DonorsChoose

DonorsChoose.org receives hundreds of thousands of project proposals each year for classroom projects in need of funding. Right now, a large number of volunteers is needed to manually screen each submission before it's approved to be posted on the DonorsChoose.org website.

Next year, DonorsChoose.org expects to receive close to 500,000 project proposals. As a result, there are three main problems they need to solve:

- How to scale current manual processes and resources to screen 500,000 projects so that they can be posted as quickly and as efficiently as possible
- How to increase the consistency of project vetting across different volunteers to improve the experience for teachers
- How to focus volunteer time on the applications that need the most assistance

The goal of the competition is to predict whether or not a DonorsChoose.org project proposal submitted by a teacher will be approved, using the text of project descriptions as well as additional metadata about the project, teacher, and school. DonorsChoose.org can then use this information to identify projects most likely to need further review before approval.

About the DonorsChoose Data Set

The `train.csv` data set provided by DonorsChoose contains the following features:

Feature	Description
`project_id`	A unique identifier for the proposed project. **Example:** `p036502`
`project_title`	Title of the project. **Examples:** • `Art Will Make You Happy!` • `First Grade Fun`
`project_grade_category`	Grade level of students for which the project is targeted. One of the following enumerated values: • `Grades PreK-2` • `Grades 3-5` • `Grades 6-8` • `Grades 9-12`
`project_subject_categories`	One or more (comma-separated) subject categories for the project from the following enumerated list of values: • `Applied Learning` • `Care & Hunger` • `Health & Sports` • `History & Civics` • `Literacy & Language` • `Math & Science` • `Music & The Arts` • `Special Needs` • `Warmth` **Examples:** • `Music & The Arts` • `Literacy & Language, Math & Science`
`school_state`	State where school is located (Two-letter U.S. postal code). **Example:** `WY`
`project_subject_subcategories`	One or more (comma-separated) subject subcategories for the project. **Examples:** • `Literacy`

Feature	Description
	• `Literature & Writing, Social Sciences`
`project_resource_summary`	An explanation of the resources needed for the project. **Example:** • `My students need hands on literacy materials to manage sensory needs!`
`project_essay_1`	First application essay*
`project_essay_2`	Second application essay*
`project_essay_3`	Third application essay*
`project_essay_4`	Fourth application essay*
`project_submitted_datetime`	Datetime when project application was submitted. **Example:** `2016-04-28 12:43:56.245`
`teacher_id`	A unique identifier for the teacher of the proposed project. **Example:** `bdf8baa8fedef6bfeec7ae4ff1c15c56`
`teacher_prefix`	Teacher's title. One of the following enumerated values: • `nan` • `Dr.` • `Mr.` • `Mrs.` • `Ms.` • `Teacher.`
`teacher_number_of_previously_posted_projects`	Number of project applications previously submitted by the same teacher. **Example:** `2`

* See the section **Notes on the Essay Data** for more details about these features.

Additionally, the `resources.csv` data set provides more data about the resources required for each project. Each line in this file represents a resource required by a project:

Feature	Description
`id`	A `project_id` value from the `train.csv` file. **Example:** `p036502`
`description`	Desciption of the resource. **Example:** `Tenor Saxophone Reeds, Box of 25`
`quantity`	Quantity of the resource required. **Example:** `3`
`price`	Price of the resource required. **Example:** `9.95`

Note: Many projects require multiple resources. The `id` value corresponds to a `project_id` in train.csv, so you use it as a key to retrieve all resources needed for a project:

The data set contains the following label (the value you will attempt to predict):

Label	Description
`project_is_approved`	A binary flag indicating whether DonorsChoose approved the project. A value of `0` indicates the project was not approved, and a value of `1` indicates the project was approved.

Notes on the Essay Data

Prior to May 17, 2016, the prompts for the essays were as follows:

- __project_essay_1:__ "Introduce us to your classroom"
- __project_essay_2:__ "Tell us more about your students"
- __project_essay_3:__ "Describe how your students will use the materials you're requesting"
- __project_essay_3:__ "Close by sharing why your project will make a difference"

Starting on May 17, 2016, the number of essays was reduced from 4 to 2, and the prompts for the first 2 essays were changed to the following:

- __project_essay_1:__ "Describe your students: What makes your students special? Specific details about their background, your neighborhood, and your school are all helpful."

your neighborhood, and your school are all helpful.

- __project_essay_2:__ "About your project: How will these materials make a difference in your students' learning and improve their school lives?"

  For all projects with project_submitted_datetime of 2016-05-17 and later, the values of project_essay_3 and project_essay_4 will be NaN.

In [1]:

```
# importing required libraries

%matplotlib inline
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.model_selection import GridSearchCV
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
# Tutorial about Python regular expressions: https://pymotw.com/2/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
from sklearn.model_selection import GridSearchCV
```

```
C:\Users\parik\AppData\Local\Continuum\anaconda3\lib\site-packages\smart_open\ssh.py:34:
UserWarning: paramiko missing, opening SSH/SCP/SFTP paths will be disabled.  `pip install
paramiko` to suppress
  warnings.warn('paramiko missing, opening SSH/SCP/SFTP paths will be disabled.  `pip install
paramiko` to suppress')
```

1.1 Reading Data

In [2]:

```
# Reading data from project and resources data file

project_data = pd.read_csv('E:/Applied AI/Donors choose
dataset/DC_data/Assignments_DonorsChoose_2018/train_data.csv')
resource_data = pd.read_csv('E:/Applied AI/Donors choose
dataset/DC_data/Assignments_DonorsChoose_2018/resources.csv')
```

In [3]:

```
# Getting basic information about the data

print("Number of data points in Project_train data", project_data.shape)
print('-'*100)
print("The attributes of Project_train data :", project_data.columns.values)
print('='*100)
print("Number of data points in Resource_train data", resource_data.shape)
print('-'*100)
print("The attributes of Resource_train data :", resource_data.columns.values)
```

```
Number of data points in Project_train data (109248, 17)
----------------------------------------------------------------------------------------------------
The attributes of Project_train data : ['Unnamed: 0' 'id' 'teacher_id' 'teacher_prefix'
'school_state'
 'project_submitted_datetime' 'project_grade_category'
 'project_subject_categories' 'project_subject_subcategories'
 'project_title' 'project_essay_1' 'project_essay_2' 'project_essay_3'
 'project_essay_4' 'project_resource_summary'
 'teacher_number_of_previously_posted_projects' 'project_is_approved']
====================================================================================================

Number of data points in Resource_train data (1541272, 4)
----------------------------------------------------------------------------------------------------
The attributes of Resource_train data : ['id' 'description' 'quantity' 'price']
```

1.2 Data Pre-Processing

In [4]:

```
# Merge two column text dataframe:
# Merge 4 essays into one:
project_data["essay"] = project_data["project_essay_1"].map(str) +\
                        project_data["project_essay_2"].map(str) + \
                        project_data["project_essay_3"].map(str) + \
                        project_data["project_essay_4"].map(str)

# Merge Price information from resource data to project data
price_data = resource_data.groupby('id').agg({'price':'sum', 'quantity':'sum'}).reset_index()
project_data = pd.merge(project_data, price_data, on='id', how='left')

# find how many digits are present in each project_resource_summary coloumn
summary = list(project_data['project_resource_summary'].values)
presence_of_numeric_data=[]
for i in summary:
    count = 0
    for j in i.split(' '):
        if j.isdigit():
            count+=1
    presence_of_numeric_data.append(count)

# Replace Text summary coloumn with new numerical coloumn presence_of_numeric_data
project_data['numerical_data_in_resource_summary'] = presence_of_numeric_data
project_data.drop(['project_resource_summary'], axis=1, inplace=True)

# how to replace elements in list python: https://stackoverflow.com/a/2582163/4084039
cols = ['Date' if x=='project_submitted_datetime' else x for x in list(project_data.columns)]

#sort dataframe based on time pandas python: https://stackoverflow.com/a/49702492/4084039
project_data['Date'] = pd.to_datetime(project_data['project_submitted_datetime'])
project_data.drop('project_submitted_datetime', axis=1, inplace=True)
project_data.sort_values(by=['Date'], inplace=True)

# how to reorder columns pandas python: https://stackoverflow.com/a/13148611/4084039
project_data = project_data[cols]

# https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.dropna.html
# Here we drop 3 rows where teacher_prefix is having np.nan value
project_data.dropna(axis=0,subset=['teacher_prefix'], inplace=True)

project_data.head(2)
```

Out[4]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_grade_cate
55660	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	Mrs.	CA	2016-04-27 00:27:36	Grades PreK-2
76127	37728	p043609	3f60494c61921b3b43ab61bdde2904df	Ms.	UT	2016-04-27 00:31:25	Grades 3-5

1.2.1 Pre-Processing Essay Text

In [5]:

```
# printing some random essays.
print(project_data['essay'].values[0])
print("="*50)
print(project_data['essay'].values[150])
print("="*50)
```

```
I have been fortunate enough to use the Fairy Tale STEM kits in my classroom as well as the STEM j
ournals, which my students really enjoyed.  I would love to implement more of the Lakeshore STEM k
its in my classroom for the next school year as they provide excellent and engaging STEM
lessons.My students come from a variety of backgrounds, including language and socioeconomic statu
s.  Many of them don't have a lot of experience in science and engineering and these kits give me
the materials to provide these exciting opportunities for my students.Each month I try to do
several science or STEM/STEAM projects.  I would use the kits and robot to help guide my science i
nstruction in engaging and meaningful ways.  I can adapt the kits to my current language arts paci
ng guide where we already teach some of the material in the kits like tall tales (Paul Bunyan) or
Johnny Appleseed.  The following units will be taught in the next school year where I will
implement these kits: magnets, motion, sink vs. float, robots.  I often get to these units and don
't know If I am teaching the right way or using the right materials.    The kits will give me
additional ideas, strategies, and lessons to prepare my students in science.It is challenging to d
evelop high quality science activities.  These kits give me the materials I need to provide my
students with science activities that will go along with the curriculum in my classroom.  Although
I have some things (like magnets) in my classroom, I don't know how to use them effectively.  The
kits will provide me with the right amount of materials and show me how to use them in an
appropriate way.
==================================================
I teach high school English to students with learning and behavioral disabilities. My students all
vary in their ability level. However, the ultimate goal is to increase all students literacy level
s. This includes their reading, writing, and communication levels.I teach a really dynamic group o
f students. However, my students face a lot of challenges. My students all live in poverty and in
a dangerous neighborhood. Despite these challenges, I have students who have the the desire to def
eat these challenges. My students all have learning disabilities and currently all are performing
below grade level. My students are visual learners and will benefit from a classroom that fulfills
their preferred learning style.The materials I am requesting will allow my students to be prepared
for the classroom with the necessary supplies.  Too often I am challenged with students who come t
o school unprepared for class due to economic challenges.  I want my students to be able to focus
on learning and not how they will be able to get school supplies.  The supplies will last all year
.  Students will be able to complete written assignments and maintain a classroom journal.  The ch
art paper will be used to make learning more visual in class and to create posters to aid students
in their learning.  The students have access to a classroom printer.  The toner will be used to pr
int student work that is completed on the classroom Chromebooks.I want to try and remove all barri
ers for the students learning and create opportunities for learning. One of the biggest barriers i
s the students not having the resources to get pens, paper, and folders. My students will be able
to increase their literacy skills because of this project.
==================================================
```

In [6]:

```
# https://stackoverflow.com/a/47091490/4084039
```

```
import re

def decontracted(phrase):
    # specific
    phrase = re.sub(r"won't", "will not", phrase)
    phrase = re.sub(r"can\'t", "can not", phrase)

    # general
    phrase = re.sub(r"n\'t", " not", phrase)
    phrase = re.sub(r"\'re", " are", phrase)
    phrase = re.sub(r"\'s", " is", phrase)
    phrase = re.sub(r"\'d", " would", phrase)
    phrase = re.sub(r"\'ll", " will", phrase)
    phrase = re.sub(r"\'t", " not", phrase)
    phrase = re.sub(r"\'ve", " have", phrase)
    phrase = re.sub(r"\'m", " am", phrase)
    return phrase
```

In [7]:

```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'not'
stopwords= ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've",
\
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his',
'himself', \
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them',
'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll",
'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having',
'do', 'does', \
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', '
while', 'of', \
            'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during',
'before', 'after',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under'
, 'again', 'further',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'e
ach', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'too', 'very', \
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll'
, 'm', 'o', 're', \
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "d
esn't", 'hadn',\
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn',
"mightn't", 'mustn',\
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn',
"wasn't", 'weren', "weren't", \
            'won', "won't", 'wouldn', "wouldn't"]
```

In [8]:

```
from tqdm import tqdm
preprocessed_essays = []
# tqdm is for printing the status bar
for sentance in tqdm(project_data['essay'].values):
    sent = decontracted(sentance)
    sent = sent.replace('\\r', ' ')
    sent = sent.replace('\\"', ' ')
    sent = sent.replace('\\n', ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e not in stopwords)
    preprocessed_essays.append(sent.lower().strip())

# Adding preprocessed_essays coloumn to our data matrix

project_data['preprocessed_essays']=preprocessed_essays
```

```
100%|██████████████████████████████████████████████████████████████████████████| 109245/109245
[01:00<00:00, 1794.30it/s]
```

In [9]:

```
# after preprocesing
preprocessed_essays[100]
```

Out[9]:

```
'a typical day campus exciting my students love learning always put smile face they big personalit
ies even bigger dedication learning they need hero someone willing change future every child
deserves champion adult never give understand power connection insists become best rita pierson we
school 610 low income students grades k 6 we eager bunch love learn we high expectations learning
low resources impact learning with new technology aim high rise occasion i would love incorporate
technology intervention time i empower students empowering students starts able give goals tools n
eed successful school my students use ipads foster love learning remediation enrichment students l
ow income families need engagement motivation succeed ipads bring closer achieving success
classroom real world my project make difference allowing students access programs reinforce classr
oom learning motivating stay focused sustained engagement'
```

1.2.2 Pre-Processing Project Title Text

In [10]:

```
from tqdm import tqdm
preprocessed_titles = []
# tqdm is for printing the status bar
for title in tqdm(project_data['project_title'].values):
    title = decontracted(title)
    title = title.replace('\\r', ' ')
    title = title.replace('\\"', ' ')
    title = title.replace('\\n', ' ')
    title = re.sub('[^A-Za-z0-9]+', ' ', title)
    # https://gist.github.com/sebleier/554280
    title = ' '.join(e for e in title.split() if e not in stopwords)
    preprocessed_titles.append(title.lower().strip())

# Adding preprocessed_titles coloumn to our data matrix

project_data['preprocessed_titles']=preprocessed_titles
preprocessed_titles[1000]
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 109245/109245
[00:02<00:00, 41730.43it/s]
```

Out[10]:

```
'empowering students through art learning about then now'
```

1.2.3 Pre-Processing Project Grades

In [11]:

```
# Remove special characters from grades
from tqdm import tqdm
preprocessed_grade_categories = []
# tqdm is for printing the status bar
for categories in tqdm(project_data['project_grade_category'].values):
    categories = decontracted(categories)
    # https://gist.github.com/sebleier/554280
    categories = '_'.join(e for e in categories.split(' ') if e not in stopwords)
    categories = '_'.join(e for e in categories.split('-') if e not in stopwords)
    preprocessed_grade_categories.append(categories.lower().strip())

# Adding preprocessed_titles coloumn to our data matrix

project_data['preprocessed_grade_category']=preprocessed_grade_categories

project_data.head(5)
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 109245/109245
[00:02<00:00, 52037.59it/s]
```

Out[11]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_grade_cate
55660	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	Mrs.	CA	2016-04-27 00:27:36	Grades PreK-2
76127	37728	p043609	3f60494c61921b3b43ab61bdde2904df	Ms.	UT	2016-04-27 00:31:25	Grades 3-5
51140	74477	p189804	4a97f3a390bfe21b99cf5e2b81981c73	Mrs.	CA	2016-04-27 00:46:53	Grades PreK-2
473	100660	p234804	cbc0e38f522143b86d372f8b43d4cff3	Mrs.	GA	2016-04-27 00:53:00	Grades PreK-2
41558	33679	p137682	06f6e62e17de34fcf81020c77549e1d5	Mrs.	WA	2016-04-27 01:05:25	Grades 3-5

5 rows × 23 columns

1.2.4 preprocessing of `project_subject_categories`

In [12]:

```
catogories = list(project_data['project_subject_categories'].values)
# remove special characters from list of strings python:
https://stackoverflow.com/a/47301924/4084039

# https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
# https://stackoverflow.com/questions/23669024/how-to-strip-a-specific-word-from-a-string
# https://stackoverflow.com/questions/8270092/remove-all-whitespace-in-a-string-in-python
cat_list = []
for i in catogories:
    temp = ""
    # consider we have text like this "Math & Science, Warmth, Care & Hunger"
    for j in i.split(','): # it will split it in three parts ["Math & Science", "Warmth", "Care & H
unger"]
        if 'The' in j.split(): # this will split each of the catogory based on space "Math & Scienc
e"=> "Math","&", "Science"
            j=j.replace('The','') # if we have the words "The" we are going to replace it with ''(i
.e removing 'The')
        j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math &
Science"=>"Math&Science"
        temp+=j.strip()+" " #" abc ".strip() will return "abc", remove the trailing spaces
        temp = temp.replace('&','_') # we are replacing the & value into
    cat_list.append(temp.strip())
```

```
project_data['clean_categories'] = cat_list
project_data.drop(['project_subject_categories'], axis=1, inplace=True)
```

1.2.5 preprocessing of `project_subject_subcategories`

In [13]:

```
sub_catogories = list(project_data['project_subject_subcategories'].values)
# remove special characters from list of strings python:
https://stackoverflow.com/a/47301924/4084039

# https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
# https://stackoverflow.com/questions/23669024/how-to-strip-a-specific-word-from-a-string
# https://stackoverflow.com/questions/8270092/remove-all-whitespace-in-a-string-in-python

sub_cat_list = []
for i in sub_catogories:
    temp = ""
    # consider we have text like this "Math & Science, Warmth, Care & Hunger"
    for j in i.split(','): # it will split it in three parts ["Math & Science", "Warmth", "Care & H
unger"]
        if 'The' in j.split(): # this will split each of the catogory based on space "Math & Scienc
e"=> "Math","&", "Science"
            j=j.replace('The','') # if we have the words "The" we are going to replace it with ''(i
.e removing 'The')
        j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math &
Science"=>"Math&Science"
        temp +=j.strip()+" "#" abc ".strip() will return "abc", remove the trailing spaces
        temp = temp.replace('&','_')
    sub_cat_list.append(temp.strip())

project_data['clean_subcategories'] = sub_cat_list
project_data.drop(['project_subject_subcategories'], axis=1, inplace=True)
```

In [14]:

```
# Drop all unnecessary featurs like project_grade_category, project_essay_1, etc.
project_data.drop(['project_grade_category'], axis=1, inplace=True)
project_data.drop(['project_essay_1'], axis=1, inplace=True)
project_data.drop(['project_essay_2'], axis=1, inplace=True)
project_data.drop(['project_essay_3'], axis=1, inplace=True)
project_data.drop(['project_essay_4'], axis=1, inplace=True)
project_data.drop(['essay'], axis=1, inplace=True)
```

In [15]:

```
project_data.head(5)
```

Out[15]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_title	teache
55660	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	Mrs.	CA	2016-04-27 00:27:36	Engineering STEAM into the Primary Classroom	53
76127	37728	p043609	3f60494c61921b3b43ab61bdde2904df	Ms.	UT	2016-04-27 00:31:25	Sensory Tools for Focus	4
51140	74477	p189804	4a97f3a390bfe21b99cf5e2b81981c73	Mrs.	CA	2016-04-27 00:46:53	Mobile Learning with a Mobile Listening Center	10
						2016	Flexible	

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_title	teache
473	100660	p234804	cbc0e38f522143b86d372f8b43d4cff3	Mrs.	CA	2016-04-27 00:53:00	Seating for Flexible Learning	2
41558	33679	p137682	06f6e62e17de34fcf81020c77549e1d5	Mrs.	WA	2016-04-27 01:05:25	Going Deep: The Art of Inner Thinking!	2

1.2.6 Add Sentiment Score of Preprocessed Essays

In [16]:

```
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer
neg_essay=[]
neu_essay=[]
pos_essay=[]
comp_essay=[]

sid = SentimentIntensityAnalyzer()

for sent in preprocessed_titles:

    ss = sid.polarity_scores(sent)
    neg_essay.append(ss.get('neg'))
    neu_essay.append(ss.get('neu'))
    pos_essay.append(ss.get('pos'))
    comp_essay.append(ss.get('compound'))

project_data['neg_essay']=neg_essay
project_data['neu_essay']=neu_essay
project_data['pos_essay']=pos_essay
project_data['comp_essay']=comp_essay


# we can use these 4 things as features/attributes (neg, neu, pos, compound)
# neg: 0.0, neu: 0.753, pos: 0.247, compound: 0.93
```

```
C:\Users\parik\AppData\Local\Continuum\anaconda3\lib\site-packages\nltk\twitter\__init__.py:20: UserWarning:

The twython library has not been installed. Some functionality from the twitter package will not be available.
```

In [17]:

```
project_data.head(5)
```

Out[17]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_title	teache
55660	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	Mrs.	CA	2016-04-27 00:27:36	Engineering STEAM into the Primary Classroom	53
76127	37728	p043609	3f60494c61921b3b43ab61bdde2904df	Ms.	UT	2016-04-27 00:31:25	Sensory Tools for Focus	4
51140	74477	p189804	4a97f3a390bfe21b99cf5e2b81981c73	Mrs.	CA	2016-04-27 00:46:53	Mobile Learning with a Mobile Listening	10

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_title	teache
473	100660	p234804	cbc0e38f522143b86d372f8b43d4cff3	Mrs.	GA	2016-04-27 00:53:00	Flexible Seating for Flexible Learning	2
41558	33679	p137682	06f6e62e17de34fcf81020c77549e1d5	Mrs.	WA	2016-04-27 01:05:25	Going Deep: The Art of Inner Thinking!	2

5 rows × 21 columns

1.2.7 Adding number of words in title and number of words in essays as two new numerical features

In [18]:

```
number_of_words_in_title=[]
for title in project_data['project_title'].values:
    list_of_words = title.split()
    number_of_words_in_title.append(len(list_of_words))

number_of_words_in_essays=[]
for title in project_data['preprocessed_essays'].values:
    list_of_words = title.split()
    number_of_words_in_essays.append(len(list_of_words))

project_data['number_of_words_in_title'] = number_of_words_in_title
project_data['number_of_words_in_essays'] = number_of_words_in_essays
```

In [19]:

```
project_data.head()
```

Out[19]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_title	teache
55660	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	Mrs.	CA	2016-04-27 00:27:36	Engineering STEAM into the Primary Classroom	53
76127	37728	p043609	3f60494c61921b3b43ab61bdde2904df	Ms.	UT	2016-04-27 00:31:25	Sensory Tools for Focus	4
51140	74477	p189804	4a97f3a390bfe21b99cf5e2b81981c73	Mrs.	CA	2016-04-27 00:46:53	Mobile Learning with a Mobile Listening Center	10
473	100660	p234804	cbc0e38f522143b86d372f8b43d4cff3	Mrs.	GA	2016-04-27 00:53:00	Flexible Seating for Flexible Learning	2
41558	33679	p137682	06f6e62e17de34fcf81020c77549e1d5	Mrs.	WA	2016-04-27 01:05:25	Going Deep: The Art of Inner Thinking!	2

5 rows × 23 columns

1.3 Sampling data for LR Assignment

In [20]:

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

Out[20]:

```
1    92703
0    16542
Name: project_is_approved, dtype: int64
```

In [21]:

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

Out[21]:

```
1    92703
0    16542
Name: project_is_approved, dtype: int64
```

In [22]:

```
data.head(5)
```

Out[22]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_title	teache
55660	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	Mrs.	CA	2016-04-27 00:27:36	Engineering STEAM into the Primary Classroom	53
76127	37728	p043609	3f60494c61921b3b43ab61bdde2904df	Ms.	UT	2016-04-27 00:31:25	Sensory Tools for Focus	4
51140	74477	p189804	4a97f3a390bfe21b99cf5e2b81981c73	Mrs.	CA	2016-04-27 00:46:53	Mobile Learning with a Mobile Listening Center	10
473	100660	p234804	cbc0e38f522143b86d372f8b43d4cff3	Mrs.	GA	2016-04-27 00:53:00	Flexible Seating for Flexible Learning	2
41558	33679	p137682	06f6e62e17de34fcf81020c77549e1d5	Mrs.	WA	2016-04-27 01:05:25	Going Deep: The Art of Inner Thinking!	2

5 rows × 23 columns

In [23]:

```
# Split the class label from data
y = data['project_is_approved'].values
X = data.drop(['project_is_approved'], axis=1)
X.head(1)
```

Out[23]:

Out[25]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_title	teache
55660	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	Mrs.	CA	2016-04-27 00:27:36	Engineering STEAM into the Primary Classroom	53

1 rows × 22 columns

2.1 Splitting data into Train and cross validation(or test): Stratified Sampling And Simple Upsampling

In [24]:

```
# train test split
# Not using CV data as it will be done by the GridsearchCV internally
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,
stratify=y_train)
```

In [25]:

```
# Simple Upsampling for negative class data points in training dataset
# https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

from sklearn.utils import resample
#df3 = pd.DataFrame(y_train,columns=['project_is_approved'],dtype = int)

#X = pd.concat([X_train,df3],axis = 1)
X_train['project_is_approved']=y_train
Accepted, Rejected = X_train.project_is_approved.value_counts()

# Divide by class
df_class_0 = X_train[X_train['project_is_approved'] == 0]
df_class_1 = X_train[X_train['project_is_approved'] == 1]

upsampled_data = df_class_0.sample(Accepted, replace=True,)
X_train = pd.concat([df_class_1, upsampled_data], axis=0)
print(X_train.project_is_approved.value_counts())
```

```
1    62111
0    62111
Name: project_is_approved, dtype: int64
```

In [26]:

```
y_train = X_train.project_is_approved
X_train = X_train.drop('project_is_approved', axis=1)
X_train.shape
```

Out[26]:

```
(124222, 22)
```

2.2 Make Data Model Ready:

2.2.1 Encoding numerical, categorical features

2.2.1.1 Encoding School State

In [27]:

```
# Encoding School State

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_ohe = vectorizer.transform(X_train['school_state'].values)
#X_cv_state_ohe = vectorizer.transform(X_cv['school_state'].values)
X_test_state_ohe = vectorizer.transform(X_test['school_state'].values)

print("After vectorizations")
print(X_train_state_ohe.shape, y_train.shape)
#print(X_cv_state_ohe.shape, y_cv.shape)
print(X_test_state_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(124222, 51) (124222,)
(36051, 51) (36051,)
['ak', 'al', 'ar', 'az', 'ca', 'co', 'ct', 'dc', 'de', 'fl', 'ga', 'hi', 'ia', 'id', 'il', 'in', 'k
s', 'ky', 'la', 'ma', 'md', 'me', 'mi', 'mn', 'mo', 'ms', 'mt', 'nc', 'nd', 'ne', 'nh', 'nj', 'nm',
'nv', 'ny', 'oh', 'ok', 'or', 'pa', 'ri', 'sc', 'sd', 'tn', 'tx', 'ut', 'va', 'vt', 'wa', 'wi', 'wv
', 'wy']
====================================================================================================
```

2.2.1.2 Encoding Teacher Prefix

In [28]:

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

# we use the fitted CountVectorizer to convert the text to vector
X_train_teacher_ohe = vectorizer.transform(X_train['teacher_prefix'].values)
#X_cv_teacher_ohe = vectorizer.transform(X_cv['teacher_prefix'].values)
X_test_teacher_ohe = vectorizer.transform(X_test['teacher_prefix'].values)

print("After vectorizations")
print(X_train_teacher_ohe.shape, y_train.shape)
#print(X_cv_teacher_ohe.shape, y_cv.shape)
print(X_test_teacher_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(124222, 5) (124222,)
(36051, 5) (36051,)
['dr', 'mr', 'mrs', 'ms', 'teacher']
====================================================================================================
```

2.2.1.3 Encoding preprocessed_grade_category

In [29]:

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

# we use the fitted CountVectorizer to convert the text to vector
X_train_grade_ohe = vectorizer.transform(X_train['preprocessed_grade_category'].values)
#X_cv_grade_ohe = vectorizer.transform(X_cv['preprocessed_grade_category'].values)
X_test_grade_ohe = vectorizer.transform(X_test['preprocessed_grade_category'].values)

print("After vectorizations")
print(X_train_grade_ohe.shape, y_train.shape)
#print(X_cv_grade_ohe.shape, y_cv.shape)
```

```
print(X_test_grade_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(124222, 4) (124222,)
(36051, 4) (36051,)
['grades_3_5', 'grades_6_8', 'grades_9_12', 'grades_prek_2']
====================================================================================================
```

2.2.1.4 Encoding numerical feature Price

In [30]:

```
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:
# array=[105.22 215.96  96.01 ... 368.98  80.53 709.67].
# Reshape your data either using
# array.reshape(-1, 1) if your data has a single feature
# array.reshape(1, -1)  if it contains a single sample.
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))

X_train_price_norm = X_train_price_norm.reshape(-1,1)
X_test_price_norm = X_test_price_norm.reshape(-1,1)


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

```
After vectorizations
(124222, 1) (124222,)
[[1.22409549e-03]
 [8.30971622e-03]
 [1.71327929e-03]
 ...
 [5.30604500e-04]
 [2.87166092e-03]
 [6.31505576e-05]]
(36051, 1) (36051,)
====================================================================================================
```

2.2.1.5 Encoding numeric feature Quantity

In [31]:

```
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:
# array=[105.22 215.96  96.01 ... 368.98  80.53 709.67].
# Reshape your data either using
# array.reshape(-1, 1) if your data has a single feature
# array.reshape(1, -1)  if it contains a single sample.
normalizer.fit(X_train['quantity'].values.reshape(1,-1))

X_train_quantity_norm = normalizer.transform(X_train['quantity'].values.reshape(1,-1))
X_train_quantity_norm = X_train_quantity_norm.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))
```

```
X_test_quantity_norm = X_test_quantity_norm.reshape(-1,1)
print(X_train_quantity_norm)
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)
```

```
[[0.00250189]
 [0.00336461]
 [0.00025882]
 ...
 [0.00069018]
 [0.00077645]
 [0.00086272]]
After vectorizations
(124222, 1) (124222,)
(36051, 1) (36051,)
====================================================================================================
```

2.2.1.6 Encoding numeric feature teacher_number_of_previously_posted_projects

In [32]:

```
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:
# array=[105.22 215.96  96.01 ... 368.98  80.53 709.67].
# Reshape your data either using
# array.reshape(-1, 1) if your data has a single feature
# array.reshape(1, -1)  if it contains a single sample.
normalizer.fit(X_train['teacher_number_of_previously_posted_projects'].values.reshape(1,-1))
#List_of_imp_features.append('teacher_number_of_previously_posted_projects')
X_train_teacher_number_of_previously_posted_projects_norm =
normalizer.transform(X_train['teacher_number_of_previously_posted_projects'].values.reshape(1,-1))
#X_cv_teacher_number_of_previously_posted_projects_norm =
normalizer.transform(X_cv['teacher_number_of_previously_posted_projects'].values.reshape(1,-1))
X_test_teacher_number_of_previously_posted_projects_norm =
normalizer.transform(X_test['teacher_number_of_previously_posted_projects'].values.reshape(1,-1))
X_train_teacher_number_of_previously_posted_projects_norm =
X_train_teacher_number_of_previously_posted_projects_norm.reshape(-1,1)
X_test_teacher_number_of_previously_posted_projects_norm =
X_test_teacher_number_of_previously_posted_projects_norm.reshape(-1,1)

print(X_test_teacher_number_of_previously_posted_projects_norm)
print("After vectorizations")
print(X_train_teacher_number_of_previously_posted_projects_norm.shape, y_train.shape)
#print(X_cv_teacher_number_of_previously_posted_projects_norm.shape, y_cv.shape)
print(X_test_teacher_number_of_previously_posted_projects_norm.shape, y_test.shape)
print("="*100)
```

```
[[0.00017362]
 [0.00017362]
 [0.00034723]
 ...
 [0.        ]
 [0.00156255]
 [0.        ]]
After vectorizations
(124222, 1) (124222,)
(36051, 1) (36051,)
====================================================================================================
```

2.2.1.7 Encoding numeric feature numerical_data_in_resource_summary

In [33]:

```
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:
# array=[105.22 215.96  96.01 ... 368.98  80.53 709.67].
# Reshape your data either using
# array.reshape(-1, 1) if your data has a single feature
# array.reshape(1, -1)  if it contains a single sample.
normalizer.fit(X_train['numerical_data_in_resource_summary'].values.reshape(1,-1))
X_train_numerical_data_in_resource_summary_norm =
normalizer.transform(X_train['numerical_data_in_resource_summary'].values.reshape(1,-1))
#X_cv_numerical_data_in_resource_summary_norm =
normalizer.transform(X_cv['numerical_data_in_resource_summary'].values.reshape(1,-1))
X_test_numerical_data_in_resource_summary_norm =
normalizer.transform(X_test['numerical_data_in_resource_summary'].values.reshape(1,-1))

X_train_numerical_data_in_resource_summary_norm = X_train_numerical_data_in_resource_summary_norm.
reshape(-1,1)
X_test_numerical_data_in_resource_summary_norm = X_test_numerical_data_in_resource_summary_norm.re
shape(-1,1)

print(X_test_numerical_data_in_resource_summary_norm)
print("After vectorizations")
print(X_train_numerical_data_in_resource_summary_norm.shape, y_train.shape)
#print(X_cv_numerical_data_in_resource_summary_norm.shape, y_cv.shape)
print(X_test_numerical_data_in_resource_summary_norm.shape, y_test.shape)
print("="*100)
```

```
[[0.]
 [0.]
 [0.]
 ...
 [0.]
 [0.]
 [0.]]
After vectorizations
(124222, 1) (124222,)
(36051, 1) (36051,)
====================================================================================================
```

2.2.1.8 Encoding numeric feature number_of_words_in_title

In [34]:

```
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:
# array=[105.22 215.96  96.01 ... 368.98  80.53 709.67].
# Reshape your data either using
# array.reshape(-1, 1) if your data has a single feature
# array.reshape(1, -1)  if it contains a single sample.
normalizer.fit(X_train['number_of_words_in_title'].values.reshape(1,-1))

X_train_number_of_words_in_title = normalizer.transform(X_train['number_of_words_in_title'].values
.reshape(1,-1))
#X_cv_price_norm = normalizer.transform(X_cv['price'].values.reshape(1,-1))
X_test_number_of_words_in_title =
normalizer.transform(X_test['number_of_words_in_title'].values.reshape(1,-1))

X_train_number_of_words_in_title = X_train_number_of_words_in_title.reshape(-1,1)
X_test_number_of_words_in_title = X_test_number_of_words_in_title.reshape(-1,1)


print("After vectorizations")
print(X_train_number_of_words_in_title.shape, y_train.shape)
print(X_train_number_of_words_in_title)
#print(X_cv_price_norm.shape, y_cv.shape)
print(X_test_number_of_words_in_title.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(124222, 1) (124222,)
[[0.00154322]
```

```
 [0.00257203]
 [0.00102881]
 ...
 [0.00154322]
 [0.00205762]
 [0.00154322]]
(36051, 1) (36051,)
==========================================================================================================
```

2.2.1.9 Encoding numeric feature number_of_words_in_essay

In [35]:

```
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:
# array=[105.22 215.96  96.01 ... 368.98  80.53 709.67].
# Reshape your data either using
# array.reshape(-1, 1) if your data has a single feature
# array.reshape(1, -1)  if it contains a single sample.
normalizer.fit(X_train['number_of_words_in_essays'].values.reshape(1,-1))

X_train_number_of_words_in_essay = normalizer.transform(X_train['number_of_words_in_essays'].value
s.reshape(1,-1))
#X_cv_price_norm = normalizer.transform(X_cv['price'].values.reshape(1,-1))
X_test_number_of_words_in_essay = normalizer.transform(X_test['number_of_words_in_essays'].values.
reshape(1,-1))

X_train_number_of_words_in_essay = X_train_number_of_words_in_essay.reshape(-1,1)
X_test_number_of_words_in_essay = X_test_number_of_words_in_essay.reshape(-1,1)


print("After vectorizations")
print(X_train_number_of_words_in_essay.shape, y_train.shape)
print(X_train_number_of_words_in_essay)
#print(X_cv_price_norm.shape, y_cv.shape)
print(X_test_number_of_words_in_essay.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(124222, 1) (124222,)
[[0.00215044]
 [0.00233583]
 [0.00228021]
 ...
 [0.00281782]
 [0.00224313]
 [0.00344812]]
(36051, 1) (36051,)
==========================================================================================================
```

2.2.1.10 Encoding numeric features of sentiment Score

In [36]:

```
train_neg_essay = X_train['neg_essay'].values.reshape(-1,1)
test_neg_essay = X_test['neg_essay'].values.reshape(-1,1)

train_neu_essay = X_train['neu_essay'].values.reshape(-1,1)
test_neu_essay = X_test['neu_essay'].values.reshape(-1,1)

train_pos_essay = X_train['pos_essay'].values.reshape(-1,1)
test_pos_essay = X_test['pos_essay'].values.reshape(-1,1)

train_comp_essay = X_train['comp_essay'].values.reshape(-1,1)
test_comp_essay = X_test['comp_essay'].values.reshape(-1,1)
```

2.3 Appling Logistic Regression on different kind of featurization as mentioned in the instructions

In [37]:

```
#Setting the values for alpha
import math
a=[]
for i in range(-7,-1):
    a.append(10**i)

log_alpha = []
for i in a:
    log_alpha.append(math.log(i,10))

print(a)
```

```
[1e-07, 1e-06, 1e-05, 0.0001, 0.001, 0.01]
```

In [42]:

```
# Define Functions for Train LR model, Test LR Model and Plot the graphs for different featurizati
ons

import matplotlib.pyplot as plt
from sklearn import linear_model
from sklearn.metrics import roc_auc_score


def train_LR(X_tr,y_train):

    train_score=[]
    test_score=[]

    lr = linear_model.SGDClassifier(loss='log',class_weight="balanced")
    #create a dictionary of all values we want to test for alpha values
    param_grid = {'alpha': a}
#use gridsearch to test all values for alpha
    LR_gscv = GridSearchCV(lr, param_grid, cv=2, scoring='roc_auc', return_train_score=True)
    LR_gscv.fit(X_tr, y_train)
    print(LR_gscv.best_params_)

    print(LR_gscv.cv_results_.keys())
    for key, value in LR_gscv.cv_results_.items():
        if key == "mean_train_score":
            train_score = value
        if key == "mean_test_score":
            test_score = value


    plt.plot(log_alpha, train_score, label='Train AUC')
    plt.plot(log_alpha, test_score, label='CV AUC')

    plt.scatter(log_alpha, train_score, label='Train AUC points')
    plt.scatter(log_alpha, test_score, label='CV AUC points')

    plt.legend()
    plt.xlabel("alpha: hyperparameter")
    plt.ylabel("AUC")
    plt.title("ERROR PLOTS")
    plt.grid()
    plt.show()

# Test the model with optimal alpha found out using training data. Plot FPR vs TPR(ROC curves) for
training and testing data

def test_LR(X_tr,X_te,best_a):

    y_train_pred=[]
    y_test_pred=[]

    from sklearn.metrics import roc_curve, auc

    lr = linear_model.SGDClassifier(loss='log',class_weight="balanced")
```

```
    lr=lr.fit(X_tr, y_train)

    y_train_pred_raw = lr.predict_proba(X_tr)

    y_test_pred_raw = lr.predict_proba(X_te)


    for i in y_train_pred_raw:
        y_train_pred.append(i[1])
    for i in y_test_pred_raw:
        y_test_pred.append(i[1])




    train_fpr, train_tpr, tr_thresholds = roc_curve(y_train, y_train_pred)
    test_fpr, test_tpr, te_thresholds = roc_curve(y_test, y_test_pred)

    plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tpr)))
    plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
    plt.legend()
    plt.xlabel("tpr")
    plt.ylabel("fpr")
    plt.title("ERROR PLOTS")
    plt.grid()
    plt.show()
    return train_fpr,train_tpr,tr_thresholds,y_train_pred,y_test_pred

# we are writing our own function for predict, with defined thresould
# we will pick a threshold that will give the least fpr
def find_best_threshold(threshould, fpr, tpr):
    t = threshould[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.round(t,3))
    return t

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

2.3.1 Applying Logistic Regression on BOW encoding eassay, and project_title

2.3.1.1 Encoding preprocessed_essays BoW (Bigrams, min_df = 10 and Max_features =5k)

In [39]:

```
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=(2,2), max_features=5000)
vectorizer.fit(X_train['preprocessed_essays'].values) # fit has to happen only on train data

# we use the fitted 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)

print("After vectorizations")
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)
```

```
# print("YOU SHOULD NOT DO SOMETHING LIKE THIS")
# vectorizer = CountVectorizer()
# x_train_bow = vectorizer.fit_transform(X_train['essay'].values)
# x_cv_bow = vectorizer.fit_transform(X_cv['essay'].values)
# x_test_bow = vectorizer.fit_transform(X_test['essay'].values)

# print(x_train_bow.shape, y_train.shape)
# print(x_cv_bow.shape, y_cv.shape)
# print(x_test_bow.shape, y_test.shape)

print("NOTE: THE NUMBER OF COLUMNS IN EACH OF THE VECTOR WONT BE SAME")
```

```
(124222, 22) (124222,)
(36051, 22) (36051,)
==========================================================================================================

After vectorizations
(124222, 5000) (124222,)
(36051, 5000) (36051,)
==========================================================================================================

NOTE: THE NUMBER OF COLUMNS IN EACH OF THE VECTOR WONT BE SAME
```

2.3.1.2 Encoding preprocessed_titles BoW

In [40]:

```
vectorizer = CountVectorizer(min_df=10,ngram_range=(1,4), max_features=10000)
vectorizer.fit(X_train['preprocessed_titles'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_titles_bow = vectorizer.transform(X_train['preprocessed_titles'].values)
#X_cv_titles_bow = vectorizer.transform(X_cv['preprocessed_titles'].values)
X_test_titles_bow = vectorizer.transform(X_test['preprocessed_titles'].values)

print("After vectorizations")
print(X_train_titles_bow.shape, y_train.shape)
#print(X_cv_titles_bow.shape, y_cv.shape)
print(X_test_titles_bow.shape, y_test.shape)
print("="*100)


# print("YOU SHOULD NOT DO SOMETHING LIKE THIS")
# vectorizer = CountVectorizer()
# x_train_bow = vectorizer.fit_transform(X_train['essay'].values)
# x_cv_bow = vectorizer.fit_transform(X_cv['essay'].values)
# x_test_bow = vectorizer.fit_transform(X_test['essay'].values)

# print(x_train_bow.shape, y_train.shape)
# print(x_cv_bow.shape, y_cv.shape)
# print(x_test_bow.shape, y_test.shape)

print("NOTE: THE NUMBER OF COLUMNS IN EACH OF THE VECTOR WONT BE SAME")
```

```
After vectorizations
(124222, 10000) (124222,)
(36051, 10000) (36051,)
==========================================================================================================

NOTE: THE NUMBER OF COLUMNS IN EACH OF THE VECTOR WONT BE SAME
```

2.3.1.3 Merge all the features and obtain final data matrix

In [41]:

```
# Merge all the features:
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
# Matrix are merged in such a way that the order is preserved in # Matrix are merged in such a way
that the order is preserved in List_of_imp_features list
```

```
from scipy.sparse import hstack
X_tr = hstack((X_train_state_ohe, X_train_teacher_ohe, X_train_grade_ohe,
X_train_price_norm,X_train_quantity_norm,
X_train_teacher_number_of_previously_posted_projects_norm,X_train_numerical_data_in_resource_summar
_norm,X_train_essay_bow,X_train_titles_bow )).tocsr()
#X_cr = hstack((X_cv_titles_bow,X_cv_essay_bow, X_cv_state_ohe, X_cv_teacher_ohe, X_cv_grade_ohe,
X_cv_price_norm, X_cv_numerical_data_in_resource_summary_norm,
X_cv_teacher_number_of_previously_posted_projects_norm, X_cv_quantity_norm)).tocsr()
X_te = hstack((X_test_state_ohe, X_test_teacher_ohe, X_test_grade_ohe,
X_test_price_norm,X_test_quantity_norm, X_test_teacher_number_of_previously_posted_projects_norm,X
_test_numerical_data_in_resource_summary_norm,X_test_essay_bow,X_test_titles_bow )).tocsr()

print("Final Data matrix")
print(X_tr.shape, y_train.shape)
#print(X_cr.shape, y_cv.shape)
print(X_te.shape, y_test.shape)
print("="*100)
```

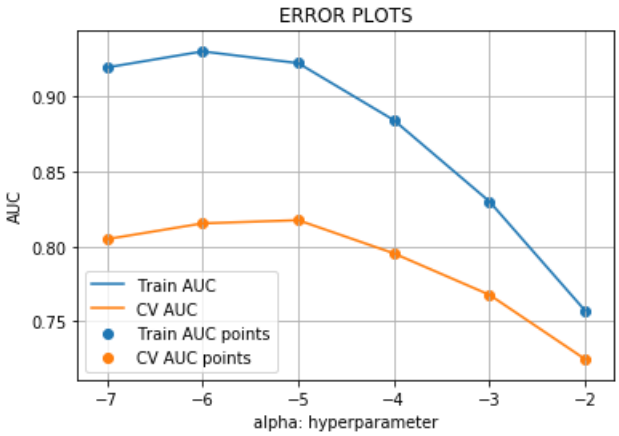
```
Final Data matrix
(124222, 15064) (124222,)
(36051, 15064) (36051,)
====================================================================================================
```

2.3.1.4 Training the data model and find best hyperparameter using ROC-AUC

In [43]:

```
# Call train_LR function on above data

train_LR(X_tr,y_train)
```

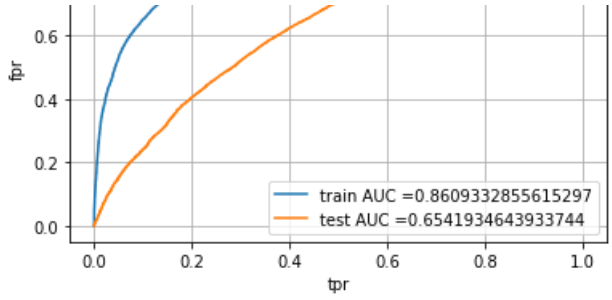
```
{'alpha': 1e-05}
dict_keys(['mean_fit_time', 'std_fit_time', 'mean_score_time', 'std_score_time', 'param_alpha', 'p
arams', 'split0_test_score', 'split1_test_score', 'mean_test_score', 'std_test_score',
'rank_test_score', 'split0_train_score', 'split1_train_score', 'mean_train_score',
'std_train_score'])
```



2.3.1.5 Testing the performance of the model on test data, plotting ROC Curves

In [44]:

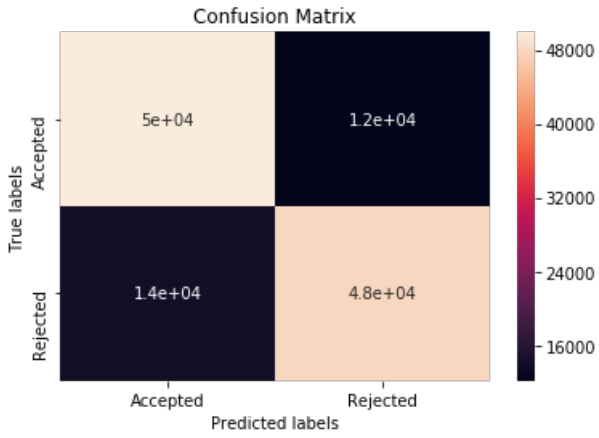
```
# Call test_LR for a obtained by training the data

best_a=0.00001
train_fpr,train_tpr,tr_thresholds,y_train_pred,y_test_pred=test_LR(X_tr,X_te,best_a)
```



In [45]:

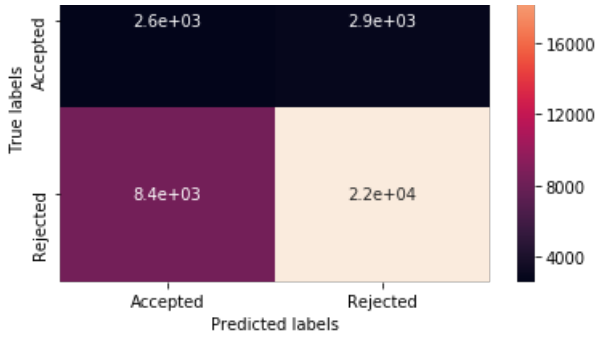
```
print("="*100)
from sklearn.metrics import confusion_matrix
best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")
print(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)))
print("Test confusion matrix")
print(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)))

ax= plt.subplot()
cm=confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t))
sns.heatmap(cm, annot=True, ax = ax); #annot=True to annotate cells

# labels, title and ticks
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
ax.xaxis.set_ticklabels(['Accepted', 'Rejected']); ax.yaxis.set_ticklabels(['Accepted', 'Rejected'
]);
```

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

the maximum value of tpr*(1-fpr) 0.6204636769556241 for threshold 0.519
Train confusion matrix
[[49947 12164]
 [14188 47923]]
Test confusion matrix
[[ 2595  2864]
 [ 8404 22188]]
```



In [46]:

```
ax= plt.subplot()
cm=confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t))
sns.heatmap(cm, annot=True, ax = ax); #annot=True to annotate cells

# labels, title and ticks
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
ax.xaxis.set_ticklabels(['Accepted', 'Rejected']); ax.yaxis.set_ticklabels(['Accepted', 'Rejected'
]);
```



2.3.2 Applying LR on TFIDF encoding eassay, and project_title

2.3.2.1 Encoding preprocessed_titles TFIDF

In [47]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(min_df=10)

#vectorizer = CountVectorizer(min_df=10,ngram_range=(1,4), max_features=10000)
vectorizer.fit(X_train['preprocessed_titles'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_titles_tfidf = vectorizer.transform(X_train['preprocessed_titles'].values)
#X_cv_titles_tfidf = vectorizer.transform(X_cv['preprocessed_titles'].values)
X_test_titles_tfidf = vectorizer.transform(X_test['preprocessed_titles'].values)

print("After vectorizations")
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 vectorizations
(124222, 3720) (124222,)
(36051, 3720) (36051,)
====================================================================================================
```

2.3.2.2 Encoding preprocessed_essays TFIDF (Bigrams, Min_df=10, max_features = 5k)

In [48]:

```
vectorizer = TfidfVectorizer(min_df=10,ngram_range=(2,2), max_features=5000)
vectorizer.fit(X_train['preprocessed_essays'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
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)

print("After vectorizations")
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 vectorizations
(124222, 5000) (124222,)
(36051, 5000) (36051,)
====================================================================================================
```

2.3.2.3 Merge all the features and obtain final data matrix

In [49]:

```
from scipy.sparse import hstack
X_tr = hstack((X_train_titles_tfidf,X_train_essay_tfidf, X_train_state_ohe, X_train_teacher_ohe,
X_train_grade_ohe, X_train_price_norm, X_train_numerical_data_in_resource_summary_norm,
X_train_teacher_number_of_previously_posted_projects_norm, X_train_quantity_norm)).tocsr()
from scipy.sparse import hstack
X_tr = hstack((X_train_titles_tfidf,X_train_essay_tfidf, X_train_state_ohe, X_train_teacher_ohe,
X_train_grade_ohe, X_train_price_norm, X_train_numerical_data_in_resource_summary_norm,
X_train_teacher_number_of_previously_posted_projects_norm, X_train_quantity_norm)).tocsr()
#X_cr = hstack((X_cv_titles_tfidf,X_cv_essay_tfidf, X_cv_state_ohe, X_cv_teacher_ohe,
X_cv_grade_ohe, X_cv_price_norm, X_cv_numerical_data_in_resource_summary_norm,
X_cv_teacher_number_of_previously_posted_projects_norm, X_cv_quantity_norm)).tocsr()
X_te = hstack((X_test_titles_tfidf,X_test_essay_tfidf, X_test_state_ohe, X_test_teacher_ohe, X_test
_grade_ohe, X_test_price_norm,X_test_numerical_data_in_resource_summary_norm,
X_test_teacher_number_of_previously_posted_projects_norm, X_test_quantity_norm)).tocsr()

print("Final Data matrix")
print(X_tr.shape, y_train.shape)
#print(X_cr.shape, y_cv.shape)
print(X_te.shape, y_test.shape)
print("="*100)
```

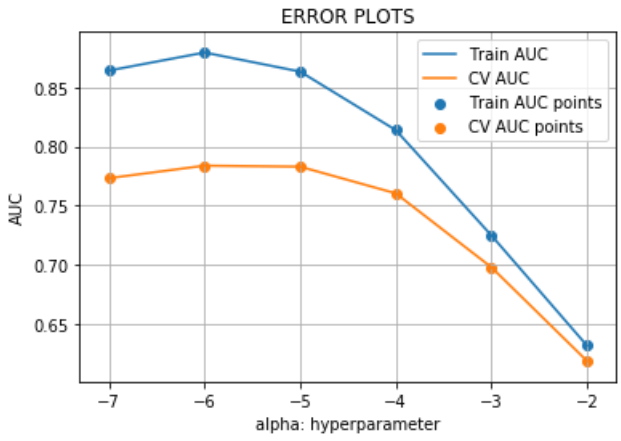
```
Final Data matrix
(124222, 8784) (124222,)
(36051, 8784) (36051,)
====================================================================================================
```

2.3.2.4 Training the data model and find best hyperparameter using ROC-AUC

In [50]:

```
# Call train_LR function on above data

train_LR(X_tr,y_train)
```

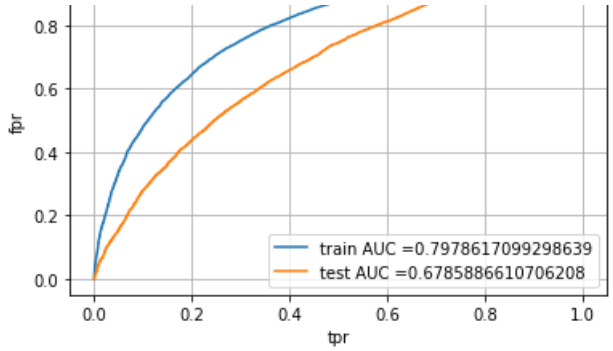
```
{'alpha': 1e-06}
dict_keys(['mean_fit_time', 'std_fit_time', 'mean_score_time', 'std_score_time', 'param_alpha', 'p
arams', 'split0_test_score', 'split1_test_score', 'mean_test_score', 'std_test_score',
'rank_test_score', 'split0_train_score', 'split1_train_score', 'mean_train_score',
'std_train_score'])
```



2.3.2.5 Testing the performance of the model on test data, plotting ROC Curves

In [51]:

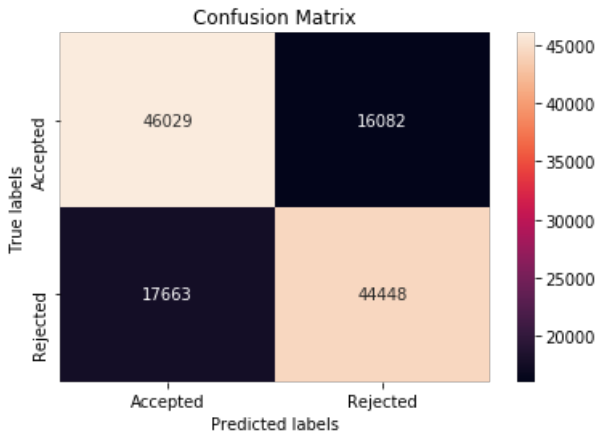
```
best_a=0.00001
train_fpr,train_tpr,tr_thresholds,y_train_pred,y_test_pred=test_LR(X_tr,X_te,best_a)
```



In [52]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")

ax= plt.subplot()
cm=confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t))
print(cm)
sns.heatmap(cm, annot=True, ax = ax,fmt='d'); #annot=True to annotate cells

# labels, title and ticks
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
ax.xaxis.set_ticklabels(['Accepted', 'Rejected']); ax.yaxis.set_ticklabels(['Accepted', 'Rejected'
]);
```

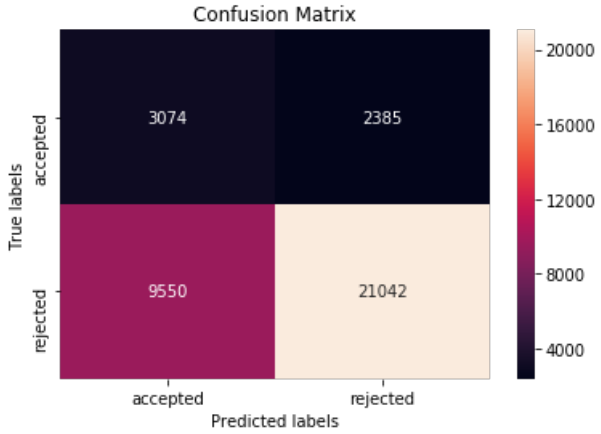
```
====================================================================================================

the maximum value of tpr*(1-fpr) 0.5303306417386245 for threshold 0.511
Train confusion matrix
[[46029 16082]
 [17663 44448]]
```



In [54]:

```
print("Test confusion matrix")


cm_test = confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t))
print(cm_test)
ax= plt.subplot()
sns.heatmap(cm_test, annot=True, ax = ax,fmt='d'); #annot=True to annotate cells

# labels, title and ticks
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
ax.xaxis.set_ticklabels(['accepted', 'rejected']); ax.yaxis.set_ticklabels(['accepted', 'rejected'
]);
```

```
Test confusion matrix
[[ 3074  2385]
 [ 9550 21042]]
```



2.3.3 Applying LR on AVG W2V

2.3.3.1 Encoding preprocessed_essays AVG W2V

In [55]:

```
with open('E:/Applied AI/Donors choose
dataset/DC_data/Assignments_DonorsChoose_2018/glove_vectors', 'rb') as f:
    model = pickle.load(f)
    glove_words =  set(model.keys())
```

In [56]:

```
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_essays_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_train['preprocessed_essays'].values): # 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_vectors_essays_train.append(vector)

print(len(avg_w2v_vectors_essays_train))
print(len(avg_w2v_vectors_essays_train[0]))
```

```
100%|██████████████████████████████████████████████████████████████████████████| 124222/124222
[00:34<00:00, 3560.25it/s]
```

```
124222
300
```

In [57]:

```
avg_w2v_vectors_essays_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_test['preprocessed_essays'].values): # 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_vectors_essays_test.append(vector)
print(len(avg_w2v_vectors_essays_test))
```

```
100%|██████████████████████████████████████████████████████████████████████████| 36051/36051
```

```
100%|████████████████████████████████████████████████████████████████| 36051/36051
[00:10<00:00, 3478.87it/s]
```

```
36051
```

2.3.3.2 Encoding preprocessed_titles AVG W2V

In [58]:

```
avg_w2v_vectors_titles_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_train['preprocessed_titles'].values): # 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_vectors_titles_train.append(vector)
print(len(avg_w2v_vectors_titles_train))
```

```
100%|████████████████████████████████████████████████████████████████| 124222/124222
[00:01<00:00, 70388.83it/s]
```

```
124222
```

In [59]:

```
avg_w2v_vectors_titles_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_test['preprocessed_titles'].values): # 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_vectors_titles_test.append(vector)
```

```
100%|████████████████████████████████████████████████████████████████| 36051/36051
[00:00<00:00, 62323.17it/s]
```

2.3.3.3 Merge all the features and obtain final data matrix

In [60]:

```
from scipy.sparse import hstack
X_tr = hstack((avg_w2v_vectors_titles_train,avg_w2v_vectors_essays_train, X_train_state_ohe,
X_train_teacher_ohe, X_train_grade_ohe, X_train_price_norm,
X_train_numerical_data_in_resource_summary_norm,
X_train_teacher_number_of_previously_posted_projects_norm, X_train_quantity_norm)).tocsr()
#X_cr = hstack((avg_w2v_vectors_titles_cv,avg_w2v_vectors_cv, X_cv_state_ohe, X_cv_teacher_ohe, X_
cv_grade_ohe, X_cv_price_norm, X_cv_numerical_data_in_resource_summary_norm,
X_cv_teacher_number_of_previously_posted_projects_norm, X_cv_quantity_norm)).tocsr()
X_te = hstack((avg_w2v_vectors_titles_test,avg_w2v_vectors_essays_test, X_test_state_ohe,
X_test_teacher_ohe, X_test_grade_ohe,
X_test_price_norm,X_test_numerical_data_in_resource_summary_norm,
X_test_teacher_number_of_previously_posted_projects_norm, X_test_quantity_norm)).tocsr()

print("Final Data matrix")
print(X_tr.shape, y_train.shape)
#print(X_cr.shape, y_cv.shape)
print(X_te.shape, y_test.shape)
print("="*100)
```

```
Final Data matrix
(124222, 664) (124222,)
```

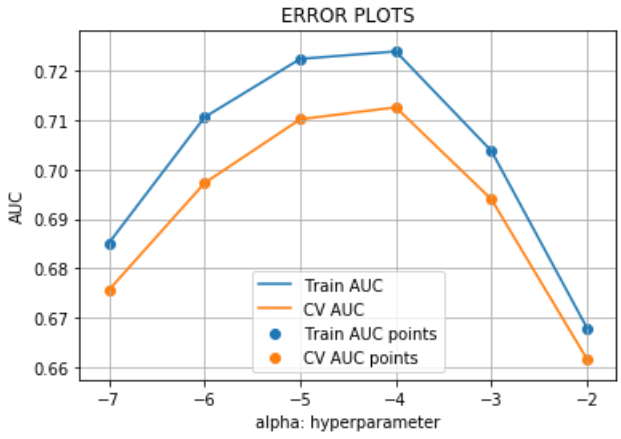
```
(36051, 664) (36051,)
==========================================================================================================
```

2.3.3.4 Training the data model and find best hyperparameter using ROC-AUC

In [61]:

```
# Call train_LR function on above data

train_LR(X_tr,y_train)
```

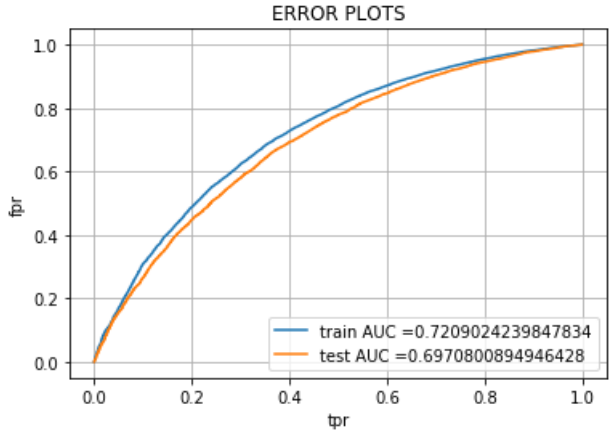
```
{'alpha': 0.0001}
dict_keys(['mean_fit_time', 'std_fit_time', 'mean_score_time', 'std_score_time', 'param_alpha', 'p
arams', 'split0_test_score', 'split1_test_score', 'mean_test_score', 'std_test_score',
'rank_test_score', 'split0_train_score', 'split1_train_score', 'mean_train_score',
'std_train_score'])
```



2.3.3.5 Testing the performance of the model on test data, plotting ROC Curves

In [62]:

```
best_a=0.0001
train_fpr,train_tpr,tr_thresholds,y_train_pred,y_test_pred=test_LR(X_tr,X_te,best_a)
```

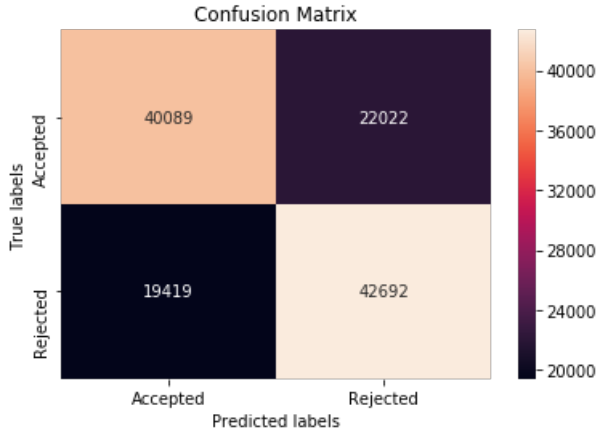


In [63]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")

ax= plt.subplot()
cm=confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t))
print(cm)
sns.heatmap(cm, annot=True, ax = ax,fmt='d'); #annot=True to annotate cells
```

```
# labels, title and ticks
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
ax.xaxis.set_ticklabels(['Accepted', 'Rejected']); ax.yaxis.set_ticklabels(['Accepted', 'Rejected'
]);
```

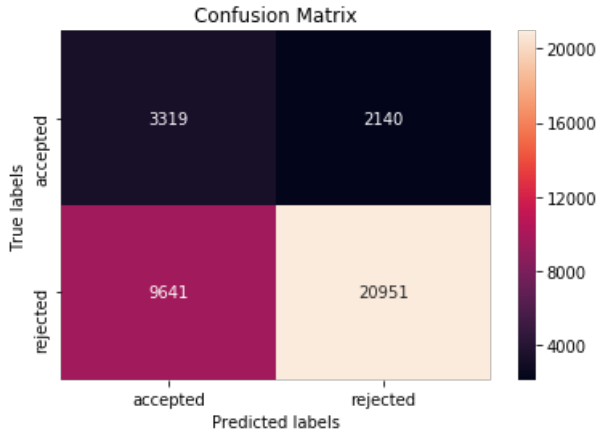
```
==========================================================================================================

the maximum value of tpr*(1-fpr) 0.4436440699486579 for threshold 0.495
Train confusion matrix
[[40089 22022]
 [19419 42692]]
```



In [64]:

```
print("Test confusion matrix")


cm_test = confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t))
print(cm_test)
ax= plt.subplot()
sns.heatmap(cm_test, annot=True, ax = ax,fmt='d'); #annot=True to annotate cells

# labels, title and ticks
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
ax.xaxis.set_ticklabels(['accepted', 'rejected']); ax.yaxis.set_ticklabels(['accepted', 'rejected'
]);
```

```
Test confusion matrix
[[ 3319  2140]
 [ 9641 20951]]
```



2.3.4 Applying LR on TFIDF W2V

2.3.4.1 Encoding preprocessed_titles tfidf W2V

In [65]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train['preprocessed_titles'])
# 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 = set(tfidf_model.get_feature_names())

# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_titles_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_train['preprocessed_titles']): # 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):
            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 tf
idf 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_vectors_titles_train.append(vector)

print(len(tfidf_w2v_vectors_titles_train))
print(len(tfidf_w2v_vectors_titles_train[0]))
```

```
100%|████████████████████████████████████████████████████████████████| 124222/124222
[00:05<00:00, 22170.24it/s]
```

```
124222
300
```

In [66]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_titles_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_test['preprocessed_titles']): # 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):
            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 tf
idf 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_vectors_titles_test.append(vector)

print(len(tfidf_w2v_vectors_titles_test))
print(len(tfidf_w2v_vectors_titles_test[0]))
```

```
100%|████████████████████████████████████████████████████████████████| 36051/36051
[00:01<00:00, 30182.09it/s]
```

```
36051
300
```

2.3.4.2 Encoding preprocessed_essays tfidf W2V

In [67]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
```

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
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 = set(tfidf_model.get_feature_names())

# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_essays_train = []; # 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):
            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 tf
idf 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_vectors_essays_train.append(vector)

print(len(tfidf_w2v_vectors_essays_train))
print(len(tfidf_w2v_vectors_essays_train[0]))
```

```
100%|████████████████████████████████████████████████████████████████| 124222/124222
[04:08<00:00, 500.85it/s]
```

```
124222
300
```

In [68]:

```
tfidf_w2v_vectors_essays_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
    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):
            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 tf
idf 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_vectors_essays_test.append(vector)

print(len(tfidf_w2v_vectors_essays_test))
print(len(tfidf_w2v_vectors_essays_test[0]))
```

```
100%|████████████████████████████████████████████████████████████████| 36051/36051 [01:
12<00:00, 498.68it/s]
```

```
36051
300
```

2.3.4.3 Merge all the features and obtain final data matrix

In [69]:

```
from scipy.sparse import hstack
X_tr = hstack((tfidf_w2v_vectors_titles_train,tfidf_w2v_vectors_essays_train, X_train_state_ohe, X
_train_teacher_ohe, X_train_grade_ohe, X_train_price_norm,
X_train_numerical_data_in_resource_summary_norm,
X_train_teacher_number_of_previously_posted_projects_norm, X_train_quantity_norm)).tocsr()
```

```
#X_cr = hstack((tfidf_w2v_vectors_titles_cv,tfidf_w2v_vectors_essays_cv, X_cv_state_ohe,
X_cv_teacher_ohe, X_cv_grade_ohe, X_cv_price_norm, X_cv_numerical_data_in_resource_summary_norm, X
_cv_teacher_number_of_previously_posted_projects_norm, X_cv_quantity_norm)).tocsr()
X_te = hstack((tfidf_w2v_vectors_titles_test,tfidf_w2v_vectors_essays_test, X_test_state_ohe,
X_test_teacher_ohe, X_test_grade_ohe,
X_test_price_norm,X_test_numerical_data_in_resource_summary_norm,
X_test_teacher_number_of_previously_posted_projects_norm, X_test_quantity_norm)).tocsr()

print("Final Data matrix")
print(X_tr.shape, y_train.shape)
#print(X_cr.shape, y_cv.shape)
print(X_te.shape, y_test.shape)
print("="*100)
```

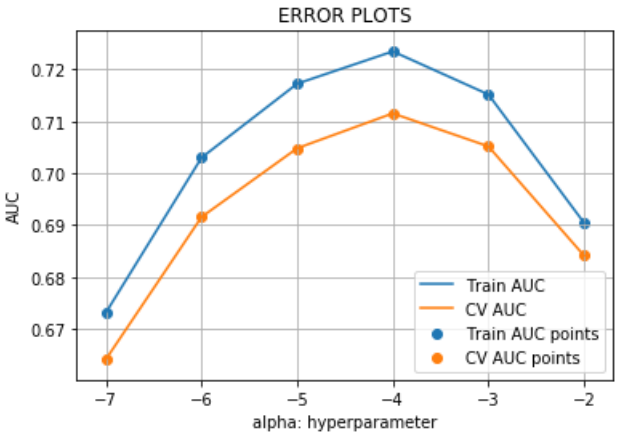
```
Final Data matrix
(124222, 664) (124222,)
(36051, 664) (36051,)
====================================================================================================
```

2.3.4.4 Training the data model and find best hyperparameter using ROC-AUC

In [70]:

```
# Call train_LR function on above data

train_LR(X_tr,y_train)
```

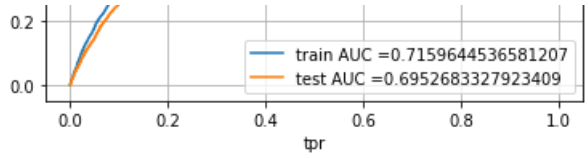
```
{'alpha': 0.0001}
dict_keys(['mean_fit_time', 'std_fit_time', 'mean_score_time', 'std_score_time', 'param_alpha', 'p
arams', 'split0_test_score', 'split1_test_score', 'mean_test_score', 'std_test_score',
'rank_test_score', 'split0_train_score', 'split1_train_score', 'mean_train_score',
'std_train_score'])
```



2.3.4.5 Testing the performance of the model on test data, plotting ROC Curves

In [93]:

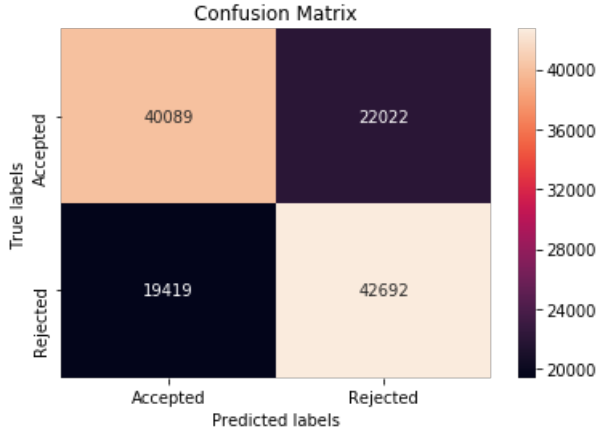
```
# Call test_LR for a obtained by training the data

best_a=0.0001
train_fpr,train_tpr,tr_thresholds,y_train_pred,y_test_pred=test_LR(X_tr,X_te,best_a)
```



In [71]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")

ax= plt.subplot()
cm=confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t))
print(cm)
sns.heatmap(cm, annot=True, ax = ax,fmt='d'); #annot=True to annotate cells

# labels, title and ticks
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
ax.xaxis.set_ticklabels(['Accepted', 'Rejected']); ax.yaxis.set_ticklabels(['Accepted', 'Rejected'
]);
```

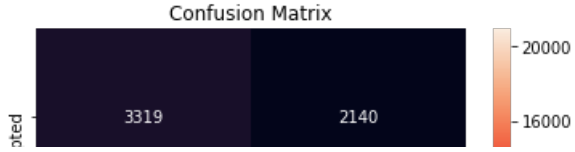
```
====================================================================================================

the maximum value of tpr*(1-fpr) 0.4436440699486579 for threshold 0.495
Train confusion matrix
[[40089 22022]
 [19419 42692]]
```

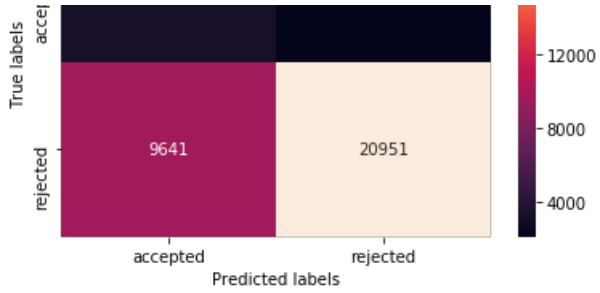


In [72]:

```
print("Test confusion matrix")


cm_test = confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t))
print(cm_test)
ax= plt.subplot()
sns.heatmap(cm_test, annot=True, ax = ax,fmt='d'); #annot=True to annotate cells

# labels, title and ticks
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
ax.xaxis.set_ticklabels(['accepted', 'rejected']); ax.yaxis.set_ticklabels(['accepted', 'rejected'
]);
```

```
Test confusion matrix
[[ 3319  2140]
 [ 9641 20951]]
```





2.3.5 Applying LR on only numerical data

In [73]:

```
import nltk
nltk.downloader.download('vader_lexicon')
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\parik\AppData\Roaming\nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

Out[73]:

```
True
```

2.3.5.1 Merging all the non_text Features

In [74]:

```
from scipy.sparse import hstack
X_tr = hstack((X_train_state_ohe, X_train_teacher_ohe, X_train_grade_ohe, X_train_price_norm,
X_train_teacher_number_of_previously_posted_projects_norm,
X_train_quantity_norm,X_train_number_of_words_in_title,X_train_number_of_words_in_essay,train_neg_e
ssay,train_neu_essay,train_pos_essay,train_comp_essay)).tocsr()
#X_cr = hstack((X_cv_titles_tfidf,X_cv_essay_tfidf, X_cv_state_ohe, X_cv_teacher_ohe,
X_cv_grade_ohe, X_cv_price_norm, X_cv_numerical_data_in_resource_summary_norm,
X_cv_teacher_number_of_previously_posted_projects_norm, X_cv_quantity_norm)).tocsr()
X_te = hstack((X_test_state_ohe, X_test_teacher_ohe, X_test_grade_ohe, X_test_price_norm,
X_test_teacher_number_of_previously_posted_projects_norm,
X_test_quantity_norm,X_test_number_of_words_in_title,X_test_number_of_words_in_essay,test_neg_essay
,test_neu_essay,test_pos_essay,test_comp_essay)).tocsr()

print("Final Data matrix")
print(X_tr.shape, y_train.shape)
#print(X_cr.shape, y_cv.shape)
print(X_te.shape, y_test.shape)
print("="*100)


'''
neg_essay= np.asarray(neg_essay)
neg_essay = neg_essay.reshape(-1,1)

neu_essay=np.asarray(neu_essay)
neu_essay = neu_essay.reshape(-1,1)

pos_essay=np.asarray(pos_essay)
pos_essay = pos_essay.reshape(-1,1)

comp_essay=np.asarray(comp_essay)
comp_essay = comp_essay.reshape(-1,1)

print(comp_essay)
'''
```

```
Final Data matrix
(124222, 69) (124222,)
(36051, 69) (36051,)
====================================================================================================
```

Out[74]:

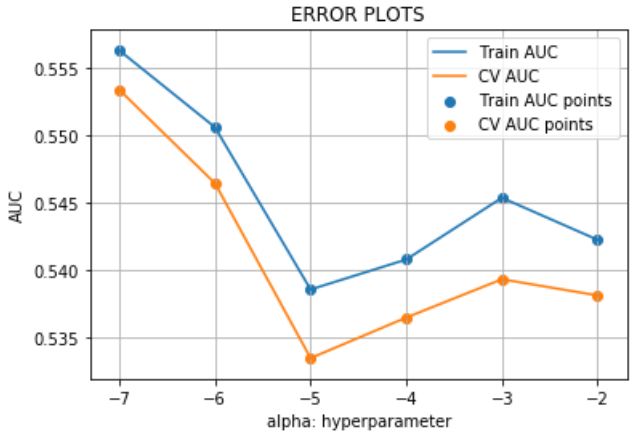
```
'\nneg_essay= np.asarray(neg_essay)\nneg_essay = neg_essay.reshape(-
1,1)\n\nneu_essay=np.asarray(neu_essay)\nneu_essay = neu_essay.reshape(-
1,1)\n\npos_essay=np.asarray(pos_essay)\npos_essay = pos_essay.reshape(-
1,1)\n\ncomp_essay=np.asarray(comp_essay)\ncomp_essay = comp_essay.reshape(-
1,1)\n\nprint(comp_essay)\n'
```

2.3.5.2 Training the data model and find best hyperparameter using ROC-AUC

In [75]:

```
# Call train_LR function on above data

train_LR(X_tr,y_train)
```

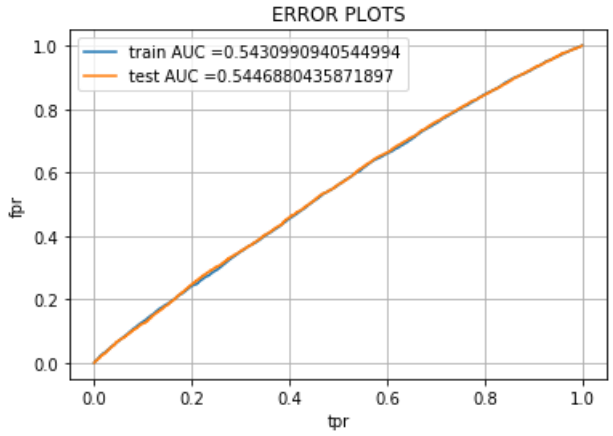
```
{'alpha': 1e-07}
dict_keys(['mean_fit_time', 'std_fit_time', 'mean_score_time', 'std_score_time', 'param_alpha', 'p
arams', 'split0_test_score', 'split1_test_score', 'mean_test_score', 'std_test_score',
'rank_test_score', 'split0_train_score', 'split1_train_score', 'mean_train_score',
'std_train_score'])
```



2.3.5.3 Testing the performance of the model on test data, plotting ROC Curves

In [76]:

```
# Call test_LR for a obtained by training the data

best_a=1e-07
train_fpr,train_tpr,tr_thresholds,y_train_pred,y_test_pred=test_LR(X_tr,X_te,best_a)
```



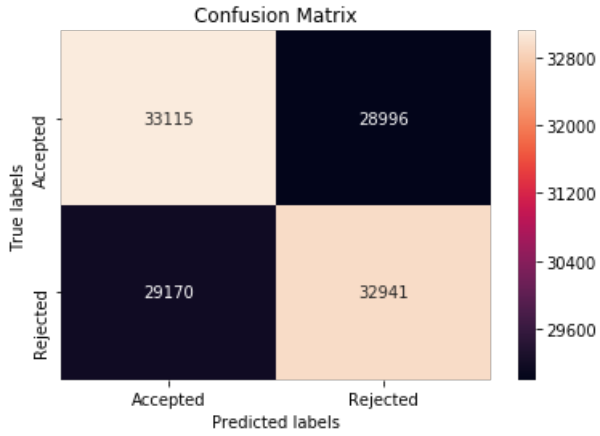
In [77]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
```

```
print("Train confusion matrix")

ax= plt.subplot()
cm=confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t))
print(cm)
sns.heatmap(cm, annot=True, ax = ax,fmt='d'); #annot=True to annotate cells

# labels, title and ticks
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
ax.xaxis.set_ticklabels(['Accepted', 'Rejected']); ax.yaxis.set_ticklabels(['Accepted', 'Rejected'
]);
```

==========

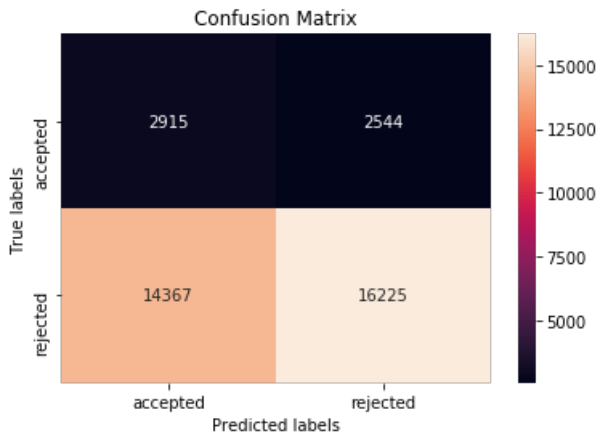
```
the maximum value of tpr*(1-fpr) 0.28276424661065774 for threshold 0.486
Train confusion matrix
[[33115 28996]
 [29170 32941]]
```



In [78]:

```
print("Test confusion matrix")


cm_test = confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t))
print(cm_test)
ax= plt.subplot()
sns.heatmap(cm_test, annot=True, ax = ax,fmt='d'); #annot=True to annotate cells

# labels, title and ticks
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
ax.xaxis.set_ticklabels(['accepted', 'rejected']); ax.yaxis.set_ticklabels(['accepted', 'rejected'
]);
```

```
Test confusion matrix
[[ 2915  2544]
 [14367 16225]]
```



2.4 Summary

In [79]:

```
# To summarize the results:
# summary table in jupyter notebook
# http://zetcode.com/python/prettytable/
# https://stackoverflow.com/questions/35160256/how-do-i-output-lists-as-a-table-in-jupyter-notebook

from prettytable import PrettyTable

x = PrettyTable(header_color='\033[40m')

x.field_names = ["Vectorizer", "Model", "Hyperparameter", "Train_AUC","Test_AUC"]

x.add_row(["Bag of Words", "LR",0.00001 ,0.86, 0.65])
x.add_row(["TF-IDF", "LR", 0.00001,0.79,0.67])
x.add_row(["Avg_W2V", "LR", 0.0001,0.72,0.69])
x.add_row(["TF-IDF W2V", "LR" , 0.0001,0.71,0.69])
x.add_row(["Numerical Features", "LR", 0.0000001,0.54,0.54])


print(x)
```

```
+--------------------+-------+----------------+-----------+----------+
|     Vectorizer     | Model | Hyperparameter | Train_AUC | Test_AUC |
+--------------------+-------+----------------+-----------+----------+
|    Bag of Words    |   LR  |     1e-05      |    0.86   |   0.65   |
|       TF-IDF       |   LR  |     1e-05      |    0.79   |   0.67   |
|      Avg_W2V       |   LR  |     0.0001     |    0.72   |   0.69   |
|     TF-IDF W2V     |   LR  |     0.0001     |    0.71   |   0.69   |
| Numerical Features |   LR  |     1e-07      |    0.54   |   0.54   |
+--------------------+-------+----------------+-----------+----------+
```