4.ML_models

April 18, 2019

```
In [1]: #System:
        import os

        #Data structures for in memory:
        import csv
        import pandas as pd
        import numpy as np
        from scipy.sparse import hstack
        import math

        #Database store:
        from sqlalchemy import create_engine # database connection
        import sqlite3

        #Date and time:
        import time
        import datetime as dt

        #Plotting:
        import matplotlib.pyplot as plt
        import seaborn as sns
        from sklearn.manifold import TSNE

        #Data transformations:
        import re
        from nltk.corpus import stopwords
        from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
        from sklearn.decomposition import TruncatedSVD
        from sklearn.preprocessing import normalize

        #Parameter tuning:
        from sklearn.cross_validation import StratifiedKFold
        from collections import Counter, defaultdict
        from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score, Ra

        #Models:
        from sklearn.neighbors import KNeighborsClassifier
        from sklearn.naive_bayes import MultinomialNB
        from sklearn.naive_bayes import GaussianNB
        from sklearn.linear_model import LogisticRegression, SGDClassifier
        from sklearn.svm import LinearSVC
        from sklearn.multiclass import OneVsRestClassifier
        from sklearn.calibration import CalibratedClassifierCV
        from mlxtend.classifier import StackingClassifier
        from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
        import xgboost as xgb

        #Model evaluation metrics:
        from sklearn.metrics import precision_recall_curve, auc, roc_curve, confusion_matrix, r

        #Switch off warnings:
        import warnings
        warnings.filterwarnings("ignore")

        from sklearn.externals import joblib
        from prettytable import PrettyTable
```

```
c:\users\byron\applications\pythonmaster\lib\site-packages\sklearn\cross_validation.py:41: Depr
  "This module will be removed in 0.20.", DeprecationWarning)
c:\users\byron\applications\pythonmaster\lib\site-packages\sklearn\ensemble\weight_boosting.py
  from numpy.core.umath_tests import inner1d
```

4. Machine Learning Models

4.1 Reading data from file and storing into sql table

```
In [2]: #Creating db file from csv
        if not os.path.isfile('train.db'):
            disk_engine = create_engine('sqlite:///train.db')
            start = dt.datetime.now()
            chunksize = 50000
            j = 0
            index_start = 1
            for df in pd.read_csv('final_features.csv', names=['Unnamed: 0','id','is_duplicate
                df.index += index_start
                j+=1
                print('{} rows'.format(j*chunksize))
                df.to_sql('data', disk_engine, if_exists='append')
                index_start = df.index[-1] + 1
```

```
In [3]: #http://www.sqlitetutorial.net/sqlite-python/create-tables/
        def create_connection(db_file):
            """ create a database connection to the SQLite database
                specified by db_file
            :param db_file: database file
            :return: Connection object or None
            """
            try:
                conn = sqlite3.connect(db_file)
                return conn
            except Error as e:
                print(e)

            return None


        def checkTableExists(dbcon):
            cursr = dbcon.cursor()
            str = "select name from sqlite_master where type='table'"
            table_names = cursr.execute(str)
            print("Tables in the databse:")
            tables =table_names.fetchall()
            print(tables[0][0])
            return(len(tables))
```

```
In [4]: read_db = 'train.db'
        conn_r = create_connection(read_db)
        checkTableExists(conn_r)
        conn_r.close()
```

```
Tables in the databse:
data
```

```
In [5]: # try to sample data according to the computing power you have
        if os.path.isfile(read_db):
            conn_r = create_connection(read_db)
            if conn_r is not None:
                # for selecting random points
                data = pd.read_sql_query("SELECT * From data ORDER BY RANDOM() LIMIT 100001;",
                conn_r.close()
```

```
In [6]: # remove the first row
        data.drop(data.index[0], inplace=True)
        y_true = data['is_duplicate']
        data.drop(['Unnamed: 0', 'id','index','is_duplicate'], axis=1, inplace=True)
```

```
In [7]: data.head()
```

```
Out[7]:              cwc_min             cwc_max             csc_min             csc_max  \
        1   0.66664444518516    0.66664444518516   0.999975000624984   0.999975000624984
        2  0.499987500312492   0.285710204139941    0.66664444518516   0.399992000159997
        3   0.66664444518516   0.399992000159997   0.749981250468738   0.333329629670781
        4  0.749981250468738   0.499991666805553   0.749981250468738   0.749981250468738
        5  0.499987500312492       0.19999800002   0.999980000399992   0.454541322351615
```

```
             ctc_min            ctc_max last_word_eq first_word_eq  \
1  0.857130612419823  0.857130612419823          0.0           1.0
2  0.571420408279882  0.307689940846609          0.0           1.0
3  0.714275510349852  0.312498046887207          1.0           0.0
4  0.749990625117186  0.599994000059999          1.0           1.0
5   0.77776913589849  0.259258299043337          1.0           0.0

  abs_len_diff mean_len        ...                      290_y  \
1          0.0      7.0        ...           -13.684094414115
2          6.0     10.0        ...          -2.89921551942825
3          9.0     11.5        ...          -32.1513776183128
4          2.0      9.0        ...           1.11949726939201
5         18.0     18.0        ...          -33.0604563355446

               291_y              292_y              293_y              294_y  \
1   5.18617536127567   3.69274061173201  -1.06995718181133  -3.38792563974857
2   2.25030846148729  -4.55699910968542   3.22107343003154   4.69975774548948
3  -6.48564624227583  -10.6156985536218  -8.61111462116241  -2.86723747849464
4   -5.4107170291245  -3.75332200527191  -4.40629441710189  0.106489285826683
5   30.6366586647928   10.6500630229712  -10.6027148663998   8.85900411009789

               295_y              296_y               297_y  \
1   1.86694558337331  -1.14174094796181   -3.97690352797508
2   1.60277144983411  -4.12365251034498   -7.01728013157845
3   15.2251101061702  -9.93901033699512   -3.61792010068893
4     3.812221378088  -5.15402545034885  -0.789197444915772
5  -24.4780361577868  -4.69883567839861   -13.0958931222558

               298_y              299_y
1  -2.48735983669758   4.12184119224548
2  0.270279049873352   8.08513672836125
3  -3.56169393658638  -3.51276577170938
4   1.25951708108187   3.11145649943501
5   26.8136106580496  -25.8301425874233

[5 rows x 626 columns]
```

```
In [8]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 100000 entries, 1 to 100000
Columns: 626 entries, cwc_min to 299_y
dtypes: object(626)
memory usage: 478.4+ MB
```

```
In [9]: col = data.columns.values
```

4.2 Converting strings to numerics

```
In [10]: data = pd.DataFrame(np.array(data).astype(float),columns = col)
```

```
In [11]: # https://stackoverflow.com/questions/7368789/convert-all-strings-in-a-list-to-int
         y_true = list(map(int, y_true.values))
```

4.3 Random train test split(70:30)

```
In [12]: X_train,X_test, y_train, y_test = train_test_split(data, y_true, stratify=y_true, tes
```

```
In [13]: print("Number of data points in train data :",X_train.shape)
         print("Number of data points in test data :",X_test.shape)
```

```
Number of data points in train data : (70000, 626)
Number of data points in test data : (30000, 626)
```

```
In [14]: print("-"*10, "Distribution of output variable in train data", "-"*10)
         train_distr = Counter(y_train)
         train_len = len(y_train)
         print("Class 0: ",int(train_distr[0])/train_len,"Class 1: ", int(train_distr[1])/trai
         print("-"*10, "Distribution of output variable in test data", "-"*10)
         test_distr = Counter(y_test)
         test_len = len(y_test)
         print("Class 0: ",int(test_distr[1])/test_len, "Class 1: ",int(test_distr[1])/test_le
```

```
---------- Distribution of output variable in train data ----------
Class 0:  0.6283285714285715 Class 1:  0.3716714285714286
---------- Distribution of output variable in test data ----------
Class 0:  0.37166666666666665 Class 1:  0.37166666666666665
```

```
In [15]: # 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 pred

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

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

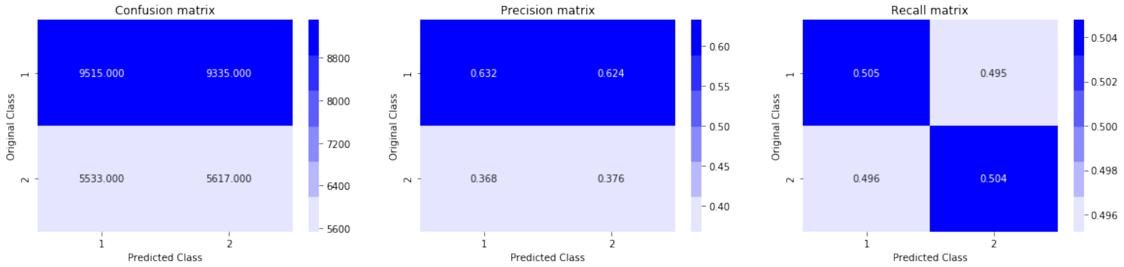
```
    # ((C.T)/(C.sum(axis=1))).T = [[1/3, 2/3]
    #                            [3/7, 4/7]]
    # sum of row elements = 1

    B =(C/C.sum(axis=0))
    #divid each element of the confusion matrix with the sum of elements in that row
    # C = [[1, 2],
    #     [3, 4]]
    # C.sum(axis = 0)  axis=0 corresonds to columns and axis=1 corresponds to rows in
    # C.sum(axix =0) = [[4, 6]]
    # (C/C.sum(axis=0)) = [[1/4, 2/6],
    #                      [3/4, 4/6]]
    plt.figure(figsize=(20,4))

    labels = [1,2]
    # representing A in heatmap format
    cmap=sns.light_palette("blue")
    plt.subplot(1, 3, 1)
    sns.heatmap(C, annot=True, cmap=cmap, fmt=".3f", xticklabels=labels, yticklabels=l
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.title("Confusion matrix")

    plt.subplot(1, 3, 2)
    sns.heatmap(B, annot=True, cmap=cmap, fmt=".3f", xticklabels=labels, yticklabels=l
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.title("Precision matrix")

    plt.subplot(1, 3, 3)
    # representing B in heatmap format
    sns.heatmap(A, annot=True, cmap=cmap, fmt=".3f", xticklabels=labels, yticklabels=l
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.title("Recall matrix")

    plt.show()
```

4.4 Building a random model (Finding worst-case log-loss)

In [16]:
```
# 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
# we create a output array that has exactly same size as the CV data
predicted_y = np.zeros((test_len,2))
for i in range(test_len):
    rand_probs = np.random.rand(1,2)
```

```
        predicted_y[i] = ((rand_probs/sum(sum(rand_probs)))[0])
    print("Log loss on Test Data using Random Model",log_loss(y_test, predicted_y, eps=1e

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

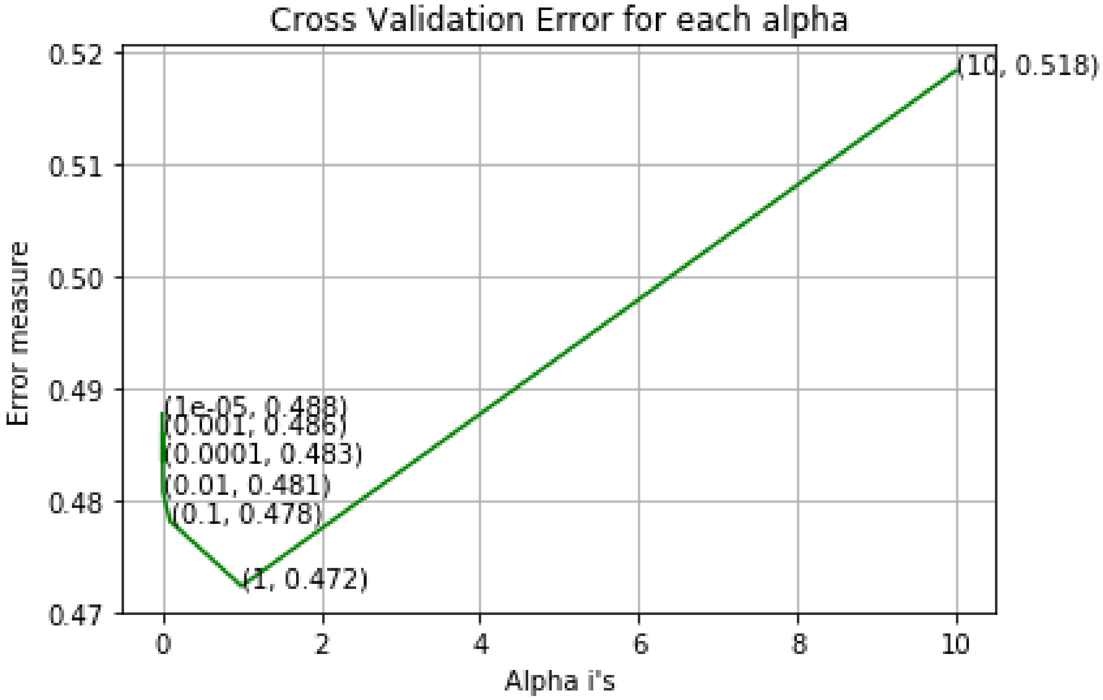
```
Log loss on Test Data using Random Model 0.885340305050038
```



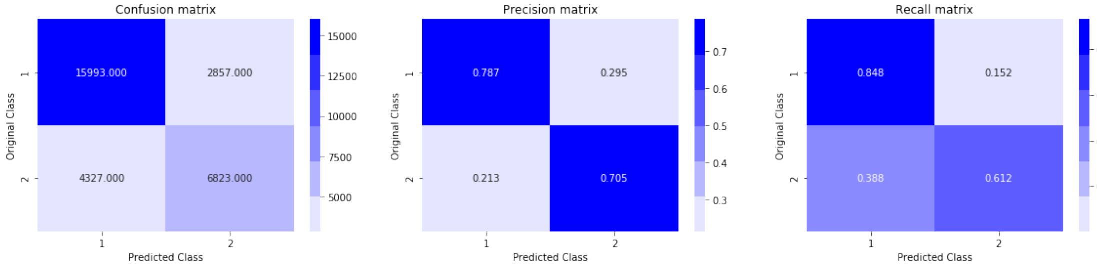
4.4 Logistic Regression with hyperparameter tuning

```
In [17]: alpha = [10 ** x for x in range(-5, 2)] # hyperparam for SGD classifier.

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

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

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


         log_error_array=[]
         for i in alpha:
             clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
             clf.fit(X_train, y_train)
             sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
             sig_clf.fit(X_train, y_train)
             predict_y = sig_clf.predict_proba(X_test)
             log_error_array.append(log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15)
             print('For values of alpha = ', i, "The log loss is:",log_loss(y_test, predict_y,
         fig, ax = plt.subplots()
         ax.plot(alpha, log_error_array,c='g')
         for i, txt in enumerate(np.round(log_error_array,3)):
             ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],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(log_error_array)
         clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=4
         clf.fit(X_train, y_train)
         sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
         sig_clf.fit(X_train, y_train)

         predict_y = sig_clf.predict_proba(X_train)
         print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_
         predict_y = sig_clf.predict_proba(X_test)
         print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_l
         predicted_y =np.argmax(predict_y,axis=1)
         print("Total number of data points :", len(predicted_y))
         plot_confusion_matrix(y_test, predicted_y)
```

```
For values of alpha =  1e-05 The log loss is: 0.48771178018319705
For values of alpha =  0.0001 The log loss is: 0.4834804577479922
For values of alpha =  0.001 The log loss is: 0.4860223659213464
For values of alpha =  0.01 The log loss is: 0.48074674780307397
For values of alpha =  0.1 The log loss is: 0.4781428484211463
For values of alpha =  1 The log loss is: 0.472357294213177
For values of alpha =  10 The log loss is: 0.5182490126857469
```



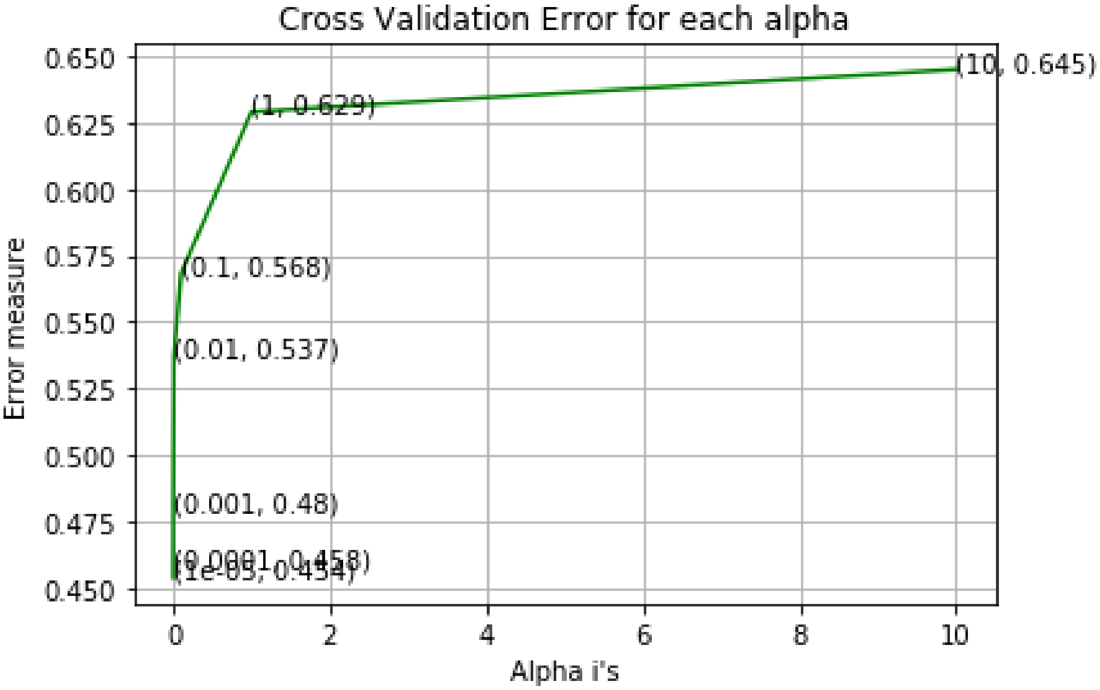
```
For values of best alpha =  1 The train log loss is: 0.46529034161165644
For values of best alpha =  1 The test log loss is: 0.472357294213177
Total number of data points : 30000
```



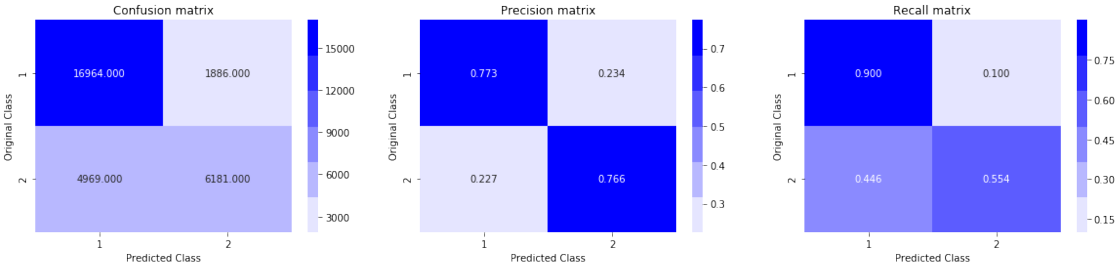
4.5 Linear SVM with hyperparameter tuning

```
In [18]: alpha = [10 ** x for x in range(-5, 2)] # hyperparam for SGD classifier.

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

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

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


         log_error_array=[]
         for i in alpha:
             clf = SGDClassifier(alpha=i, penalty='l1', loss='hinge', random_state=42)
             clf.fit(X_train, y_train)
             sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
             sig_clf.fit(X_train, y_train)
             predict_y = sig_clf.predict_proba(X_test)
             log_error_array.append(log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15
             print('For values of alpha = ', i, "The log loss is:",log_loss(y_test, predict_y,

         fig, ax = plt.subplots()
         ax.plot(alpha, log_error_array,c='g')
         for i, txt in enumerate(np.round(log_error_array,3)):
             ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],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(log_error_array)
         clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l1', loss='hinge', random_state
         clf.fit(X_train, y_train)
         sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
         sig_clf.fit(X_train, y_train)

         predict_y = sig_clf.predict_proba(X_train)
         print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_
         predict_y = sig_clf.predict_proba(X_test)
         print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_l
         predicted_y =np.argmax(predict_y,axis=1)
         print("Total number of data points :", len(predicted_y))
         plot_confusion_matrix(y_test, predicted_y)
```

```
For values of alpha =  1e-05 The log loss is: 0.4539013938818281
For values of alpha =  0.0001 The log loss is: 0.4582849827404135
For values of alpha =  0.001 The log loss is: 0.47961404331751945
For values of alpha =  0.01 The log loss is: 0.5372885487844338
For values of alpha =  0.1 The log loss is: 0.5682885200407735
For values of alpha =  1 The log loss is: 0.6289676976712478
For values of alpha =  10 The log loss is: 0.6448495605353364
```



```
For values of best alpha =  1e-05 The train log loss is: 0.4485062220124629
For values of best alpha =  1e-05 The test log loss is: 0.4539013938818281
Total number of data points : 30000
```



4.6 XGBoost

```
In [19]: params = {}
         params['objective'] = 'binary:logistic'
         params['eval_metric'] = 'logloss'
         params['eta'] = 0.02
         params['max_depth'] = 4

         d_train = xgb.DMatrix(X_train, label=y_train)
         d_test = xgb.DMatrix(X_test, label=y_test)

         watchlist = [(d_train, 'train'), (d_test, 'valid')]

         bst = xgb.train(params, d_train, 400, watchlist, early_stopping_rounds=20, verbose_eva

         xgdmat = xgb.DMatrix(X_train,y_train)
         predict_y = bst.predict(d_test)
         print("The test log loss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e
```

```
[10:31:16] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[0]	train-logloss:0.684597	valid-logloss:0.684678
Multiple eval metrics have been passed: 'valid-logloss' will be used for early stopping.

Will train until valid-logloss hasn't improved in 20 rounds.
[10:31:18] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:19] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:20] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:21] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:22] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:24] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:25] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:26] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:27] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10]	train-logloss:0.614357	valid-logloss:0.615202
[10:31:28] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:29] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:30] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:31] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:32] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:33] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:34] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:35] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:36] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:37] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[20]	train-logloss:0.562945	valid-logloss:0.564105
[10:31:38] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:40] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:41] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:31:42] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
```

```
[10:31:43] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:44] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:45] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:46] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:47] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:48] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[30]       train-logloss:0.524599      valid-logloss:0.525964
[10:31:49] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:50] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:51] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:52] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:53] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:54] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:55] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:56] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:57] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:31:58] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[40]       train-logloss:0.495618      valid-logloss:0.497216
[10:32:00] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:01] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:02] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:03] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:04] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:05] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:06] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:07] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:08] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:09] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[50]       train-logloss:0.472726      valid-logloss:0.474311
[10:32:10] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:11] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:13] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:14] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:15] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:16] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:17] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:18] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:19] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:20] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[60]       train-logloss:0.454394      valid-logloss:0.456073
[10:32:21] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:22] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:24] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:25] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:26] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:27] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:28] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:32:29] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
```

```
[10:32:30] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:32] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[70]        train-logloss:0.439625        valid-logloss:0.441484
[10:32:33] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:34] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:35] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:36] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:37] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:39] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:40] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:41] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:42] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:43] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[80]        train-logloss:0.42751        valid-logloss:0.429424
[10:32:44] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:45] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:46] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:48] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:49] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:50] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:51] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:52] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:53] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:55] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[90]        train-logloss:0.417708        valid-logloss:0.419707
[10:32:56] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:57] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:58] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:32:59] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:00] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:01] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:03] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:04] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:05] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:06] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[100]        train-logloss:0.409574        valid-logloss:0.411677
[10:33:07] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:08] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:10] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:11] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:12] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:13] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:14] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:15] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:16] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:18] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[110]        train-logloss:0.402842        valid-logloss:0.405011
[10:33:19] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
```

```
[10:33:20] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:21] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:22] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:25] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:26] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:27] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:28] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:29] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[120]        train-logloss:0.39732        valid-logloss:0.399618
[10:33:30] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:32] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:33] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:34] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:35] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:36] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:37] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:39] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:40] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:41] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[130]        train-logloss:0.392631        valid-logloss:0.395056
[10:33:42] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:43] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:45] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:46] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:47] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:48] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:49] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:50] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:52] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:53] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[140]        train-logloss:0.388547        valid-logloss:0.391233
[10:33:54] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:55] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:56] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:58] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:33:59] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:00] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:01] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:02] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:04] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:05] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[150]        train-logloss:0.38489        valid-logloss:0.387811
[10:34:06] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:07] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:08] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:10] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:11] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
```

```
[10:34:12] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:13] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:14] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:16] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:17] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[160]        train-logloss:0.38175        valid-logloss:0.384898
[10:34:18] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:19] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:20] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:21] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:24] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:25] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:26] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:27] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:29] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[170]        train-logloss:0.378976        valid-logloss:0.382312
[10:34:30] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:31] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:32] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:34] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:35] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:36] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:37] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:38] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:40] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:41] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[180]        train-logloss:0.376378        valid-logloss:0.379909
[10:34:42] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:44] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:45] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:46] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:48] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:49] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:50] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:51] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:52] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:54] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[190]        train-logloss:0.374156        valid-logloss:0.377896
[10:34:55] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:56] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:57] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:34:58] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:00] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:01] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:02] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:03] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:05] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
```

```
[10:35:06] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[200]        train-logloss:0.371804         valid-logloss:0.375778
[10:35:07] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:08] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:10] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:11] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:12] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:13] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:14] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:16] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:17] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:18] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[210]        train-logloss:0.369697         valid-logloss:0.373924
[10:35:19] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:20] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:22] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:24] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:25] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:27] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:28] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:29] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:30] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[220]        train-logloss:0.367601         valid-logloss:0.372074
[10:35:32] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:33] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:34] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:35] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:36] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:38] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:39] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:40] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:41] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:43] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[230]        train-logloss:0.365504         valid-logloss:0.370202
[10:35:44] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:45] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:46] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:48] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:49] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:50] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:51] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:53] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:54] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:55] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[240]        train-logloss:0.363688         valid-logloss:0.368722
[10:35:56] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:35:58] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
```

```
[10:35:59] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:00] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:01] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:02] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:04] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:05] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:06] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:07] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[250]        train-logloss:0.361917        valid-logloss:0.36728
[10:36:09] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:10] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:11] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:12] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:14] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:15] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:16] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:17] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:19] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:20] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[260]        train-logloss:0.360215        valid-logloss:0.365872
[10:36:21] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:22] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:24] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:25] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:26] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:27] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:28] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:30] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:31] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:32] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[270]        train-logloss:0.358696        valid-logloss:0.364659
[10:36:33] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:35] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:36] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:37] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:38] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:40] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:41] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:42] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:43] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:44] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[280]        train-logloss:0.357394        valid-logloss:0.363706
[10:36:46] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:47] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:48] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:49] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:51] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
[10:36:52] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning en
```

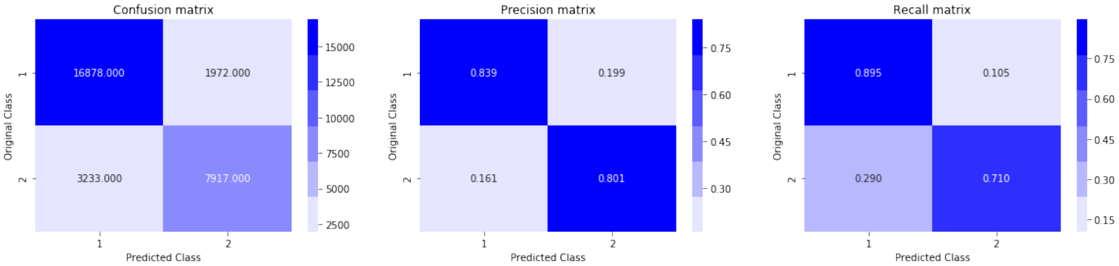
```
[10:36:53] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:36:54] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:36:56] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:36:57] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[290]        train-logloss:0.355979        valid-logloss:0.362664
[10:36:58] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:36:59] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:00] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:02] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:03] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:04] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:05] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:07] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:08] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:09] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[300]        train-logloss:0.354522        valid-logloss:0.361537
[10:37:10] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:12] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:13] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:14] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:15] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:17] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:18] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:19] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:20] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:22] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[310]        train-logloss:0.35306        valid-logloss:0.360463
[10:37:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:24] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:25] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:26] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:28] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:29] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:30] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:32] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:33] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:34] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[320]        train-logloss:0.351827        valid-logloss:0.359595
[10:37:35] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:36] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:38] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:39] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:40] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:41] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:43] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:44] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:45] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
[10:37:46] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning e
```

```
[330]        train-logloss:0.350569       valid-logloss:0.358616
[10:37:48] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:37:49] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:37:50] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:37:51] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:37:52] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:37:54] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:37:55] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:37:56] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:37:57] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:37:59] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[340]        train-logloss:0.349352       valid-logloss:0.357729
[10:38:00] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:01] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:02] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:03] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:05] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:06] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:07] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:08] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:10] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:11] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[350]        train-logloss:0.348173       valid-logloss:0.356881
[10:38:12] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:13] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:15] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:16] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:17] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:18] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:20] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:21] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:22] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:23] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[360]        train-logloss:0.347052       valid-logloss:0.356105
[10:38:25] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:26] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:27] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:28] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:29] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:31] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:32] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:33] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:35] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:36] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[370]        train-logloss:0.345888       valid-logloss:0.35527
[10:38:37] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:38] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:39] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
```

```
[10:38:41] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:42] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:43] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:44] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:46] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:47] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:48] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[380]	train-logloss:0.344861	valid-logloss:0.354599
[10:38:49] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:51] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:52] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:53] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:54] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:56] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:57] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:58] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:38:59] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:39:01] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[390]	train-logloss:0.343845	valid-logloss:0.353868
[10:39:02] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:39:03] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:39:04] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:39:06] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:39:07] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:39:08] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:39:09] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:39:10] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[10:39:12] C:\Users\Administrator\Desktop\xgboost\src\tree\updater_prune.cc:74: tree pruning er
[399]	train-logloss:0.342943	valid-logloss:0.353267
The test log loss is: 0.35326659471739547
```

```
In [20]: predicted_y =np.array(predict_y>0.5,dtype=int)
         print("Total number of data points :", len(predicted_y))
         plot_confusion_matrix(y_test, predicted_y)
```

```
Total number of data points : 30000
```



5. Assignments

1. Try out models (Logistic regression, Linear-SVM) with simple TF-IDF vectors instead of TF_IDF weighted word2Vec.
2. Hyperparameter tune XgBoost using RandomSearch to reduce the log-loss.

```
In [21]: del X_train, X_test, y_train, y_test, data, y_true
```

```
In [22]: #Read in the training data
         df = pd.read_csv("train.csv")
         df['question1'] = df['question1'].apply(lambda x: str(x))
         df['question2'] = df['question2'].apply(lambda x: str(x))
```

```
In [23]: df = df.sample(n=100000)
```

```
In [24]: df.head(5)
```

```
Out[24]:             id    qid1    qid2  \
         52066    52066   92247   92248
         348520  348520   85073  477085
         94223    94223  157392  157393
         235267  235267  345946  345947
         278508  278508  397794  397795

                                                  question1  \
         52066   Why allow refugees in Europe when most of th...
         348520                  How is CAT percentile calculated?
         94223   What does thanoo, vanno and kooi means in mala...
         235267  Which planet in our solar system is the most h...
         278508      What is a good free C compiler for Windows 7?

                                                  question2  is_duplicate
         52066            Why should we accept refugees in Europe?             0
         348520  How would one explain the percentile system in...             0
         94223             What does the Malayalam word AYYO mean?             0
         235267  To which planet in our Solar System would you ...             0
         278508      Where can I download a free Turbo C compiler?             0
```

```
In [25]: data = {'id':df['id'], 'text':df['question1'] + ' ' + df['question2'], 'is_duplicate'
         df = pd.DataFrame(data=data,index=data['id'])
         df.drop(labels=['id'],axis=1,inplace=True)
         df.head()
```

```
Out[25]:                                                       text  is_duplicate
         id
         52066   Why allow refugees in Europe when most of th...            0
         348520  How is CAT percentile calculated? How would on...             0
         94223   What does thanoo, vanno and kooi means in mala...             0
         235267  Which planet in our solar system is the most h...             0
         278508  What is a good free C compiler for Windows 7? ...             0
```

```
In [26]: del data
```

```
In [27]: X_train, X_test = train_test_split(df, stratify=df['is_duplicate'], test_size=0.3)
```

```
In [28]: X_train.shape
```

```
Out[28]: (70000, 2)
```

```
In [29]: Y_train = X_train['is_duplicate']
         X_train.drop(labels=['is_duplicate'], axis=1, inplace=True)
```

```
In [30]: X_train.shape
```

```
Out[30]: (70000, 1)
```

```
In [31]: X_test.shape
```

```
Out[31]: (30000, 2)
```

```
In [32]: Y_test = X_test['is_duplicate']
         X_test.drop(labels=['is_duplicate'], axis=1, inplace=True)
```

```
In [33]: X_test.shape
```

```
Out[33]: (30000, 1)
```

```
In [34]: tfidf = TfidfVectorizer(lowercase=True,stop_words='english',ngram_range=(1,3),use_idf
```

```
In [35]: #TRAIN
         X_tfidf_train = tfidf.fit_transform(X_train['text'])
         X_tfidf_train = pd.DataFrame(data=X_tfidf_train.toarray(),index=X_train.index.values,
         X_tfidf_train.head()
```

```
Out[35]:              000        10  100  1000  1000 notes  1000 rupee  \
         264395  0.000000  0.000000  0.0   0.0         0.0         0.0
         247334  0.000000  0.000000  0.0   0.0         0.0         0.0
         162234  0.655974  0.277419  0.0   0.0         0.0         0.0
         265526  0.000000  0.000000  0.0   0.0         0.0         0.0
         200974  0.000000  0.000000  0.0   0.0         0.0         0.0

                 1000 rupee notes  1000 rupees  1000 rupees notes   11  ...   year old  \
         264395               0.0          0.0                0.0  0.0  ...        0.0
         247334               0.0          0.0                0.0  0.0  ...        0.0
         162234               0.0          0.0                0.0  0.0  ...        0.0
         265526               0.0          0.0                0.0  0.0  ...        0.0
         200974               0.0          0.0                0.0  0.0  ...        0.0

                 year resolution  year resolutions  years  years old  yes  york  young  \
         264395              0.0               0.0    0.0        0.0  0.0   0.0    0.0
         247334              0.0               0.0    0.0        0.0  0.0   0.0    0.0
         162234              0.0               0.0    0.0        0.0  0.0   0.0    0.0
         265526              0.0               0.0    0.0        0.0  0.0   0.0    0.0
         200974              0.0               0.0    0.0        0.0  0.0   0.0    0.0

                 youtube  zero
         264395      0.0   0.0
         247334      0.0   0.0
         162234      0.0   0.0
         265526      0.0   0.0
         200974      0.0   0.0

         [5 rows x 2000 columns]

In [36]: X_tfidf_train.shape

Out[36]: (70000, 2000)

In [37]: #TEST
         X_tfidf_test = tfidf.transform(X_test['text'])
         X_tfidf_test = pd.DataFrame(data=X_tfidf_test.toarray(),index=X_test.index.values,colu
         X_tfidf_test.head()

Out[37]:         000   10  100  1000  1000 notes  1000 rupee  1000 rupee notes  \
         322177  0.0  0.0  0.0   0.0         0.0         0.0               0.0
         147359  0.0  0.0  0.0   0.0         0.0         0.0               0.0
         121243  0.0  0.0  0.0   0.0         0.0         0.0               0.0
         360595  0.0  0.0  0.0   0.0         0.0         0.0               0.0
         261909  0.0  0.0  0.0   0.0         0.0         0.0               0.0

                 1000 rupees  1000 rupees notes   11  ...   year old  year resolution  \
         322177          0.0                0.0  0.0  ...        0.0              0.0
         147359          0.0                0.0  0.0  ...        0.0              0.0
         121243          0.0                0.0  0.0  ...        0.0              0.0
         360595          0.0                0.0  0.0  ...        0.0              0.0
         261909          0.0                0.0  0.0  ...        0.0              0.0

                 year resolutions  years  years old  yes  york  young  youtube  zero
         322177               0.0    0.0        0.0  0.0   0.0    0.0      0.0   0.0
         147359               0.0    0.0        0.0  0.0   0.0    0.0      0.0   0.0
         121243               0.0    0.0        0.0  0.0   0.0    0.0      0.0   0.0
         360595               0.0    0.0        0.0  0.0   0.0    0.0      0.0   0.0
         261909               0.0    0.0        0.0  0.0   0.0    0.0      0.0   0.0

         [5 rows x 2000 columns]

In [38]: X_tfidf_test.shape

Out[38]: (30000, 2000)
```

```
In [39]: if os.path.isfile('nlp_features_train.csv'):
             dfnlp = pd.read_csv("nlp_features_train.csv",encoding='latin-1')
         else:
             print("download nlp_features_train.csv from drive or run previous notebook")

         if os.path.isfile('df_fe_without_preprocessing_train.csv'):
             dfppro = pd.read_csv("df_fe_without_preprocessing_train.csv",encoding='latin-1')
         else:
             print("download df_fe_without_preprocessing_train.csv from drive or run previous
```

```
In [40]: dfnlp.head(n=2)
```

```
Out[40]:    id  qid1  qid2                                          question1  \
         0   0     1     2  what is the step by step guide to invest in sh...
         1   1     3     4  what is the story of kohinoor  koh i noor  dia...

                                                    question2  is_duplicate   cwc_min  \
         0  what is the step by step guide to invest in sh...             0  0.999980
         1  what would happen if the indian government sto...             0  0.799984

             cwc_max   csc_min   csc_max          ...             ctc_max  last_word_eq  \
         0  0.833319  0.999983  0.999983          ...            0.785709           0.0
         1  0.399996  0.749981  0.599988          ...            0.466664           0.0

            first_word_eq  abs_len_diff  mean_len  token_set_ratio  token_sort_ratio  \
         0            1.0           2.0      13.0              100                93
         1            1.0           5.0      12.5               86                63

            fuzz_ratio  fuzz_partial_ratio  longest_substr_ratio
         0          93                 100              1.000000
         1          66                  75              0.607843

         [2 rows x 21 columns]
```

```
In [41]: dfppro.head(n=2)
```

```
Out[41]:    id  qid1  qid2                                          question1  \
         0   0     1     2  What is the step by step guide to invest in sh...
         1   1     3     4  What is the story of Kohinoor (Koh-i-Noor) Dia...

                                                    question2  is_duplicate  freq_qid1  \
         0  What is the step by step guide to invest in sh...             0          1
         1  What would happen if the Indian government sto...             0          4

            freq_qid2  q1len  q2len  q1_n_words  q2_n_words  word_Common  word_Total  \
         0          1     66     57          14          12         10.0        23.0
         1          1     51     88           8          13          4.0        20.0

            word_share  freq_q1+q2  freq_q1-q2
```

```
0    0.434783           2           0
1    0.200000           5           3
```

```
In [42]: dfnlp.drop(labels=['id','qid1','qid2','question1','question2','is_duplicate'],axis=1,i
         dfppro.drop(labels=['id','qid1','qid2','question1','question2','is_duplicate'],axis=1
```

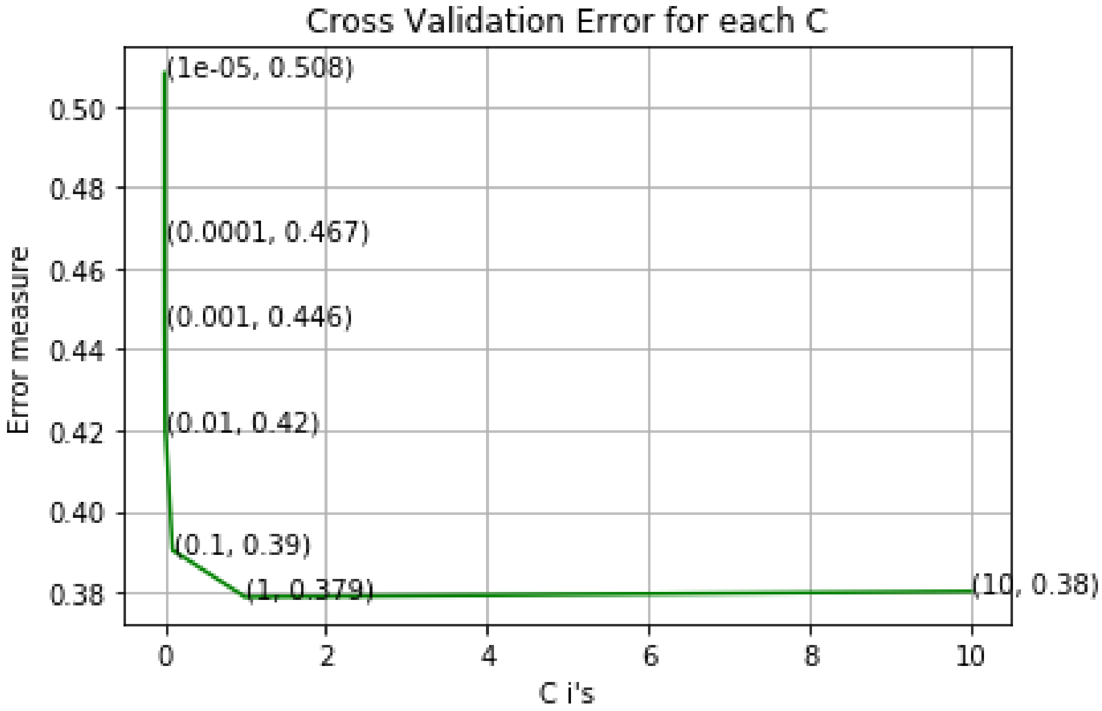
```
In [43]: X_train = np.array(X_tfidf_train.merge(dfnlp, how='inner', left_index=True, right_inde
         Y_train = np.array(Y_train)
```

```
In [44]: X_test = np.array(X_tfidf_test.merge(dfnlp, how='inner', left_index=True, right_index=
         Y_test = np.array(Y_test)
```

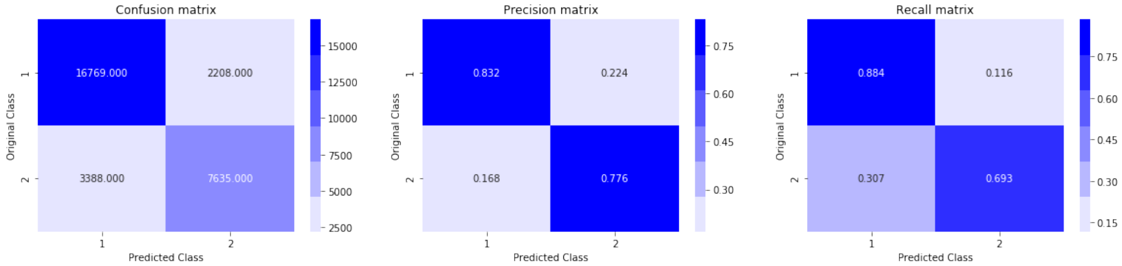
```
In [45]: del X_tfidf_train, X_tfidf_test, dfnlp, dfppro
```

```
In [64]: C = [10 ** x for x in range(-5, 2)]
         log_error_array=list()
         for i in C:
             clf = LogisticRegression(penalty='l2',C=i,random_state=42,n_jobs=-1)
             clf.fit(X_train, Y_train)
             sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
             sig_clf.fit(X_train, Y_train)
             predict_y = sig_clf.predict_proba(X_test)
             log_error_array.append(log_loss(Y_test, predict_y, labels=clf.classes_, eps=1e-15)
             print('For values of C = ', i, "The log loss is:",log_loss(Y_test, predict_y, labe

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


         best_C = np.argmin(log_error_array)
         clf = LogisticRegression(penalty='l2',C=C[best_C],random_state=42,n_jobs=-1)
         clf.fit(X_train, Y_train)
         sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
         sig_clf.fit(X_train, Y_train)

         predict_prob_train = sig_clf.predict_proba(X_train)
         print('For values of best C = ', C[best_C], "The train log loss is:",log_loss(Y_train
         predict_prob_test = sig_clf.predict_proba(X_test)
         print('For values of best C = ', C[best_C], "The test log loss is:",log_loss(Y_test, p
         predicted_class_test = sig_clf.predict(X_test)
         print("Total number of data points :", len(predicted_class_test))
         plot_confusion_matrix(Y_test, predicted_class_test)
```

```
For values of C =  1e-05 The log loss is: 0.5082266785082132
For values of C =  0.0001 The log loss is: 0.4670614577299387
For values of C =  0.001 The log loss is: 0.44647625434776567
For values of C =  0.01 The log loss is: 0.4201115739826379
For values of C =  0.1 The log loss is: 0.3904853215500971
For values of C =  1 The log loss is: 0.37886642262304765
For values of C =  10 The log loss is: 0.38025203183389183
```



```
For values of best C =  1 The train log loss is: 0.3634400307826507
For values of best C =  1 The test log loss is: 0.37886642262304765
Total number of data points : 30000
```



```
In [65]: miss_class = 1 - accuracy_score(Y_test,predicted_class_test)
         print("Number of missclassified points :",round(miss_class*100,2),'%')
```

```
Number of missclassified points : 18.65 %
```

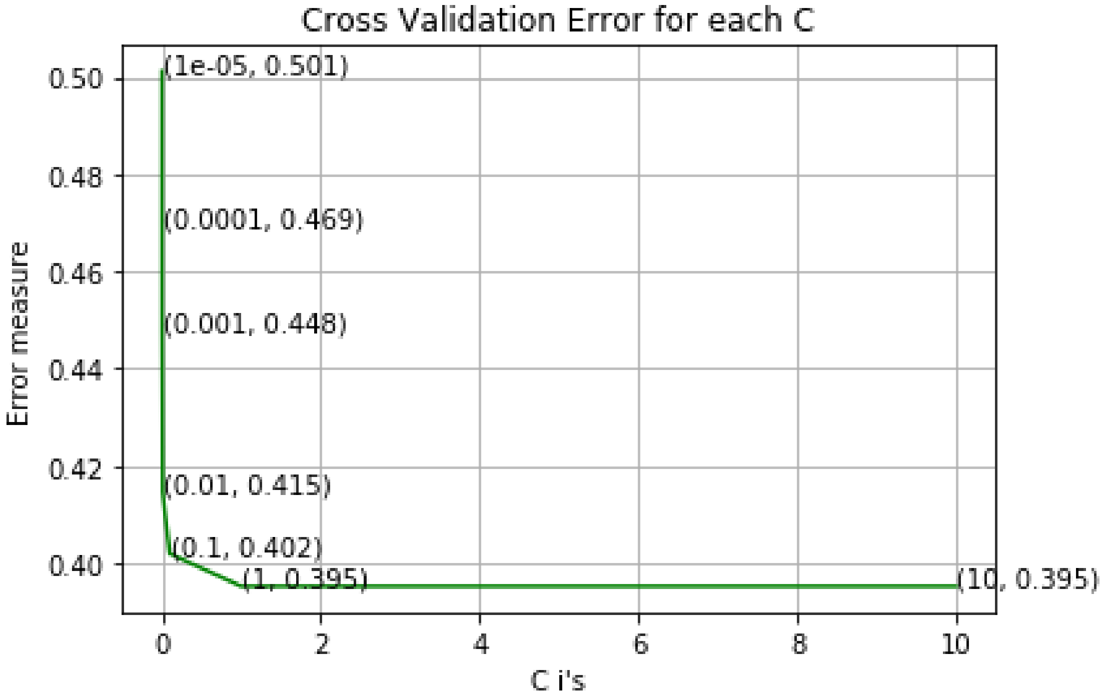
```
In [66]: auc_score = roc_auc_score(Y_test,predict_prob_test[:,1])
         print("AUC score :",round(auc_score*100,2))
```

```
AUC score : 90.06
```

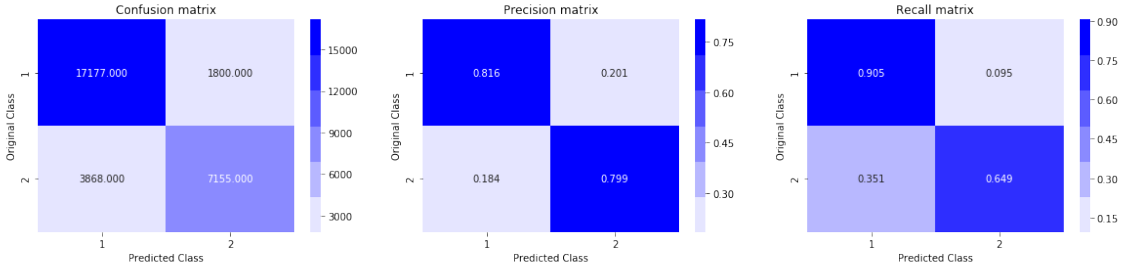
```
In [68]: fpr_LR,tpr_LR,thresholds_LR = roc_curve(Y_test, predict_prob_test[:,1])
```

```
In [69]: C = [10 ** x for x in range(-5, 2)]
         log_error_array=list()
         for i in C:
             clf = LinearSVC(penalty='l2', loss='hinge', C=i, random_state=42)
             clf.fit(X_train, Y_train)
             sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
             sig_clf.fit(X_train, Y_train)
             predict_y = sig_clf.predict_proba(X_test)
             log_error_array.append(log_loss(Y_test, predict_y, labels=clf.classes_, eps=1e-15
             print('For values of C = ', i, "The log loss is:",log_loss(Y_test, predict_y, labe

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


         best_C = np.argmin(log_error_array)
         clf = LinearSVC(penalty='l2', loss='hinge', C= C[best_C],random_state=42)
         clf.fit(X_train, Y_train)
         sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
         sig_clf.fit(X_train, Y_train)

         predict_prob_train = sig_clf.predict_proba(X_train)
         print('For values of best C = ', C[best_C], "The train log loss is:",log_loss(Y_train
         predict_prob_test = sig_clf.predict_proba(X_test)
         print('For values of best C = ', C[best_C], "The test log loss is:",log_loss(Y_test,
         predicted_class_test = sig_clf.predict(X_test)
         print("Total number of data points :", len(predicted_class_test))
         plot_confusion_matrix(Y_test, predicted_class_test)
```

```
For values of C =  1e-05 The log loss is: 0.5012534911437849
For values of C =  0.0001 The log loss is: 0.4692509702920933
For values of C =  0.001 The log loss is: 0.4481429451633474
For values of C =  0.01 The log loss is: 0.4145877516468119
For values of C =  0.1 The log loss is: 0.4019437533693079
For values of C =  1 The log loss is: 0.3952581358024728
For values of C =  10 The log loss is: 0.3952581358024728
```



```
For values of best C =  1 The train log loss is: 0.38663769406321974
For values of best C =  1 The test log loss is: 0.3952581358024728
Total number of data points : 30000
```



```
In [70]: miss_class = 1 - accuracy_score(Y_test,predicted_class_test)
         print("Number of missclassified points :",round(miss_class*100,2),'%')
```

```
Number of missclassified points : 18.89 %
```

```
In [71]: auc_score = roc_auc_score(Y_test,predict_prob_test[:,1])
         print("AUC score :",round(auc_score*100,2))
```

```
AUC score : 89.41
```

```
In [72]: fpr_SVM,tpr_SVM,thresholds_SVM = roc_curve(Y_test, predict_prob_test[:,1])
```

```
In [73]: # Sklearn version:
         # n_estimators = [3,5,20,50,80,100]
         # max_depth = [3,5,7]
         # log_error_array=list()
         # for estimator in n_estimators:
         #     for depth in max_depth:
         #         clf = GradientBoostingClassifier(n_estimators=estimator, max_depth=depth, r
         #         clf.fit(X_train, Y_train)
         #         sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
         #         sig_clf.fit(X_train, Y_train)
         #         predict_y = sig_clf.predict_proba(X_test)
         #         log_error_array.append(log_loss(Y_test, predict_y, labels=clf.classes_, eps
         #         print('For values of n_estimators = ', estimator, ' and max_depth = ', dept
```

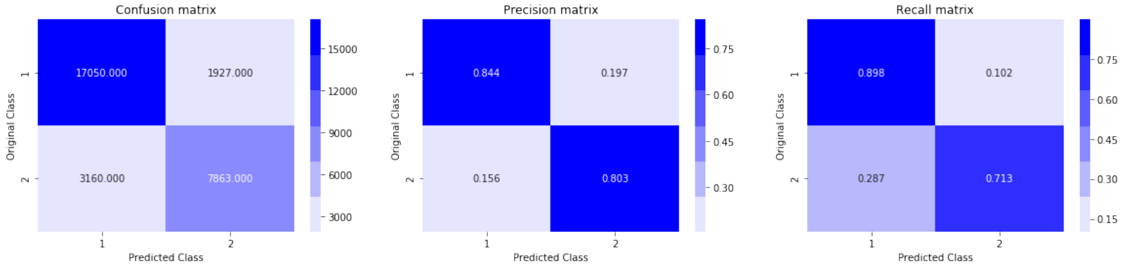
```
In [74]: # Xgboost - sklearn wrapper version:

         # Create the parameter grid: param_grid
         # model.get_params()
         param_grid = {
             'base_estimator__learning_rate': np.arange(0.05,1.05,0.05),
             'base_estimator__n_estimators': [5,10,20,40,60,80,100,120],
             'base_estimator__subsample': np.arange(0.05,1.05,0.05),
             'base_estimator__max_depth':[3,5,7,9]
         }

         if os.path.isfile('xgbClassifier.pkl') == False:
             # Model:
             base_model = xgb.XGBClassifier()
             model = CalibratedClassifierCV(base_estimator = base_model)

             # Perform random search:
             grid_search = RandomizedSearchCV(estimator=model, param_distributions=param_grid,
             grid_result = grid_search.fit(X_train,Y_train)

             # summarize results
             print("Best: %f using %s" % (grid_result.best_score_, grid_result.best_params_))
             means = grid_result.cv_results_['mean_test_score']
             stds = grid_result.cv_results_['std_test_score']
             params = grid_result.cv_results_['params']
             for mean, stdev, param in zip(means, stds, params):
                 print("%f (%f) with: %r" % (mean, stdev, param))

             joblib.dump(grid_result.best_estimator_, 'xgbClassifier.pkl')
             model = grid_result.best_estimator_
         else:
             model = joblib.load('xgbClassifier.pkl')
```

```
In [75]: predict_prob_train = model.predict_proba(X_train)
         print("The train log loss is:",log_loss(Y_train, predict_prob_train))
         predict_prob_test = model.predict_proba(X_test)
         print("The test log loss is:",log_loss(Y_test, predict_prob_test))
         predicted_class_test = model.predict(X_test)
```

```
The train log loss is: 0.3549992799848697
The test log loss is: 0.35377406307865034
```

The xgbclassifier model did really well on both train and test data which implies that model is stable and does not over or under fit.

```
In [76]: plot_confusion_matrix(Y_test, predicted_class_test)
```



```
In [77]: miss_class = 1 - accuracy_score(Y_test,predicted_class_test)
         print("Number of missclassified points :",round(miss_class*100,2),'%')
```

```
Number of missclassified points : 16.96 %
```

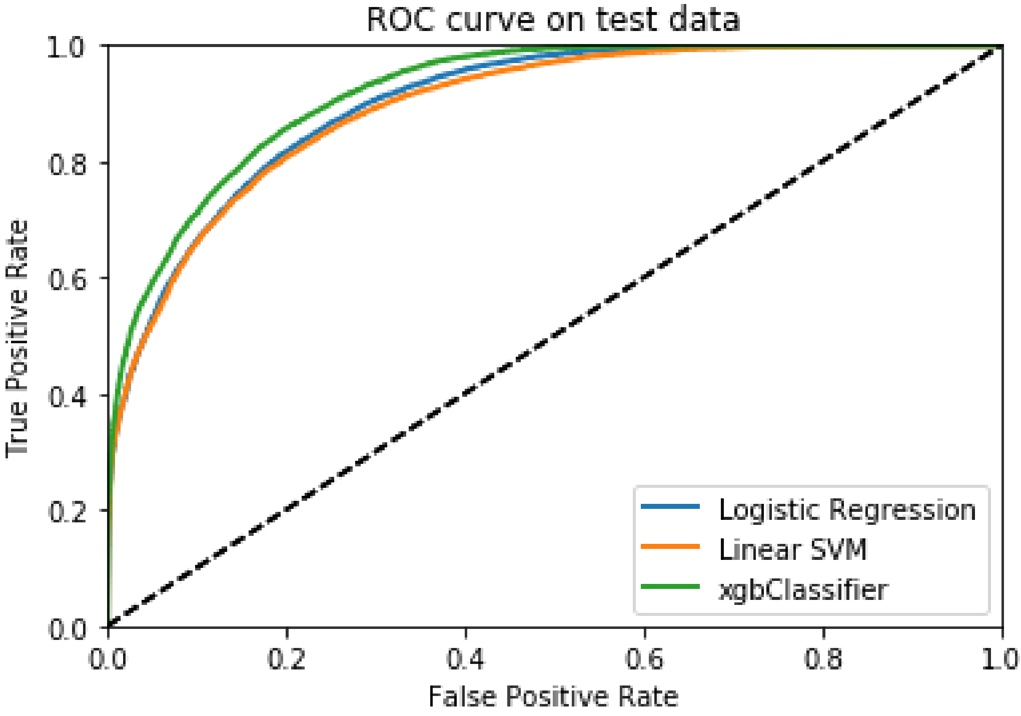
```
In [78]: auc_score = roc_auc_score(Y_test,predict_prob_test[:,1])
         print("AUC score :",round(auc_score*100,2))
```

```
AUC score : 91.99
```

```
In [79]: fpr_BT,tpr_BT,thresholds_BT = roc_curve(Y_test, predict_prob_test[:,1])
```

```
In [83]: def plot_roc_curve(fpr,tpr,label=None):
             plt.plot(fpr,tpr,linewidth=2,label=label)
             plt.plot([0,1],[0,1],"k--")
             plt.xlabel("False Positive Rate")
             plt.ylabel("True Positive Rate")
             plt.axis([0,1,0,1])
             plt.legend()
             plt.title("ROC curve on test data")
```

```
In [84]: plot_roc_curve(fpr_LR,tpr_LR,label='Logistic Regression')
         plot_roc_curve(fpr_SVM,tpr_SVM,label='Linear SVM')
         plot_roc_curve(fpr_BT,tpr_BT,label='xgbClassifier')
         plt.show()
```



```
In [85]: table = PrettyTable(["Model", "Encoding", "Train Log Loss", "Test Log Loss", "Miss cla
```

```
In [86]: table.add_row(['LR','TF-IDF','0.369','0.385','18.65','90.06'])
         table.add_row(['Linear SVM','TF-IDF','0.387','0.395','18.89','89.41'])
         table.add_row(['xgb Classifier','TF-IDF','0.265','0.346','16.29','92.34'])
```

```
In [87]: print(table)
```

```
+----------------+----------+----------------+---------------+--------------+-----------+
|     Model      | Encoding | Train Log Loss | Test Log Loss | Miss class % | AUC score |
+----------------+----------+----------------+---------------+--------------+-----------+
|       LR       |  TF-IDF  |     0.369      |     0.385     |    18.65     |   90.06   |
|   Linear SVM   |  TF-IDF  |     0.387      |     0.395     |    18.89     |   89.41   |
| xgb Classifier |  TF-IDF  |     0.265      |     0.346     |    16.29     |   92.34   |
+----------------+----------+----------------+---------------+--------------+-----------+
```