

Quora Question Pairs

In [1]:

```
import warnings
warnings.filterwarnings("ignore")
```

Business Problem:

Quora is a place to gain and share knowledge—about anything. It's a platform to ask questions and connect with people who contribute unique insights and quality answers. This empowers people to learn from each other and to better understand the world.

Over 100 million people visit Quora every month, so it's no surprise that many people ask similarly worded questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same question. Quora values canonical questions because they provide a better experience to active seekers and writers, and offer more value to both of these groups in the long term.

Problem Statement:

- Identify which questions asked on Quora are duplicates of questions that have already been asked.
- This could be useful to instantly provide answers to questions that have already been answered.
- We are tasked with predicting whether a pair of questions are duplicates or not.

Business Objectives and constraints

1. The cost of a mis-classification can be very high.
2. You would want a probability of a pair of questions to be duplicates so that you can choose any threshold of choice.
3. No strict latency concerns.
4. Interpretability is partially important.

Data

- Size of Train.csv - 60MB
- Number of rows in Train.csv = 404,290
- Size of Train.csv - 60MBTrain.csv contains 5 columns : qid1, qid2, question1, question2, is_duplicate
 - qid{1, 2}: The unique ID of each question in the pair
 - question{1, 2}: The actual textual contents of the questions.
 - is_duplicate: The label that we are trying to predict - whether the two questions are duplicates of each other.

Machine Learning Problem:

It is a binary classification problem, for a given pair of questions we need to predict if they are duplicate or not.

Performance Metric:

- Confusion Matrix
- Log-Loss

Load Data

In [2]:

```
import pandas as pd
import os
```

```
from matplotlib import pyplot as plt
import seaborn as sns
import numpy as np
import sqlite3
```

In [3]:

```
path = r'C:\Users\Friend\AI\AI_datasets\Quora'
```

In [4]:

```
data = pd.read_csv(os.path.join(path, "train.csv"))
print(data.shape)
data.head()
```

(404290, 6)

Out[4]:

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when 23^{24} i...	0
4	4	9	10	Which one dissolve in water quickly sugar, salt...	Which fish would survive in salt water?	0

Exploratory Data Analysis without pre-processing:

In [5]:

```
#Basic Info
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 404290 entries, 0 to 404289
Data columns (total 6 columns):
id                404290 non-null int64
qid1              404290 non-null int64
qid2              404290 non-null int64
question1         404289 non-null object
question2         404288 non-null object
is_duplicate      404290 non-null int64
dtypes: int64(4), object(2)
memory usage: 18.5+ MB
```

In [6]:

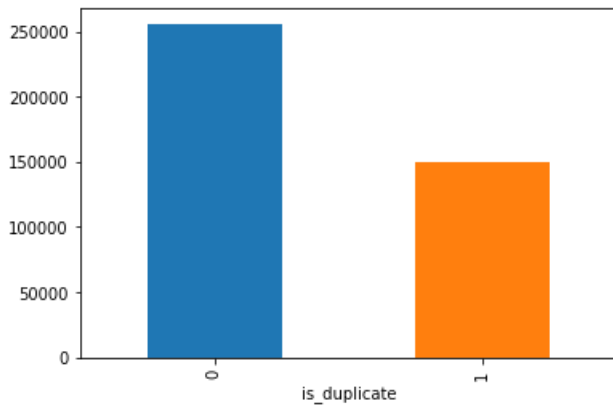
```
# Distribution of class labels:
```

```
print('Questions that are not similar ; with class_label = 0', (100 - round(data['is_duplicate'].mean() * 100, 2)))
print('Questions that are similar ; with class_label = 1', round(data['is_duplicate'].mean() * 100, 2))
data.groupby("is_duplicate")["id"].count().plot.bar()
```

Questions that are not similar ; with class_label = 0 63.08
 Questions that are similar ; with class_label = 1 36.92

Out[6]:

<matplotlib.axes._subplots.AxesSubplot at 0x2b68ebc2438>



In [7]:

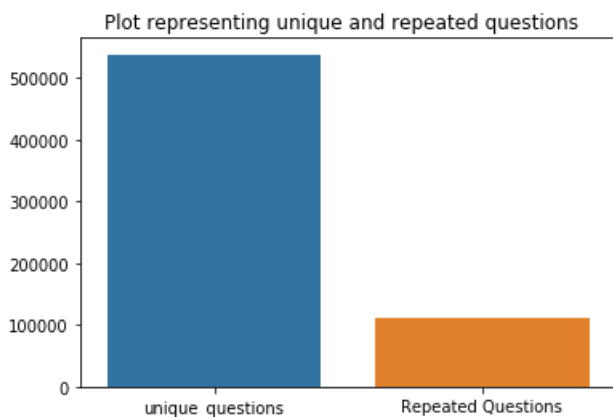
```
# Unique questions

qids = pd.Series(data['qid1'].tolist() + data['qid2'].tolist())
print('unique questions from total pairs', len(np.unique(qids)))
print('number of questions appeared more than once', np.sum(qids.value_counts() > 1))
print('Max number of times a single question is repeated', max(qids.value_counts()))

#plot unique questions vs repeated questions
x = ["unique_questions", "Repeated Questions"]
y = [len(np.unique(qids)), np.sum(qids.value_counts() > 1)]

plt.figure()
plt.title ("Plot representing unique and repeated questions ")
sns.barplot(x,y)
plt.show()
```

unique questions from total pairs 537933
number of questions appeared more than once 111780
Max number of times a single question is repeated 157



In [8]:

```
# Checking for duplicates

pair_duplicates = data[['qid1', 'qid2', 'is_duplicate']].groupby(['qid1', 'qid2']).count().reset_index()
print('number of duplicates', data.shape[0]-pair_duplicates.shape[0])
```

number of duplicates 0

In [9]:

```
# Checking for null values
```

```
null_rows = data[data.isnull().any(1)]
print (null_rows)
```

	id	qid1	qid2	question1 \	question2	is_duplicate
105780	105780	174363	174364	How can I develop android app?		
201841	201841	303951	174364	How can I create an Android app?		
363362	363362	493340	493341		NaN	
105780					NaN	0
201841					NaN	0
363362				My Chinese name is Haichao Yu. What English na...		0

In [10]:

```
# filling null values with empty strings

data = data.fillna('')
null_rows = data[data.isnull().any(1)]
print (null_rows)
```

Empty DataFrame
Columns: [id, qid1, qid2, question1, question2, is_duplicate]
Index: []

Feature Extraction

In [11]:

```
if os.path.isfile(os.path.join(path, 'data_fe_without_preprocessing_train.csv')):
    data = pd.read_csv(os.path.join(path, "data_fe_without_preprocessing_train.csv"), encoding='latin-1')
else:
    # Frequency of qid1's
    data['freq_qid1'] = data.groupby('qid1')['qid1'].transform('count')
    # Frequency of qid2's
    data['freq_qid2'] = data.groupby('qid2')['qid2'].transform('count')
    # Length of q1
    data['q1len'] = data['question1'].str.len()
    # Length of q2
    data['q2len'] = data['question2'].str.len()
    # Number of words in Question 1
    data['q1_n_words'] = data['question1'].apply(lambda row: len(row.split(" ")))
    # Number of words in Question 2
    data['q2_n_words'] = data['question2'].apply(lambda row: len(row.split(" ")))

    # Number of common unique words in Question 1 and Question 2
    def normalized_word_Common(row):
        w1 = set(map(lambda word: word.lower().strip(), row['question1'].split(" ")))
        w2 = set(map(lambda word: word.lower().strip(), row['question2'].split(" ")))
        return 1.0 * len(w1 & w2)
    data['word_Common'] = data.apply(normalized_word_Common, axis=1)

    # Total num of words in Question 1 + Total num of words in Question 2
    def normalized_word_Total(row):
        w1 = set(map(lambda word: word.lower().strip(), row['question1'].split(" ")))
        w2 = set(map(lambda word: word.lower().strip(), row['question2'].split(" ")))
        return 1.0 * (len(w1) + len(w2))
    data['word_Total'] = data.apply(normalized_word_Total, axis=1)

    # (word_common)/(word_Total)
    def normalized_word_share(row):
        w1 = set(map(lambda word: word.lower().strip(), row['question1'].split(" ")))
        w2 = set(map(lambda word: word.lower().strip(), row['question2'].split(" ")))
        return 1.0 * len(w1 & w2) / (len(w1) + len(w2))
    data['word_share'] = data.apply(normalized_word_share, axis=1)

    # sum total of frequency of qid1 and qid2
    data['freq_q1+q2'] = data['freq_qid1'] + data['freq_qid2']
    # absolute difference of frequency of qid1 and qid2
    data['freq_q1-q2'] = abs(data['freq_qid1'] - data['freq_qid2'])
```

```
data['freq_q1_q2'] = abs(data['freq_q1a'] - data['freq_q1b'])
```

```
data.to_csv(os.path.join(path, "data_fe_without_preprocessing_train.csv"), index=False)
```

```
data.head()
```

Out[11]:

	id	qid1	qid2	question1	question2	is_duplicate	freq_qid1	freq_qid2	q1len	q2len	q1_n_words	q2_n_words	v
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0	1	1	66	57	14	12	1
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0	4	1	51	88	8	13	4
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0	1	1	73	59	14	10	4
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when 23^{24} $[/math> i...$	0	1	1	50	65	11	9	0
4	4	9	10	Which one dissolve in water quickly sugar, salt...	Which fish would survive in salt water?	0	3	1	76	39	13	7	2

In [12]:

```
print ("Minimum length of the questions in question1 : " , min(data['q1_n_words']))
print ("Minimum length of the questions in question2 : " , min(data['q2_n_words']))
print ("Number of Questions with minimum length in question1 :", data[data['q1_n_words']== 1].shape[0])
print ("Number of Questions with minimum length in question2 :", data[data['q2_n_words']== 1].shape[0])
```

```
Minimum length of the questions in question1 : 1
Minimum length of the questions in question2 : 1
Number of Questions with minimum length in question1 : 67
Number of Questions with minimum length in question2 : 24
```

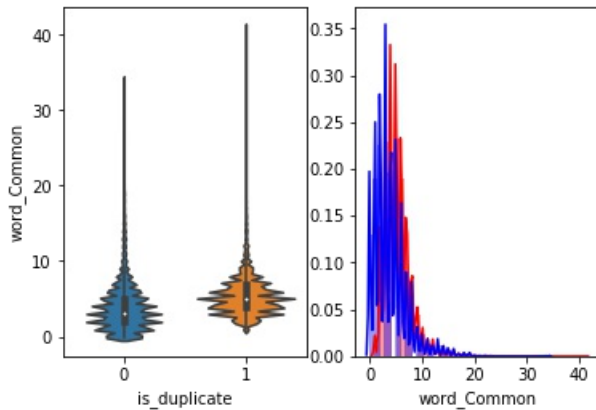
In [13]:

```
# word Common
# The distributions of the word_Common feature in similar and non-similar questions are highly overlapping

plt.figure()
plt.subplot(1,2,1)
sns.violinplot(x = 'is_duplicate', y = 'word_Common', data = data[0:])
plt.subplot(1,2,2)
sns.distplot(data[data['is_duplicate'] == 1.0]['word_Common'][0:], label = "1", color = 'red')
sns.distplot(data[data['is_duplicate'] == 0.0]['word_Common'][0:], label = "0", color = 'blue' )
```

```
plt.show()
```

```
C:\Users\Friend\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kw
arg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\Friend\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kw
arg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

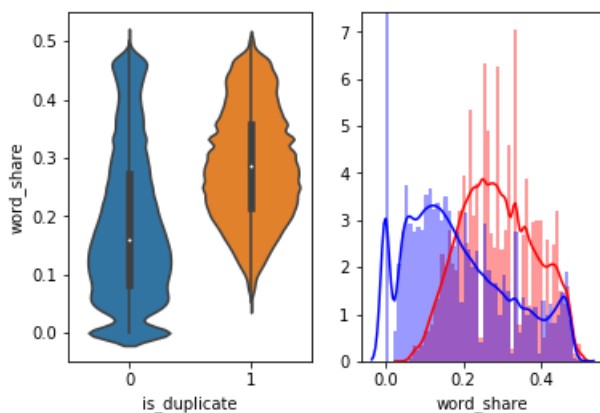


In [14]:

```
# word_share
# The distributions for word_share have some overlap on the far right-hand side,i.e., there are quite a
lot of questions with high word similarity.The average word share and Common no. of words of qid1 and q
id2 is more when they are duplicate(Similar)

plt.figure()
plt.subplot(1,2,1)
sns.violinplot(x = 'is_duplicate', y = 'word_share', data = data[0:])
plt.subplot(1,2,2)
sns.distplot(data[data['is_duplicate'] == 1.0]['word_share'][0:] , label = "1", color = 'red')
sns.distplot(data[data['is_duplicate'] == 0.0]['word_share'][0:] , label = "0" , color = 'blue' )
plt.show()
```

```
C:\Users\Friend\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kw
arg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\Friend\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kw
arg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



Text Pre-processing

In [15]:

```
import re
from nltk.corpus import stopwords
```

```

from nltk.stem import PorterStemmer
from fuzzywuzzy import fuzz
from bs4 import BeautifulSoup

```

In [16]:

```

data = pd.read_csv(os.path.join(path, "train.csv"))

# To get the results in 4 decimal points
SAFE_DIV = 0.0001

STOP_WORDS = stopwords.words("english")

def preprocess(x):
    x = str(x).lower()
    x = x.replace(",000,000", "m").replace(",000", "k").replace("'", "").replace('"', "'")\
        .replace("won't", "will not").replace("cannot", "can not").replace("can't",
"can not")\
        .replace("n't", " not").replace("what's", "what is").replace("it's", "it is"
)\
        .replace("'ve", " have").replace("i'm", "i am").replace("'re", " are")\
        .replace("he's", "he is").replace("she's", "she is").replace("'s", " own")\
        .replace("%", " percent ").replace("₹", " rupee ").replace("$", " dollar ")\
        .replace("€", " euro ").replace("'ll", " will")

    x = re.sub(r"([0-9]+)000000", r"\1m", x)
    x = re.sub(r"([0-9]+)000", r"\1k", x)

    porter = PorterStemmer()
    pattern = re.compile('\W')

    if type(x) == type(''):
        x = re.sub(pattern, ' ', x)

    if type(x) == type(''):
        x = porter.stem(x)
        example1 = BeautifulSoup(x)
        x = example1.get_text()

    return x

```

Advanced Feature Extraction

In [17]:

```

def get_token_features(q1, q2):
    token_features = [0.0]*10

    # Converting the Sentence into Tokens:
    q1_tokens = q1.split()
    q2_tokens = q2.split()

    if len(q1_tokens) == 0 or len(q2_tokens) == 0:
        return token_features

    # Get the non-stopwords in Questions
    q1_words = set([word for word in q1_tokens if word not in STOP_WORDS])
    q2_words = set([word for word in q2_tokens if word not in STOP_WORDS])

    #Get the stopwords in Questions
    q1_stops = set([word for word in q1_tokens if word in STOP_WORDS])
    q2_stops = set([word for word in q2_tokens if word in STOP_WORDS])

    # Get the common non-stopwords from Question pair
    common_word_count = len(q1_words.intersection(q2_words))
    # Get the common stopwords from Question pair
    common_stop_count = len(q1_stops.intersection(q2_stops))
    # Get the common Tokens from Question pair
    common_token_count = len(set(q1_tokens).intersection(set(q2_tokens)))

    token_features[0] = common_word_count / (min(len(q1_words), len(q2_words)) + SAFE_DIV)
    token_features[1] = common_word_count / (max(len(q1_words), len(q2_words)) + SAFE_DIV)
    token_features[2] = common_stop_count / (min(len(q1_stops), len(q2_stops)) + SAFE_DIV)

```

```

token_features[3] = common_stop_count / (max(len(q1_stops), len(q2_stops)) + SAFE_DIV)
token_features[4] = common_token_count / (min(len(q1_tokens), len(q2_tokens)) + SAFE_DIV)
token_features[5] = common_token_count / (max(len(q1_tokens), len(q2_tokens)) + SAFE_DIV)
# Last word of both question is same or not
token_features[6] = int(q1_tokens[-1] == q2_tokens[-1])
# First word of both question is same or not
token_features[7] = int(q1_tokens[0] == q2_tokens[0])
token_features[8] = abs(len(q1_tokens) - len(q2_tokens))
#Average Token Length of both Questions
token_features[9] = (len(q1_tokens) + len(q2_tokens))/2
return token_features

```

get the Longest Common sub string

```

def get_longest_substr_ratio(a, b):
    strs = list(distance.lcs substrings(a, b))
    if len(strs) == 0:
        return 0
    else:
        return len(strs[0]) / (min(len(a), len(b)) + 1)

```

In [18]:

```

def extract_features(df):

    # preprocessing each question

    df["question1"] = df["question1"].fillna("").apply(preprocess)
    df["question2"] = df["question2"].fillna("").apply(preprocess)

    # Merging Features with dataset

    token_features = df.apply(lambda x: get_token_features(x["question1"], x["question2"]), axis=1)
    # Ratio of common word count to min length of word count of Q1 and Q2
    df["cwc_min"] = list(map(lambda x: x[0], token_features))
    # Ratio of common word count to max length of word count of Q1 and Q2
    df["cwc_max"] = list(map(lambda x: x[1], token_features))
    # Ratio of common stop count to min length of stop count of Q1 and Q2
    df["csc_min"] = list(map(lambda x: x[2], token_features))
    # Ratio of common stop count to max length of stop count of Q1 and Q2
    df["csc_max"] = list(map(lambda x: x[3], token_features))
    # Ratio of common token count to min length of token count of Q1 and Q2
    df["ctc_min"] = list(map(lambda x: x[4], token_features))
    # Ratio of common token count to max length of token count of Q1 and Q2
    df["ctc_max"] = list(map(lambda x: x[5], token_features))
    # Check if Last word of both questions is equal or not
    df["last_word_eq"] = list(map(lambda x: x[6], token_features))
    # Check if First word of both questions is equal or not
    df["first_word_eq"] = list(map(lambda x: x[7], token_features))
    # Abs. length difference
    df["abs_len_diff"] = list(map(lambda x: x[8], token_features))
    # Average Token Length of both Questions
    df["mean_len"] = list(map(lambda x: x[9], token_features))
    # Ratio of length longest common substring to min length of token count of Q1 and Q2
    df["longest_substr_ratio"] = df.apply(lambda x: get_longest_substr_ratio(x["question1"], x["question2"]), axis=1)

    # Fuzzy features

    # measurement of edit distance-compare two strings
    df["fuzz_ratio"] = df.apply(lambda x: fuzz.QRatio(x["question1"], x["question2"]), axis=1)
    # measurement of edit distance-compare two strings partially
    df["fuzz_partial_ratio"] = df.apply(lambda x: fuzz.partial_ratio(x["question1"], x["question2"]), axis=1)
    # tokenize both strings.split the tokens into two groups: intersection and remainder.
    # use those sets to build up a comparison string.
    df["token_set_ratio"] = df.apply(lambda x: fuzz.token_set_ratio(x["question1"], x["question2"]), axis=1)
    # The token sort approach involves tokenizing the string in question, sorting the tokens alphabetically, and
    # then joining them back into a string We then compare the transformed strings with a simple ratio()
    df["token_sort_ratio"] = df.apply(lambda x: fuzz.token_sort_ratio(x["question1"], x["question2"]), axis=1)
    return df

```


In [19]:

```
if os.path.isfile(os.path.join(path, 'nlp_features_train.csv')):
    data = pd.read_csv(os.path.join(path, "nlp_features_train.csv"), encoding='latin-1')
    data.fillna('')
else:
    data = pd.read_csv(os.path.join(path, "train.csv"))
    data = extract_features(data)
    data.to_csv(os.path.join(path, "nlp_features_train.csv"), index=False)
data.head()
```

Out[19]:

	id	qid1	qid2	question1	question2	is_duplicate	cwc_min	cwc_max	csc_min	csc_max	...	ctc_max	last_word_e
0	0	1	2	what is the step by step guide to invest in sh...	what is the step by step guide to invest in sh...	0	0.999980	0.833319	0.999983	0.999983	...	0.785709	0.0
1	1	3	4	what is the story of kohinoor koh i noor dia...	what would happen if the indian government sto...	0	0.799984	0.399996	0.749981	0.599988	...	0.466664	0.0
2	2	5	6	how can i increase the speed of my internet co...	how can internet speed be increased by hacking...	0	0.399992	0.333328	0.399992	0.249997	...	0.285712	0.0
3	3	7	8	why am i mentally very lonely how can i solve...	find the remainder when math 23 24 math i...	0	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0
4	4	9	10	which one dissolve in water quikly sugar salt...	which fish would survive in salt water	0	0.399992	0.199998	0.999950	0.666644	...	0.307690	0.0

5 rows × 21 columns



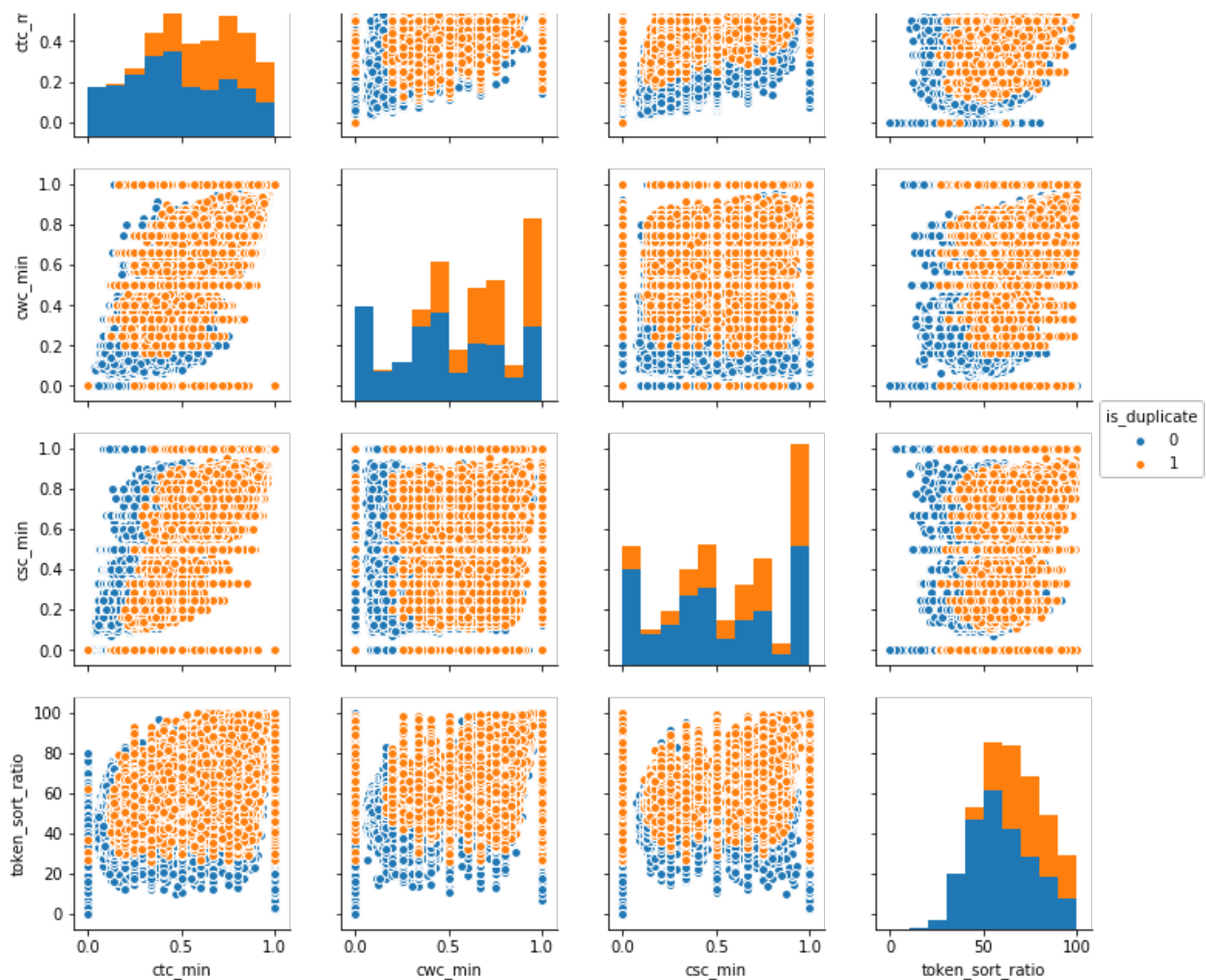
Exploratory Data Analysis on Advanced Features

In [17]:

```
#EDA on advanced features
# From the plots given below it is very clear tat certain features gives us a clear cut value on decidi
ng similarity.

n = data.shape[0]
sns.pairplot(data[['ctc_min', 'cwc_min', 'csc_min', 'token_sort_ratio', 'is_duplicate']][0:n], hue='is_
duplicate', vars=['ctc_min', 'cwc_min', 'csc_min', 'token_sort_ratio'])
plt.show()
```





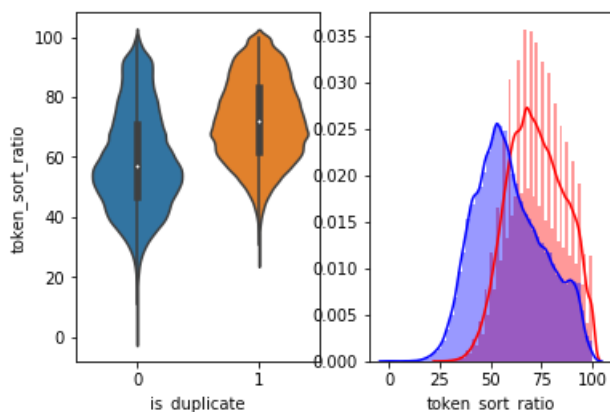
In [18]:

```
# Distribution of the token_sort_ratio
plt.figure()

plt.subplot(1,2,1)
sns.violinplot(x = 'is_duplicate', y = 'token_sort_ratio', data = data[0:] , )

plt.subplot(1,2,2)
sns.distplot(data[data['is_duplicate'] == 1.0]['token_sort_ratio'][0:] , label = "1", color = 'red')
sns.distplot(data[data['is_duplicate'] == 0.0]['token_sort_ratio'][0:] , label = "0" , color = 'blue' )
plt.show()
```

C:\Users\Friend\Anaconda3\lib\site-packages\matplotlib\axes_axes.py:6462: UserWarning: The 'normed' kw arg is deprecated, and has been replaced by the 'density' kwarg.
 warnings.warn("The 'normed' kwarg is deprecated, and has been "
 C:\Users\Friend\Anaconda3\lib\site-packages\matplotlib\axes_axes.py:6462: UserWarning: The 'normed' kw arg is deprecated, and has been replaced by the 'density' kwarg.
 warnings.warn("The 'normed' kwarg is deprecated, and has been "



In [19]:

```
# Distribution of the fuzz_ratio

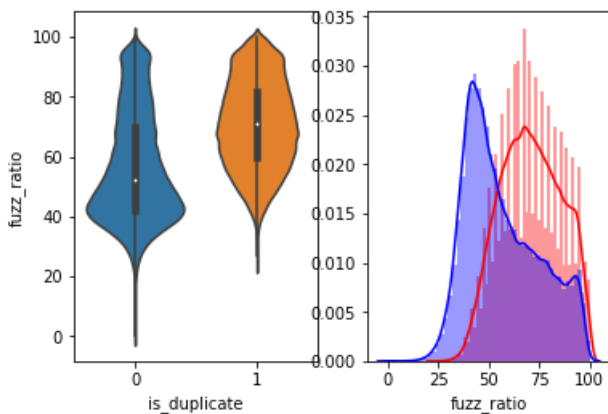
plt.figure()

plt.subplot(1,2,1)
sns.violinplot(x = 'is_duplicate', y = 'fuzz_ratio', data = data[0:] , )

plt.subplot(1,2,2)
sns.distplot(data[data['is_duplicate'] == 1.0]['fuzz_ratio'][0:] , label = "1", color = 'red')
sns.distplot(data[data['is_duplicate'] == 0.0]['fuzz_ratio'][0:] , label = "0" , color = 'blue' )

plt.show()
```

C:\Users\Friend\Anaconda3\lib\site-packages\matplotlib\axes_axes.py:6462: UserWarning: The 'normed' kw arg is deprecated, and has been replaced by the 'density' kwarg.
warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\Friend\Anaconda3\lib\site-packages\matplotlib\axes_axes.py:6462: UserWarning: The 'normed' kw arg is deprecated, and has been replaced by the 'density' kwarg.
warnings.warn("The 'normed' kwarg is deprecated, and has been "



Featurizations

In [22]:

```
from sklearn.preprocessing import normalize
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [23]:

```
data = pd.read_csv(os.path.join(path, "train.csv"))
data['question1'] = data['question1'].apply(lambda x: str(x))
data['question2'] = data['question2'].apply(lambda x: str(x))

# merge texts
question1 = list(data['question1'])
question1_train = question1[0:int(len(question1)*0.8)+1]
question1_test = question1[int(len(question1)*0.8)+1:len(question1)]

question2 = list(data['question2'])
question2_train = question2[0:int(len(question2)*0.8)+1]
question2_test = question2[int(len(question2)*0.8)+1:len(question2)]
```

In [24]:

```
from sklearn.feature_extraction.text import TfidfVectorizer

tf_idf_vect = TfidfVectorizer(lowercase=False, max_features = 100)
vocabulary = tf_idf_vect.fit(question1)
tf_idf_train = tf_idf_vect.transform(question1_train)
```

```
tf_idf_train = tf_idf_vect.transform(question1_train)
tf_idf_test = tf_idf_vect.transform(question1_test)

df3_train = pd.DataFrame(tf_idf_train.todense(), columns=tf_idf_vect.get_feature_names())
df3_test = pd.DataFrame(tf_idf_test.todense(), columns=tf_idf_vect.get_feature_names())
```

In [25]:

```
from sklearn.feature_extraction.text import TfidfVectorizer

tf_idf_vect = TfidfVectorizer(lowercase=False,max_features = 100)
vocabulary = tf_idf_vect.fit(question2_train)
tf_idf_train = tf_idf_vect.transform(question2_train)
tf_idf_test = tf_idf_vect.transform(question2_test)

df4_train = pd.DataFrame(tf_idf_train.todense(), columns=tf_idf_vect.get_feature_names())
df4_test = pd.DataFrame(tf_idf_test.todense(), columns=tf_idf_vect.get_feature_names())
```

Data Preparation

In [26]:

```
#nlp_features_train.csv (NLP Features)

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

df1 = dfnlp.drop(['qid1', 'qid2', 'question1', 'question2'], axis=1)
df1.shape
```

Out[26]:

(404290, 17)

In [27]:

```
#prepro_features_train.csv (Simple Preprocessing Features)

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

df2 = dfppro.drop(['qid1', 'qid2', 'question1', 'question2', 'is_duplicate'], axis=1)
df2.shape
```

Out[27]:

(404290, 12)

In [36]:

```
#Train Data
df12_train = df12.head(323433)

df3_train['id'] = [x for x in range(0, df12_train.shape[0])]
df4_train['id'] = [x for x in range(0, df12_train.shape[0])]

df34_train = df3_train.merge(df4_train, on='id', how='left')
finaldata_train = df12_train.merge(df34_train, on='id', how='left')

y_train = finaldata_train['is_duplicate']
finaldata_train.drop(['id', 'is_duplicate'], axis=1, inplace=True)

X_train = finaldata_train[0:80000]
y_train = y_train[0:80000]
```

```
y_train = y_train[0:80000]

print(X_train.shape, y_train.shape)
```

```
(80000, 226) (80000,)
```

In [37]:

```
#Test Data
df12_test = df12.tail(80857)

df3_test['id'] = [x for x in range(323433, 323433+df12_test.shape[0])]
df4_test['id'] = [x for x in range(323433, 323433+df12_test.shape[0])]

df34_test = df3_test.merge(df4_test, on='id', how='left')
finaldata_test = df12_test.merge(df34_test, on='id', how='left')

y_test = finaldata_test['is_duplicate']
finaldata_test.drop(['id', 'is_duplicate'], axis=1, inplace=True)

X_test = finaldata_test[0:20000]
y_test = y_test[0:20000]

print(X_test.shape, y_test.shape)
```

```
(20000, 226) (20000,)
```

Standardize data

In [46]:

```
from sklearn.preprocessing import StandardScaler
```

In [49]:

```
scaler = StandardScaler()
vocab = scaler.fit(X_train)
X_train = vocab.transform(X_train)
X_test = vocab.transform(X_test)
```

Machine Learning Models

In [38]:

```
from sklearn.metrics.classification import accuracy_score, log_loss
from sklearn.metrics import confusion_matrix
from sklearn.metrics import precision_recall_curve, auc, roc_curve
from sklearn.linear_model import LogisticRegression
from sklearn.multiclass import OneVsRestClassifier
from sklearn.svm import SVC
from sklearn.cross_validation import StratifiedKFold
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import SGDClassifier
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
from sklearn.calibration import CalibratedClassifierCV
import xgboost as xgb
```

C:\Users\Friend\Anaconda3\lib\site-packages\sklearn\cross_validation.py:41: DeprecationWarning: This module was deprecated in version 0.18 in favor of the model_selection module into which all the refactored classes and functions are moved. Also note that the interface of the new CV iterators are different from that of this module. This module will be removed in 0.20.

"This module will be removed in 0.20.", DeprecationWarning)

C:\Users\Friend\Anaconda3\lib\site-packages\sklearn\ensemble\weight_boosting.py:29: DeprecationWarning: numpy.core.umath_tests is an internal NumPy module and should not be imported. It will be removed in a future NumPy release.

from numpy.core.umath_tests import inner1d

In [51]:

```
from collections import Counter

train_distr = Counter(y_train)
train_len = len(y_train)
test_distr = Counter(y_test)
test_len = len(y_test)
```

In [52]:

```
# 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)
    B = (C/C.sum(axis=0))
    A = ((C.T)/(C.sum(axis=1))).T)
    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=labels)
    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=labels)
    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=labels)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.title("Recall matrix")

    plt.show()
```

In [53]:

```
# Logistic Regression

from sklearn.linear_model import SGDClassifier

alpha = [10 ** x for x in range(-5, 2)]

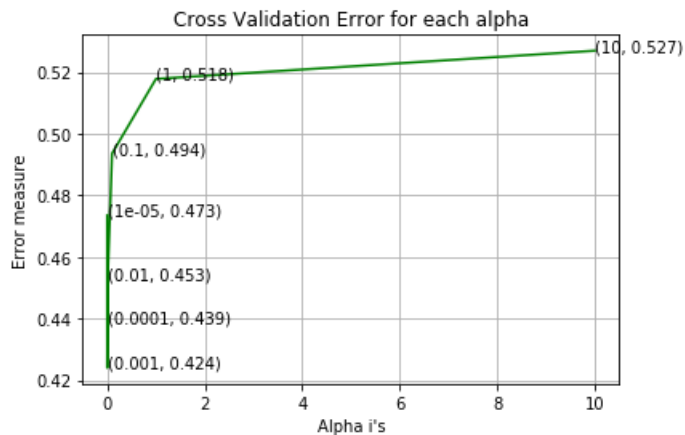
log_error_array=[]
for i in alpha:
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42, class_weight = {0:.1, 1:.9})
    clf.fit(X_train, y_train)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(X_train, y_train)
    predict_y = sig_clf.predict_proba(X_test)
    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, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha = 1e-05 The log loss is: 0.47349575992134885
For values of alpha = 0.0001 The log loss is: 0.4388588184144344
For values of alpha = 0.001 The log loss is: 0.42396911815444327
For values of alpha = 0.01 The log loss is: 0.4525858652857807
For values of alpha = 0.1 The log loss is: 0.4935840951245068
For values of alpha = 1 The log loss is: 0.5178047797661722
For values of alpha = 10 The log loss is: 0.5269573682426553
```

In [54]:

```
#plot errors

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()
```



In [55]:

```
#Logistic regression for best alpha

best_alpha = np.argmin(log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42,class_weight = {
0:.1, 1:.9})
clf.fit(X_train, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(X_train, y_train)
```

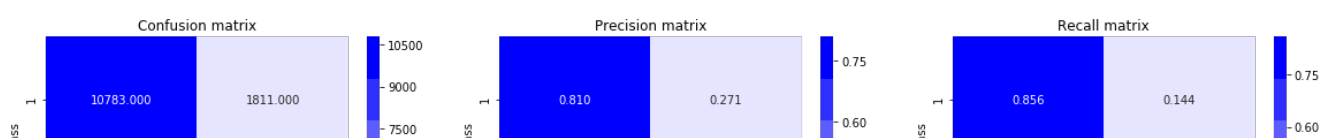
Out[55]:

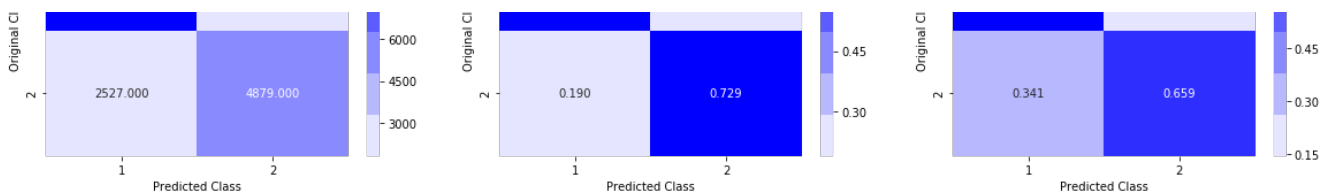
```
CalibratedClassifierCV(base_estimator=SGDClassifier(alpha=0.001, average=False, class_weight={0: 0.1, 1
: 0.9},
    epsilon=0.1, eta0=0.0, fit_intercept=True, l1_ratio=0.15,
    learning_rate='optimal', loss='log', max_iter=None, n_iter=None,
    n_jobs=1, penalty='l2', power_t=0.5, random_state=42, shuffle=True,
    tol=None, verbose=0, warm_start=False),
    cv=3, method='sigmoid')
```

In [56]:

```
predict_y = sig_clf.predict_proba(X_train)
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(X_test)
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))
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 best alpha = 0.001 The train log loss is: 0.4163257224039521
For values of best alpha = 0.001 The test log loss is: 0.42396911815444327
Total number of data points : 20000





In [57]:

```
#Linear-SVM

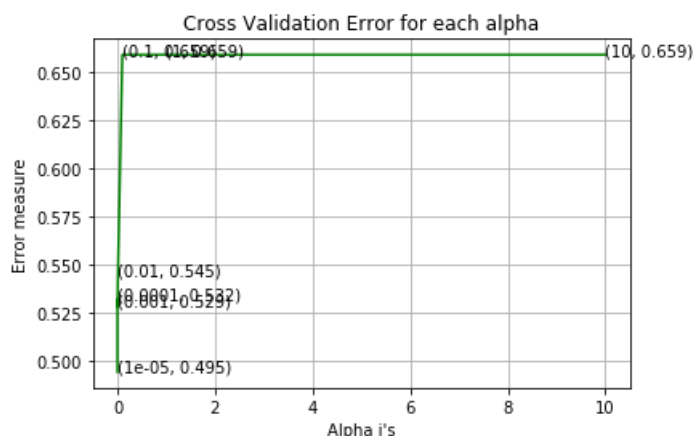
log_error_array=[]

for i in alpha:
    clf = SGDClassifier(alpha=i, penalty='l1', loss='hinge', random_state=42, class_weight = {0:.1, 1:.9})
    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, labels=clf.classes_, eps=1e-15))
```

For values of alpha = 1e-05 The log loss is: 0.4945100447427286
 For values of alpha = 0.0001 The log loss is: 0.5322143762461328
 For values of alpha = 0.001 The log loss is: 0.5286363028921687
 For values of alpha = 0.01 The log loss is: 0.5450504484161294
 For values of alpha = 0.1 The log loss is: 0.6591257509095056
 For values of alpha = 1 The log loss is: 0.6591257509095058
 For values of alpha = 10 The log loss is: 0.6591257509095038

In [58]:

```
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()
```



In [59]:

```
best_alpha = np.argmin(log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l1', loss='hinge', random_state=42, class_weight = {0:.1, 1:.9})
clf.fit(X_train, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(X_train, y_train)
```

Out [59]:

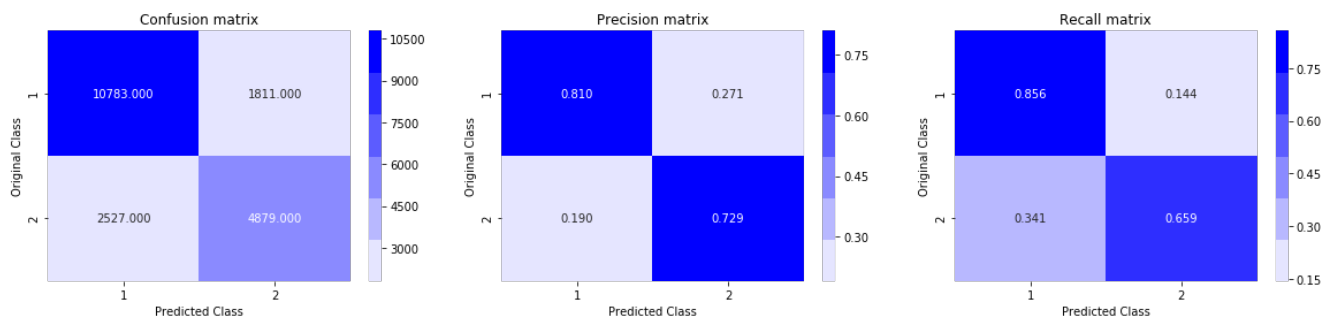

```
calib[0].
```

```
CalibratedClassifierCV(base_estimator=SGDClassifier(alpha=1e-05, average=False, class_weight={0: 0.1, 1: 0.9},
epsilon=0.1, eta0=0.0, fit_intercept=True, l1_ratio=0.15,
learning_rate='optimal', loss='hinge', max_iter=None, n_iter=None,
n_jobs=1, penalty='l1', power_t=0.5, random_state=42, shuffle=True,
tol=None, verbose=0, warm_start=False),
cv=3, method='sigmoid')
```

In [56]:

```
predict_y = sig_clf.predict_proba(X_train)
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(X_test)
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))
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 best alpha = 0.001 The train log loss is: 0.4163257224039521
For values of best alpha = 0.001 The test log loss is: 0.42396911815444327
Total number of data points : 20000



In [62]:

```
max_depth = [4,8,10,50,100]
log_error_array = []

params = {}
params['objective'] = 'binary:logistic'
params['eval_metric'] = 'logloss'
params['eta'] = 0.02

for depth in max_depth:
    params['max_depth'] = depth
    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=10, verbose_eval=10)
    xgdmatrix = xgb.DMatrix(X_train, y_train)
    predict_y = bst.predict(d_test)
    log_error_array.append(log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

[10:33:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max_depth=4

[0] train-logloss:0.684834 valid-logloss:0.684929

Multiple eval metrics have been passed: 'valid-logloss' will be used for early stopping.

Will train until valid-logloss hasn't improved in 10 rounds.

[10:33:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max_depth=4

[10:33:50] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max_depth=4

[10:33:50] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max_depth=4

[10:33:50] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max_depth=4

[10:33:50] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max_depth=4

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```

[10:35:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 24 extra nodes, 0 pruned nodes, max
depth=4
[10:35:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 28 extra nodes, 0 pruned nodes, max
depth=4
[10:35:40] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:40] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:40] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 24 extra nodes, 0 pruned nodes, max
depth=4
[10:35:40] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 28 extra nodes, 0 pruned nodes, max
depth=4
[380] train-logloss:0.353791 valid-logloss:0.359805
[10:35:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 28 extra nodes, 0 pruned nodes, max
depth=4
[10:35:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 20 extra nodes, 0 pruned nodes, max
depth=4
[10:35:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 28 extra nodes, 0 pruned nodes, max
depth=4
[10:35:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:43] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 24 extra nodes, 0 pruned nodes, max
depth=4
[10:35:43] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 18 extra nodes, 0 pruned nodes, max
depth=4
[10:35:43] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:43] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:44] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[390] train-logloss:0.353152 valid-logloss:0.359295
[10:35:44] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:44] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 22 extra nodes, 0 pruned nodes, max
depth=4
[10:35:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 28 extra nodes, 0 pruned nodes, max
depth=4
[10:35:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 30 extra nodes, 0 pruned nodes, max
depth=4
[10:35:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 18 extra nodes, 0 pruned nodes, max
depth=4
[399] train-logloss:0.352435 valid-logloss:0.358713
[10:35:48] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 414 extra nodes, 0 pruned nodes, ma
x_depth=8
[0] train-logloss:0.683053 valid-logloss:0.683333
Multiple eval metrics have been passed: 'valid-logloss' will be used for early stopping.

```

Will train until valid-logloss hasn't improved in 10 rounds.

```

[10:35:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 406 extra nodes, 0 pruned nodes, ma
x_depth=8
[10:35:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 406 extra nodes, 0 pruned nodes, ma
x_depth=8
[10:35:50] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 406 extra nodes, 0 pruned nodes, ma
x_depth=8
[10:35:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 402 extra nodes, 0 pruned nodes, ma
x_depth=8
[10:35:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 402 extra nodes, 0 pruned nodes, ma
x_depth=8
[10:35:52] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 418 extra nodes, 0 pruned nodes, ma
x_depth=8

```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[10:39:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 154 extra nodes, 0 pruned nodes, max_depth=8
[10:39:32] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 188 extra nodes, 0 pruned nodes, max_depth=8
[10:39:32] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 146 extra nodes, 0 pruned nodes, max_depth=8
[10:39:33] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 174 extra nodes, 0 pruned nodes, max_depth=8
[10:39:34] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 178 extra nodes, 0 pruned nodes, max_depth=8
[10:39:34] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 168 extra nodes, 0 pruned nodes, max_depth=8
[10:39:35] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 146 extra nodes, 0 pruned nodes, max_depth=8
[10:39:35] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 102 extra nodes, 0 pruned nodes, max_depth=8
[380] train-logloss:0.298475 valid-logloss:0.337418
[10:39:36] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 126 extra nodes, 0 pruned nodes, max_depth=8
[10:39:36] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 178 extra nodes, 0 pruned nodes, max_depth=8
[10:39:37] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 68 extra nodes, 0 pruned nodes, max_depth=8
[10:39:37] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 150 extra nodes, 0 pruned nodes, max_depth=8
[10:39:38] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 118 extra nodes, 0 pruned nodes, max_depth=8
[10:39:38] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 134 extra nodes, 0 pruned nodes, max_depth=8
[10:39:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 102 extra nodes, 0 pruned nodes, max_depth=8
[10:39:40] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 126 extra nodes, 0 pruned nodes, max_depth=8
[10:39:40] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 162 extra nodes, 0 pruned nodes, max_depth=8
[10:39:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 240 extra nodes, 0 pruned nodes, max_depth=8
[390] train-logloss:0.297587 valid-logloss:0.337141
[10:39:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 176 extra nodes, 0 pruned nodes, max_depth=8
[10:39:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 244 extra nodes, 0 pruned nodes, max_depth=8
[10:39:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 200 extra nodes, 0 pruned nodes, max_depth=8
[10:39:43] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 134 extra nodes, 0 pruned nodes, max_depth=8
[10:39:43] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 70 extra nodes, 0 pruned nodes, max_depth=8
[10:39:44] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 182 extra nodes, 0 pruned nodes, max_depth=8
[10:39:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 138 extra nodes, 0 pruned nodes, max_depth=8
[10:39:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 164 extra nodes, 0 pruned nodes, max_depth=8
[10:39:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 164 extra nodes, 0 pruned nodes, max_depth=8
[399] train-logloss:0.296692 valid-logloss:0.336874
[10:39:48] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 978 extra nodes, 0 pruned nodes, max_depth=10
[0] train-logloss:0.68246 valid-logloss:0.683042
Multiple eval metrics have been passed: 'valid-logloss' will be used for early stopping.
```

Will train until valid-logloss hasn't improved in 10 rounds.

```
[10:39:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 952 extra nodes, 0 pruned nodes, max_depth=10
[10:39:50] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 922 extra nodes, 0 pruned nodes, max_depth=10
[10:39:50] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 940 extra nodes, 0 pruned nodes, max_depth=10
[10:39:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 966 extra nodes, 0 pruned nodes, max_depth=10
[10:39:52] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 936 extra nodes, 0 pruned nodes, max_depth=10
[10:39:53] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 948 extra nodes, 0 pruned nodes, max_depth=10
[10:39:54] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 938 extra nodes, 0 pruned nodes, max_depth=10
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[10:44:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 290 extra nodes, 0 pruned nodes, max_depth=10
[10:44:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 480 extra nodes, 0 pruned nodes, max_depth=10
[10:44:47] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 298 extra nodes, 0 pruned nodes, max_depth=10
[10:44:47] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 314 extra nodes, 0 pruned nodes, max_depth=10
[10:44:48] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 218 extra nodes, 0 pruned nodes, max_depth=10
[10:44:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 358 extra nodes, 0 pruned nodes, max_depth=10
[10:44:50] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 268 extra nodes, 0 pruned nodes, max_depth=10
[380] train-logloss:0.264245 valid-logloss:0.33384
[10:44:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 200 extra nodes, 0 pruned nodes, max_depth=10
[10:44:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 350 extra nodes, 0 pruned nodes, max_depth=10
[10:44:52] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 234 extra nodes, 0 pruned nodes, max_depth=10
[10:44:53] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 244 extra nodes, 0 pruned nodes, max_depth=10
[10:44:53] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 272 extra nodes, 0 pruned nodes, max_depth=10
[10:44:54] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 218 extra nodes, 0 pruned nodes, max_depth=10
[10:44:55] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 436 extra nodes, 0 pruned nodes, max_depth=10
[10:44:56] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 268 extra nodes, 0 pruned nodes, max_depth=10
[10:44:57] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 176 extra nodes, 0 pruned nodes, max_depth=10
[10:44:58] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 306 extra nodes, 0 pruned nodes, max_depth=10
[390] train-logloss:0.263075 valid-logloss:0.333626
[10:44:58] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 204 extra nodes, 0 pruned nodes, max_depth=10
[10:44:59] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 250 extra nodes, 0 pruned nodes, max_depth=10
[10:45:00] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 384 extra nodes, 0 pruned nodes, max_depth=10
[10:45:01] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 274 extra nodes, 0 pruned nodes, max_depth=10
[10:45:01] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 214 extra nodes, 0 pruned nodes, max_depth=10
[10:45:02] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 398 extra nodes, 0 pruned nodes, max_depth=10
[10:45:03] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 268 extra nodes, 0 pruned nodes, max_depth=10
[10:45:04] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 162 extra nodes, 0 pruned nodes, max_depth=10
[10:45:05] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 236 extra nodes, 0 pruned nodes, max_depth=10
[399] train-logloss:0.26204 valid-logloss:0.333445
[10:45:10] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9772 extra nodes, 0 pruned nodes, max_depth=50
[0] train-logloss:0.678703 valid-logloss:0.682639
Multiple eval metrics have been passed: 'valid-logloss' will be used for early stopping.
```

Will train until valid-logloss hasn't improved in 10 rounds.

```
[10:45:13] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8120 extra nodes, 0 pruned nodes, max_depth=39
[10:45:17] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8018 extra nodes, 0 pruned nodes, max_depth=50
[10:45:20] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8134 extra nodes, 0 pruned nodes, max_depth=48
[10:45:24] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8188 extra nodes, 0 pruned nodes, max_depth=48
[10:45:27] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8334 extra nodes, 0 pruned nodes, max_depth=47
[10:45:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8246 extra nodes, 0 pruned nodes, max_depth=50
[10:45:35] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8442 extra nodes, 0 pruned nodes, max_depth=50
[10:45:38] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8484 extra nodes, 0 pruned nodes, max_depth=42
[10:45:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8488 extra nodes, 0 pruned nodes, max_depth=42
```

```
[10:45:42] src/tree/updater_prune.cc:14: tree pruning end, 1 roots, 8488 extra nodes, 0 pruned nodes, max_depth=50
[10:45:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8570 extra nodes, 0 pruned nodes, max_depth=50
[10] train-logloss:0.561066 valid-logloss:0.597536
[10:45:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8556 extra nodes, 0 pruned nodes, max_depth=50
[10:45:52] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8694 extra nodes, 0 pruned nodes, max_depth=50
[10:45:56] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8714 extra nodes, 0 pruned nodes, max_depth=50
[10:46:00] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8770 extra nodes, 0 pruned nodes, max_depth=50
[10:46:03] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8716 extra nodes, 0 pruned nodes, max_depth=49
[10:46:07] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8706 extra nodes, 0 pruned nodes, max_depth=50
[10:46:10] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8778 extra nodes, 0 pruned nodes, max_depth=50
[10:46:14] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8954 extra nodes, 0 pruned nodes, max_depth=50
[10:46:18] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8816 extra nodes, 0 pruned nodes, max_depth=50
[10:46:21] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8954 extra nodes, 0 pruned nodes, max_depth=50
[20] train-logloss:0.472563 valid-logloss:0.537901
[10:46:25] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8924 extra nodes, 0 pruned nodes, max_depth=50
[10:46:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8908 extra nodes, 0 pruned nodes, max_depth=50
[10:46:32] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8948 extra nodes, 0 pruned nodes, max_depth=50
[10:46:36] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9020 extra nodes, 0 pruned nodes, max_depth=50
[10:46:40] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9066 extra nodes, 0 pruned nodes, max_depth=50
[10:46:44] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9092 extra nodes, 0 pruned nodes, max_depth=49
[10:46:47] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9084 extra nodes, 0 pruned nodes, max_depth=50
[10:46:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9134 extra nodes, 0 pruned nodes, max_depth=50
[10:46:54] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9110 extra nodes, 0 pruned nodes, max_depth=46
[10:46:58] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9168 extra nodes, 0 pruned nodes, max_depth=50
[30] train-logloss:0.403127 valid-logloss:0.494293
[10:47:02] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9238 extra nodes, 0 pruned nodes, max_depth=50
[10:47:05] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9178 extra nodes, 0 pruned nodes, max_depth=50
[10:47:09] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9122 extra nodes, 0 pruned nodes, max_depth=50
[10:47:13] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9138 extra nodes, 0 pruned nodes, max_depth=50
[10:47:16] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9284 extra nodes, 0 pruned nodes, max_depth=47
[10:47:20] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9264 extra nodes, 0 pruned nodes, max_depth=48
[10:47:24] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9282 extra nodes, 0 pruned nodes, max_depth=44
[10:47:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9332 extra nodes, 0 pruned nodes, max_depth=50
[10:47:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9192 extra nodes, 0 pruned nodes, max_depth=47
[10:47:35] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9222 extra nodes, 0 pruned nodes, max_depth=44
[40] train-logloss:0.347413 valid-logloss:0.461728
[10:47:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9294 extra nodes, 0 pruned nodes, max_depth=50
[10:47:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9196 extra nodes, 0 pruned nodes, max_depth=45
[10:47:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9168 extra nodes, 0 pruned nodes, max_depth=50
[10:47:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9292 extra nodes, 0 pruned nodes, max_depth=47
[10:47:53] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9238 extra nodes, 0 pruned nodes, max_depth=46
```

```
ax_depth=46
[10:47:57] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9210 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:48:00] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9178 extra nodes, 0 pruned nodes, m
ax_depth=47
[10:48:04] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9174 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:48:08] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9192 extra nodes, 0 pruned nodes, m
ax_depth=46
[10:48:11] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9104 extra nodes, 0 pruned nodes, m
ax_depth=50
[50] train-logloss:0.301931 valid-logloss:0.436443
[10:48:15] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9132 extra nodes, 0 pruned nodes, m
ax_depth=49
[10:48:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9138 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:48:23] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9094 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:48:26] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9104 extra nodes, 0 pruned nodes, m
ax_depth=47
[10:48:30] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9120 extra nodes, 0 pruned nodes, m
ax_depth=45
[10:48:34] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9086 extra nodes, 0 pruned nodes, m
ax_depth=49
[10:48:37] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9064 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:48:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9110 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:48:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9002 extra nodes, 0 pruned nodes, m
ax_depth=49
[10:48:48] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8942 extra nodes, 0 pruned nodes, m
ax_depth=44
[60] train-logloss:0.264433 valid-logloss:0.417245
[10:48:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8934 extra nodes, 0 pruned nodes, m
ax_depth=45
[10:48:55] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8888 extra nodes, 0 pruned nodes, m
ax_depth=42
[10:48:58] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8944 extra nodes, 0 pruned nodes, m
ax_depth=47
[10:49:02] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8788 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:49:06] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8768 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:49:10] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8792 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:49:14] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8760 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:49:18] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8778 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:49:22] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8736 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:49:25] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8678 extra nodes, 0 pruned nodes, m
ax_depth=50
[70] train-logloss:0.233225 valid-logloss:0.40195
[10:49:29] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8628 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:49:33] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8614 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:49:36] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8554 extra nodes, 0 pruned nodes, m
ax_depth=47
[10:49:40] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8600 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:49:44] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8592 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:49:47] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8492 extra nodes, 0 pruned nodes, m
ax_depth=44
[10:49:52] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8406 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:49:56] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8456 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:50:00] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8466 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:50:04] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8500 extra nodes, 0 pruned nodes, m
ax_depth=48
[80] train-logloss:0.207187 valid-logloss:0.390139
[10:50:07] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8412 extra nodes, 0 pruned nodes, m
ax_depth=46
```

```
[10:50:11] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8264 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:50:15] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8274 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:50:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8260 extra nodes, 0 pruned nodes, m
ax_depth=45
[10:50:22] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8254 extra nodes, 0 pruned nodes, m
ax_depth=46
[10:50:26] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8140 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:50:29] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8144 extra nodes, 0 pruned nodes, m
ax_depth=47
[10:50:33] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8154 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:50:37] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8102 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:50:40] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8074 extra nodes, 0 pruned nodes, m
ax_depth=45
[90] train-logloss:0.185059 valid-logloss:0.380851
[10:50:44] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8012 extra nodes, 0 pruned nodes, m
ax_depth=42
[10:50:47] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8042 extra nodes, 0 pruned nodes, m
ax_depth=44
[10:50:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8000 extra nodes, 0 pruned nodes, m
ax_depth=49
[10:50:54] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7908 extra nodes, 0 pruned nodes, m
ax_depth=49
[10:50:58] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7900 extra nodes, 0 pruned nodes, m
ax_depth=49
[10:51:01] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7934 extra nodes, 0 pruned nodes, m
ax_depth=44
[10:51:05] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7836 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:51:09] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7818 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:51:13] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7886 extra nodes, 0 pruned nodes, m
ax_depth=45
[10:51:17] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7878 extra nodes, 0 pruned nodes, m
ax_depth=50
[100] train-logloss:0.16622 valid-logloss:0.373735
[10:51:20] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7776 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:51:24] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7688 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:51:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7756 extra nodes, 0 pruned nodes, m
ax_depth=46
[10:51:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7734 extra nodes, 0 pruned nodes, m
ax_depth=42
[10:51:35] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7628 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:51:38] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7500 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:51:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7520 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:51:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7548 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:51:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7354 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:51:53] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7354 extra nodes, 0 pruned nodes, m
ax_depth=50
[110] train-logloss:0.150195 valid-logloss:0.368138
[10:51:57] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7396 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:52:00] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7362 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:52:04] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7368 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:52:08] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7368 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:52:12] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7380 extra nodes, 0 pruned nodes, m
ax_depth=42
[10:52:15] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7292 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:52:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7298 extra nodes, 0 pruned nodes, m
ax_depth=47
[10:52:22] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7280 extra nodes, 0 pruned nodes, m
ax_depth=43
```

[10:52:26] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7144 extra nodes, 0 pruned nodes, max_depth=50
[10:52:30] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7198 extra nodes, 0 pruned nodes, max_depth=50
[120] train-logloss:0.136455 valid-logloss:0.363799
[10:52:34] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7162 extra nodes, 0 pruned nodes, max_depth=48
[10:52:37] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7166 extra nodes, 0 pruned nodes, max_depth=50
[10:52:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7062 extra nodes, 0 pruned nodes, max_depth=50
[10:52:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7130 extra nodes, 0 pruned nodes, max_depth=49
[10:52:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7088 extra nodes, 0 pruned nodes, max_depth=50
[10:52:53] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6942 extra nodes, 0 pruned nodes, max_depth=50
[10:52:56] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6974 extra nodes, 0 pruned nodes, max_depth=50
[10:53:00] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6782 extra nodes, 0 pruned nodes, max_depth=50
[10:53:04] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6954 extra nodes, 0 pruned nodes, max_depth=45
[10:53:08] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6984 extra nodes, 0 pruned nodes, max_depth=49
[130] train-logloss:0.12458 valid-logloss:0.360367
[10:53:11] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6818 extra nodes, 0 pruned nodes, max_depth=50
[10:53:15] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6902 extra nodes, 0 pruned nodes, max_depth=50
[10:53:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6886 extra nodes, 0 pruned nodes, max_depth=47
[10:53:23] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6844 extra nodes, 0 pruned nodes, max_depth=50
[10:53:26] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6904 extra nodes, 0 pruned nodes, max_depth=48
[10:53:30] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6732 extra nodes, 0 pruned nodes, max_depth=50
[10:53:34] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6756 extra nodes, 0 pruned nodes, max_depth=50
[10:53:37] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6750 extra nodes, 0 pruned nodes, max_depth=49
[10:53:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6752 extra nodes, 0 pruned nodes, max_depth=50
[10:53:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6710 extra nodes, 0 pruned nodes, max_depth=50
[140] train-logloss:0.114254 valid-logloss:0.358117
[10:53:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6568 extra nodes, 0 pruned nodes, max_depth=50
[10:53:52] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6544 extra nodes, 0 pruned nodes, max_depth=48
[10:53:56] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6544 extra nodes, 0 pruned nodes, max_depth=50
[10:53:59] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6544 extra nodes, 0 pruned nodes, max_depth=50
[10:54:03] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6564 extra nodes, 0 pruned nodes, max_depth=46
[10:54:06] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6450 extra nodes, 0 pruned nodes, max_depth=50
[10:54:10] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6542 extra nodes, 0 pruned nodes, max_depth=48
[10:54:13] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6538 extra nodes, 0 pruned nodes, max_depth=50
[10:54:17] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6386 extra nodes, 0 pruned nodes, max_depth=44
[10:54:20] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6468 extra nodes, 0 pruned nodes, max_depth=44
[150] train-logloss:0.105152 valid-logloss:0.356266
[10:54:24] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6308 extra nodes, 0 pruned nodes, max_depth=50
[10:54:27] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6458 extra nodes, 0 pruned nodes, max_depth=49
[10:54:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6350 extra nodes, 0 pruned nodes, max_depth=50
[10:54:35] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6306 extra nodes, 0 pruned nodes, max_depth=50
[10:54:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5828 extra nodes, 0 pruned nodes, max_depth=50

```
ax_depth=50
[10:54:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6194 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:54:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6218 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:54:50] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6232 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:54:54] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6174 extra nodes, 0 pruned nodes, m
ax_depth=49
[10:54:58] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6096 extra nodes, 0 pruned nodes, m
ax_depth=50
[160] train-logloss:0.097246 valid-logloss:0.354884
[10:55:01] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6050 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:55:05] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6062 extra nodes, 0 pruned nodes, m
ax_depth=44
[10:55:09] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6060 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:55:12] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6044 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:55:16] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5908 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:55:20] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6080 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:55:24] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5982 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:55:27] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5960 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:55:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5846 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:55:35] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5860 extra nodes, 0 pruned nodes, m
ax_depth=50
[170] train-logloss:0.090221 valid-logloss:0.353763
[10:55:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5868 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:55:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5650 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:55:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5836 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:55:50] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5834 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:55:54] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5688 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:55:58] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5716 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:56:01] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5806 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:56:05] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5726 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:56:09] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5614 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:56:12] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5698 extra nodes, 0 pruned nodes, m
ax_depth=46
[180] train-logloss:0.084025 valid-logloss:0.353163
[10:56:16] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5536 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:56:20] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5614 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:56:24] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5574 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:56:29] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5542 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:56:33] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5446 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:56:37] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5564 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:56:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5370 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:56:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5480 extra nodes, 0 pruned nodes, m
ax_depth=49
[10:56:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5402 extra nodes, 0 pruned nodes, m
ax_depth=49
[10:56:52] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5366 extra nodes, 0 pruned nodes, m
ax_depth=49
[190] train-logloss:0.078444 valid-logloss:0.3525
[10:56:56] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5372 extra nodes, 0 pruned nodes, m
ax_depth=50
```

```
[10:57:00] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5330 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:04] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5344 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:08] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4984 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:12] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5108 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:16] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5206 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:20] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5292 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:24] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5254 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4806 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5052 extra nodes, 0 pruned nodes, m
ax_depth=50
[200] train-logloss:0.073586 valid-logloss:0.352224
[10:57:35] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5202 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5120 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:43] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5104 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:47] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4924 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5018 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:57:56] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5072 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:58:00] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5130 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:58:03] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5000 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:58:07] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4990 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:58:11] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4968 extra nodes, 0 pruned nodes, m
ax_depth=50
[210] train-logloss:0.069184 valid-logloss:0.352262
[10:58:15] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4956 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:58:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5014 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:58:23] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5026 extra nodes, 0 pruned nodes, m
ax_depth=50
[10:58:27] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4762 extra nodes, 0 pruned nodes, m
ax_depth=50
Stopping. Best iteration:
[204] train-logloss:0.071771 valid-logloss:0.35216

[10:58:33] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9776 extra nodes, 0 pruned nodes, m
ax_depth=52
[0] train-logloss:0.678703 valid-logloss:0.682641
Multiple eval metrics have been passed: 'valid-logloss' will be used for early stopping.

Will train until valid-logloss hasn't improved in 10 rounds.
[10:58:36] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8120 extra nodes, 0 pruned nodes, m
ax_depth=39
[10:58:40] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8084 extra nodes, 0 pruned nodes, m
ax_depth=70
[10:58:44] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8140 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:58:47] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8194 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:58:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8332 extra nodes, 0 pruned nodes, m
ax_depth=47
[10:58:55] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8298 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:58:59] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8418 extra nodes, 0 pruned nodes, m
ax_depth=60
[10:59:02] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8492 extra nodes, 0 pruned nodes, m
ax_depth=43
[10:59:06] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8472 extra nodes, 0 pruned nodes, m
ax_depth=48
[10:59:10] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8570 extra nodes, 0 pruned nodes, m
ax_depth=50
```



```
[10] train-logloss:0.56104 valid-logloss:0.597545
[10:59:15] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8568 extra nodes, 0 pruned nodes, m
ax_depth=52
[10:59:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8730 extra nodes, 0 pruned nodes, m
ax_depth=55
[10:59:23] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8700 extra nodes, 0 pruned nodes, m
ax_depth=52
[10:59:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8752 extra nodes, 0 pruned nodes, m
ax_depth=62
[10:59:32] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8680 extra nodes, 0 pruned nodes, m
ax_depth=58
[10:59:36] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8756 extra nodes, 0 pruned nodes, m
ax_depth=54
[10:59:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8812 extra nodes, 0 pruned nodes, m
ax_depth=53
[10:59:43] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8976 extra nodes, 0 pruned nodes, m
ax_depth=56
[10:59:48] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8908 extra nodes, 0 pruned nodes, m
ax_depth=62
[10:59:52] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8960 extra nodes, 0 pruned nodes, m
ax_depth=56
[20] train-logloss:0.472533 valid-logloss:0.537902
[10:59:56] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8942 extra nodes, 0 pruned nodes, m
ax_depth=54
[11:00:00] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8924 extra nodes, 0 pruned nodes, m
ax_depth=48
[11:00:04] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8970 extra nodes, 0 pruned nodes, m
ax_depth=49
[11:00:08] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8980 extra nodes, 0 pruned nodes, m
ax_depth=54
[11:00:12] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9098 extra nodes, 0 pruned nodes, m
ax_depth=58
[11:00:16] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9102 extra nodes, 0 pruned nodes, m
ax_depth=50
[11:00:20] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9144 extra nodes, 0 pruned nodes, m
ax_depth=57
[11:00:24] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9146 extra nodes, 0 pruned nodes, m
ax_depth=59
[11:00:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9118 extra nodes, 0 pruned nodes, m
ax_depth=46
[11:00:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9200 extra nodes, 0 pruned nodes, m
ax_depth=50
[30] train-logloss:0.403054 valid-logloss:0.494231
[21:48:15] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9242 extra nodes, 0 pruned nodes, m
ax_depth=50
[21:48:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9208 extra nodes, 0 pruned nodes, m
ax_depth=52
[21:48:21] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9212 extra nodes, 0 pruned nodes, m
ax_depth=50
[21:48:23] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9162 extra nodes, 0 pruned nodes, m
ax_depth=54
[21:48:26] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9282 extra nodes, 0 pruned nodes, m
ax_depth=46
[21:48:29] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9246 extra nodes, 0 pruned nodes, m
ax_depth=52
[21:48:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9228 extra nodes, 0 pruned nodes, m
ax_depth=48
[21:48:34] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9254 extra nodes, 0 pruned nodes, m
ax_depth=62
[21:48:36] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9196 extra nodes, 0 pruned nodes, m
ax_depth=47
[21:48:38] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9290 extra nodes, 0 pruned nodes, m
ax_depth=53
[40] train-logloss:0.3473 valid-logloss:0.461656
[21:48:40] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9296 extra nodes, 0 pruned nodes, m
ax_depth=53
[21:48:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9220 extra nodes, 0 pruned nodes, m
ax_depth=47
[21:48:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9258 extra nodes, 0 pruned nodes, m
ax_depth=57
[21:48:47] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9236 extra nodes, 0 pruned nodes, m
ax_depth=54
[21:48:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9222 extra nodes, 0 pruned nodes, m
ax_depth=49
[21:48:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9182 extra nodes, 0 pruned nodes, m
ax_depth=51
[21:48:53] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9214 extra nodes, 0 pruned nodes, m
```

ax_depth=47
[21:48:55] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9240 extra nodes, 0 pruned nodes, m
ax_depth=48
[21:48:57] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9152 extra nodes, 0 pruned nodes, m
ax_depth=55
[21:48:59] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9136 extra nodes, 0 pruned nodes, m
ax_depth=41
[50] train-logloss:0.301814 valid-logloss:0.43639
[21:49:01] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9110 extra nodes, 0 pruned nodes, m
ax_depth=47
[21:49:03] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9178 extra nodes, 0 pruned nodes, m
ax_depth=44
[21:49:04] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9158 extra nodes, 0 pruned nodes, m
ax_depth=49
[21:49:06] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9084 extra nodes, 0 pruned nodes, m
ax_depth=60
[21:49:08] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9194 extra nodes, 0 pruned nodes, m
ax_depth=44
[21:49:10] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9034 extra nodes, 0 pruned nodes, m
ax_depth=45
[21:49:12] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9084 extra nodes, 0 pruned nodes, m
ax_depth=45
[21:49:13] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9086 extra nodes, 0 pruned nodes, m
ax_depth=42
[21:49:15] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8932 extra nodes, 0 pruned nodes, m
ax_depth=49
[21:49:17] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 9028 extra nodes, 0 pruned nodes, m
ax_depth=43
[60] train-logloss:0.264338 valid-logloss:0.41698
[21:49:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8936 extra nodes, 0 pruned nodes, m
ax_depth=60
[21:49:21] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8956 extra nodes, 0 pruned nodes, m
ax_depth=43
[21:49:23] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8882 extra nodes, 0 pruned nodes, m
ax_depth=46
[21:49:25] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8800 extra nodes, 0 pruned nodes, m
ax_depth=54
[21:49:27] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8780 extra nodes, 0 pruned nodes, m
ax_depth=50
[21:49:29] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8818 extra nodes, 0 pruned nodes, m
ax_depth=48
[21:49:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8752 extra nodes, 0 pruned nodes, m
ax_depth=48
[21:49:33] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8798 extra nodes, 0 pruned nodes, m
ax_depth=50
[21:49:35] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8818 extra nodes, 0 pruned nodes, m
ax_depth=50
[21:49:37] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8670 extra nodes, 0 pruned nodes, m
ax_depth=50
[70] train-logloss:0.233139 valid-logloss:0.401549
[21:49:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8624 extra nodes, 0 pruned nodes, m
ax_depth=57
[21:49:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8584 extra nodes, 0 pruned nodes, m
ax_depth=53
[21:49:43] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8598 extra nodes, 0 pruned nodes, m
ax_depth=51
[21:49:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8552 extra nodes, 0 pruned nodes, m
ax_depth=51
[21:49:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8502 extra nodes, 0 pruned nodes, m
ax_depth=45
[21:49:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8470 extra nodes, 0 pruned nodes, m
ax_depth=49
[21:49:52] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8472 extra nodes, 0 pruned nodes, m
ax_depth=51
[21:49:56] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8454 extra nodes, 0 pruned nodes, m
ax_depth=57
[21:50:00] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8486 extra nodes, 0 pruned nodes, m
ax_depth=50
[21:50:04] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8412 extra nodes, 0 pruned nodes, m
ax_depth=50
[80] train-logloss:0.207055 valid-logloss:0.389777
[21:50:08] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8318 extra nodes, 0 pruned nodes, m
ax_depth=53
[21:50:12] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8280 extra nodes, 0 pruned nodes, m
ax_depth=59
[21:50:17] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8274 extra nodes, 0 pruned nodes, m
ax_depth=65

[21:50:21] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8232 extra nodes, 0 pruned nodes, max_depth=59
[21:50:26] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8252 extra nodes, 0 pruned nodes, max_depth=57
[21:50:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8164 extra nodes, 0 pruned nodes, max_depth=69
[21:50:35] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8136 extra nodes, 0 pruned nodes, max_depth=56
[21:50:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8170 extra nodes, 0 pruned nodes, max_depth=51
[21:50:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8146 extra nodes, 0 pruned nodes, max_depth=46
[21:50:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 8154 extra nodes, 0 pruned nodes, max_depth=43
[90] train-logloss:0.184941 valid-logloss:0.380157
[21:50:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7988 extra nodes, 0 pruned nodes, max_depth=42
[21:50:53] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7970 extra nodes, 0 pruned nodes, max_depth=42
[21:50:56] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7974 extra nodes, 0 pruned nodes, max_depth=45
[21:51:00] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7906 extra nodes, 0 pruned nodes, max_depth=51
[21:51:04] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7918 extra nodes, 0 pruned nodes, max_depth=46
[21:51:08] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7958 extra nodes, 0 pruned nodes, max_depth=64
[21:51:12] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7918 extra nodes, 0 pruned nodes, max_depth=47
[21:51:16] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7888 extra nodes, 0 pruned nodes, max_depth=54
[21:51:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7792 extra nodes, 0 pruned nodes, max_depth=44
[21:51:23] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7872 extra nodes, 0 pruned nodes, max_depth=51
[100] train-logloss:0.16611 valid-logloss:0.372839
[21:51:27] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7796 extra nodes, 0 pruned nodes, max_depth=53
[21:51:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7656 extra nodes, 0 pruned nodes, max_depth=61
[21:51:36] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7750 extra nodes, 0 pruned nodes, max_depth=65
[21:51:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7754 extra nodes, 0 pruned nodes, max_depth=48
[21:51:43] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7750 extra nodes, 0 pruned nodes, max_depth=49
[21:51:47] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7562 extra nodes, 0 pruned nodes, max_depth=45
[21:51:50] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7542 extra nodes, 0 pruned nodes, max_depth=51
[21:51:55] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7532 extra nodes, 0 pruned nodes, max_depth=63
[21:51:58] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7604 extra nodes, 0 pruned nodes, max_depth=50
[21:52:02] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7528 extra nodes, 0 pruned nodes, max_depth=47
[110] train-logloss:0.150074 valid-logloss:0.36736
[21:52:06] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7474 extra nodes, 0 pruned nodes, max_depth=45
[21:52:10] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7486 extra nodes, 0 pruned nodes, max_depth=52
[21:52:13] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7384 extra nodes, 0 pruned nodes, max_depth=50
[21:52:17] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7424 extra nodes, 0 pruned nodes, max_depth=49
[21:52:21] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7396 extra nodes, 0 pruned nodes, max_depth=63
[21:52:25] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7338 extra nodes, 0 pruned nodes, max_depth=47
[21:52:29] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7308 extra nodes, 0 pruned nodes, max_depth=54
[21:52:33] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7308 extra nodes, 0 pruned nodes, max_depth=52
[21:52:36] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7212 extra nodes, 0 pruned nodes, max_depth=52
[21:52:40] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7226 extra nodes, 0 pruned nodes, max_depth=53

```
[120] train-logloss:0.136308 valid-logloss:0.363204
[21:52:44] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7206 extra nodes, 0 pruned nodes, m
ax_depth=51
[21:52:48] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7196 extra nodes, 0 pruned nodes, m
ax_depth=57
[21:52:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7154 extra nodes, 0 pruned nodes, m
ax_depth=49
[21:52:55] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7054 extra nodes, 0 pruned nodes, m
ax_depth=53
[21:52:59] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7024 extra nodes, 0 pruned nodes, m
ax_depth=56
[21:53:03] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7014 extra nodes, 0 pruned nodes, m
ax_depth=54
[21:53:07] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 7012 extra nodes, 0 pruned nodes, m
ax_depth=62
[21:53:12] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6928 extra nodes, 0 pruned nodes, m
ax_depth=55
[21:53:15] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6958 extra nodes, 0 pruned nodes, m
ax_depth=46
[21:53:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6966 extra nodes, 0 pruned nodes, m
ax_depth=50
[130] train-logloss:0.124389 valid-logloss:0.359928
[21:53:23] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6894 extra nodes, 0 pruned nodes, m
ax_depth=61
[21:53:26] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6960 extra nodes, 0 pruned nodes, m
ax_depth=46
[21:53:30] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6882 extra nodes, 0 pruned nodes, m
ax_depth=49
[21:53:34] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6870 extra nodes, 0 pruned nodes, m
ax_depth=45
[21:53:38] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6926 extra nodes, 0 pruned nodes, m
ax_depth=50
[21:53:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6882 extra nodes, 0 pruned nodes, m
ax_depth=59
[21:53:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6810 extra nodes, 0 pruned nodes, m
ax_depth=59
[21:53:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6786 extra nodes, 0 pruned nodes, m
ax_depth=50
[21:53:53] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6754 extra nodes, 0 pruned nodes, m
ax_depth=50
[21:53:57] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6618 extra nodes, 0 pruned nodes, m
ax_depth=45
[140] train-logloss:0.114098 valid-logloss:0.35749
[21:54:01] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6586 extra nodes, 0 pruned nodes, m
ax_depth=55
[21:54:04] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6640 extra nodes, 0 pruned nodes, m
ax_depth=49
[21:54:08] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6580 extra nodes, 0 pruned nodes, m
ax_depth=46
[21:54:12] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6544 extra nodes, 0 pruned nodes, m
ax_depth=54
[21:54:15] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6572 extra nodes, 0 pruned nodes, m
ax_depth=53
[21:54:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6528 extra nodes, 0 pruned nodes, m
ax_depth=53
[21:54:23] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6626 extra nodes, 0 pruned nodes, m
ax_depth=49
[21:54:27] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6450 extra nodes, 0 pruned nodes, m
ax_depth=55
[21:54:30] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6410 extra nodes, 0 pruned nodes, m
ax_depth=42
[21:54:33] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6438 extra nodes, 0 pruned nodes, m
ax_depth=48
[150] train-logloss:0.105032 valid-logloss:0.35578
[21:54:37] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6410 extra nodes, 0 pruned nodes, m
ax_depth=58
[21:54:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6364 extra nodes, 0 pruned nodes, m
ax_depth=46
[21:54:44] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6370 extra nodes, 0 pruned nodes, m
ax_depth=43
[21:54:48] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6318 extra nodes, 0 pruned nodes, m
ax_depth=49
[21:54:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6256 extra nodes, 0 pruned nodes, m
ax_depth=46
[21:54:55] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6330 extra nodes, 0 pruned nodes, m
ax_depth=57
[21:54:59] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6238 extra nodes, 0 pruned nodes, m
```

ax_depth=47
[21:55:03] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6144 extra nodes, 0 pruned nodes, m
ax_depth=55
[21:55:07] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6136 extra nodes, 0 pruned nodes, m
ax_depth=64
[21:55:11] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6206 extra nodes, 0 pruned nodes, m
ax_depth=49
[160] train-logloss:0.097096 valid-logloss:0.354354
[21:55:15] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6102 extra nodes, 0 pruned nodes, m
ax_depth=56
[21:55:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6138 extra nodes, 0 pruned nodes, m
ax_depth=51
[21:55:23] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6028 extra nodes, 0 pruned nodes, m
ax_depth=56
[21:55:27] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6042 extra nodes, 0 pruned nodes, m
ax_depth=50
[21:55:30] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6080 extra nodes, 0 pruned nodes, m
ax_depth=52
[21:55:35] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6034 extra nodes, 0 pruned nodes, m
ax_depth=63
[21:55:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6006 extra nodes, 0 pruned nodes, m
ax_depth=59
[21:55:43] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5960 extra nodes, 0 pruned nodes, m
ax_depth=70
[21:55:47] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5894 extra nodes, 0 pruned nodes, m
ax_depth=48
[21:55:51] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5878 extra nodes, 0 pruned nodes, m
ax_depth=58
[170] train-logloss:0.090081 valid-logloss:0.353348
[21:55:55] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5780 extra nodes, 0 pruned nodes, m
ax_depth=52
[21:55:59] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5800 extra nodes, 0 pruned nodes, m
ax_depth=63
[21:56:03] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5832 extra nodes, 0 pruned nodes, m
ax_depth=55
[21:56:08] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5832 extra nodes, 0 pruned nodes, m
ax_depth=60
[21:56:11] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5784 extra nodes, 0 pruned nodes, m
ax_depth=53
[21:56:16] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5806 extra nodes, 0 pruned nodes, m
ax_depth=71
[21:56:20] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5784 extra nodes, 0 pruned nodes, m
ax_depth=56
[21:56:25] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5714 extra nodes, 0 pruned nodes, m
ax_depth=64
[21:56:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5664 extra nodes, 0 pruned nodes, m
ax_depth=49
[21:56:33] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5574 extra nodes, 0 pruned nodes, m
ax_depth=58
[180] train-logloss:0.083874 valid-logloss:0.352549
[21:56:37] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5578 extra nodes, 0 pruned nodes, m
ax_depth=54
[21:56:41] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5624 extra nodes, 0 pruned nodes, m
ax_depth=53
[21:56:45] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5648 extra nodes, 0 pruned nodes, m
ax_depth=51
[21:56:49] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5590 extra nodes, 0 pruned nodes, m
ax_depth=55
[21:56:53] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5466 extra nodes, 0 pruned nodes, m
ax_depth=51
[21:56:56] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5516 extra nodes, 0 pruned nodes, m
ax_depth=51
[21:57:01] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5556 extra nodes, 0 pruned nodes, m
ax_depth=60
[21:57:05] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5482 extra nodes, 0 pruned nodes, m
ax_depth=60
[21:57:09] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5564 extra nodes, 0 pruned nodes, m
ax_depth=54
[21:57:13] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5340 extra nodes, 0 pruned nodes, m
ax_depth=51
[190] train-logloss:0.078335 valid-logloss:0.352058
[21:57:17] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5358 extra nodes, 0 pruned nodes, m
ax_depth=55
[21:57:21] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5330 extra nodes, 0 pruned nodes, m
ax_depth=47
[21:57:25] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5306 extra nodes, 0 pruned nodes, m
ax_depth=57

```

[21:57:29] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5306 extra nodes, 0 pruned nodes, m
ax_depth=64
[21:57:33] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5238 extra nodes, 0 pruned nodes, m
ax_depth=58
[21:57:37] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5306 extra nodes, 0 pruned nodes, m
ax_depth=52
[21:57:42] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5316 extra nodes, 0 pruned nodes, m
ax_depth=65
[21:57:46] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5234 extra nodes, 0 pruned nodes, m
ax_depth=55
[21:57:50] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5244 extra nodes, 0 pruned nodes, m
ax_depth=48
[21:57:54] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5264 extra nodes, 0 pruned nodes, m
ax_depth=61
[200] train-logloss:0.073377 valid-logloss:0.351677
[21:57:58] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5176 extra nodes, 0 pruned nodes, m
ax_depth=48
[21:58:02] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5214 extra nodes, 0 pruned nodes, m
ax_depth=53
[21:58:07] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5246 extra nodes, 0 pruned nodes, m
ax_depth=69
[21:58:11] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5240 extra nodes, 0 pruned nodes, m
ax_depth=47
[21:58:15] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5124 extra nodes, 0 pruned nodes, m
ax_depth=53
[21:58:19] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5136 extra nodes, 0 pruned nodes, m
ax_depth=60
[21:58:23] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5128 extra nodes, 0 pruned nodes, m
ax_depth=54
[21:58:27] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5116 extra nodes, 0 pruned nodes, m
ax_depth=57
[21:58:31] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5108 extra nodes, 0 pruned nodes, m
ax_depth=56
[21:58:35] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4974 extra nodes, 0 pruned nodes, m
ax_depth=55
[210] train-logloss:0.068932 valid-logloss:0.351435
[21:58:39] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5042 extra nodes, 0 pruned nodes, m
ax_depth=59
[21:58:44] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5012 extra nodes, 0 pruned nodes, m
ax_depth=57
[21:58:48] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4962 extra nodes, 0 pruned nodes, m
ax_depth=58
[21:58:52] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5036 extra nodes, 0 pruned nodes, m
ax_depth=60
[21:58:57] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 5030 extra nodes, 0 pruned nodes, m
ax_depth=70
[21:59:01] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4916 extra nodes, 0 pruned nodes, m
ax_depth=50
[21:59:05] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4956 extra nodes, 0 pruned nodes, m
ax_depth=57
[21:59:09] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4874 extra nodes, 0 pruned nodes, m
ax_depth=67
Stopping. Best iteration:
[208] train-logloss:0.069792 valid-logloss:0.351432

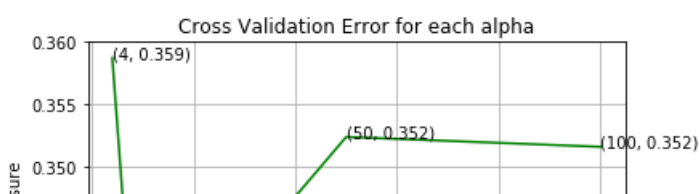
```

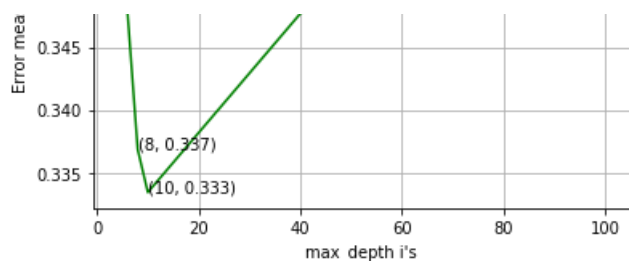
In [63]:

```

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

```





In []:

```
params = {}
params['objective'] = 'binary:logistic'
params['eval_metric'] = 'logloss'
params['eta'] = 0.02
params['max_depth'] = max_depth[np.argmin(log_error_array)]

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')]

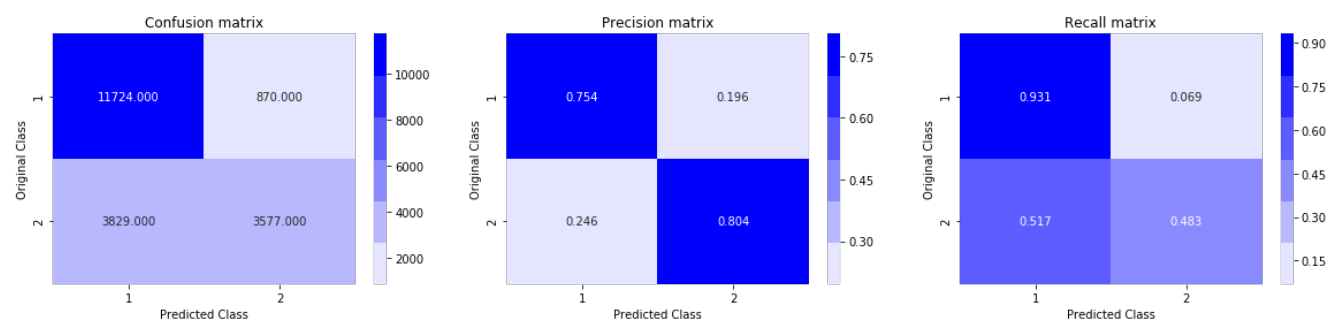
bst = xgb.train(params, d_train, 400, watchlist, early_stopping_rounds=20, verbose_eval=10)

xgdmatrix = xgb.DMatrix(X_train, y_train)
predict_y_train = bst.predict(d_train)
predict_y_test = bst.predict(d_test)
```

In [65]:

```
print('For values of best depth = ', max_depth[np.argmin(log_error_array)] , "The train log loss is:", log_loss(y_train, predict_y_train, labels=clf.classes_, eps=1e-15))
print('For values of best depth = ', max_depth[np.argmin(log_error_array)] , "The test log loss is:", log_loss(y_test, predict_y_test, labels=clf.classes_, eps=1e-15))
plot_confusion_matrix(y_test, predicted_y)
```

For values of best depth = 10 The train log loss is: 0.2620397457373918
For values of best depth = 10 The test log loss is: 0.3334446800083824



Summary:

1. Understand business problem
2. Jot down Business objectives and constraints
3. Data Information
4. Mapping to ML.
5. Deciding performance metrics
6. Load data
7. Exploratory Data Analysis without pre-processing
 - Basic data info
 - Distribution of class labels
 - Unique questions
 - Checking for duplicates
 - Checking for null values
 - filling null values with empty strings

8. Feature Extraction

- freq_qid1 = Frequency of qid1's
- freq_qid2= Frequency of qid2's
- q1len = Length of q1
- q2len = Length of q2
- q1_n_words = Number of words in Question 1
- q2_n_words = Number of words in Question 2
- word_Common = (Number of common unique words in Question 1 and Question 2)
- word_Total =(Total num of words in Question 1 + Total num of words in Question 2)
- word_share = (word_common)/(word_Total)
- freq_q1+freq_q2 = sum total of frequency of qid1 and qid2
- freq_q1-freq_q2 = absolute difference of frequency of qid1 and qid2

9. Text Pre-processing

- Removing html tags
- Removing Punctuations
- Performing stemming
- Removing Stopwords
- Expanding contractions

10. Advanced Feature Extraction

- cwc_min : Ratio of common_word_count to min length of word count of Q1 and Q2
- cwc_max : Ratio of common_word_count to max length of word count of Q1 and Q2
- csc_min : Ratio of common_stop_count to min length of stop count of Q1 and Q2
- csc_max : Ratio of common_stop_count to max length of stop count of Q1 and Q2
- ctc_min : Ratio of common_token_count to min length of token count of Q1 and Q2
- ctc_max : Ratio of common_token_count to max length of token count of Q1 and Q2
- last_word_eq : Check if First word of both questions is equal or not
- first_word_eq : Check if First word of both questions is equal or not
- abs_len_diff : Abs. length difference
- mean_len : Average Token Length of both Questions
- fuzz_ratio : measurement of edit distance-compare two strings
- fuzz_partial_ratio : measurement of edit distance-compare two strings partially
- token_set_ratio : tokenize both strings.split the tokens into two groups: intersection and remainder.use those sets to build up a comparison string.
- token_sort_ratio : tokenizing the string in question, sorting the tokens alphabetically, and then joining them back into a string We then compare the transformed strings with a simple ratio().

11. Exploratory Data Analysis on Advanced Features.

- pair plot on four features(ctc_min', 'cwc_min', 'csc_min', 'token_sort_ratio).From the plots given below it is very clear tat certain features gives us a clear cut value on deciding similarity.
- Univariate Analysis on token_sort_ratio and fuzz_ratio

12. Featurization(tf-idf weighted word to vec)

- tf-idf vectorization question1
- tf-idf vectorization question2

13. Data Preparation(Load final_features.csv - 226)

- Load nlp_features_train.csv (NLP Features - 17)
- Load prepro_features_train.csv (Simple Preprocessing Featrures - 12)
- tf-idf vectorization question1(100)
- tf-idf vectorization question2(100)

14. ML models

- Logistic Regression
- Linear SVM
- XGBoost

15. Conclusion:

Model	train log-loss	test log-loss
Logistic Regression	0.4342389101511353	0.4421663977505779
Linear SVM	0.46955106918493666	0.4800812551095535
XGBoost	0.2620397457373918	0.33344952850797854