Assignment-5 Apply Logistic Regression on Donors Choose dataset.

In [1]:

```
import warnings
warnings.filterwarnings("ignore")
import sqlite3
import pandas as pd
import numpy as np
import nltk
import string
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import roc_curve, auc
from nltk.stem.porter import PorterStemmer
import re
import string
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer
from gensim.models import Word2Vec
from gensim.models import KeyedVectors
import pickle
from tqdm import tqdm
import os
from plotly import plotly
import plotly.offline as offline
import plotly.graph_objs as go
offline.init_notebook_mode()
from collections import Counter
```

1.1 Loading Data

In [2]:

```
data = pd.read_csv('preprocessed_data.csv', nrows=50000)
data.head(2)
```

Out[2]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_s
0	160221	p253737	c90749f5d961ff158d4b4d1e7dc665fc	Mrs.	IN
1	140945	p258326	897464ce9ddc600bced1151f324dd63a	Mr.	FL

2 rows × 29 columns

In [4]:

```
data['project_is_approved'].value_counts()
```

Out[4]:

```
1    42286
0     7714
Name: project_is_approved, dtype: int64
```

In [5]:

```
y = data['project_is_approved'].values
X = data.drop(['project_is_approved'], axis=1)
X.head(1)
```

Out[5]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_sta
0	160221	p253737	c90749f5d961ff158d4b4d1e7dc665fc	Mrs.	IN

1 rows × 28 columns

1.2 Splitting data into Train and cross validation(or test): Stratified Sampling

In [6]:

```
# train test split
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, stratify=y)
X_train, X_cv, y_train, y_cv = train_test_split(X_train, y_train, test_size=0.33, strat
ify=y_train)
```

1.3 Make Data Model Ready: encoding essay, and project_title

1.3.1 Vectorizing preprocessed essays & project_title using BOW

In [7]:

```
# preprocessed essays
print(X_train.shape, y_train.shape)
print(X_cv.shape, y_cv.shape)
print(X_test.shape, y_test.shape)

print("="*100)


vectorizer = CountVectorizer(min_df=10,ngram_range=(1,2), max_features=5000)
vectorizer.fit(X_train['preprocessed_essays'].values)  # fit has to happen only on trai
n data

# we use the fit CountVectorizer to convert the text to vector
X_train_essay_bow = vectorizer.transform(X_train['preprocessed_essays'].values)
X_cv_essay_bow = vectorizer.transform(X_cv['preprocessed_essays'].values)
X_test_essay_bow = vectorizer.transform(X_test['preprocessed_essays'].values)
```

```
(22445, 28) (22445,)
(11055, 28) (11055,)
(16500, 28) (16500,)
==========================================================================
==========================
```

In [8]:

```
print("After vectorization")
print(X_train_essay_bow.shape, y_train.shape)
print(X_cv_essay_bow.shape, y_cv.shape)
print(X_test_essay_bow.shape, y_test.shape)
print("="*100)
```

```
After vectorization
(22445, 5000) (22445,)
(11055, 5000) (11055,)
(16500, 5000) (16500,)
==========================================================================
==========================
```

In [12]:

```
#project_title
vectorizer = CountVectorizer(min_df=10,ngram_range=(1,2), max_features=5000)
vectorizer.fit(X_train['preprocessed_titles'].values.astype('U'))

X_train_title_bow = vectorizer.transform(X_train['preprocessed_titles'].values.astype(
'U'))
X_cv_title_bow = vectorizer.transform(X_cv['preprocessed_titles'].values.astype('U'))
X_test_title_bow = vectorizer.transform(X_test['preprocessed_titles'].values.astype('U'
))
```

In [13]:

```
print("After vectorization")
print(X_train_title_bow.shape, y_train.shape)
print(X_cv_title_bow.shape, y_cv.shape)
print(X_test_title_bow.shape, y_test.shape)
print("="*100)
```

```
After vectorization
(22445, 1589) (22445,)
(11055, 1589) (11055,)
(16500, 1589) (16500,)
==========================================================================
==========================
```

1.3.2 Vectorizing preprocessed essays & project_title using TFIDF

In [14]:

```
#TFIDF for preprocessed_essays
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(min_df=10,ngram_range=(1,2), max_features=5000)
vectorizer.fit(X_train['preprocessed_essays'].values)

X_train_essay_tfidf = vectorizer.transform(X_train['preprocessed_essays'].values)
X_cv_essay_tfidf = vectorizer.transform(X_cv['preprocessed_essays'].values)
X_test_essay_tfidf = vectorizer.transform(X_test['preprocessed_essays'].values)
```

In [15]:

```
print("After vectorization")
print(X_train_essay_tfidf.shape, y_train.shape)
print(X_cv_essay_tfidf.shape, y_cv.shape)
print(X_test_essay_tfidf.shape, y_test.shape)
print("="*100)
```

```
After vectorization
(22445, 5000) (22445,)
(11055, 5000) (11055,)
(16500, 5000) (16500,)
==========================================================================
==========================
```

In [16]:

```
#TFIDF for preprocessed_titles
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(min_df=10,ngram_range=(1,2), max_features=5000)
vectorizer.fit(X_train['preprocessed_titles'].values.astype('U'))

X_train_titles_tfidf = vectorizer.transform(X_train['preprocessed_titles'].values.astyp
e('U'))
X_cv_titles_tfidf = vectorizer.transform(X_cv['preprocessed_titles'].values.astype('U'
))
X_test_titles_tfidf = vectorizer.transform(X_test['preprocessed_titles'].values.astype(
'U'))
```

In [17]:

```
print("After vectorization")
print(X_train_titles_tfidf.shape, y_train.shape)
print(X_cv_titles_tfidf.shape, y_cv.shape)
print(X_test_titles_tfidf.shape, y_test.shape)
print("="*100)
```

```
After vectorization
(22445, 1589) (22445,)
(11055, 1589) (11055,)
(16500, 1589) (16500,)
==========================================================================
==========================
```

1.3.3 Vectorizing preprocessed essays & project_title using Avg W2V

1.3.3.1 For preprocessed_titles

In [19]:

```
#Avg W2V for preprocessed_titles
#Train your own Word2Vec model using your own text corpus
import warnings
warnings.filterwarnings("ignore")
#train data
w2v_data= X_train['preprocessed_titles']
split_title_train=[]
for row in w2v_data:
    split_title_train.append([word for word in str(row).split()])     #splitting words

#train your W2v
train_w2v = Word2Vec(split_title_train,min_count=1,size=50, workers=4)
word_vectors_train = train_w2v.wv
w2v_words_train =list(word_vectors_train.vocab)
print(len(w2v_words_train ))
```

```
7953
```

In [20]:

```
# compute average word2vec for each title.
sent_vectors_train = [] # the avg-w2v for each title is stored in this list
for sent in tqdm(split_title_train):   # for each title
    sent_vec = np.zeros(50)  # as word vectors are of zero length 50
    cnt_words =0   # num of words with a valid vector in the title
    for word in sent:   # for each word in a title
        if word in w2v_words_train:
            vec = word_vectors_train[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
        sent_vectors_train.append(sent_vec)
print(len(sent_vectors_train))
print(len(sent_vectors_train[3]))
```

```
100%|
| 22445/22445 [00:02<
00:00, 8820.50it/s]


22445
50
```

In [28]:

```
# For CV data

# compute average word2vec for each title.
sent_vectors_cv = [] # the avg-w2v for each title is stored in this list
for sent in tqdm(X_cv['preprocessed_titles']):   # for each title
    sent_vec = np.zeros(50)  # as word vectors are of zero length 50
    #cnt_words =0   # num of words with a valid vector in the title
    for word in str(sent):   # for each word in a title
        if word in w2v_words_train:
            vec = word_vectors_train[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
        sent_vectors_cv.append(sent_vec)
print(len(sent_vectors_cv))
print(len(sent_vectors_cv[3]))
```

```
100%|
| 11055/11055 [00:28
<00:00, 390.90it/s]

11055
50
```

In [27]:

```
# For test data

# compute average word2vec for each title.
sent_vectors_test = [] # the avg-w2v for each title is stored in this list
for sent in tqdm(X_test['preprocessed_titles']):   # for each title
    sent_vec = np.zeros(50)  # as word vectors are of zero length 50
    #cnt_words =0   # num of words with a valid vector in the title
    for word in str(sent):   # for each word in a title
        if word in w2v_words_train:
            vec = word_vectors_train[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
        sent_vectors_test.append(sent_vec)
print(len(sent_vectors_test))
print(len(sent_vectors_test[3]))
```

```
100%|████████████████████████████████████████████████████████████████████████████████████████████████████| 16500/16500 [00:41<00:00, 394.87it/s]

16500
50
```

1.3.3.2 For preprocessed_essays

Using Pretrained Models: Avg W2V

In [29]:

```
# stronging variables into pickle files python: http://www.jessicayung.com/how-to-use-pickle-to-save-and-load-variables-in-python/
# make sure you have the glove_vectors file
with open('glove_vectors', 'rb') as f:
    model = pickle.load(f)
    glove_words =  set(model.keys())
print ("Done.",len(model)," words loaded!")
```

```
Done. 51510  words loaded!
```

In [30]:

```
# Avg W2V for train data
# compute average word2vec for each review.
avg_w2v_essay_train = []   # the avg-w2v for each sentence/review is stored in this lis
t
for sentence in tqdm(X_train['preprocessed_essays']):   # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    cnt_words =0 # num of words with a valid vector in the sentence/review
    for word in sentence.split():  # for each word in a review/sentence
        if word in glove_words:
            vector += model[word]
            cnt_words += 1
    if cnt_words != 0:
        vector /= cnt_words
    avg_w2v_essay_train.append(vector)
print(len(avg_w2v_essay_train))
print(len(avg_w2v_essay_train[0]))
```

```
100%|██████████████████████████████████████████████████████████████████████████████████████████████████| 22445/22445 [00:08<00:00, 2563.46it/s]

22445
300
```

In [31]:

```
# Avg W2V for cv data

avg_w2v_essay_cv = []   # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_cv['preprocessed_essays']):   # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    cnt_words =0 # num of words with a valid vector in the sentence/review
    for word in sentence.split():  # for each word in a review/sentence
        if word in glove_words:
            vector += model[word]
            cnt_words += 1
    if cnt_words != 0:
        vector /= cnt_words
    avg_w2v_essay_cv.append(vector)
print(len(avg_w2v_essay_cv))
print(len(avg_w2v_essay_cv[0]))
```

```
100%|██████████████████████████████████████████████████████████████████████████████████████████████████| 11055/11055 [00:05<00:00, 2127.66it/s]

11055
300
```

In [32]:

```
# Avg W2V for test data

avg_w2v_essay_test = []   # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_test['preprocessed_essays']):   # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    cnt_words =0 # num of words with a valid vector in the sentence/review
    for word in sentence.split():  # for each word in a review/sentence
        if word in glove_words:
            vector += model[word]
            cnt_words += 1
    if cnt_words != 0:
        vector /= cnt_words
    avg_w2v_essay_test.append(vector)
print(len(avg_w2v_essay_test))
print(len(avg_w2v_essay_test[0]))
```

```
100%|██████████████████████████████████████████████████████████████████████
████████████████████████████████████████████████████| 16500/16500 [00:07<
00:00, 2220.17it/s]

16500
300
```

1.3.4 Vectorizing preprocessed essays & project_title using TFIDF weighted W2V

1.3.4.1 For preprocessed essays

In [33]:

```
# For train data

tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train['preprocessed_essays'])
#we are converting a dictionary with word as a key, and the idf as a value
dictionary = dict(zip(tfidf_model.get_feature_names(), list(tfidf_model.idf_)))
tfidf_words_essays = set(tfidf_model.get_feature_names())
```

In [34]:

```
# average Word2Vec using pretrained models
# compute average word2vec for each review.
tfidf_w2v_train_essay = [] # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_train['preprocessed_essays']): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =0; # num of words with a valid vector in the sentence/review
    for word in sentence.split(): # for each word in a review/sentence
        if (word in glove_words) and (word in tfidf_words_essays):
            vec = model[word]  # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tfidf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
        tfidf_w2v_train_essay.append(vector)
print(len(tfidf_w2v_train_essay))
print(len(tfidf_w2v_train_essay[0]))
```

```
100%|██████████| 22445/22445 [01:03<00:00, 354.18it/s]

22445
300
```

In [35]:

```
#For cv data

tfidf_w2v_cv_essay = [] # the avg-w2v for each sentence/review is stored in this list
for sentence1 in tqdm(X_cv['preprocessed_essays']): # for each review/sentence
    vector1 = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight1 =0; # num of words with a valid vector in the sentence/review
    for word1 in sentence1.split(): # for each word in a review/sentence
        if (word1 in glove_words) and (word1 in tfidf_words_essays):
            vec1 = model[word1]  # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf value((sentence.count(word)/len(sentence.split())))
            tf_idf1 = dictionary[word1]*(sentence1.count(word1)/len(sentence1.split())) # getting the tfidf value for each word
            vector1 += (vec1 * tf_idf1) # calculating tfidf weighted w2v
            tf_idf_weight1 += tf_idf1
    if tf_idf_weight1 != 0:
        vector1 /= tf_idf_weight1
        tfidf_w2v_cv_essay.append(vector1)
print(len(tfidf_w2v_cv_essay))
print(len(tfidf_w2v_cv_essay[0]))
```

```
100%|██████████| 11055/11055 [00:31<00:00, 353.27it/s]

11055
300
```

In [36]:

```
# For test data

tfidf_w2v_test_essay = [] # the avg-w2v for each sentence/review is stored in this list
for sentence2 in tqdm(X_test['preprocessed_essays']): # for each review/sentence
    vector2 = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight2 =0; # num of words with a valid vector in the sentence/review
    for word2 in sentence2.split(): # for each word in a review/sentence
        if (word2 in glove_words) and (word2 in tfidf_words_essays):
            vec2 = model[word2]  # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf value((sen
tence.count(word)/len(sentence.split())))
            tf_idf2 = dictionary[word2]*(sentence2.count(word2)/len(sentence2.split()))
# getting the tfidf value for each word
            vector2 += (vec2 * tf_idf2) # calculating tfidf weighted w2v
            tf_idf_weight2 += tf_idf2
    if tf_idf_weight2 != 0:
        vector2 /= tf_idf_weight2
        tfidf_w2v_test_essay.append(vector2)
print(len(tfidf_w2v_test_essay))
print(len(tfidf_w2v_test_essay[0]))
```

```
100%|██████████████████████████████████████████████████████████████████████| 16500/16500 [00:47
<00:00, 345.95it/s]

16500
300
```

1.3.4.2 For preprocessed titles

Using pretrained models

In [38]:

```
# For train data

tfidf_model1 = TfidfVectorizer()
tfidf_model1.fit(X_train['preprocessed_titles'].values.astype('U'))
#we are converting a dictionary with word as a key, and the idf as a value
dictionary_title = dict(zip(tfidf_model1.get_feature_names(), list(tfidf_model1.idf_)))
tfidf_words_titles = set(tfidf_model1.get_feature_names())
```

In [40]:

```
# average Word2Vec using pretrained models
# compute average word2vec for each review.
tfidf_w2v_train_title = [] # the avg-w2v for each sentence/review is stored in this lis
t
for sentence_title in tqdm(X_train['preprocessed_titles']): # for each review/sentence
    vector3 = np.zeros(300) # as word vectors are of zero length
    #tf_idf_weight3=0; # num of words with a valid vector in the sentence/review
    for word3 in str(sentence_title).split(): # for each word in a review/sentence
        if (word3 in glove_words) and (word3 in tfidf_words_titles):
            vec4 = model[word3]  # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf value((sen
tence.count(word)/len(sentence.split())))
            tf_idf3 = dictionary_title[word3]*(sentence_title.count(word3)/len(str(sent
ence_title).split())) # getting the tfidf value for each word
            vector3 += (vec4 * tf_idf3) # calculating tfidf weighted w2v
            tf_idf_weight3 += tf_idf3
    if tf_idf_weight3 != 0:
        vector3 /= tf_idf_weight3
        tfidf_w2v_train_title.append(vector3)
print(len(tfidf_w2v_train_title))
print(len(tfidf_w2v_train_title[0]))
```

```
100%|██████████████████████████████████████████████████████████████████████████| 22445/22445 [00:00<0
0:00, 23214.06it/s]

22445
300
```

In [43]:

```
# For cv data
tfidf_w2v_cv_title = [] # the avg-w2v for each sentence/review is stored in this list
for sentence_cv in tqdm(X_cv['preprocessed_titles']): # for each review/sentence
    vector4 = np.zeros(300) # as word vectors are of zero length
    #tf_idf_weight4 =0; # num of words with a valid vector in the sentence/review
    for word4 in str(sentence_cv).split(): # for each word in a review/sentence
        if (word4 in glove_words) and (word4 in tfidf_words_titles):
            vec5 = model[word4]  # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf value((sen
tence.count(word)/len(sentence.split())))
            tf_idf4 = dictionary_title[word4]*(sentence_cv.count(word4)/len(str(sentenc
e_cv).split())) # getting the tfidf value for each word
            vector4 += (vec5 * tf_idf4) # calculating tfidf weighted w2v
            tf_idf_weight4 += tf_idf4
    if tf_idf_weight4 != 0:
        vector4 /= tf_idf_weight4
        tfidf_w2v_cv_title.append(vector4)
print(len(tfidf_w2v_cv_title))
print(len(tfidf_w2v_cv_title[0]))
```

```
100%|██████████████████████████████████████████████████████████████████████████| 11055/11055 [00:00<0
0:00, 23351.13it/s]

11055
300
```

In [45]:

```
# For test data
tfidf_w2v_test_title = [] # the avg-w2v for each sentence/review is stored in this list
for sentence_test in tqdm(X_test['preprocessed_titles']): # for each review/sentence
    vector5 = np.zeros(300) # as word vectors are of zero length
    #tf_idf_weight5 =0; # num of words with a valid vector in the sentence/review
    for word5 in str(sentence_test).split(): # for each word in a review/sentence
        if (word5 in glove_words) and (word5 in tfidf_words_titles):
            vec6 = model[word5]  # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf value((sen
tence.count(word)/len(sentence.split())))
            tf_idf5 = dictionary_title[word5]*(sentence_test.count(word5)/len(str(sente
nce_test).split())) # getting the tfidf value for each word
            vector5 += (vec6 * tf_idf5) # calculating tfidf weighted w2v
            tf_idf_weight5 += tf_idf5
    if tf_idf_weight5 != 0:
        vector5 /= tf_idf_weight5
        tfidf_w2v_test_title.append(vector5)
print(len(tfidf_w2v_test_title))
print(len(tfidf_w2v_test_title[0]))
```

```
100%|███████████████████████████████████
██████████████████████████| 16500/16500 [00:00<0
0:00, 23608.35it/s]

16500
300
```

1.4 Make Data Model Ready: encoding numerical, categorical features

1.4.1 Encoding categorical features: School State

In [46]:

```
vectorizer = CountVectorizer()
vectorizer.fit(X_train['school_state'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_state = vectorizer.transform(X_train['school_state'].values)
X_cv_state = vectorizer.transform(X_cv['school_state'].values)
X_test_state = vectorizer.transform(X_test['school_state'].values)

print("After vectorizations")
print(X_train_state.shape, y_train.shape)
print(X_cv_state.shape, y_cv.shape)
print(X_test_state.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(22445, 51) (22445,)
(11055, 51) (11055,)
(16500, 51) (16500,)
['ak', 'al', 'ar', 'az', 'ca', 'co', 'ct', 'dc', 'de', 'fl', 'ga', 'hi',
'ia', 'id', 'il', 'in', 'ks', 'ky', 'la', 'ma', 'md', 'me', 'mi', 'mn', 'm
o', 'ms', 'mt', 'nc', 'nd', 'ne', 'nh', 'nj', 'nm', 'nv', 'ny', 'oh', 'o
k', 'or', 'pa', 'ri', 'sc', 'sd', 'tn', 'tx', 'ut', 'va', 'vt', 'wa', 'w
i', 'wv', 'wy']
===========================================================================
=========================
```

1.4.2 Encoding categorical features: teacher_prefix

In [47]:

```
vectorizer = CountVectorizer()
vectorizer.fit(X_train['teacher_prefix'].values)

X_train_teacher = vectorizer.transform(X_train['teacher_prefix'].values)
X_cv_teacher = vectorizer.transform(X_cv['teacher_prefix'].values)
X_test_teacher = vectorizer.transform(X_test['teacher_prefix'].values)

print("After vectorizations")
print(X_train_teacher.shape, y_train.shape)
print(X_cv_teacher.shape, y_cv.shape)
print(X_test_teacher.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(22445, 6) (22445,)
(11055, 6) (11055,)
(16500, 6) (16500,)
['dr', 'mr', 'mrs', 'ms', 'none', 'teacher']
===========================================================================
=========================
```

1.4.3 Encoding categorical features: project_grade_category

In [48]:

```
#This step is to intialize a vectorizer with vocab from train data
#Ref: https://www.kaggle.com/shashank49/donors-choose-knn#Concatinating-all-features-(T
FIDF)
from collections import Counter
my_counter = Counter()
for word in X_train['project_grade_category'].values:
    my_counter.update([word[i:i+14] for i in range(0, len(word),14)]) #https://www.geek
sforgeeks.org/python-string-split/

# dict sort by value python: https://stackoverflow.com/a/613218/4084039
project_grade_category_dict = dict(my_counter)
sorted_project_grade_category_dict = dict(sorted(project_grade_category_dict.items(), k
ey=lambda kv: kv[1]))
```

In [49]:

```
vectorizer = CountVectorizer(vocabulary=list(sorted_project_grade_category_dict.keys
()), lowercase=False, binary=True,max_features=4)
vectorizer.fit(X_train['project_grade_category'].values) # fit has to happen only on tr
ain data

# we use the fitted CountVectorizer to convert the text to vector
X_train_grade = vectorizer.transform(X_train['project_grade_category'].values)
X_cv_grade = vectorizer.transform(X_cv['project_grade_category'].values)
X_test_grade = vectorizer.transform(X_test['project_grade_category'].values)

print("After vectorizations")
print(X_train_grade.shape, y_train.shape)
print(X_cv_grade.shape, y_cv.shape)
print(X_test_grade.shape, y_test.shape)
print(vectorizer.get_feature_names())
```

```
After vectorizations
(22445, 4) (22445,)
(11055, 4) (11055,)
(16500, 4) (16500,)
['Grades 9-12', 'Grades 6-8', 'Grades 3-5', 'Grades PreK-2']
```

1.4.4 Encoding categorical features: clean_categories

In [50]:

```
vectorizer = CountVectorizer()
vectorizer.fit(X_train['clean_categories'].values) # fit has to happen only on train da
ta

# we use the fitted CountVectorizer to convert the text to vector
X_train_cat = vectorizer.transform(X_train['clean_categories'].values)
X_cv_cat = vectorizer.transform(X_cv['clean_categories'].values)
X_test_cat = vectorizer.transform(X_test['clean_categories'].values)

print("After vectorizations")
print(X_train_cat.shape, y_train.shape)
print(X_cv_cat.shape, y_cv.shape)
print(X_test_cat.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(22445, 9) (22445,)
(11055, 9) (11055,)
(16500, 9) (16500,)
['appliedlearning', 'care_hunger', 'health_sports', 'history_civics', 'lit
eracy_language', 'math_science', 'music_arts', 'specialneeds', 'warmth']
==========================================================================
==========================
```

1.4.5 Encoding categorical features: clean_subcategories

In [51]:

```
vectorizer = CountVectorizer()
vectorizer.fit(X_train['clean_subcategories'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_subcat = vectorizer.transform(X_train['clean_subcategories'].values)
X_cv_subcat = vectorizer.transform(X_cv['clean_subcategories'].values)
X_test_subcat = vectorizer.transform(X_test['clean_subcategories'].values)

print("After vectorizations")
print(X_train_subcat.shape, y_train.shape)
print(X_cv_subcat.shape, y_cv.shape)
print(X_test_subcat.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(22445, 30) (22445,)
(11055, 30) (11055,)
(16500, 30) (16500,)
['appliedsciences', 'care_hunger', 'charactereducation', 'civics_governmen
t', 'college_careerprep', 'communityservice', 'earlydevelopment', 'economi
cs', 'environmentalscience', 'esl', 'extracurricular', 'financialliterac
y', 'foreignlanguages', 'gym_fitness', 'health_lifescience', 'health_welln
ess', 'history_geography', 'literacy', 'literature_writing', 'mathematic
s', 'music', 'nutritioneducation', 'other', 'parentinvolvement', 'performi
ngarts', 'socialsciences', 'specialneeds', 'teamsports', 'visualarts', 'wa
rmth']
==========================================================================
==========================
```

1.4.6 Encoding numerical features: Price

In [52]:

```
from sklearn.preprocessing import Normalizer
normalizer = Normalizer()
# normalizer.fit(X_train['price'].values)
#this will rise an error Expected 2D array, got 1D array instead:
normalizer.fit(X_train['price'].values.reshape(1,-1))

X_train_price_norm = normalizer.transform(X_train['price'].values.reshape(1,-1))
X_cv_price_norm = normalizer.transform(X_cv['price'].values.reshape(1,-1))
X_test_price_norm = normalizer.transform(X_test['price'].values.reshape(1,-1))

print("After vectorizations")
print(X_train_price_norm.shape, y_train.shape)
print(X_cv_price_norm.shape, y_cv.shape)
print(X_test_price_norm.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(22445, 1) (22445,)
(11055, 1) (11055,)
(16500, 1) (16500,)
==========================================================================
==========================
```

1.4.7 Encoding numerical features: Quantity

In [53]:

```
from sklearn.preprocessing import Normalizer
normalizer = Normalizer()
normalizer.fit(X_train['quantity'].values.reshape(1,-1))

X_train_quantity_norm = normalizer.transform(X_train['quantity'].values.reshape(1,-1))
X_cv_quantity_norm = normalizer.transform(X_cv['quantity'].values.reshape(1,-1))
X_test_quantity_norm = normalizer.transform(X_test['quantity'].values.reshape(1,-1))

print("After vectorizations")
print(X_train_quantity_norm.shape, y_train.shape)
print(X_cv_quantity_norm.shape, y_cv.shape)
print(X_test_quantity_norm.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(22445, 1) (22445,)
(11055, 1) (11055,)
(16500, 1) (16500,)
==========================================================================
==========================
```

1.4.8 Encoding numerical features: teacher_number_of_previously_posted_projects

In [54]:

```
from sklearn.preprocessing import Normalizer
normalizer = Normalizer()
normalizer.fit(X_train['teacher_number_of_previously_posted_projects'].values.reshape(1
,-1))

X_train_projects_norm = normalizer.transform(X_train['teacher_number_of_previously_post
ed_projects'].values.reshape(1,-1))
X_cv_projects_norm = normalizer.transform(X_cv['teacher_number_of_previously_posted_pro
jects'].values.reshape(1,-1))
X_test_projects_norm = normalizer.transform(X_test['teacher_number_of_previously_posted
_projects'].values.reshape(1,-1))

print("After vectorizations")
print(X_train_projects_norm.shape, y_train.shape)
print(X_cv_projects_norm.shape, y_cv.shape)
print(X_test_projects_norm.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(22445, 1) (22445,)
(11055, 1) (11055,)
(16500, 1) (16500,)
==========================================================================
==========================
```

1.4.9 Encoding numerical features: sentimental_score

In [56]:

```
from sklearn.preprocessing import Normalizer
normalizer = Normalizer()
normalizer.fit(X_train['sentimental_score'].values.reshape(1,-1))

X_train_senti_norm = normalizer.transform(X_train['sentimental_score'].values.reshape(1
,-1))
X_cv_senti_norm = normalizer.transform(X_cv['sentimental_score'].values.reshape(1,-1))
X_test_senti_norm = normalizer.transform(X_test['sentimental_score'].values.reshape(1,-
1))

print("After vectorizations")
print(X_train_senti_norm.shape, y_train.shape)
print(X_cv_senti_norm.shape, y_cv.shape)
print(X_test_senti_norm.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(22445, 1) (22445,)
(11055, 1) (11055,)
(16500, 1) (16500,)
==========================================================================
==========================
```

In [55]:

```
X_train.head(1)
```

Out[55]:

	Unnamed: 0	id	teacher_id	teacher_prefix	scho
3044	180222	p156306	c6740b16a6c2158dc21450d595ec3b91	Ms.	LA

1 rows × 28 columns

1.4.10 Encoding numerical features: preprocessed_essay_word_count

In [57]:

```
from sklearn.preprocessing import Normalizer
normalizer = Normalizer()
normalizer.fit(X_train['preprocessed_essay_word_count'].values.reshape(1,-1))

X_train_ewc_norm = normalizer.transform(X_train['preprocessed_essay_word_count'].values
.reshape(1,-1))
X_cv_ewc_norm = normalizer.transform(X_cv['preprocessed_essay_word_count'].values.resha
pe(1,-1))
X_test_ewc_norm = normalizer.transform(X_test['preprocessed_essay_word_count'].values.r
eshape(1,-1))

print("After vectorization")
print(X_train_ewc_norm.shape, y_train.shape)
print(X_cv_ewc_norm.shape, y_cv.shape)
print(X_test_ewc_norm.shape, y_test.shape)
print("="*100)
```

```
After vectorization
(22445, 1) (22445,)
(11055, 1) (11055,)
(16500, 1) (16500,)
==========================================================================
==========================
```

1.4.11 Encoding numerical features: preprocessed_title_word_count

In [58]:

```
from sklearn.preprocessing import Normalizer
normalizer = Normalizer()
normalizer.fit(X_train['preprocessed_title_word_count'].values.reshape(1,-1))

X_train_twc_norm = normalizer.transform(X_train['preprocessed_title_word_count'].values
.reshape(1,-1))
X_cv_twc_norm = normalizer.transform(X_cv['preprocessed_title_word_count'].values.resha
pe(1,-1))
X_test_twc_norm = normalizer.transform(X_test['preprocessed_title_word_count'].values.r
eshape(1,-1))

print("After vectorization")
print(X_train_twc_norm.shape, y_train.shape)
print(X_cv_twc_norm.shape, y_cv.shape)
print(X_test_twc_norm.shape, y_test.shape)
print("="*100)
```

```
After vectorization
(22445, 1) (22445,)
(11055, 1) (11055,)
(16500, 1) (16500,)
==========================================================================
==========================
```

1.4.5 Concatinating all the features

1.4.5.1 Set 1: Using categorical features + numerical features + preprocessed_titles(BOW) + preprocessed_essays(BOW)

In [59]:

```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack

X_tr_bow = hstack((X_train_essay_bow, X_train_title_bow, X_train_state, X_train_teacher
, X_train_grade, X_train_cat, X_train_subcat, X_train_price_norm, X_train_quantity_norm
, X_train_projects_norm )).tocsr()

X_cv_bow = hstack((X_cv_essay_bow, X_cv_title_bow, X_cv_state, X_cv_teacher, X_cv_grade
, X_cv_cat, X_cv_subcat, X_cv_price_norm, X_cv_quantity_norm, X_cv_projects_norm )).toc
sr()

X_test_bow = hstack((X_test_essay_bow, X_test_title_bow, X_test_state, X_test_teacher,
X_test_grade, X_test_cat, X_test_subcat, X_test_price_norm, X_test_quantity_norm, X_tes
t_projects_norm )).tocsr()

print("Final Data Matrix")
print(X_tr_bow.shape, y_train.shape)
print(X_cv_bow.shape, y_train.shape)
print(X_test_bow.shape, y_train.shape)
```

```
Final Data Matrix
(22445, 6692) (22445,)
(11055, 6692) (22445,)
(16500, 6692) (22445,)
```

1.4.5.2 Set 2: Using categorical features + numerical features + preprocessed_titles(TFIDF) + preprocessed_essays(TFIDF)

In [60]:

```
X_tr_tfidf = hstack((X_train_essay_tfidf, X_train_titles_tfidf, X_train_state, X_train_teacher, X_train_grade, X_train_cat, X_train_subcat, X_train_price_norm, X_train_quantity_norm, X_train_projects_norm )).tocsr()

X_cv_tfidf = hstack((X_cv_essay_tfidf, X_cv_titles_tfidf, X_cv_state, X_cv_teacher, X_cv_grade, X_cv_cat, X_cv_subcat, X_cv_price_norm, X_cv_quantity_norm, X_cv_projects_norm )).tocsr()

X_test_tfidf = hstack((X_test_essay_tfidf, X_test_titles_tfidf, X_test_state, X_test_teacher, X_test_grade, X_test_cat, X_test_subcat, X_test_price_norm, X_test_quantity_norm, X_test_projects_norm )).tocsr()

print("Final Data Matrix")
print(X_tr_tfidf.shape, y_train.shape)
print(X_cv_tfidf.shape, y_train.shape)
print(X_test_tfidf.shape, y_train.shape)
```

```
Final Data Matrix
(22445, 6692) (22445,)
(11055, 6692) (22445,)
(16500, 6692) (22445,)
```

1.4.5.3 Set 3: Using categorical features + numerical features + preprocessed_titles(Avg W2V) + preprocessed_essays(Avg W2V)

In [61]:

```
X_tr_avgw2v = hstack((sent_vectors_train, avg_w2v_essay_train, X_train_state, X_train_teacher, X_train_grade, X_train_cat, X_train_subcat, X_train_price_norm, X_train_quantity_norm, X_train_projects_norm )).tocsr()

X_cv_avgw2v = hstack((sent_vectors_cv, avg_w2v_essay_cv, X_cv_state, X_cv_teacher, X_cv_grade, X_cv_cat, X_cv_subcat, X_cv_price_norm, X_cv_quantity_norm, X_cv_projects_norm )).tocsr()

X_test_avgw2v = hstack((sent_vectors_test, avg_w2v_essay_test, X_test_state, X_test_teacher, X_test_grade, X_test_cat, X_test_subcat, X_test_price_norm, X_test_quantity_norm, X_test_projects_norm )).tocsr()

print("Final Data Matrix")
print(X_tr_avgw2v.shape, y_train.shape)
print(X_cv_avgw2v.shape, y_train.shape)
print(X_test_avgw2v.shape, y_train.shape)
```

```
Final Data Matrix
(22445, 453) (22445,)
(11055, 453) (22445,)
(16500, 453) (22445,)
```

1.4.5.4 Set 4: Using categorical features + numerical features + preprocessed_titles(TFIDF W2V) + preprocessed_essays(TFIDF W2V)

In [63]:

```
X_tr_tfidf_w2v = hstack((tfidf_w2v_train_essay, tfidf_w2v_train_title, X_train_state, X
_train_teacher, X_train_grade, X_train_cat, X_train_subcat, X_train_price_norm, X_train
_quantity_norm, X_train_projects_norm )).tocsr()

X_cv_tfidf_w2v = hstack((tfidf_w2v_cv_essay, tfidf_w2v_cv_title, X_cv_state, X_cv_teach
er, X_cv_grade, X_cv_cat, X_cv_subcat, X_cv_price_norm, X_cv_quantity_norm, X_cv_projec
ts_norm )).tocsr()

X_test_tfidf_w2v = hstack((tfidf_w2v_test_essay, tfidf_w2v_test_title, X_test_state, X_
test_teacher, X_test_grade, X_test_cat, X_test_subcat, X_test_price_norm, X_test_quanti
ty_norm, X_test_projects_norm )).tocsr()

print("Final Data Matrix")
print(X_tr_tfidf_w2v.shape, y_train.shape)
print(X_cv_tfidf_w2v.shape, y_train.shape)
print(X_test_tfidf_w2v.shape, y_train.shape)
```

```
Final Data Matrix
(22445, 703) (22445,)
(11055, 703) (22445,)
(16500, 703) (22445,)
```

1.4.5.5 Set 5: Using all categorical features & numerical features.

In [65]:

```
from scipy.sparse import hstack

X_tr_cn = hstack((X_train_state, X_train_teacher, X_train_grade, X_train_cat, X_train_s
ubcat, X_train_price_norm, X_train_quantity_norm, X_train_projects_norm, X_train_senti_
norm, X_train_ewc_norm, X_train_twc_norm)).tocsr()

X_cv_cn = hstack((X_cv_state, X_cv_teacher, X_cv_grade, X_cv_cat, X_cv_subcat, X_cv_pri
ce_norm, X_cv_quantity_norm, X_cv_projects_norm, X_cv_senti_norm, X_cv_ewc_norm, X_cv_t
wc_norm )).tocsr()

X_test_cn = hstack((X_test_state, X_test_teacher, X_test_grade, X_test_cat, X_test_subc
at, X_test_price_norm, X_test_quantity_norm, X_test_projects_norm, X_test_senti_norm, X
_test_ewc_norm, X_test_twc_norm )).tocsr()

print("Final Data Matrix")
print(X_tr_cn.shape, y_train.shape)
print(X_cv_cn.shape, y_train.shape)
print(X_test_cn.shape, y_train.shape)
```

```
Final Data Matrix
(22445, 106) (22445,)
(11055, 106) (22445,)
(16500, 106) (22445,)
```

1.5 Applying Logistic regression

1.5.1 Set 1: BOW featurization

1.5.1.1 Hyper parameter tuning

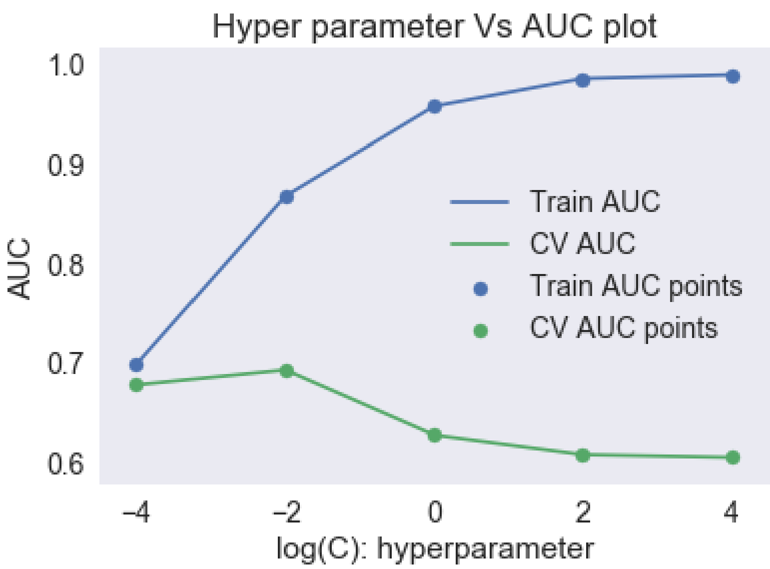
In [71]:

```
def batch_predict(clf, data):
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates
 of the positive class
    # not the predicted outputs
    y_data_pred = []
    tr_loop = data.shape[0] - data.shape[0]%1000;
    # consider you X_tr shape is 49041, then your tr_loop will be 49041 - 49041%1000 =
 49000
    # in this for loop we will iterate until the last 1000 multiplier
    for i in range(0, tr_loop, 1000):
        y_data_pred.extend(clf.predict_proba(data[i:i+1000])[:,1]) # we will be predict
ing for the last data points
    if data.shape[0]%1000 !=0:
        y_data_pred.extend(clf.predict_proba(data[tr_loop:])[:,1])

    return y_data_pred
```

In [127]:

```
import warnings
warnings.filterwarnings("ignore")
import matplotlib.pyplot as plt
from sklearn.metrics import roc_auc_score
from sklearn.linear_model import LogisticRegression

# Simple CV using for loops.

train_auc_bow = []
cv_auc_bow = []
parameters = [10**-4, 10**-2, 10**0, 10**2, 10**4] #values of C
for i in tqdm(parameters):
    clf1=LogisticRegression(C=i, penalty='l2', n_jobs=-1,class_weight='balanced')
    clf1.fit(X_tr_bow, y_train)
    y_train_pred = batch_predict(clf1, X_tr_bow)
    y_cv_pred = batch_predict(clf1, X_cv_bow)
    train_auc_bow.append(roc_auc_score(y_train,y_train_pred))
    cv_auc_bow.append(roc_auc_score(y_cv, y_cv_pred))

plt.plot(np.log10(parameters), train_auc_bow, label='Train AUC')
plt.plot(np.log10(parameters), cv_auc_bow, label='CV AUC')

plt.scatter(np.log10(parameters), train_auc_bow, label='Train AUC points')
plt.scatter(np.log10(parameters), cv_auc_bow, label='CV AUC points')
plt.legend()
plt.xlabel("log(C): hyperparameter")
plt.ylabel("AUC")
plt.title("Hyper parameter Vs AUC plot")
plt.grid()
plt.show()
```

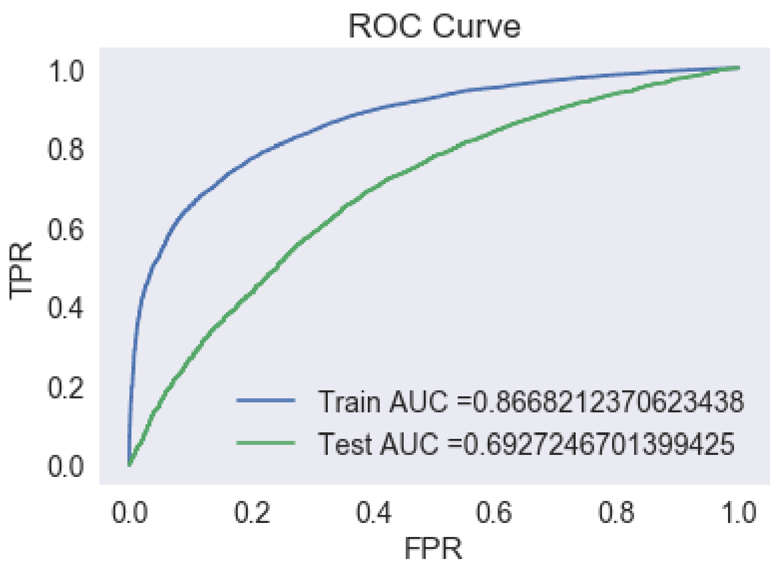
```
100%|████████████████████████████████████████
█████████████████████████████████████| 5/5 [03:0
0<00:00, 44.58s/it]
```



1.5.1.2 Testing the performance of the model on test data & plotting ROC Curves for train & test data

In [128]:

```
best_c = 0.01 #equivalent to lambda=100

clf2= LogisticRegression(C=best_c, penalty='l2', n_jobs=-1,class_weight='balanced')
clf2.fit(X_tr_bow, y_train)
y_train_pred_bow_best = batch_predict(clf2, X_tr_bow)
y_test_pred_bow_best = batch_predict(clf2, X_test_bow)

train_tpr_bow, train_fpr_bow, tr_thresholds_bow = roc_curve(y_train, y_train_pred_bow_b
est)
test_tpr_bow, test_fpr_bow, te_thresholds_bow = roc_curve(y_test, y_test_pred_bow_best)

plt.plot(train_tpr_bow, train_fpr_bow,label="Train AUC ="+str(auc(train_tpr_bow, train_
fpr_bow)))
plt.plot(test_tpr_bow, test_fpr_bow, label="Test AUC ="+str(auc(test_tpr_bow, test_fpr_
bow)))
plt.legend()
plt.xlabel("FPR")
plt.ylabel("TPR")
plt.title("ROC Curve")
plt.grid()
plt.show()
```



- From AUC vs C graph it is seen that difference between train & Cv scores is less for values of "C" in range 10^-4-10^-2.
- After experimenting with values of the above mentioned range, I found C=0.01 (Lambda=100) as my optimum value & the test AUC score for the optimal value was 0.69(69.27%).

In [96]:

```
## we will pick a threshold that will give the least fpr

def find_best_threshold(threshold, fpr, tpr):
    t = threshold[np.argmax(tpr*(1-fpr))]
    # (tpr*(1-fpr)) will be maximum if your fpr is very low and tpr is very high
    print("The maximum value of tpr*(1-fpr)", max(tpr*(1-fpr)),"for threshold", np.roun
d(t,3))
    return t

def predict_with_best_t(proba, threshold):
    predictions = []
    for i in proba:
        if i>=threshold:
            predictions.append(1)
        else:
            predictions.append(0)
    return predictions

print("="*100)
```

```
===========================================================================
==========================
```

In [97]:

```
#function to get heatmap of confusion matrix
# Reference: https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-ma
trix

def cm_heatmap(cm):
    #y_pred = clf.predict(X_te)
    df_cm = pd.DataFrame(cm, range(2),range(2))
    df_cm.columns = ['Predicted NO','Predicted YES']
    df_cm = df_cm.rename({0: 'Actual NO', 1: 'Actual YES'})
    sns.set(font_scale=1.4)#for label size
    sns.heatmap(df_cm, annot=True,annot_kws={"size": 16}, fmt='d')
```

1.5.1.3 Confusion matrices: For best C

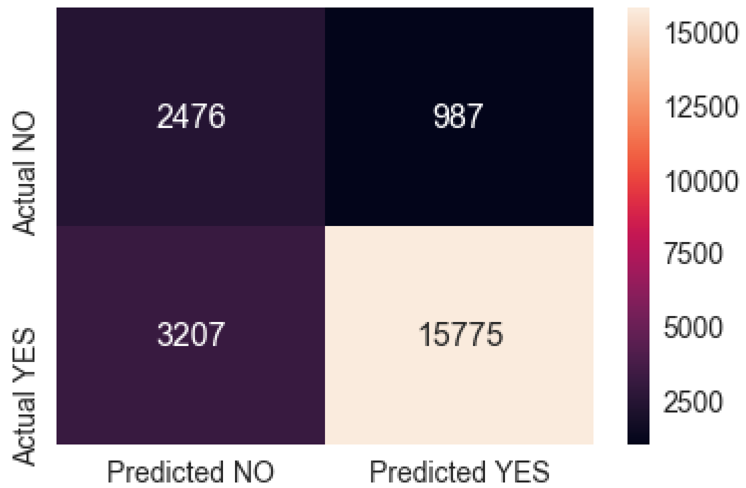
In [129]:

```
from sklearn.metrics import confusion_matrix
best_t_bow = find_best_threshold(tr_thresholds_bow, train_fpr_bow, train_tpr_bow)
print("Train confusion matrix")
cm_train_bow=confusion_matrix(y_train, predict_with_best_t(y_train_pred_bow_best, best_
t_bow))
print(cm_train_bow)
print("Test confusion matrix")
cm_test_bow=confusion_matrix(y_test, predict_with_best_t(y_test_pred_bow_best, best_t_b
ow))
print(cm_test_bow)
```

```
The maximum value of tpr*(1-fpr) 0.048152811790357325 for threshold 0.441
Train confusion matrix
[[ 2476   987]
 [ 3207 15775]]
Test confusion matrix
[[ 1193  1353]
 [ 2835 11119]]
```

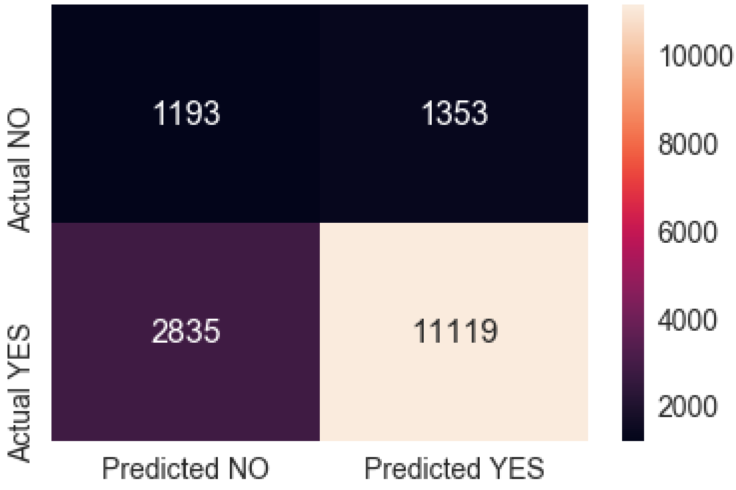
In [130]:

```
# confusion matrix heatmap for train data
cm_heatmap(cm_train_bow)
```



In [131]:

```
# confusion matrix heatmap for test data
cm_heatmap(cm_test_bow)
```



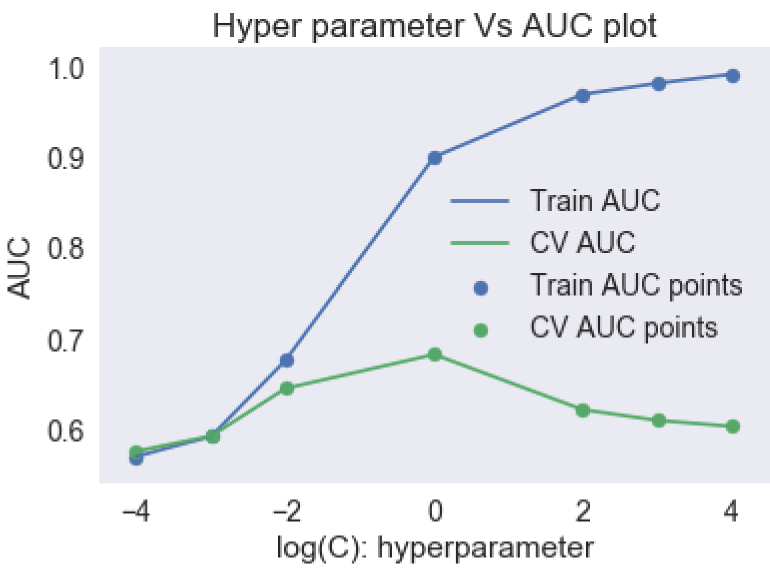
1.5.2 Set 2: TFIDF featurization

1.5.2.1 Hyper parameter tuning

In [132]:

```
# Simple CV using for loops.

train_auc_tfidf = []
cv_auc_tfidf = []
parameters = [10**-4,10**-3, 10**-2, 10**0, 10**2, 10**3, 10**4] #values of C
for i in tqdm(parameters):
    clf3=LogisticRegression(C=i, penalty='l2', n_jobs=-1,class_weight='balanced')
    clf3.fit(X_tr_tfidf, y_train)
    y_train_pred_tfidf = batch_predict(clf3, X_tr_tfidf)
    y_cv_pred_tfidf = batch_predict(clf3, X_cv_tfidf)
    train_auc_tfidf.append(roc_auc_score(y_train,y_train_pred_tfidf))
    cv_auc_tfidf.append(roc_auc_score(y_cv, y_cv_pred_tfidf))

plt.plot(np.log10(parameters), train_auc_tfidf, label='Train AUC')
plt.plot(np.log10(parameters), cv_auc_tfidf, label='CV AUC')

plt.scatter(np.log10(parameters), train_auc_tfidf, label='Train AUC points')
plt.scatter(np.log10(parameters), cv_auc_tfidf, label='CV AUC points')
plt.legend()
plt.xlabel("log(C): hyperparameter")
plt.ylabel("AUC")
plt.title("Hyper parameter Vs AUC plot")
plt.grid()
plt.show()
```

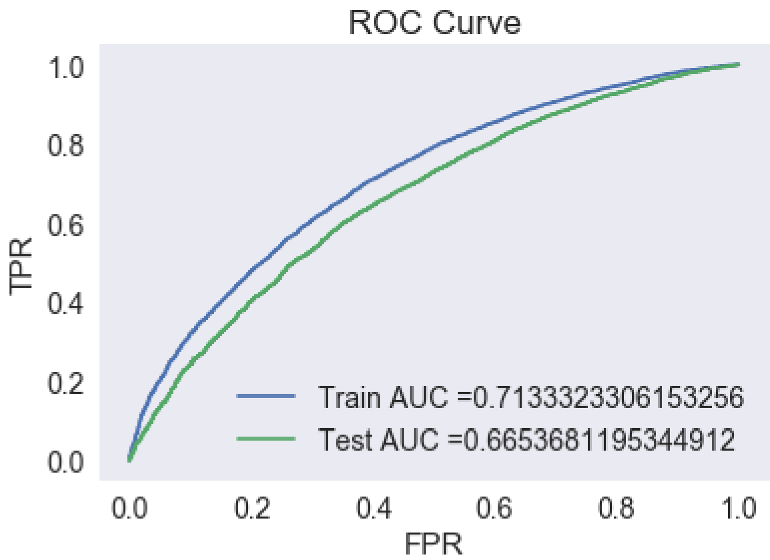
```
100%|████████████████████████████████████████
████████████████████████████████████| 7/7 [03:0
1<00:00, 47.33s/it]
```



1.5.2.2 Testing the performance of the model on test data & plotting ROC Curves for train & test data

In [143]:

```
best_c = 0.02 #equivalent to lambda=50

clf4= LogisticRegression(C=best_c, penalty='l2', n_jobs=-1,class_weight='balanced')
clf4.fit(X_tr_tfidf, y_train)
y_train_pred_tfidf_best = batch_predict(clf4, X_tr_tfidf)
y_test_pred_tfidf_best = batch_predict(clf4, X_test_tfidf)

train_tpr_tfidf, train_fpr_tfidf, tr_thresholds_tfidf = roc_curve(y_train, y_train_pred
_tfidf_best)
test_tpr_tfidf, test_fpr_tfidf, te_thresholds_tfidf = roc_curve(y_test, y_test_pred_tfi
df_best)

plt.plot(train_tpr_tfidf, train_fpr_tfidf,label="Train AUC ="+str(auc(train_tpr_tfidf,
train_fpr_tfidf)))
plt.plot(test_tpr_tfidf, test_fpr_tfidf, label="Test AUC ="+str(auc(test_tpr_tfidf, tes
t_fpr_tfidf)))
plt.legend()
plt.xlabel("FPR")
plt.ylabel("TPR")
plt.title("ROC Curve")
plt.grid()
plt.show()
```



- From AUC vs C graph it is seen that difference between train & Cv scores is less for values of "C" in range 10^-4-10^-1.5.
- After experimenting with values of the above mentioned range, I found C=0.02 (Lambda=50) as my optimum value & the test AUC score for the optimal value was 0.67(67%).

1.5.2.3 Confusion matrices: For best C

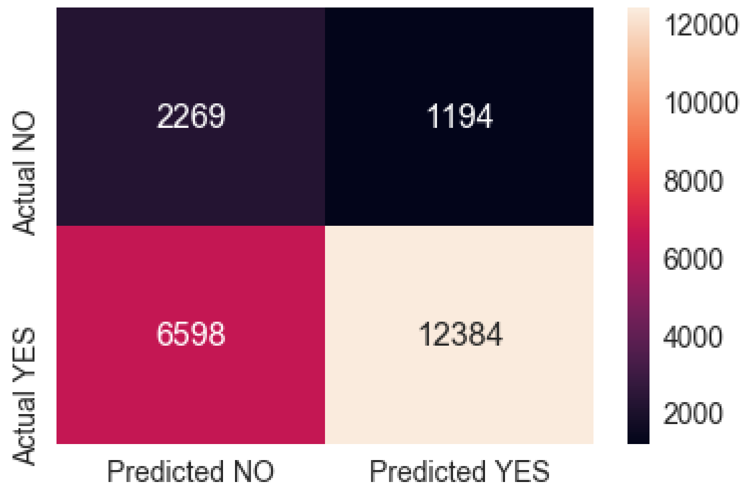
In [144]:

```
from sklearn.metrics import confusion_matrix
best_t_tfidf = find_best_threshold(tr_thresholds_tfidf, train_fpr_tfidf, train_tpr_tfid
f)
print("Train confusion matrix")
cm_train_tfidf=confusion_matrix(y_train, predict_with_best_t(y_train_pred_tfidf_best, b
est_t_tfidf))
print(cm_train_tfidf)
print("Test confusion matrix")
cm_test_tfidf=confusion_matrix(y_test, predict_with_best_t(y_test_pred_tfidf_best, best
_t_tfidf))
print(cm_test_tfidf)
```

```
The maximum value of tpr*(1-fpr) 0.11984562300810961 for threshold 0.5
Train confusion matrix
[[ 2269  1194]
 [ 6598 12384]]
Test confusion matrix
[[1533 1013]
 [5002 8952]]
```

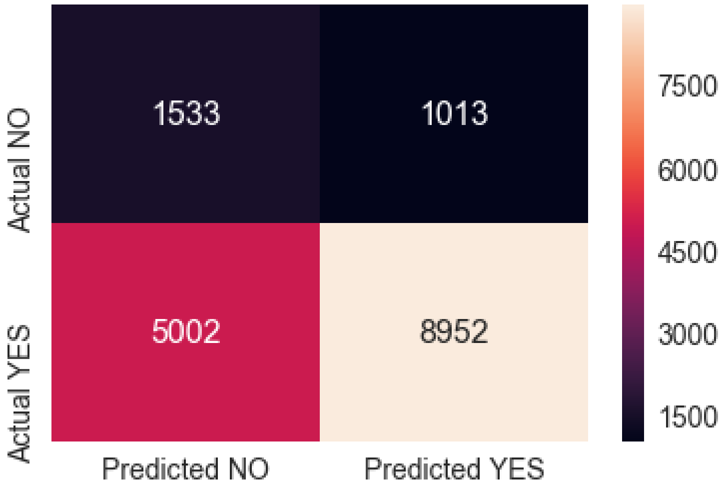
In [145]:

```
# confusion matrix heatmap for train data
cm_heatmap(cm_train_tfidf)
```



In [146]:

```
# confusion matrix heatmap for test data
cm_heatmap(cm_test_tfidf)
```



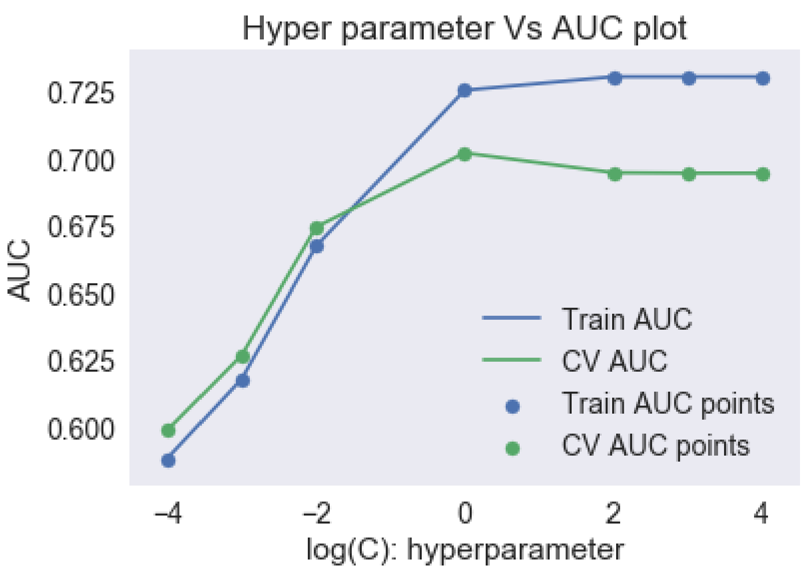
1.5.3 Set 3: AvgW2V featurization

1.5.3.1 Hyper parameter tuning

In [147]:

```
# Simple CV using for loops.

train_auc_avg = []
cv_auc_avg = []
parameters = [10**-4,10**-3, 10**-2, 10**0, 10**2, 10**3, 10**4] #values of C
for i in tqdm(parameters):
    clf5=LogisticRegression(C=i, penalty='l2', n_jobs=-1,class_weight='balanced')
    clf5.fit(X_tr_avgw2v, y_train)
    y_train_pred_avg = batch_predict(clf5, X_tr_avgw2v)
    y_cv_pred_avg = batch_predict(clf5, X_cv_avgw2v)
    train_auc_avg.append(roc_auc_score(y_train,y_train_pred_avg))
    cv_auc_avg.append(roc_auc_score(y_cv, y_cv_pred_avg))

plt.plot(np.log10(parameters), train_auc_avg, label='Train AUC')
plt.plot(np.log10(parameters), cv_auc_avg, label='CV AUC')

plt.scatter(np.log10(parameters), train_auc_avg, label='Train AUC points')
plt.scatter(np.log10(parameters), cv_auc_avg, label='CV AUC points')
plt.legend()
plt.xlabel("log(C): hyperparameter")
plt.ylabel("AUC")
plt.title("Hyper parameter Vs AUC plot")
plt.grid()
plt.show()
```

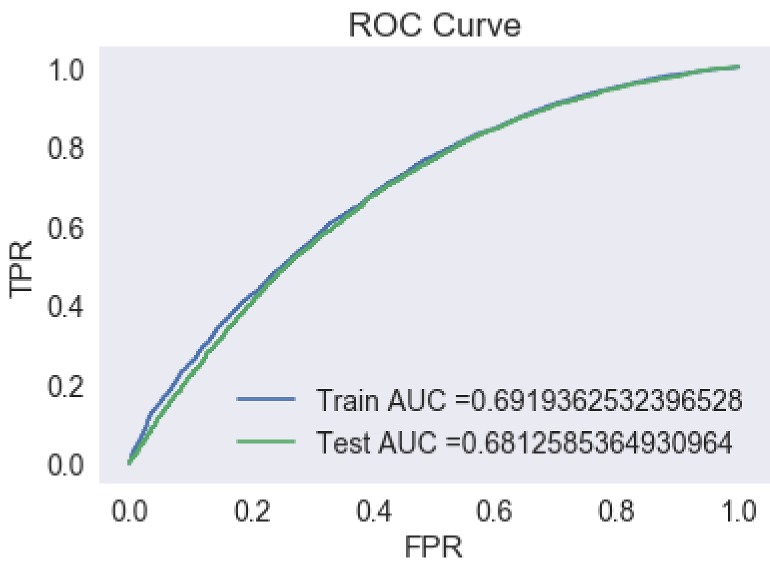
```
100%|████████████████████████████████████████████████████████████████████
████████████████████████████████████████████████████████████████| 7/7 [01:5
8<00:00, 24.29s/it]
```



1.5.3.2 Testing the performance of the model on test data & plotting ROC Curves for train & test data

In [151]:

```
best_c = 0.04 #equivalent to lambda=25(approx)

clf6= LogisticRegression(C=best_c, penalty='l2', n_jobs=-1,class_weight='balanced')
clf6.fit(X_tr_avgw2v, y_train)
y_train_pred_avg_best = batch_predict(clf6, X_tr_avgw2v)
y_test_pred_avg_best = batch_predict(clf6, X_test_avgw2v)

train_tpr_avg, train_fpr_avg, tr_thresholds_avg = roc_curve(y_train, y_train_pred_avg_b
est)
test_tpr_avg, test_fpr_avg, te_thresholds_avg = roc_curve(y_test, y_test_pred_avg_best)

plt.plot(train_tpr_avg, train_fpr_avg,label="Train AUC ="+str(auc(train_tpr_avg, train_
fpr_avg)))
plt.plot(test_tpr_avg, test_fpr_avg, label="Test AUC ="+str(auc(test_tpr_avg, test_fpr_
avg)))
plt.legend()
plt.xlabel("FPR")
plt.ylabel("TPR")
plt.title("ROC Curve")
plt.grid()
plt.show()
```



- From AUC vs C graph it is seen that difference between train & Cv scores is less for values of "C" in range 10^-2-10^-1.0.
- After experimenting with values of the above mentioned range, I found C=0.04 (Lambda=25) as my optimum value & the test AUC score for the optimal value was 0.68(68%).

1.5.3.3 Confusion matrices: For best C

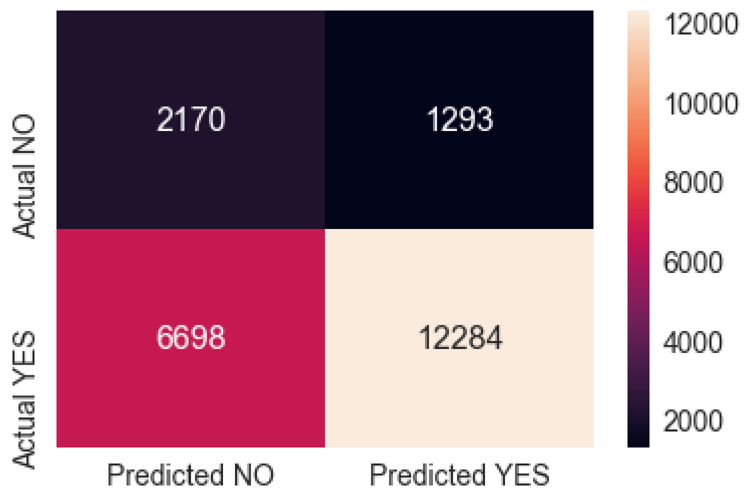
In [152]:

```
from sklearn.metrics import confusion_matrix
best_t_avg = find_best_threshold(tr_thresholds_avg, train_fpr_avg, train_tpr_avg)
print("Train confusion matrix")
cm_train_avg=confusion_matrix(y_train, predict_with_best_t(y_train_pred_avg_best, best_
t_avg))
print(cm_train_avg)
print("Test confusion matrix")
cm_test_avg=confusion_matrix(y_test, predict_with_best_t(y_test_pred_avg_best, best_t_a
vg))
print(cm_test_avg)
```

```
The maximum value of tpr*(1-fpr) 0.1317495703104356 for threshold 0.494
Train confusion matrix
[[ 2170  1293]
 [ 6698 12284]]
Test confusion matrix
[[1564  982]
 [4756 9198]]
```

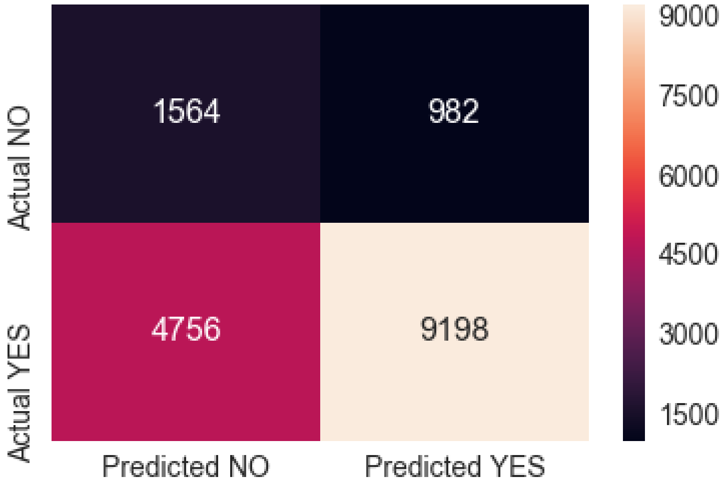
In [153]:

```
# confusion matrix heatmap for train data
cm_heatmap(cm_train_avg)
```



In [154]:

```
# confusion matrix heatmap for test data
cm_heatmap(cm_test_avg)
```



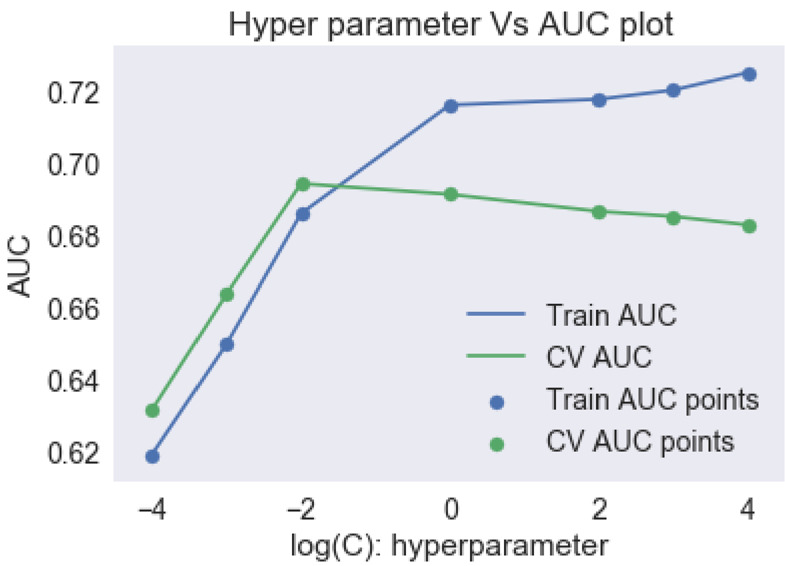
1.5.4 Set 4: TFIDFW2V featurization

1.5.4.1 Hyper parameter tuning

In [155]:

```
# Simple CV using for loops.

train_auc_tw = []
cv_auc_tw = []
parameters = [10**-4,10**-3, 10**-2, 10**0, 10**2, 10**3, 10**4] #values of C
for i in tqdm(parameters):
    clf7=LogisticRegression(C=i, penalty='l2', n_jobs=-1,class_weight='balanced')
    clf7.fit(X_tr_tfidf_w2v, y_train)
    y_train_pred_tw = batch_predict(clf7, X_tr_tfidf_w2v)
    y_cv_pred_tw = batch_predict(clf7, X_cv_tfidf_w2v)
    train_auc_tw.append(roc_auc_score(y_train,y_train_pred_tw))
    cv_auc_tw.append(roc_auc_score(y_cv, y_cv_pred_tw))

plt.plot(np.log10(parameters), train_auc_tw, label='Train AUC')
plt.plot(np.log10(parameters), cv_auc_tw, label='CV AUC')

plt.scatter(np.log10(parameters), train_auc_tw, label='Train AUC points')
plt.scatter(np.log10(parameters), cv_auc_tw, label='CV AUC points')
plt.legend()
plt.xlabel("log(C): hyperparameter")
plt.ylabel("AUC")
plt.title("Hyper parameter Vs AUC plot")
plt.grid()
plt.show()
```

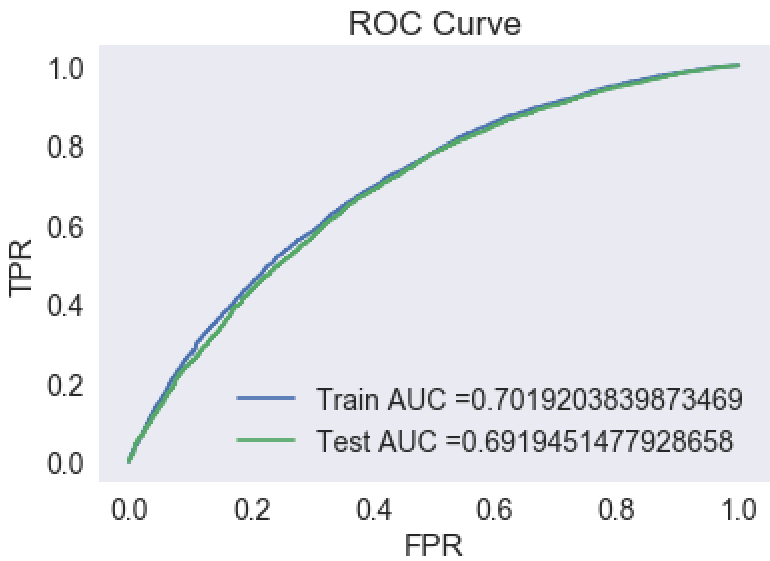
```
100%|████████████████████████████████████████████| 7/7 [16:47<00:00, 260.93s/it]
```



1.5.4.2 Testing the performance of the model on test data & plotting ROC Curves for train & test data

In [156]:

```
best_c = 0.04 #equivalent to lambda=25(approx)

clf8= LogisticRegression(C=best_c, penalty='l2', n_jobs=-1,class_weight='balanced')
clf8.fit(X_tr_tfidf_w2v, y_train)
y_train_pred_tw_best = batch_predict(clf8, X_tr_tfidf_w2v)
y_test_pred_tw_best = batch_predict(clf8, X_test_tfidf_w2v)

train_tpr_tw, train_fpr_tw, tr_thresholds_tw = roc_curve(y_train, y_train_pred_tw_best)
test_tpr_tw, test_fpr_tw, te_thresholds_tw = roc_curve(y_test, y_test_pred_tw_best)

plt.plot(train_tpr_tw, train_fpr_tw,label="Train AUC ="+str(auc(train_tpr_tw, train_fpr
_tw)))
plt.plot(test_tpr_tw, test_fpr_tw, label="Test AUC ="+str(auc(test_tpr_tw, test_fpr_tw
)))
plt.legend()
plt.xlabel("FPR")
plt.ylabel("TPR")
plt.title("ROC Curve")
plt.grid()
plt.show()
```



- From AUC vs C graph it is seen that difference between train & Cv scores is less for values of "C" in range 10^-2-10^-1.0.
- After experimenting with values of the above mentioned range, I found C=0.04 (Lambda=25) as my optimum value & the test AUC score for the optimal value was 0.69(69%).

1.5.4.3 Confusion matrices: For best C

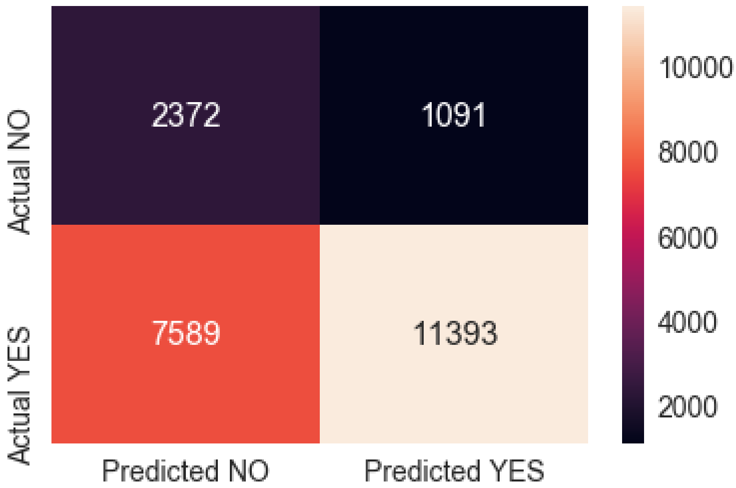
In [157]:

```
from sklearn.metrics import confusion_matrix
best_t_tw = find_best_threshold(tr_thresholds_tw, train_fpr_tw, train_tpr_tw)
print("Train confusion matrix")
cm_train_tw=confusion_matrix(y_train, predict_with_best_t(y_train_pred_tw_best, best_t_
tw))
print(cm_train_tw)
print("Test confusion matrix")
cm_test_tw=confusion_matrix(y_test, predict_with_best_t(y_test_pred_tw_best, best_t_tw
))
print(cm_test_tw)
```

```
The maximum value of tpr*(1-fpr) 0.12595483485076198 for threshold 0.506
Train confusion matrix
[[ 2372  1091]
 [ 7589 11393]]
Test confusion matrix
[[1752  794]
 [5758 8196]]
```

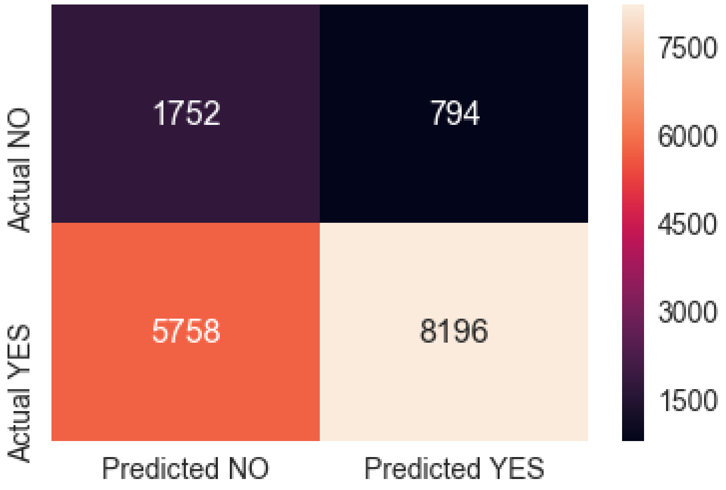
In [158]:

```
# confusion matrix heatmap for train data
cm_heatmap(cm_train_tw)
```



In [159]:

```
# confusion matrix heatmap for test data
cm_heatmap(cm_test_tw)
```



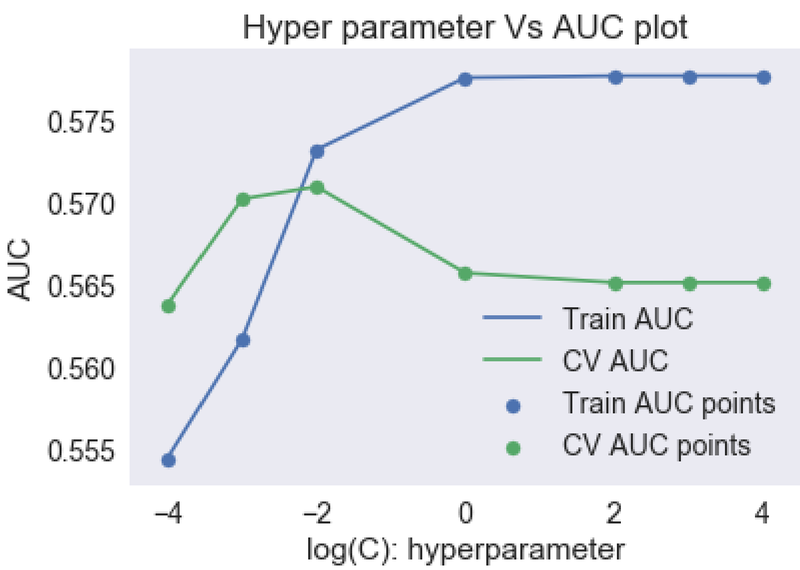
1.5.5 Set 5: Categorical & numerical featurization

1.5.5.1 Hyper parameter tuning

In [160]:

```
# Simple CV using for loops.

train_auc_cn = []
cv_auc_cn = []
parameters = [10**-4,10**-3, 10**-2, 10**0, 10**2, 10**3, 10**4] #values of C
for i in tqdm(parameters):
    clf9=LogisticRegression(C=i, penalty='l2', n_jobs=-1,class_weight='balanced')
    clf9.fit(X_tr_cn, y_train)
    y_train_pred_cn = batch_predict(clf9, X_tr_cn)
    y_cv_pred_cn = batch_predict(clf9, X_cv_cn)
    train_auc_cn.append(roc_auc_score(y_train,y_train_pred_cn))
    cv_auc_cn.append(roc_auc_score(y_cv, y_cv_pred_cn))

plt.plot(np.log10(parameters), train_auc_cn, label='Train AUC')
plt.plot(np.log10(parameters), cv_auc_cn, label='CV AUC')

plt.scatter(np.log10(parameters), train_auc_cn, label='Train AUC points')
plt.scatter(np.log10(parameters), cv_auc_cn, label='CV AUC points')
plt.legend()
plt.xlabel("log(C): hyperparameter")
plt.ylabel("AUC")
plt.title("Hyper parameter Vs AUC plot")
plt.grid()
plt.show()
```

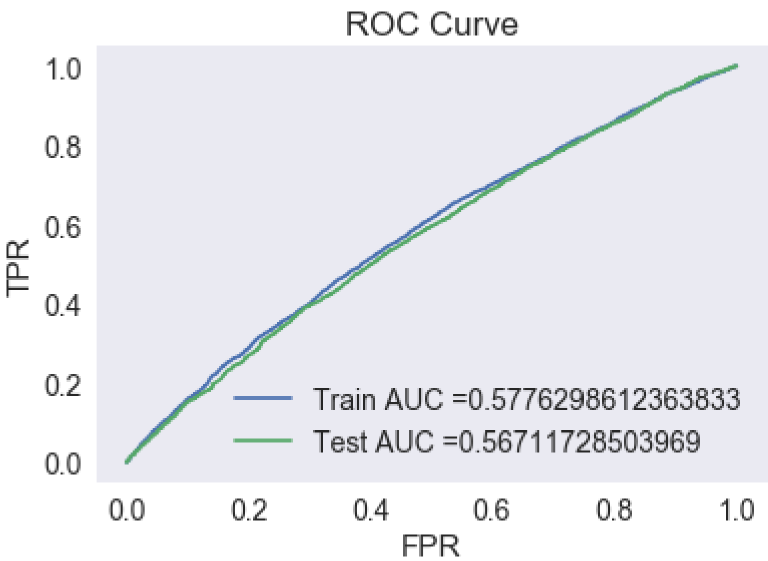
```
100%|██████████████████████████████████████████████████████
██████████████████████████████████████████████████| 7/7 [00:0
3<00:00,  1.35it/s]
```



1.5.5.2 Testing the performance of the model on test data & plotting ROC Curves for train & test data

In [162]:

```
best_c = 0.01 #equivalent to lambda=100

clf10= LogisticRegression(C=best_c, penalty='l2', n_jobs=-1,class_weight='balanced')
clf10.fit(X_tr_cn, y_train)
y_train_pred_cn_best = batch_predict(clf9, X_tr_cn)
y_test_pred_cn_best = batch_predict(clf9, X_test_cn)

train_tpr_cn, train_fpr_cn, tr_thresholds_cn = roc_curve(y_train, y_train_pred_cn_best)
test_tpr_cn, test_fpr_cn, te_thresholds_cn = roc_curve(y_test, y_test_pred_cn_best)

plt.plot(train_tpr_cn, train_fpr_cn,label="Train AUC ="+str(auc(train_tpr_cn, train_fpr
_cn)))
plt.plot(test_tpr_cn, test_fpr_cn, label="Test AUC ="+str(auc(test_tpr_cn, test_fpr_cn
)))
plt.legend()
plt.xlabel("FPR")
plt.ylabel("TPR")
plt.title("ROC Curve")
plt.grid()
plt.show()
```



- From AUC vs C graph it is seen that difference between train & Cv scores is least for value of "C" as 0.01.
- Hence C=0.01 (Lambda=100) is the optimum value with test AUC score = 0.57(57%).

1.5.5.3 Confusion matrices: For best C

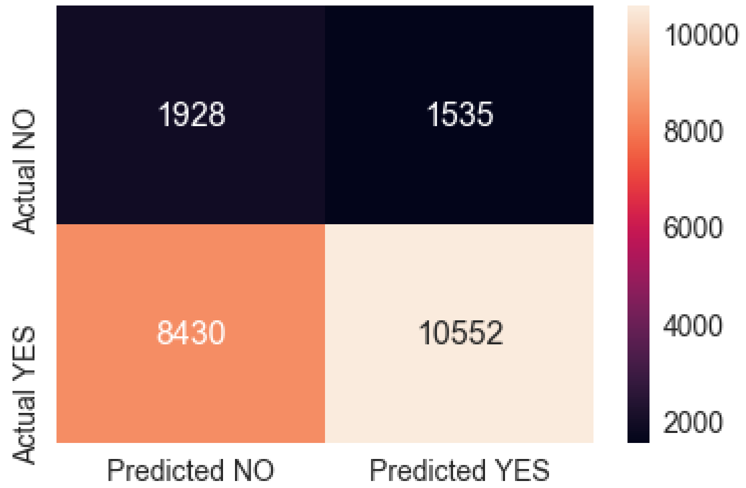
In [163]:

```
from sklearn.metrics import confusion_matrix
best_t_cn = find_best_threshold(tr_thresholds_cn, train_fpr_cn, train_tpr_cn)
print("Train confusion matrix")
cm_train_cn=confusion_matrix(y_train, predict_with_best_t(y_train_pred_cn_best, best_t_
cn))
print(cm_train_cn)
print("Test confusion matrix")
cm_test_cn=confusion_matrix(y_test, predict_with_best_t(y_test_pred_cn_best, best_t_cn
))
print(cm_test_cn)
```

```
The maximum value of tpr*(1-fpr) 0.19685275346192524 for threshold 0.503
Train confusion matrix
[[ 1928  1535]
 [ 8430 10552]]
Test confusion matrix
[[1386 1160]
 [6207 7747]]
```

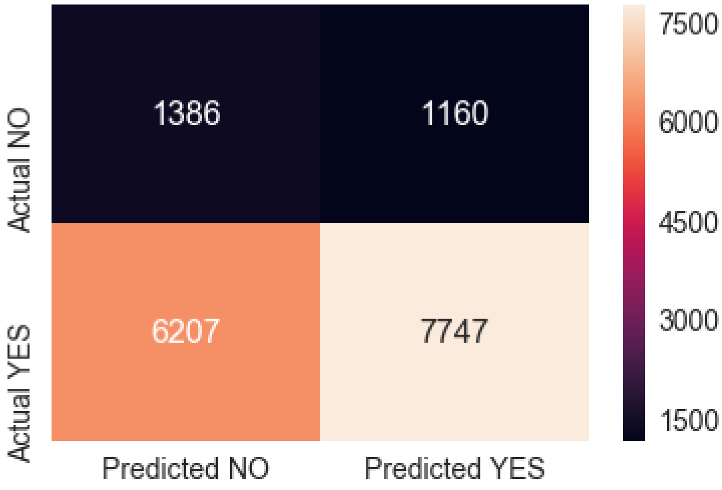
In [164]:

```
# confusion matrix heatmap for train data
cm_heatmap(cm_train_cn)
```



In [165]:

```
# confusion matrix heatmap for test data
cm_heatmap(cm_test_cn)
```



2.0 Summary

In [166]:

```
#Ref: http://zetcode.com/python/prettytable/
from prettytable import PrettyTable

x = PrettyTable()
x.field_names = ["Vectorizer","Hyperparameter(Lambda)", "Test AUC"]
x.add_row(["BOW",100, 0.69])
x.add_row(["TFIDF",50,0.67])
x.add_row(["Avg W2V", 25,0.68])
x.add_row(["TFIDF W2V",25,0.69])
x.add_row(["Categorical & numerical features: Set-5",100,0.57])

print(x)
```

```
+-----------------------------------------+------------------------+------
----+
|                Vectorizer               | Hyperparameter(Lambda) | Test
AUC |
+-----------------------------------------+------------------------+------
----+
|                   BOW                   |          100           |   0.6
9   |
|                  TFIDF                  |           50           |   0.6
7   |
|                 Avg W2V                 |           25           |   0.6
8   |
|                TFIDF W2V                |           25           |   0.6
9   |
| Categorical & numerical features: Set-5 |          100           |   0.5
7   |
+-----------------------------------------+------------------------+------
----+
```

- It can be observed that the test AUC score of the LR model using only categorical & numerical features was 0.57 which was significantly less when compared to LR models using text features encoded in the form of BOW, TFIDF, AvgW2V & TFIDF W2V.
- Hence text features has a lot of impact on the performance of logistic regression models.