -------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='g
ini', max_depth=None, min_samples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='aut
o', max_leaf_nodes=None, min_impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, r
andom_state=None, verbose=0, warm_start=False,
# class_weight=None)

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

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

# -------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/random-forest-and-their-construction-2/
# -------------------------------


clf1 = SGDClassifier(alpha=0.001, penalty='l2', loss='log', class_weigh
t='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_cl
f1.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.predic
t_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_proba(cv_x_onehotCoding))))
    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: 1.01
Support vector machines : Log Loss: 1.76
Naive Bayes : Log Loss: 1.17
--------------------------------------------------
Stacking Classifer : for the value of alpha: 0.000100 Log Loss: 2.177
Stacking Classifer : for the value of alpha: 0.001000 Log Loss: 2.026
Stacking Classifer : for the value of alpha: 0.010000 Log Loss: 1.474
Stacking Classifer : for the value of alpha: 0.100000 Log Loss: 1.152
Stacking Classifer : for the value of alpha: 1.000000 Log Loss: 1.404
Stacking Classifer : for the value of alpha: 10.000000 Log Loss: 1.897
```

4.7.2 testing the model with the best hyper parameters

In [100]:
```python
lr = LogisticRegression(C=0.1)
sclf = StackingClassifier(classifiers=[sig_clf1, sig_clf2, sig_clf3], m
eta_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.predic
t(test_x_onehotCoding)- test_y))/test_y.shape[0])
plot_confusion_matrix(test_y=test_y, predict_y=sclf.predict(test_x_oneh
otCoding))
```

```
Log loss (train) on the stacking classifier : 0.5221250491119382
Log loss (CV) on the stacking classifier : 1.152242089217523
Log loss (test) on the stacking classifier : 1.1955601096930255
Number of missclassified point : 0.3819548872180451
------------------- Confusion matrix --------------------
```
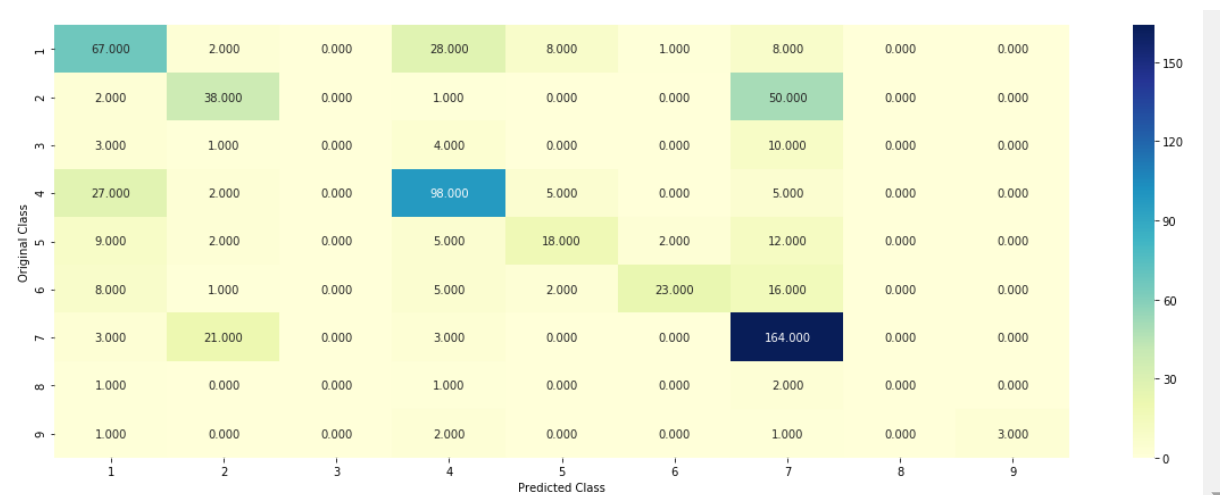
-------------------- Precision matrix (Columm Sum=1) -------------------
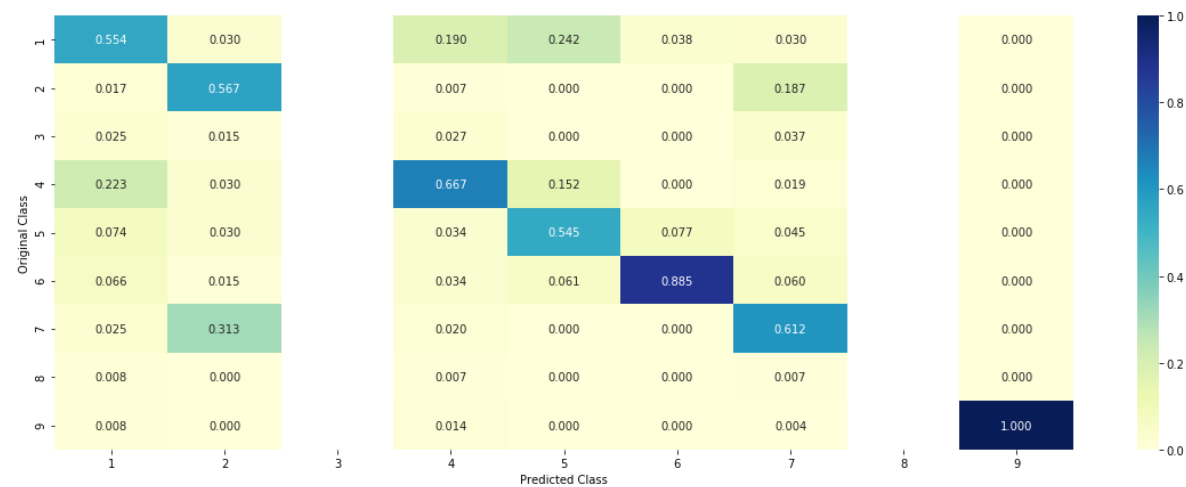
-------------------- Recall matrix (Row sum=1) --------------------

4.7.3 Maximum Voting classifier

```
In [101]: #Refer:http://scikit-learn.org/stable/modules/generated/sklearn.ensembl
          e.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, v
          clf.predict_proba(train_x_onehotCoding)))
          print("Log loss (CV) on the VotingClassifier :", log_loss(cv_y, vclf.pr
          edict_proba(cv_x_onehotCoding)))
          print("Log loss (test) on the VotingClassifier :", log_loss(test_y, vcl
          f.predict_proba(test_x_onehotCoding)))
          print("Number of missclassified point :", np.count_nonzero((vclf.predic
          t(test_x_onehotCoding)- test_y))/test_y.shape[0])
          plot_confusion_matrix(test_y=test_y, predict_y=vclf.predict(test_x_oneh
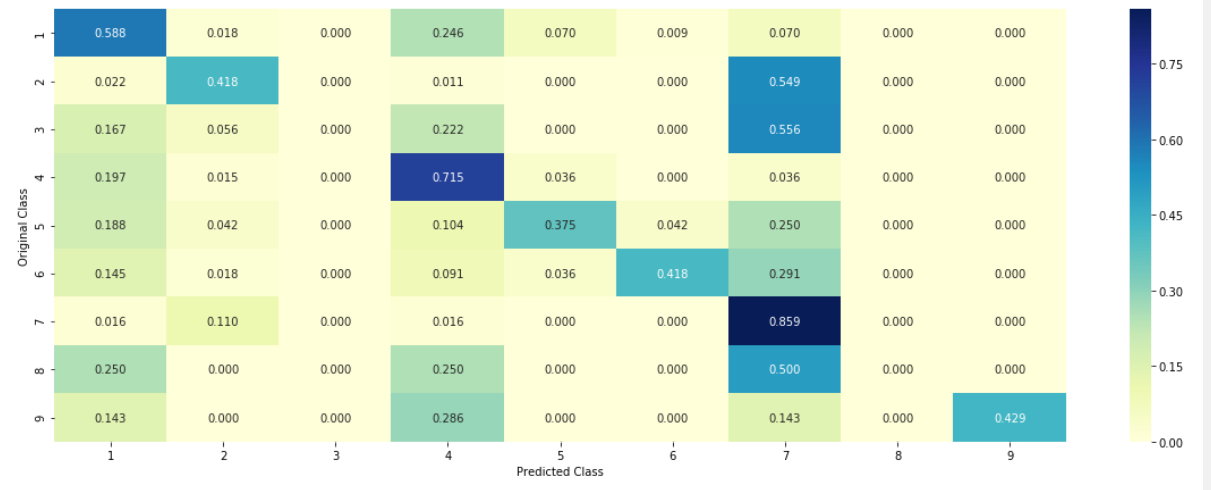          otCoding))
```

```
Log loss (train) on the VotingClassifier : 0.8225217966778062
Log loss (CV) on the VotingClassifier : 1.1612041182635207
Log loss (test) on the VotingClassifier : 1.2224910872322219
```

Number of missclassified point : 0.3699248120300752
------------------- Confusion matrix --------------------

------------------- Precision matrix (Columm Sum=1) -----------------
--
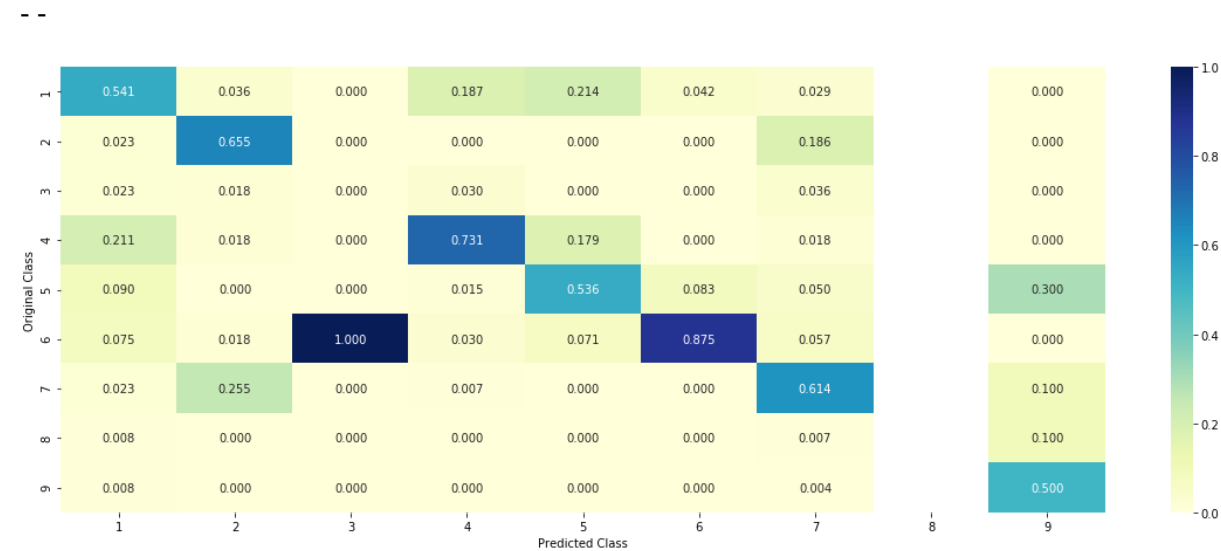
------------------- Recall matrix (Row sum=1) -------------------
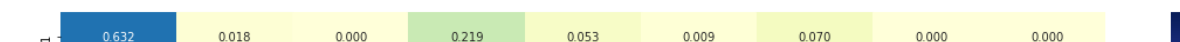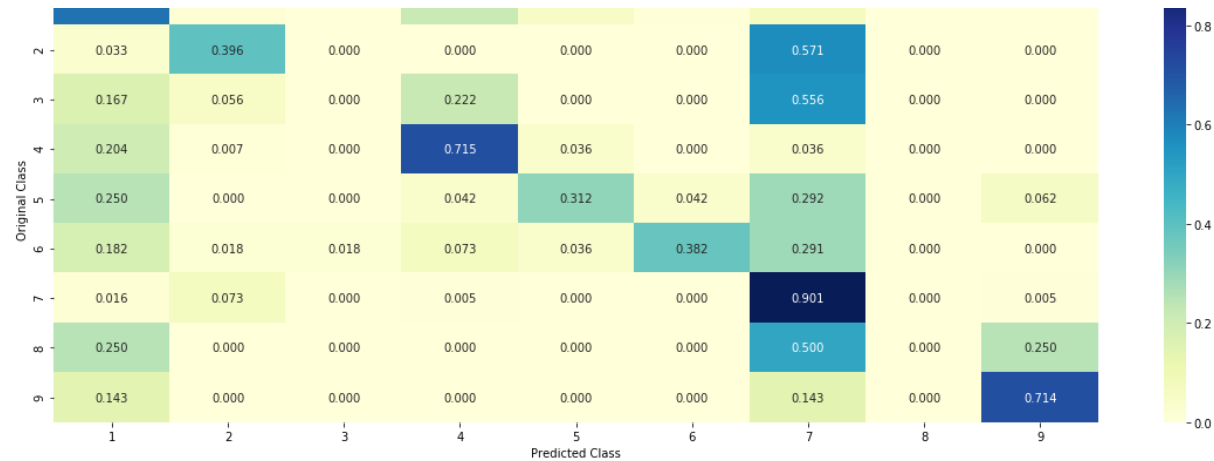
Summarizing the above loss into tabular format

In [104]:
```python
from prettytable import PrettyTable
x = PrettyTable()
#Adding Field-names
x.field_names = ["Model Name","TrainLoss","CVLoss","TestLoss","% Miscla
ssified Points"]
x.add_row(["Naive Bayes","0.50","1.16","1.22","41"])
x.add_row(["KNN","0.81","1.00","1.11","37"])
x.add_row(["Logistic Regression With Class balancing","0.72","1.11","1.
10","37"])
x.add_row(["Logistic Regression Without Class balancing","0.43","1.04",
"1.11","38"])
x.add_row(["Linear SVM","0.38","1.04","1.10","38"])
x.add_row(["Random Forest Classifier With One hot Encoding","0.84","1.1
8","1.22","40"])
x.add_row(["Random Forest Classifier With Response Coding","0.07","1.2
7","1.38","47"])
x.add_row(["Stack Models:LR+NB+SVM","1.01","1.17","1.76","32"])
```

```
x.add_row(["Maximum Voting classifier","0.82","1.16","1.22","33"])
print(x)
```

```
+---------------------------------------------------+-----------+--------+
----------+------------------------+

|                  Model Name                       | TrainLoss | CVLoss |
TestLoss | % Misclassified Points |
+---------------------------------------------------+-----------+--------+
----------+------------------------+
|                  Naive Bayes                      |   0.50    |  1.16  |
   1.22    |          41            |
|                      KNN                          |   0.81    |  1.00  |
   1.11    |          37            |
|    Logistic Regression With Class balancing       |   0.72    |  1.11  |
   1.10    |          37            |
|   Logistic Regression Without Class balancing     |   0.43    |  1.04  |
   1.11    |          38            |
|                  Linear SVM                       |   0.38    |  1.04  |
   1.10    |          38            |
| Random Forest Classifier With One hot Encoding    |   0.84    |  1.18  |
   1.22    |          40            |
| Random Forest Classifier With Response Coding     |   0.07    |  1.27  |
   1.38    |          47            |
|            Stack Models:LR+NB+SVM                 |   1.01    |  1.17  |
   1.76    |          32            |
|            Maximum Voting classifier              |   0.82    |  1.16  |
   1.22    |          33            |
+---------------------------------------------------+-----------+--------+
----------+------------------------+
```

Observation:

- From the above table we can say that LogisticRegression with class-balancing is perforing better compare to other models.
- Further more implementing with unigram-bigram let's reduce the log-loss further.

LogisticRegression including both unigrams and bigrams

In [105]:
```python
#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, "Gen
e", train_df))
# test gene feature
test_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gen
e", test_df))
# cross validation gene feature
cv_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene",
 cv_df))
```

In [108]:
```python
# one-hot encoding of Gene feature.
gene_vectorizer = CountVectorizer(ngram_range=(1,2))
train_gene_feature_onehotCoding = gene_vectorizer.fit_transform(train_d
f['Gene'])
test_gene_feature_onehotCoding = gene_vectorizer.transform(test_df['Gen
e'])
cv_gene_feature_onehotCoding = gene_vectorizer.transform(cv_df['Gene'])

# don't forget to normalize every feature
train_gene_feature_onehotCoding = normalize(train_gene_feature_onehotCo
ding, axis=0)
test_gene_feature_onehotCoding = normalize(test_gene_feature_onehotCodi
ng, axis=0)
cv_gene_feature_onehotCoding = normalize(cv_gene_feature_onehotCoding,
axis=0)
```

Variation Feature

In [109]:
```python
# 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, "V
ariation", cv_df))
```

In [110]:
```
# one-hot encoding of variation feature.
variation_vectorizer = CountVectorizer(ngram_range=(1,2))
train_variation_feature_onehotCoding = variation_vectorizer.fit_transfo
rm(train_df['Variation'])
test_variation_feature_onehotCoding = variation_vectorizer.transform(te
st_df['Variation'])
cv_variation_feature_onehotCoding = variation_vectorizer.transform(cv_d
f['Variation'])

# don't forget to normalize every feature
train_variation_feature_onehotCoding = normalize(train_variation_featur
e_onehotCoding, axis=0)
test_variation_feature_onehotCoding = normalize(test_variation_feature_
onehotCoding, axis=0)
cv_variation_feature_onehotCoding = normalize(cv_variation_feature_oneh
otCoding, axis=0)
```

Text Feature

In [111]:
```
# building a CountVectorizer with all the words that occured minimum 3
 times in train data
text_vectorizer = CountVectorizer(min_df=3,ngram_range=(1,2))
train_text_feature_onehotCoding = text_vectorizer.fit_transform(train_d
f['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 : 781894

In [112]:
```
#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)

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

In [113]:
```
# 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)
```

Stacking the three types of features

In [114]:
```python
# 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,t
rain_variation_feature_onehotCoding))
test_gene_var_onehotCoding = hstack((test_gene_feature_onehotCoding,tes
t_variation_feature_onehotCoding))
cv_gene_var_onehotCoding = hstack((cv_gene_feature_onehotCoding,cv_vari
ation_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_fea
ture_onehotCoding)).tocsr()
test_y = np.array(list(test_df['Class']))

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


train_gene_var_responseCoding = np.hstack((train_gene_feature_responseC
```

```
oding,train_variation_feature_responseCoding))
test_gene_var_responseCoding = np.hstack((test_gene_feature_responseCod
ing,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, trai
n_text_feature_responseCoding))
test_x_responseCoding = np.hstack((test_gene_var_responseCoding, test_t
ext_feature_responseCoding))
cv_x_responseCoding = np.hstack((cv_gene_var_responseCoding, cv_text_fe
ature_responseCoding))
```

In [115]:
```
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 = ", t
est_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, 78
4184)
(number of data points * number of features) in test data =  (665, 7841
84)
(number of data points * number of features) in cross validation data =
(532, 784184)
```

In [116]:
```
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 = ", t
est_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, 2
```

```
7)
(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)
```

Logistic Regression with Class Balancing

In [117]:
```python
# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_rate='optimal', eta0=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, …])     Fit linear model with S
tochastic 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.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
tml
# ------------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight])     Fit the calibrated model
```

```python
# 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], p
enalty='l2', loss='log', random_state=42)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
```

```python
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 vali
dation 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 l
oss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.5672345718641811
for alpha = 1e-05
Log Loss : 1.5367694719918228
for alpha = 0.0001
Log Loss : 1.5033990032933264
for alpha = 0.001
Log Loss : 1.1887588982549553
for alpha = 0.01
Log Loss : 1.192860952213491
for alpha = 0.1
Log Loss : 1.2469270974758326
for alpha = 1
Log Loss : 1.3132183430619164
for alpha = 10
Log Loss : 1.3764110949915342
for alpha = 100
Log Loss : 1.3908005111225688
```

Cross Validation Error for each alpha
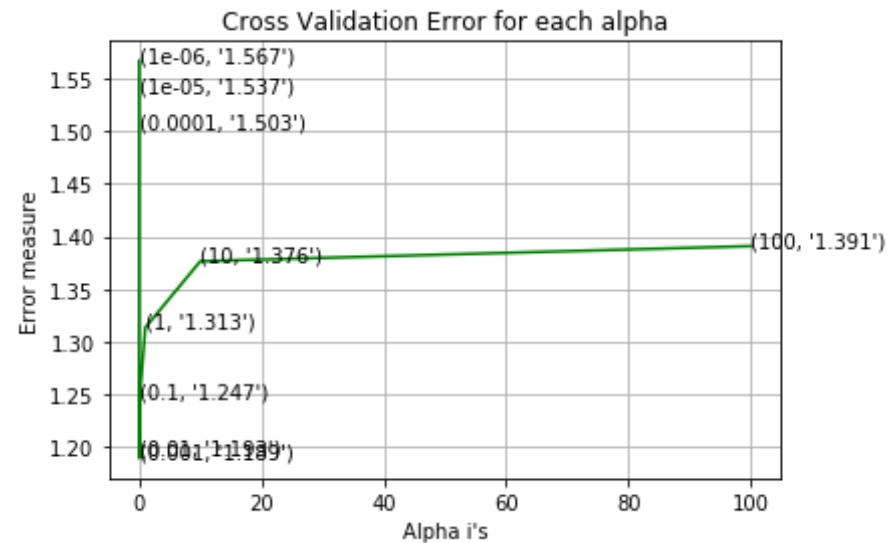
```
For values of best alpha =  0.001 The train log loss is: 0.793864902343
2437
For values of best alpha =  0.001 The cross validation log loss is: 1.1
887588982549553
For values of best alpha =  0.001 The test log loss is: 1.2042815457415
175
```

Testing model with hyperparameter

In [118]:
```python
# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_rate='optimal', eta0=0.0, power_t=0.5,
```

```
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, …])     Fit linear model with S
tochastic 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], p
enalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_o
nehotCoding, cv_y, clf)
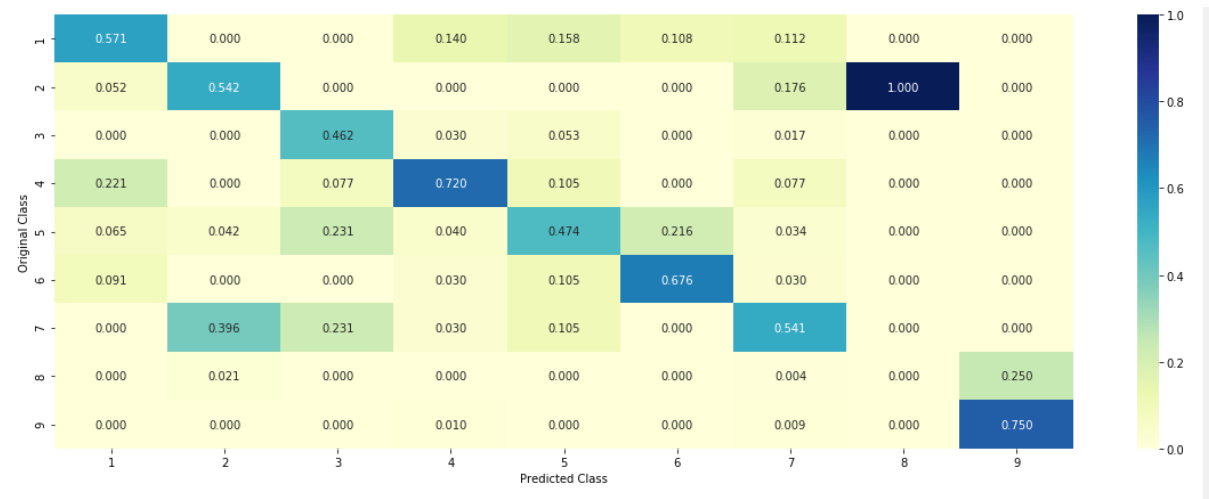```

```
Log loss : 1.1887588982549553
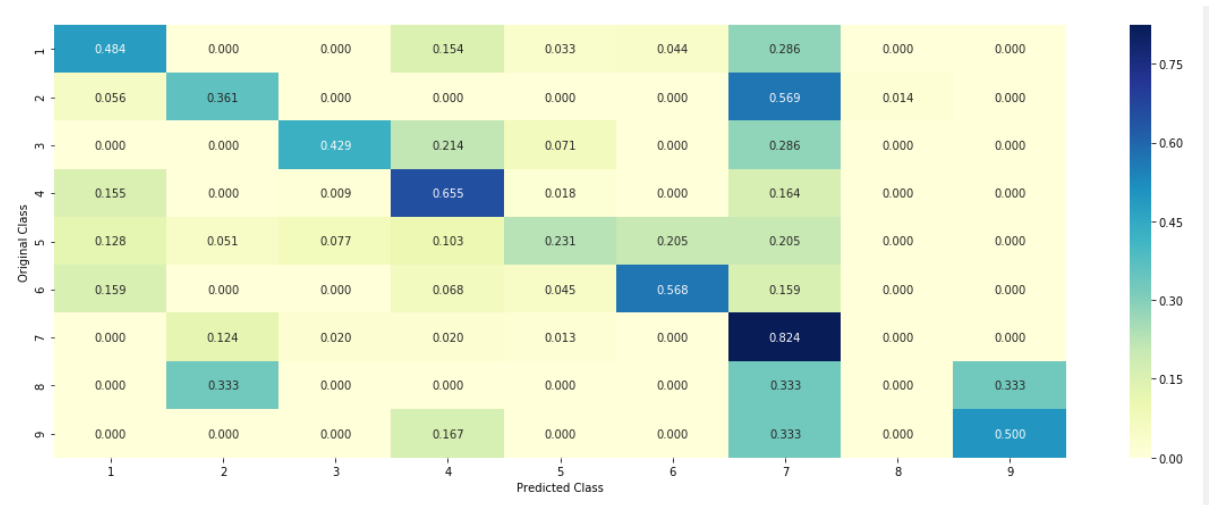Number of mis-classified points : 0.41541353383458646
------------------- Confusion matrix -------------------
```



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

Original Class \ Predicted Class	1	2	3	4	5	6	7	8	9
1	0.571	0.000	0.000	0.140	0.158	0.108	0.112	0.000	0.000
2	0.052	0.542	0.000	0.000	0.000	0.000	0.176	1.000	0.000
3	0.000	0.000	0.462	0.030	0.053	0.000	0.017	0.000	0.000
4	0.221	0.000	0.077	0.720	0.105	0.000	0.077	0.000	0.000
5	0.065	0.042	0.231	0.040	0.474	0.216	0.034	0.000	0.000
6	0.091	0.000	0.000	0.030	0.105	0.676	0.030	0.000	0.000
7	0.000	0.396	0.231	0.030	0.105	0.000	0.541	0.000	0.000
8	0.000	0.021	0.000	0.000	0.000	0.000	0.004	0.000	0.250
9	0.000	0.000	0.000	0.010	0.000	0.000	0.009	0.000	0.750

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

Original Class \ Predicted Class	1	2	3	4	5	6	7	8	9
1	0.484	0.000	0.000	0.154	0.033	0.044	0.286	0.000	0.000
2	0.056	0.361	0.000	0.000	0.000	0.000	0.569	0.014	0.000
3	0.000	0.000	0.429	0.214	0.071	0.000	0.286	0.000	0.000
4	0.155	0.000	0.009	0.655	0.018	0.000	0.164	0.000	0.000
5	0.128	0.051	0.077	0.103	0.231	0.205	0.205	0.000	0.000
6	0.159	0.000	0.000	0.068	0.045	0.568	0.159	0.000	0.000
7	0.000	0.124	0.020	0.020	0.013	0.000	0.824	0.000	0.000
8	0.000	0.333	0.000	0.000	0.000	0.000	0.333	0.000	0.333
9	0.000	0.000	0.000	0.167	0.000	0.000	0.333	0.000	0.500

Observation:

- Key take away from above confusion matrix is still we can't reduced log-loss after using CountVectorizer(unigrams, bi-grams).

Feature Engineering

Try Merge Features and Apply Logistic Regression.

```
In [159]: #merging both gene_variations and text data based on ID
          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+', '_')

          # split the data into test and train by maintaining same distribution o
          f output varaible 'y_true' [stratify=y_true]
          x_train, x_test, y_train, y_test = train_test_split(result, y_true, str
          atify=y_true, test_size=0.2)
          # split the train data into train and cross validation by maintaining s
          ame distribution of output varaible 'y_train' [stratify=y_train]
          x_train, x_cv, y_train, y_cv = train_test_split(x_train, y_train, strat
          ify=y_train, test_size=0.2)
```

```
In [160]: # get_gv_fea_dict: Get Gene varaition Feature Dict
          # get_gv_fea_dict: Get Gene varaition Feature Dict
          def get_gv_fea_dict(alpha, feature, df):
              value_count = x_train[feature].value_counts()
              gv_dict = dict()
              for i, denominator in value_count.items():
                  vec = []
                  for k in range(1,10):
                      cls_cnt = x_train.loc[(x_train['Class']==k) & (x_train[feat
          ure]==i)]
                      vec.append((cls_cnt.shape[0] + alpha*10)/ (denominator + 90
          *alpha))
                  gv_dict[i]=vec
              return gv_dict

          # Get Gene variation feature
          def get_gv_feature(alpha, feature, df):
```

```
        gv_dict = get_gv_fea_dict(alpha, feature, df)
        value_count = x_train[feature].value_counts()
        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])
        return gv_fea
```

Gene Feature With Tfidf Vectorizer

In [161]:
```python
#response-coding of the Gene feature
# alpha is used for laplace smoothing
alpha = 1
# train gene feature
train_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_train))
# test gene feature
test_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_test))
# cross validation gene feature
cv_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_cv))
```

In [162]:
```python
# one-hot encoding of 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'])
```

Variation feature

```
In [163]:  # alpha is used for laplace smoothing
           alpha = 1
           # train variation feature
           train_variation_feature_responseCoding = np.array(get_gv_feature(alpha,
            "Variation", x_train))
           # test variation feature
           test_variation_feature_responseCoding = np.array(get_gv_feature(alpha,
           "Variation", x_test))
           # cross validation variation feature
           cv_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "V
           ariation", x_cv))
```

```
In [164]:  # one-hot encoding of variation feature.
           variation_vectorizer = TfidfVectorizer()
           train_variation_feature_onehotCoding = variation_vectorizer.fit_transfo
           rm(x_train['Variation'])
           test_variation_feature_onehotCoding = variation_vectorizer.transform(x_
           test['Variation'])
           cv_variation_feature_onehotCoding = variation_vectorizer.transform(x_cv
           ['Variation'])
```

Text Feature

```
In [165]:  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
```

```python
        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 )/(t
otal_dict.get(word,0)+90)))
            text_feature_responseCoding[row_index][i] = math.exp(sum_pr
ob/len(row['TEXT'].split()))
            row_index += 1
    return text_feature_responseCoding
```

In [167]:
```python
# building a TfidfVectorizer with all the words that occured minimum 3
 times in train data
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 num
ber of times it occured
text_fea_dict = dict(zip(list(train_text_features),train_text_fea_count
s))


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

Total number of unique words in train data : 124496

In [168]:
```python
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))
```

```python
        # 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 [169]:
```python
#response coding of text features
train_text_feature_responseCoding  = get_text_responsecoding(x_train)
test_text_feature_responseCoding  = get_text_responsecoding(x_test)
cv_text_feature_responseCoding  = get_text_responsecoding(x_cv)

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

In [170]:
```python
test_text_feature_onehotCoding = text_vectorizer.transform(x_test['TEX
T'])
cv_text_feature_onehotCoding = text_vectorizer.transform(x_cv['TEXT'])
```

Features after Feature Engineering

In [171]:
```python
#Creating Gene_variation into SinglelIst
```

```
gene_variation = []

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

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

In [172]:
```
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'])
```

Stacking the three types of features

In [173]:
```
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))

# Adding the train_text feature
train_x_onehotCoding = hstack((train_gene_var_onehotCoding, train_text))
train_x_onehotCoding = hstack((train_x_onehotCoding, train_text_feature_onehotCoding)).tocsr()
train_y = np.array(list(x_train['Class']))

# Adding the test_text feature
test_x_onehotCoding = hstack((test_gene_var_onehotCoding, test_text))
test_x_onehotCoding = hstack((test_x_onehotCoding, test_text_feature_onehotCoding)).tocsr()
test_y = np.array(list(x_test['Class']))
```

```python
# Adding the cv_text feature
cv_x_onehotCoding = hstack((cv_gene_var_onehotCoding, cv_text))
cv_x_onehotCoding = hstack((cv_x_onehotCoding, cv_text_feature_onehotCo
ding)).tocsr()
cv_y = np.array(list(x_cv['Class']))


train_gene_var_responseCoding = np.hstack((train_gene_feature_responseC
oding,train_variation_feature_responseCoding))
test_gene_var_responseCoding = np.hstack((test_gene_feature_responseCod
ing,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, trai
n_text_feature_responseCoding))
test_x_responseCoding = np.hstack((test_gene_var_responseCoding, test_t
ext_feature_responseCoding))
cv_x_responseCoding = np.hstack((cv_gene_var_responseCoding, cv_text_fe
ature_responseCoding))
```

In [174]:
```python
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 = ", t
est_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, 12
7682)
(number of data points * number of features) in test data =  (665, 1276
82)
(number of data points * number of features) in cross validation data =
(532, 127682)
```

In [175]:
```python
print("Respons 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 = ", t
est_x_responseCoding.shape)
print("(number of data points * number of features) in cross validation
 data =", cv_x_responseCoding.shape)
```

```
Respons encoding features :
(number of data points * number of features) in train data =  (2124, 2
7)
(number of data points * number of features) in test data =  (665, 27)
(number of data points * number of features) in cross validation data =
(532, 27)
```

In [176]:
```python
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()
```

```python
best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], p
enalty='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, 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 : 1.0574841176687122
for alpha = 1e-05
Log Loss : 1.0085005460427279
for alpha = 0.0001
Log Loss : 0.9556567965185517
for alpha = 0.001
Log Loss : 1.0227703660353653
for alpha = 0.01
Log Loss : 1.2411521498523874
for alpha = 0.1
Log Loss : 1.5932755665423082
for alpha = 1
Log Loss : 1.7067161061483258
```

```
for alpha = 10
Log Loss : 1.718617109627257
for alpha = 100
Log Loss : 1.7198828741214127
```

Cross Validation Error for each alpha
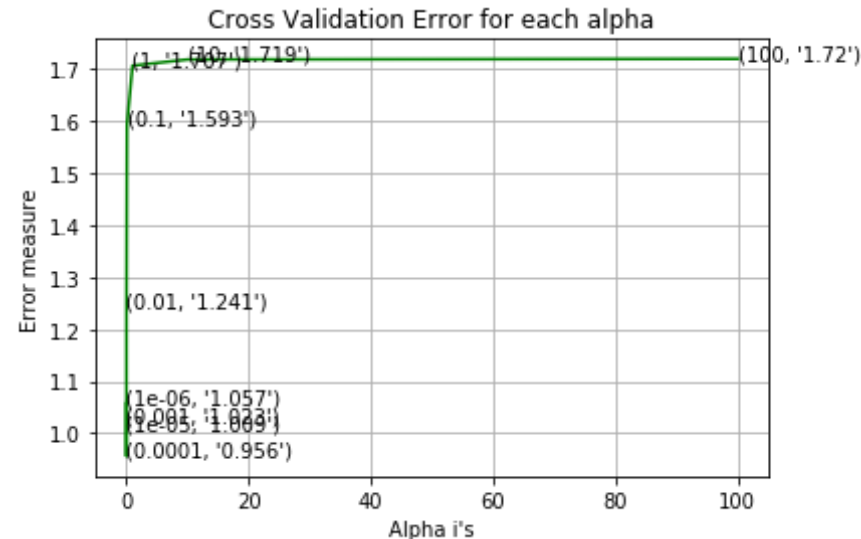

```
For values of best alpha =  0.0001 The train log loss is: 0.42677493359
188695
For values of best alpha =  0.0001 The cross validation log loss is: 0.
9556567965185517
For values of best alpha =  0.0001 The test log loss is: 0.997723710100
6537
```

In [177]:
```python
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], p
enalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_o
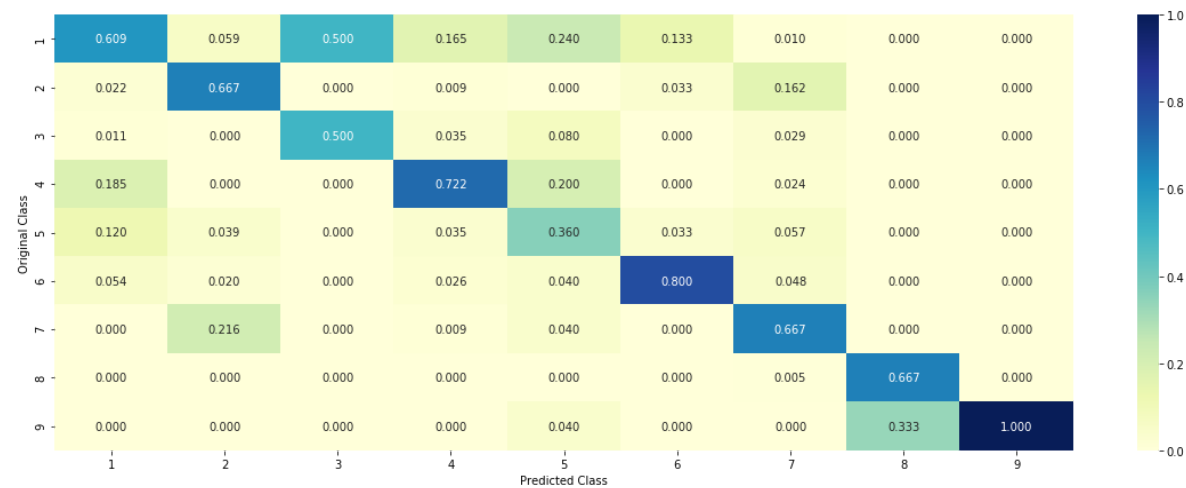nehotCoding, cv_y, clf)
```

```
Log loss : 0.9556567965185517
Number of mis-classified points : 0.33646616541353386
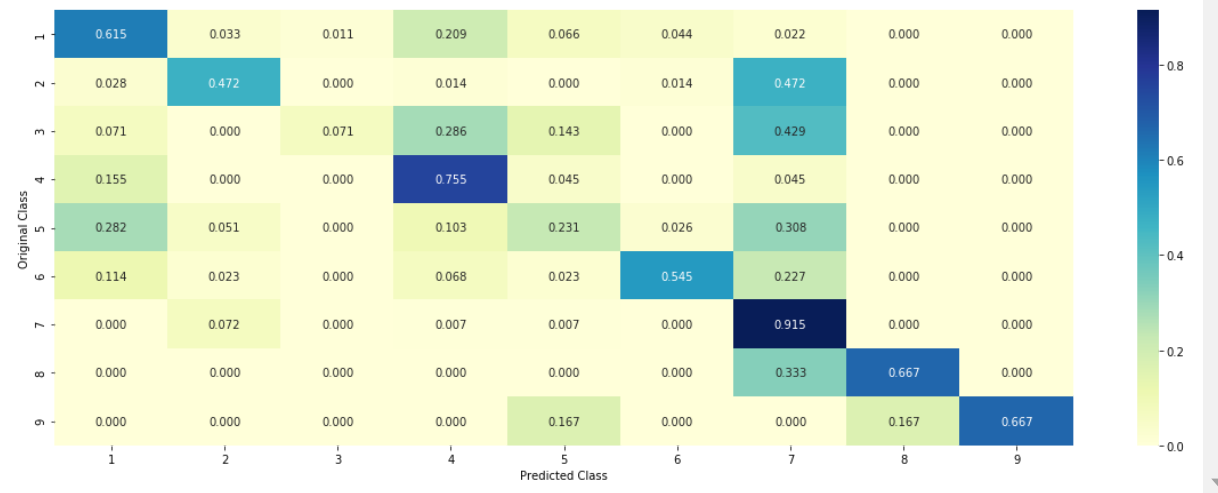------------------- Confusion matrix --------------------
```
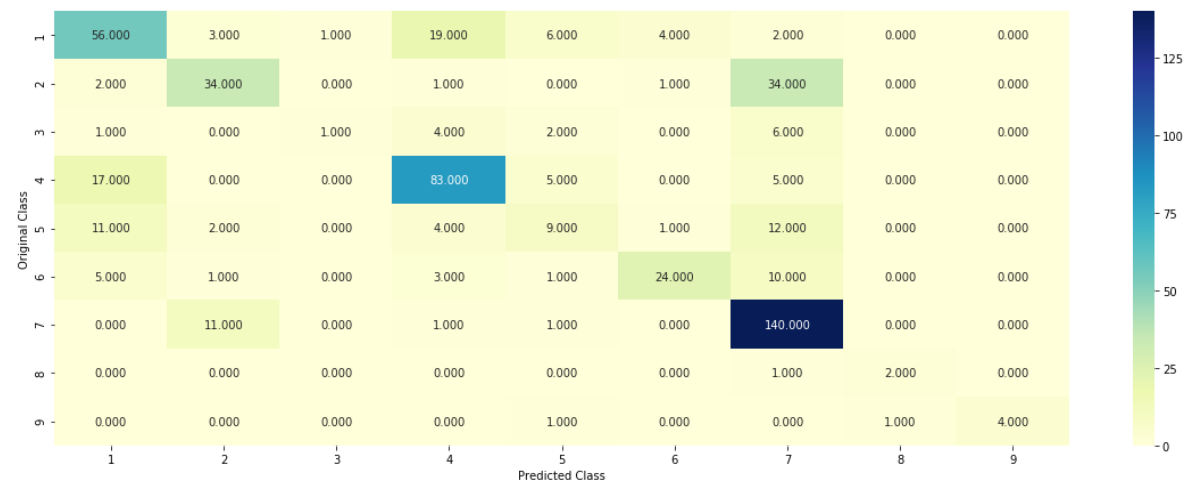
-------------------- Precision matrix (Columm Sum=1) ---------------
----

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

Observation:

- Keytake away from the above model after feature engineering is we reduced log-loss < 1, prior to the model of logistic regression with unigrams, bi-grams.
- Applied Tfidf Vectorizer instead of CountVectroizer by merging features of Gene and Variation.
- Observed log-loss of > 1 when applied with CountVectorizer.

Conclusions

- Data Contains two files which is Trainig Variation, Training Text downloaded from kaggle.
- Training Variation contains Different gene Mutations.
- Training text contains Clinical evidences.
- Random Split od Data to 64% Train, 16%CV, 20% Test data sets.
- Applied tf-idf features for all the models above.
- Instead of using all the words , used only top 1000 words based of tf-idf values.

- Applied logistic regression with class='balancing' performed better compared to without class 'balancing'.
- In order to improve the log-loss implemented CountVectorizer with uni-grams and bi-grams, observed log-loss < 1.
- By implementing feature engineering technique of merging both gene and variation into single list.
- Using Stacked Classifiers also observed high log-loss.
- After implemented feature engineering observed log-loss < 1 compared to other models.