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 [3]:

```
%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.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 chart_studio.plotly import plotly
import plotly.offline as offline
import plotly.graph_objs as go
offline.init_notebook_mode()
from collections import Counter
```

1.1 Reading Data

In [41]:

```
project_data = pd.read_csv('train_data.csv')
resource_data = pd.read_csv('resources.csv')
```

In [42]:

```
print("Number of data points in train data", project_data.shape)
print('-'*50)
print("The attributes of data :", project_data.columns.values)
```

```
Number of data points in train data (109248, 17)
--------------------------------------------------
The attributes of 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']
```

In [43]:

```
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]


project_data.head(2)
```

Out[43]:

	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

In [44]:

```
print("Number of data points in train data", resource_data.shape)
print(resource_data.columns.values)
resource_data.head(2)
```

```
Number of data points in train data (1541272, 4)
['id' 'description' 'quantity' 'price']
```

Out[44]:

	id	description	quantity	price
0	p233245	LC652 - Lakeshore Double-Space Mobile Drying Rack	1	149.00
1	p069063	Bouncy Bands for Desks (Blue support pipes)	3	14.95

1.2 preprocessing of `project_subject_categories`

In [45]:

```
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)

from collections import Counter
my_counter = Counter()
for word in project_data['clean_categories'].values:
    my_counter.update(word.split())

cat_dict = dict(my_counter)
sorted_cat_dict = dict(sorted(cat_dict.items(), key=lambda kv: kv[1]))
```

1.3 preprocessing of `project_subject_subcategories`

In [46]:

```
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)

# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
my_counter = Counter()
for word in project_data['clean_subcategories'].values:
    my_counter.update(word.split())

sub_cat_dict = dict(my_counter)
sorted_sub_cat_dict = dict(sorted(sub_cat_dict.items(), key=lambda kv: kv[1]))
```

1.3 Text preprocessing

In [47]:

```
# merge two column text dataframe:
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)
```

In [48]:

```
project_data.head(2)
```

Out[48]:

	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

In [49]:

```
#### 1.4.2.3 Using Pretrained Models: TFIDF weighted W2V
```

In [50]:

```
# printing some random reviews
print(project_data['essay'].values[0])
print("="*50)
print(project_data['essay'].values[150])
print("="*50)
print(project_data['essay'].values[1000])
print("="*50)
print(project_data['essay'].values[20000])
print("="*50)
print(project_data['essay'].values[99999])
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.
==========
\"Life moves pretty fast. If you don't stop and look around once in awhile, you could miss it.\"
from the movie, Ferris Bueller's Day Off. Think back...what do you remember about your
grandparents? How amazing would it be to be able to flip through a book to see a day in their
lives?My second graders are voracious readers! They love to read both fiction and nonfiction books
. Their favorite characters include Pete the Cat, Fly Guy, Piggie and Elephant, and Mercy Watson.
They also love to read about insects, space and plants. My students are hungry bookworms! My stude
nts are eager to learn and read about the world around them. My kids love to be at school and are
like little sponges absorbing everything around them. Their parents work long hours and usually do
not see their children. My students are usually cared for by their grandparents or a family
friend. Most of my students do not have someone who speaks English at home. Thus it is difficult f
or my students to acquire language.Now think forward... wouldn't it mean a lot to your kids,
nieces or nephews or grandchildren, to be able to see a day in your life today 30 years from now?
Memories are so precious to us and being able to share these memories with future generations will
be a rewarding experience. As part of our social studies curriculum, students will be learning ab
out changes over time. Students will be studying photos to learn about how their community has ch
anged over time. In particular, we will look at photos to study how the land, buildings,
clothing, and schools have changed over time. As a culminating activity, my students will capture
a slice of their history and preserve it through scrap booking. Key important events in their
young lives will be documented with the date, location, and names. Students will be using photos
from home and from school to create their second grade memories. Their scrap books will preserve
their unique stories for future generations to enjoy.Your donation to this project will provide my
second graders with an opportunity to learn about social studies in a fun and creative manner. Th
rough their scrapbooks, children will share their story with others and have a historical document
for the rest of their lives.
==========
\"A person's a person, no matter how small.\" (Dr.Seuss) I teach the smallest students with the bi
ggest enthusiasm for learning. My students learn in many different ways using all of our senses an
d multiple intelligences. I use a wide range of techniques to help all my students succeed. \r\nSt
udents in my class come from a variety of different backgrounds which makes for wonderful sharing
of experiences and cultures, including Native Americans.\r\nOur school is a caring community of su
ccessful learners which can be seen through collaborative student project based learning in and ou
t of the classroom. Kindergarteners in my class love to work with hands-on materials and have many
different opportunities to practice a skill before it is mastered. Having the social skills to wor
k cooperatively with friends is a crucial aspect of the kindergarten curriculum.Montana is the
perfect place to learn about agriculture and nutrition. My students love to role play in our
pretend kitchen in the early childhood classroom. I have had several kids ask me, \"Can we try coo
king with REAL food?\" I will take their idea and create \"Common Core Cooking Lessons\" where we
learn important math and writing concepts while cooking delicious healthy food for snack time. My
students will have a grounded appreciation for the work that went into making the food and knowled
ge of where the ingredients came from as well as how it's healthy for their bodies. This project w
ould expand our learning of nutrition and agricultural cooking recipes by having us peel our own a
pples to make homemade applesauce, make our own bread, and mix up healthy plants from our classroo
m garden in the spring. We will also create our own cookbooks to be printed and shared with famili
es. \r\nStudents will gain math and literature skills as well as a life long enjoyment for healthy
cooking.nannan
==========
My classroom consists of twenty-two amazing sixth graders from different cultures and backgrounds.
They are a social bunch who enjoy working in partners and working with groups. They are hard-worki
ng and eager to head to middle school next year. My job is to get them ready to make this
transition and make it as smooth as possible. In order to do this, my students need to come to
school every day and feel safe and ready to learn. Because they are getting ready to head to
middle school, I give them lots of choice- choice on where to sit and work, the order to complete
assignments, choice of projects, etc. Part of the students feeling safe is the ability for them to
come into a welcoming, encouraging environment. My room is colorful and the atmosphere is casual.
I want them to take ownership of the classroom because we ALL share it together. Because my time w
ith them is limited, I want to ensure they get the most of this time and enjoy it to the best of t
heir abilities.Currently, we have twenty-two desks of differing sizes, yet the desks are similar t
o the ones the students will use in middle school. We also have a kidney table with crates for sea
ting. I allow my students to choose their own spots while they are working independently or in
groups. More often than not, most of them move out of their desks and onto the crates. Believe it
or not, this has proven to be more successful than making them stay at their desks! It is because
of this that I am looking toward the “Flexible Seating” option for my classroom.\r\n The students
look forward to their work time so they can move around the room. I would like to get rid of the c
onstricting desks and move toward more “fun” seating options. I am requesting various seating so m
y students have more options to sit. Currently, I have a stool and a papasan chair I inherited fro
m the previous sixth-grade teacher as well as five milk crate seats I made, but I would like to gi
ve them more options and reduce the competition for the “good seats”. I am also requesting two rug
s as not only more seating options but to make the classroom more welcoming and appealing. In orde
r for my students to be able to write and complete work without desks, I am requesting a class set
of clipboards. Finally, due to curriculum that requires groups to work together, I am requesting t
ables that we can fold up when we are not using them to leave more room for our flexible seating o
ptions.\r\nI know that with more seating options, they will be that much more excited about coming

to school! Thank you for your support in making my classroom one students will remember forever!nannan
==========

In [51]:

```
# 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 [52]:

```
sent = decontracted(project_data['essay'].values[20000])
print(sent)
print("="*50)
```

\"A person is a person, no matter how small.\" (Dr.Seuss) I teach the smallest students with the b
iggest enthusiasm for learning. My students learn in many different ways using all of our senses a
nd multiple intelligences. I use a wide range of techniques to help all my students succeed. \r\nS
tudents in my class come from a variety of different backgrounds which makes for wonderful sharing
of experiences and cultures, including Native Americans.\r\nOur school is a caring community of su
ccessful learners which can be seen through collaborative student project based learning in and ou
t of the classroom. Kindergarteners in my class love to work with hands-on materials and have many
different opportunities to practice a skill before it is mastered. Having the social skills to wor
k cooperatively with friends is a crucial aspect of the kindergarten curriculum.Montana is the
perfect place to learn about agriculture and nutrition. My students love to role play in our
pretend kitchen in the early childhood classroom. I have had several kids ask me, \"Can we try coo
king with REAL food?\" I will take their idea and create \"Common Core Cooking Lessons\" where we
learn important math and writing concepts while cooking delicious healthy food for snack time. My
students will have a grounded appreciation for the work that went into making the food and knowled
ge of where the ingredients came from as well as how it is healthy for their bodies. This project
would expand our learning of nutrition and agricultural cooking recipes by having us peel our own
apples to make homemade applesauce, make our own bread, and mix up healthy plants from our classro
om garden in the spring. We will also create our own cookbooks to be printed and shared with famil
ies. \r\nStudents will gain math and literature skills as well as a life long enjoyment for health
y cooking.nannan
==========

In [53]:

```
# \r \n \t remove from string python: http://texthandler.com/info/remove-line-breaks-python/
sent = sent.replace('\\r', ' ')
sent = sent.replace('\\"', ' ')
sent = sent.replace('\\n', ' ')
print(sent)
```

 A person is a person, no matter how small. (Dr.Seuss) I teach the smallest students with the big
gest enthusiasm for learning. My students learn in many different ways using all of our senses and
multiple intelligences. I use a wide range of techniques to help all my students succeed.
Students in my class come from a variety of different backgrounds which makes for wonderful
sharing of experiences and cultures, including Native Americans. Our school is a caring community
of successful learners which can be seen through collaborative student project based learning in a
nd out of the classroom. Kindergarteners in my class love to work with hands-on materials and have
many different opportunities to practice a skill before it is mastered. Having the social skills t
o work cooperatively with friends is a crucial aspect of the kindergarten curriculum.Montana is
the perfect place to learn about agriculture and nutrition. My students love to role play in our p
retend kitchen in the early childhood classroom. I have had several kids ask me, Can we try cooki
ng with REAL food? I will take their idea and create Common Core Cooking Lessons where we learn

ng with REAL food. I will take their idea and create "Common Core Cooking Lessons" where we learn important math and writing concepts while cooking delicious healthy food for snack time. My students will have a grounded appreciation for the work that went into making the food and knowled ge of where the ingredients came from as well as how it is healthy for their bodies. This project would expand our learning of nutrition and agricultural cooking recipes by having us peel our own apples to make homemade applesauce, make our own bread, and mix up healthy plants from our classro om garden in the spring. We will also create our own cookbooks to be printed and shared with famil ies. Students will gain math and literature skills as well as a life long enjoyment for healthy cooking.nannan

In [54]:

```
#remove spacial character: https://stackoverflow.com/a/5843547/4084039
sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
print(sent)
```

A person is a person no matter how small Dr Seuss I teach the smallest students with the biggest enthusiasm for learning My students learn in many different ways using all of our senses and multi ple intelligences I use a wide range of techniques to help all my students succeed Students in my class come from a variety of different backgrounds which makes for wonderful sharing of experiences and cultures including Native Americans Our school is a caring community of successful learners which can be seen through collaborative student project based learning in and out of the classroom Kindergarteners in my class love to work with hands on materials and have many different opportunities to practice a skill before it is mastered Having the social skills to work cooperatively with friends is a crucial aspect of the kindergarten curriculum Montana is the perfect place to learn about agriculture and nutrition My students love to role play in our pretend kitchen in the early childhood classroom I have had several kids ask me Can we try cooking with REAL food I will take their idea and create Common Core Cooking Lessons where we learn important math and writing concepts while cooking delicious healthy food for snack time My students will have a grounded appreciation for the work that went into making the food and knowled ge of where the ingredients came from as well as how it is healthy for their bodies This project w ould expand our learning of nutrition and agricultural cooking recipes by having us peel our own a pples to make homemade applesauce make our own bread and mix up healthy plants from our classroom garden in the spring We will also create our own cookbooks to be printed and shared with families Students will gain math and literature skills as well as a life long enjoyment for healthy cooking nannan

In [55]:

```
# 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 [19]:

```
# Combining all the above stundents
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())
```

```
100%|██████████████████████████████████████████████████████████████████| 109248/109248
[01:06<00:00, 1654.56it/s]
```

In [56]:

```
# after preprocesing
project_data['preprocessed_essays'] = preprocessed_essays

preprocessed_essays[20000]
```

Out[56]:

'a person person no matter small dr seuss i teach smallest students biggest enthusiasm learning my students learn many different ways using senses multiple intelligences i use wide range techniques help students succeed students class come variety different backgrounds makes wonderful sharing ex periences cultures including native americans our school caring community successful learners seen collaborative student project based learning classroom kindergarteners class love work hands materials many different opportunities practice skill mastered having social skills work cooperatively friends crucial aspect kindergarten curriculum montana perfect place learn agriculture nutrition my students love role play pretend kitchen early childhood classroom i several kids ask can try cooking real food i take idea create common core cooking lessons learn im portant math writing concepts cooking delicious healthy food snack time my students grounded appre ciation work went making food knowledge ingredients came well healthy bodies this project would ex pand learning nutrition agricultural cooking recipes us peel apples make homemade applesauce make bread mix healthy plants classroom garden spring we also create cookbooks printed shared families students gain math literature skills well life long enjoyment healthy cooking nannan'

1.4 Preprocessing of `project_title`

In [57]:

```
preprocessed_titles = []
for sentence in project_data['project_title'].values:
    snt= decontracted(sentence)
    snt= snt.replace('\\r', ' ')
    snt= snt.replace('\\"', ' ')
    snt= snt.replace('\\n', ' ')
    snt= re.sub('[^A-Za-z0-9]+', ' ', snt)

    # https://gist.github.com/sebleier/554280
    snt = ' '.join(e for e in snt.split() if e not in stopwords)
    preprocessed_titles.append(snt.lower().strip())

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

Out[57]:

'empowering students through art learning about then now'

1.5 Preparing data for models

In [58]:

```
project_data.columns
```

Out[58]:

```
Index(['Unnamed: 0', 'id', 'teacher_id', 'teacher_prefix', 'school_state',
       'Date', 'project_grade_category', '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',
       'clean_categories', 'clean_subcategories', 'essay',
       'preprocessed_essays', 'preprocessed_titles'],
      dtype='object')
```

we are going to consider

```
       - school_state : categorical data
       - clean_categories : categorical data
       - clean_subcategories : categorical data
       - project_grade_category : categorical data
       - teacher_prefix : categorical data

       - project_title : text data
       - text : text data
       - project_resource_summary: text data (optinal)

       - quantity : numerical (optinal)
       - teacher_number_of_previously_posted_projects : numerical
       - price : numerical
```

In [59]:

```
# 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())
```

Assignment 9: RF and GBDT

Response Coding: Example

> The response tabel is built only on train dataset. For a category which is not there in train data and present in test data, we will encode them with default values Ex: in our test data if have State: D then we encode it as [0.5, 0.5]

1. **Apply both Random Forrest and GBDT on these feature sets**
   - Set 1: categorical(instead of one hot encoding, try response coding: use probability values), numerical features + project_title(BOW) + preprocessed_eassay (BOW)
   - Set 2: categorical(instead of one hot encoding, try response coding: use probability values), numerical features + project_title(TFIDF)+ preprocessed_eassay (TFIDF)
   - Set 3: categorical(instead of one hot encoding, try response coding: use probability values), numerical features + project_title(AVG W2V)+ preprocessed_eassay (AVG W2V). Here for this set take **20K** datapoints only.
   - Set 4: categorical(instead of one hot encoding, try response coding: use probability values), numerical features + project_title(TFIDF W2V)+ preprocessed_eassay (TFIDF W2V). Here for this set take **20K** datapoints only.
2. **The hyper paramter tuning (Consider any two hyper parameters preferably n_estimators, max_depth)**
   - Consider the following range for hyperparameters **n_estimators** = [10, 50, 100, 150, 200, 300, 500, 1000], **max_depth** = [2, 3, 4, 5, 6, 7, 8, 9, 10]
   - Find the best hyper parameter which will give the maximum AUC value
   - Find the best hyper paramter using simple cross validation data
   - You can write your own for loops to do this task
3. **Representation of results**

- You need to plot the performance of model both on train data and cross validation data for each hyper parameter, like shown in the figure

  with X-axis as **n_estimators**, Y-axis as **max_depth**, and Z-axis as **AUC Score** , we have given the notebook which explains how to plot this 3d plot, you can find it in the same drive *3d_scatter_plot.ipynb*

or

- You need to plot the performance of model both on train data and cross validation data for each hyper parameter, like shown in the figure

  seaborn heat maps with rows as **n_estimators**, columns as **max_depth**, and values inside the cell representing **AUC Score**
- You can choose either of the plotting techniques: 3d plot or heat map
- Once after you found the best hyper parameter, you need to train your model with it, and find the AUC on test data and plot the ROC curve on both train and test.
- Along with plotting ROC curve, you need to print the confusion matrix with predicted and original labels of test data points

4. **Conclusion**

   - You need to summarize the results at the end of the notebook, summarize it in the table format. To print out a table please refer to this prettytable library link

Note: Data Leakage

1. There will be an issue of data-leakage if you vectorize the entire data and then split it into train/cv/test.
2. To avoid the issue of data-leakag, make sure to split your data first and then vectorize it.
3. While vectorizing your data, apply the method fit_transform() on you train data, and apply the method transform() on cv/test data.
4. For more details please go through this link.

2. Random Forest and GBDT

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

In [0]:

```
# please write all the code with proper documentation, and proper titles for each subsection
# go through documentations and blogs before you start coding
# first figure out what to do, and then think about how to do.
# reading and understanding error messages will be very much helpfull in debugging your code
# when you plot any graph make sure you use
    # a. Title, that describes your plot, this will be very helpful to the reader
    # b. Legends if needed
    # c. X-axis label
    # d. Y-axis label
```

In [60]:

```
project_data['teacher_prefix'].fillna(" ", inplace = True)
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')
```

In [61]:

```
from scipy import sparse
from scipy.sparse import csr_matrix
from scipy.sparse import lil_matrix

Y = project_data['project_is_approved'].values
X = project_data

Y = Y[0:60000]
```

```
X = X.head(60000)
```

In [62]:

```
X.head()
```

Out[62]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_grade_category
0	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	Mrs.	CA	2016-04-27 00:27:36	Grades PreK-2
1	37728	p043609	3f60494c61921b3b43ab61bdde2904df	Ms.	UT	2016-04-27 00:31:25	Grades 3-5
2	74477	p189804	4a97f3a390bfe21b99cf5e2b81981c73	Mrs.	CA	2016-04-27 00:46:53	Grades PreK-2
3	100660	p234804	cbc0e38f522143b86d372f8b43d4cff3	Mrs.	GA	2016-04-27 00:53:00	Grades PreK-2
4	33679	p137682	06f6e62e17de34fcf81020c77549e1d5	Mrs.	WA	2016-04-27 01:05:25	Grades 3-5

5 rows × 22 columns

In [63]:

```
teacher_prefix=[]
for s1 in X['teacher_prefix']:
    s1= s1.replace('.', '')
    teacher_prefix.append(s1.lower().strip())

X['teacher_prefix'] = teacher_prefix
X['teacher_prefix'].fillna(" ", inplace = True)
from collections import Counter
my_counter = Counter()
for word in X['teacher_prefix'].values:
    my_counter.update(word.split())
teacher_prefix = dict(my_counter)
teacher_prefix = dict(sorted(teacher_prefix.items(), key=lambda kv: kv[1]))
```

In [64]:

```
project_grade=[]
for s1 in X['project_grade_category']:
    s1= s1.replace('Grades', '')
```

```
    s1= s1.replace('-', '_')
    project_grade.append(s1.lower().strip())

X['project_grade_category'] = project_grade

from collections import Counter
my_counter = Counter()
for word in X['project_grade_category'].values:
    my_counter.update(word.split())
project_grade_category = dict(my_counter)
project_grade_category = dict(sorted(project_grade_category.items(), key=lambda kv: kv[1]))
```

In [65]:

```
l =[]
for i in X['essay']:
    words = len(i.split())
    l.append(words)
X['essay_len'] = l
```

In [313]:

```
l =[]
for i in X['project_title']:
    words = len(i.split())
    l.append(words)
X['title_len'] = l
```

In [66]:

```
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer

import nltk
nltk.download('vader_lexicon')

sid = SentimentIntensityAnalyzer()

for_sentiment = X['preprocessed_essays']
ss =[]
for i in range(0,len(for_sentiment)):
    j= sid.polarity_scores(for_sentiment[i])
    ss.append(j)

# 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

neg = []
neu=[]
pos=[]
compound=[]
for k in range(0,len(ss)):
    neg.append(ss[k]['neg'])
    neu.append(ss[k]['neu'])
    pos.append(ss[k]['pos'])
    compound.append(ss[k]['compound'])

X['neg']= neg
X['neu']= neu
X['pos']= pos
X['compound']=compound

X.head()
```

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

Out[66]:

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_grade_category

	Unnamed: 0	id	teacher_id	teacher_prefix	school_state	Date	project_grade_category
0	8393	p205479	2bf07ba08945e5d8b2a3f269b2b3cfe5	mrs	CA	2016-04-27 00:27:36	prek_2
1	37728	p043609	3f60494c61921b3b43ab61bdde2904df	ms	UT	2016-04-27 00:31:25	3_5
2	74477	p189804	4a97f3a390bfe21b99cf5e2b81981c73	mrs	CA	2016-04-27 00:46:53	prek_2
3	100660	p234804	cbc0e38f522143b86d372f8b43d4cff3	mrs	GA	2016-04-27 00:53:00	prek_2
4	33679	p137682	06f6e62e17de34fcf81020c77549e1d5	mrs	WA	2016-04-27 01:05:25	3_5

5 rows × 27 columns

In [67]:

```
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, shuffle = False)


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

```
(40200, 27) (40200,)
(19800, 27) (19800,)
```

2.2 Make Data Model Ready: encoding numerical, categorical features

In [0]:

```
# please write all the code with proper documentation, and proper titles for each subsection
# go through documentations and blogs before you start coding
# first figure out what to do, and then think about how to do.
# reading and understanding error messages will be very much helpfull in debugging your code
# make sure you featurize train and test data separatly

# when you plot any graph make sure you use
    # a. Title, that describes your plot, this will be very helpful to the reader
    # b. Legends if needed
    # c. X-axis label
    # d. Y-axis label
```

In [94]:

```
def get_fea_dict(alpha, feature, df):

    value_count = X_train[feature].value_counts()
    feature_dict = dict()

    for i, denominator in value_count.items():
        vec = []

        for k in range(0,2):

            cls_cnt = X_train.loc[(X_train['project_is_approved']==k) & (X_train[feature]==i)]

            vec.append((cls_cnt.shape[0] + alpha)/ (denominator + 2*alpha))
        feature_dict[i]=vec
    return feature_dict

def get_feature_response(alpha, feature, df):

    feature_dict = get_fea_dict(alpha, feature, df)
    value_count = X_train[feature].value_counts()

    feature_fea = []
    for index, row in df.iterrows():
        if row[feature] in dict(value_count).keys():
            feature_fea.append(feature_dict[row[feature]])
        else:
            feature_fea.append([1/2,1/2])
    return feature_fea
```

In [106]:

```
alpha = 1
X_train_cat = np.array(get_feature_response(alpha, "clean_categories", X_train))
X_test_cat = np.array(get_feature_response(alpha, "clean_categories", X_test))

print(X_train_cat.shape)
print(X_test_cat.shape)
```

```
(40200, 2)
(19800, 2)
```

In [107]:

```
print(X_train_cat)
```

```
[[0.18646187 0.81353813]
 [0.18630573 0.81369427]
 [0.1382066  0.8617934 ]
 ...
 [0.18646187 0.81353813]
 [0.14397556 0.85602444]
 [0.18823529 0.81176471]]
```

In [108]:

```
alpha = 1
X_train_subcat = np.array(get_feature_response(alpha, "clean_subcategories", X_train))
X_test_subcat = np.array(get_feature_response(alpha, "clean_subcategories", X_test))

print(X_train_subcat.shape)
print(X_test_subcat.shape)
```

```
(40200, 2)
(19800, 2)
```

In [109]:

```
alpha = 1
X_train_school = np.array(get_feature_response(alpha, "school_state", X_train))
X_test_school = np.array(get_feature_response(alpha, "school_state", X_test))
```

```
print(X_train_school.shape)
print(X_test_school.shape)
```

```
(40200, 2)
(19800, 2)
```

In [110]:

```
alpha = 1
X_train_teacher = np.array(get_feature_response(alpha, "teacher_prefix", X_train))
X_test_teacher = np.array(get_feature_response(alpha, "teacher_prefix", X_test))

print(X_train_teacher.shape)
print(X_test_teacher.shape)
```

```
(40200, 2)
(19800, 2)
```

In [111]:

```
alpha = 1
X_train_grade = np.array(get_feature_response(alpha, "project_grade_category", X_train))
X_test_grade = np.array(get_feature_response(alpha, "project_grade_category", X_test))

print(X_train_grade.shape)
print(X_test_grade.shape)
```

```
(40200, 2)
(19800, 2)
```

In [115]:

```
from sklearn.preprocessing import Normalizer
normalizer = Normalizer()
normalizer.fit(X_train['price'].values.reshape(1,-1))
X_train_price = normalizer.transform(X_train['price'].values.reshape(1,-1))
X_test_price = normalizer.transform(X_test['price'].values.reshape(1,-1))

print("After vectorizations:")
print(X_train_price.shape, y_train.shape)
print(X_test_price.shape, y_test.shape)
print("="*50)
```

```
After vectorizations:
(1, 40200) (40200,)
(1, 19800) (19800,)
==================================================
```

In [116]:

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

print("After vectorizations:")
print(X_train_quantity.shape, y_train.shape)
print(X_test_quantity.shape, y_test.shape)
print("="*50)
```

```
After vectorizations:
(1, 40200) (40200,)
(1, 19800) (19800,)
==================================================
```

In [117]:

```
from sklearn.preprocessing import Normalizer
normalizer = Normalizer()
normalizer.fit(X_train['teacher_number_of_previously_posted_projects'].values.reshape(1,-1))
X_train_previous = normalizer.transform(X_train['teacher_number_of_previously_posted_projects'].va
lues.reshape(1,-1))
X_test_previous =
normalizer.transform(X_test['teacher_number_of_previously_posted_projects'].values.reshape(1,-1))

print("After vectorizations:")
print(X_train_previous.shape, y_train.shape)
print(X_test_previous.shape, y_test.shape)
print("="*50)
```

```
After vectorizations:
(1, 40200) (40200,)
(1, 19800) (19800,)
==================================================
```

In [118]:

```
X_train_num_bow = np.hstack((X_train_price.reshape(-1,1),X_train_quantity.reshape(-1,1),X_train_pre
vious.reshape(-1,1)))
X_test_num_bow = np.hstack((X_test_price.reshape(-1,1),X_test_quantity.reshape(-1,1),X_test_previou
s.reshape(-1,1)))
```

In [126]:

```
from scipy.sparse import hstack

X_train_features = np.hstack((X_train_cat, X_train_subcat,X_train_school,X_train_teacher,X_train_gr
ade,X_train_num_bow))
X_test_features = np.hstack((X_test_cat,
X_test_subcat,X_test_school,X_test_teacher,X_test_grade,X_test_num_bow))

#X_features= X_features.tocsr()[0:10000,]


print("Final Data matrix")
print(X_train_features.shape, y_train.shape)
print(X_test_features.shape, y_test.shape)
print("="*100)
```

```
Final Data matrix
(40200, 13) (40200,)
(19800, 13) (19800,)
====================================================================================================
```

In [127]:

```
X_train_features_set = np.hstack((X_train_features,X_train['neg'].values.reshape(-1,1),X_train['neu
'].values.reshape(-1,1),X_train['pos'].values.reshape(-1,1),X_train['compound'].values.reshape(-1,1
)))
X_test_features_set = np.hstack((X_test_features,X_test['neg'].values.reshape(-1,1),X_test['neu'].v
alues.reshape(-1,1),X_test['pos'].values.reshape(-1,1),X_test['compound'].values.reshape(-1,1)))

print("Final Data matrix")
print(X_train_features_set.shape, y_train.shape)
print(X_test_features_set.shape, y_test.shape)
print("="*100)
```

```
Final Data matrix
(40200, 17) (40200,)
(19800, 17) (19800,)
====================================================================================================
```

2.3 Make Data Model Ready: encoding eassay, and project_title

BOW

In [175]:

```
X_train_text = X_train['preprocessed_essays']
X_test_text = X_test['preprocessed_essays']
```

In [176]:

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(min_df =10, max_features = 3000)
vectorizer.fit(X_train_text) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_bow = vectorizer.fit_transform(X_train_text)
X_test_bow = vectorizer.transform(X_test_text)

print("After vectorizations")
print(X_train_bow.shape, y_train.shape)
print(X_test_bow.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(40200, 3000) (40200,)
(19800, 3000) (19800,)
====================================================================================================
```

In [177]:

```
X_train_text = X_train['preprocessed_titles']
X_test_text = X_test['preprocessed_titles']
```

In [178]:

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(min_df =10)
vectorizer.fit(X_train_text) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_bow2 = vectorizer.fit_transform(X_train_text)
X_test_bow2 = vectorizer.transform(X_test_text)

print("After vectorizations")
print(X_train_bow2.shape, y_train.shape)
print(X_test_bow2.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(40200, 1777) (40200,)
(19800, 1777) (19800,)
====================================================================================================
```

In [179]:

```
X_train_features_bow = hstack((X_train_features, X_train_bow, X_train_bow2))
X_test_features_bow = hstack((X_test_features, X_test_bow, X_test_bow2))

print(X_train_features_bow.shape, y_train.shape)
print(X_test_features_bow.shape, y_test.shape)
print("="*100)
```

```
(40200, 4790) (40200,)
(19800, 4790) (19800,)
====================================================================================================
```

TFIDF

In [238]:

```
X_train_text = X_train['preprocessed_essays']
X_test_text = X_test['preprocessed_essays']

from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(min_df=10, max_features = 3000)
vectorizer.fit(X_train_text) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_tfidf = vectorizer.transform(X_train_text)
X_test_tfidf = vectorizer.transform(X_test_text)

print("After vectorizations")
print(X_train_tfidf.shape, y_train.shape)
print(X_test_tfidf.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(40200, 3000) (10000,)
(19800, 3000) (19800,)
====================================================================================================
```

In [239]:

```
X_train_text = X_train['preprocessed_titles']
X_test_text = X_test['preprocessed_titles']

from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(min_df=10)
vectorizer.fit(X_train_text) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_tfidf2 = vectorizer.transform(X_train_text)
X_test_tfidf2 = vectorizer.transform(X_test_text)

print("After vectorizations")
print(X_train_tfidf2.shape, y_train.shape)
print(X_test_tfidf2.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(40200, 1777) (10000,)
(19800, 1777) (19800,)
====================================================================================================
```

In [240]:

```
X_train_features_tfidf = hstack((X_train_features_set, X_train_tfidf, X_train_tfidf2))
X_test_features_tfidf = hstack((X_test_features_set, X_test_tfidf, X_test_tfidf2))
```

AvgW2V

In [139]:

```
from gensim.models import Word2Vec
from gensim.models import KeyedVectors
```

In [270]:

```
X_train_text = X_train['preprocessed_essays']
X_test_text = X_test['preprocessed_essays']

i=0
list_of_sentence_train=[]
for sentence in X_train_text:
```

```
for sentence in X_train_text:
    list_of_sentence_train.append(sentence.split())
```

In [271]:

```
# average Word2Vec
# compute average word2vec for each review.
sent_vectors_train = []; # the avg-w2v for each sentence/review is stored in this list
for sent in X_train_text: # 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
    sent_vectors_train.append(vector)
sent_vectors_train = np.array(sent_vectors_train)
print(sent_vectors_train.shape)
```

```
(40200, 300)
```

In [272]:

```
i=0
list_of_sentence_test=[]
for sentence in X_test_text:
    list_of_sentence_test.append(sentence.split())


sent_vectors_test = []; # the avg-w2v for each sentence/review is stored in this list
for sent in list_of_sentence_test: # 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
    sent_vectors_test.append(vector)
sent_vectors_test = np.array(sent_vectors_test)
print(sent_vectors_test.shape)
```

```
(19800, 300)
```

In [273]:

```
X_train_features_avgw2v = np.hstack((X_train_features, sent_vectors_train))
X_test_features_avgw2v = np.hstack((X_test_features, sent_vectors_test))
```

In [274]:

```
X_train_text = X_train['preprocessed_titles']
X_test_text = X_test['preprocessed_titles']

i=0
list_of_sentence_train=[]
for sentence in X_train_text:
    list_of_sentence_train.append(sentence.split())

# average Word2Vec
# compute average word2vec for each review.
sent_vectors_train = []; # the avg-w2v for each sentence/review is stored in this list
for sent in X_train_text: # 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:
```

```
    if cnt_words != 0:
        vector /= cnt_words
    sent_vectors_train.append(vector)
sent_vectors_train = np.array(sent_vectors_train)
print(sent_vectors_train.shape)
```

```
(40200, 300)
```

In [275]:

```
i=0
list_of_sentence_test=[]
for sentence in X_test_text:
    list_of_sentence_test.append(sentence.split())


sent_vectors_test = []; # the avg-w2v for each sentence/review is stored in this list
for sent in list_of_sentence_test: # 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
    sent_vectors_test.append(vector)
sent_vectors_test = np.array(sent_vectors_test)
print(sent_vectors_test.shape)
```

```
(19800, 300)
```

In [276]:

```
X_train_features_avgw2v = np.hstack((X_train_features, sent_vectors_train))
X_test_features_avgw2v = np.hstack((X_test_features, sent_vectors_test))
```

TFIDF Word2Vec

In [205]:

```
X_train_text = X_train['preprocessed_essays']
X_test_text = X_test['preprocessed_essays']
```

In [150]:

```
with open('glove_vectors', 'rb') as f:
    model = pickle.load(f)
    glove_words =  set(model.keys())
```

In [206]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train_text)
tfidf_model.transform(X_train_text)
# 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())

tfidf_w2v_vectors = [];
for sentence in tqdm(X_train_text): # 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.append(vector)

print(len(tfidf_w2v_vectors))
print(len(tfidf_w2v_vectors[0]))

X_train_features_tfidfw2v = np.hstack((X_train_features_set, tfidf_w2v_vectors))
```

```
100%|████████████████████████████████████████████████████████████| 40200/40200 [03:
49<00:00, 175.18it/s]
```

```
40200
300
```

In [207]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train_text)
tfidf_model.transform(X_test_text)
# 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())

tfidf_w2v_vectors = [];
for sentence in tqdm(X_test_text): # 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.append(vector)

print(len(tfidf_w2v_vectors))
print(len(tfidf_w2v_vectors[0]))

X_test_features_tfidfw2v = np.hstack((X_test_features_set, tfidf_w2v_vectors))
```

```
100%|████████████████████████████████████████████████████████████| 19800/19800 [00:
49<00:00, 402.80it/s]
```

```
19800
300
```

In [208]:

```
X_train_text = X_train['preprocessed_titles']
X_test_text = X_test['preprocessed_titles']
```

In [209]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train_text)
tfidf_model.transform(X_train_text)
# 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())

tfidf_w2v_vectors = [];
```

```
tfidf_w2v_vectors = [];
for sentence in tqdm(X_train_text): # 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.append(vector)

print(len(tfidf_w2v_vectors))
print(len(tfidf_w2v_vectors[0]))

X_train_features_tfidfw2v = np.hstack((X_train_features_set, tfidf_w2v_vectors))
```

```
100%|████████████████████████████████████████████████████████████████| 40200/40200
[00:03<00:00, 12787.89it/s]
```

```
40200
300
```

In [210]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train_text)
tfidf_model.transform(X_test_text)
# 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())

tfidf_w2v_vectors = [];
for sentence in tqdm(X_test_text): # 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.append(vector)

print(len(tfidf_w2v_vectors))
print(len(tfidf_w2v_vectors[0]))

X_test_features_tfidfw2v = np.hstack((X_test_features_set, tfidf_w2v_vectors))
```

```
100%|████████████████████████████████████████████████████████████████| 19800/19800
[00:00<00:00, 29499.59it/s]
```

```
19800
300
```

All the features have been encoded and vectorized.

2.4 Applying Random Forest

Apply Random Forest on different kind of featurization as mentioned in the instructions
For Every model that you work on make sure you do the step 2 and step 3 of instrucations

2.4.1 Applying Random Forests on BOW, SET 1

In [154]:

```
%matplotlib inline
import plotly.offline as offline
import plotly.graph_objs as go
offline.init_notebook_mode()
import numpy as np

def enable_plotly_in_cell():
  import IPython
  from plotly.offline import init_notebook_mode
  display(IPython.core.display.HTML('''<script src="/static/components/requirejs/require.js"></scr
ipt>'''))
  init_notebook_mode(connected=False)
```

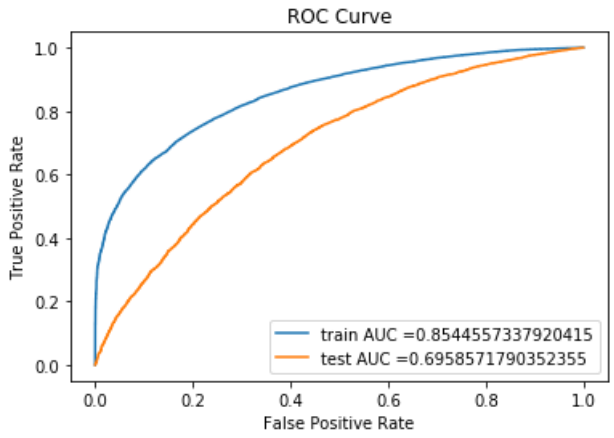
In [186]:

```
# https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt

n_estimators = [10, 50, 100, 150, 200, 300, 500, 1000]
max_depth = [2, 3, 4, 5, 6, 7, 8, 9, 10]
parameters = {'n_estimators': n_estimators, 'max_depth':max_depth}
rf = RandomForestClassifier(class_weight = 'balanced')
clf = GridSearchCV(rf, parameters, cv=3, scoring='roc_auc')
clf.fit(X_train_features_bow, y_train)
best_est = clf.best_estimator_.get_params()['n_estimators']
best_depth = clf.best_estimator_.get_params()['max_depth']
print('Best estimator:', best_est)
print('Best Depth:', best_depth)


train_auc = []
cv_auc = []
for i in range(0,72,8):
    train_auc.append(clf.cv_results_['mean_train_score'][i])
    cv_auc.append(clf.cv_results_['mean_test_score'][i])

trace1 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=train_auc, name = 'train')
trace2 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=cv_auc, name = 'Cross validation
')
data = [trace1, trace2]
enable_plotly_in_cell()

layout = go.Layout(scene = dict(
        xaxis = dict(title='n_estimators'),
        yaxis = dict(title='max_depth'),
        zaxis = dict(title='AUC'),))

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='3d-scatter-colorscale')
```

```
Best estimator: 1000
Best Depth: 10
```

In [187]:

```
from sklearn.metrics import roc_curve, auc


rf = RandomForestClassifier(class_weight = 'balanced',n_estimators= best_est, max_depth =
best_depth)
rf.fit(X_train_features_bow, y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs
y_train_pred_bow =  rf.predict_proba(X_train_features_bow)[:,1]
y_test_pred_bow = rf.predict_proba(X_test_features_bow)[:,1]
train_fpr, train_tpr, thresholds = roc_curve(y_train, y_train_pred_bow)
test_fpr, test_tpr, thresholds = roc_curve(y_test, y_test_pred_bow)

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("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve")
plt.show()

print("="*100)
```



==========

In [188]:

```
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
```

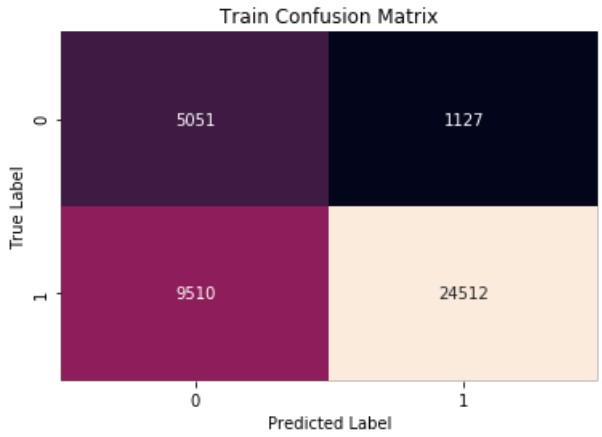
In [189]:

```
from sklearn.metrics import confusion_matrix
best_t = find_best_threshold(thresholds, train_fpr, train_tpr)

print("Train confusion matrix")
matrix_train= confusion_matrix(y_train, predict_with_best_t(y_train_pred_bow, best_t))
print(matrix_train)
sns.heatmap(matrix_train,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Train Confusion Matrix')
```

```
the maximum value of tpr*(1-fpr) 0.5905972444200835 for threshold 0.508
Train confusion matrix
[[ 5051  1127]
 [ 9510 24512]]
```

Out[189]:

```
Text(0.5,1,'Train Confusion Matrix')
```



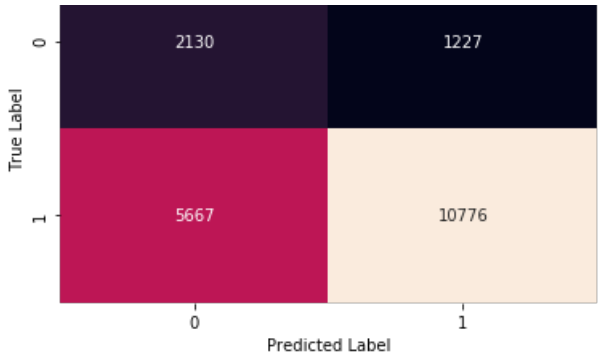
In [190]:

```
print("Test confusion matrix")
matrix_test= confusion_matrix(y_test, predict_with_best_t(y_test_pred_bow, best_t))
print(matrix_test)
sns.heatmap(matrix_test,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Test Confusion Matrix')
```

```
Test confusion matrix
[[ 2130  1227]
 [ 5667 10776]]
```

Out[190]:

```
Text(0.5,1,'Test Confusion Matrix')
```



2.4.2 Applying Random Forests on TFIDF, SET 2

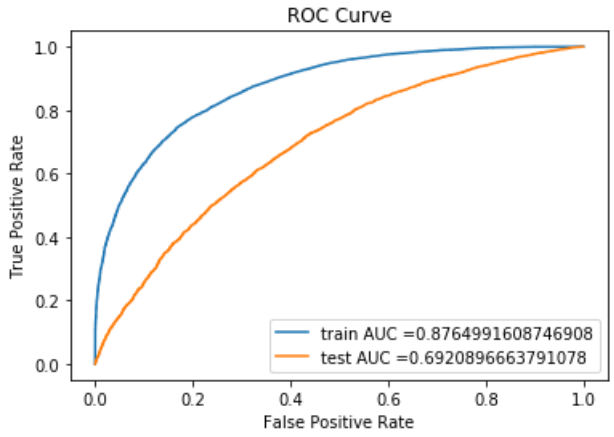
In [191]:

```
# https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt

n_estimators = [10, 50, 100, 150, 200, 300, 500, 1000]
max_depth = [2, 3, 4, 5, 6, 7, 8, 9, 10]
parameters = {'n_estimators': n_estimators, 'max_depth':max_depth}
rf = RandomForestClassifier(class_weight = 'balanced')
clf = GridSearchCV(rf, parameters, cv=3, scoring='roc_auc')
clf.fit(X_train_features_tfidf, y_train)
best_est = clf.best_estimator_.get_params()['n_estimators']
best_depth = clf.best_estimator_.get_params()['max_depth']
print('Best estimator:', best_est)
print('Best Depth:', best_depth)


train_auc = []
cv_auc = []
for i in range(0,72,8):
    train_auc.append(clf.cv_results_['mean_train_score'][i])
    cv_auc.append(clf.cv_results_['mean_test_score'][i])

trace1 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=train_auc, name = 'train')
trace2 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=cv_auc, name = 'Cross validation
')
data = [trace1, trace2]
enable_plotly_in_cell()

layout = go.Layout(scene = dict(
        xaxis = dict(title='n_estimators'),
        yaxis = dict(title='max_depth'),
        zaxis = dict(title='AUC'),))

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='3d-scatter-colorscale')
```

```
Best estimator: 1000
Best Depth: 10
```

In [192]:

```
from sklearn.metrics import roc_curve, auc


rf = RandomForestClassifier(class_weight = 'balanced',n_estimators= best_est, max_depth =
best_depth)
rf.fit(X_train_features_tfidf, y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs
y_train_pred_tfidf =  rf.predict_proba(X_train_features_tfidf)[:,1]
y_test_pred_tfidf = rf.predict_proba(X_test_features_tfidf)[:,1]
train_fpr, train_tpr, thresholds = roc_curve(y_train, y_train_pred_tfidf)
test_fpr, test_tpr, thresholds = roc_curve(y_test, y_test_pred_tfidf)

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("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve")
plt.show()

print("="*100)
```



==========

In [193]:

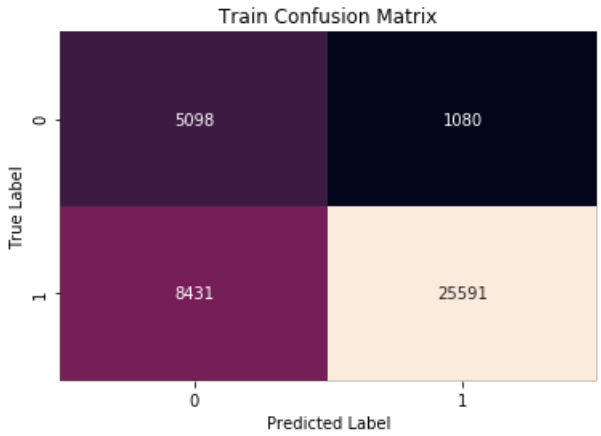
```
best_t = find_best_threshold(thresholds, train_fpr, train_tpr)

print("Train confusion matrix")
matrix_train= confusion_matrix(y_train, predict_with_best_t(y_train_pred_tfidf, best_t))
print(matrix_train)
sns.heatmap(matrix_train,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
```

```
plt.title('Train Confusion Matrix')
```

```
the maximum value of tpr*(1-fpr) 0.6229294789715695 for threshold 0.514
Train confusion matrix
[[ 5098  1080]
 [ 8431 25591]]
```

Out[193]:

```
Text(0.5,1,'Train Confusion Matrix')
```



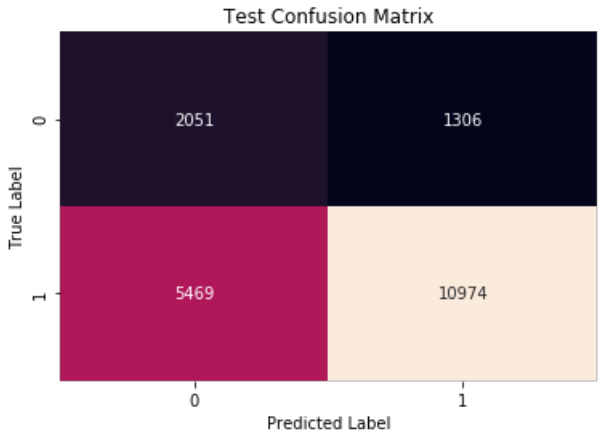
In [194]:

```
print("Test confusion matrix")
matrix_test= confusion_matrix(y_test, predict_with_best_t(y_test_pred_tfidf, best_t))
print(matrix_test)
sns.heatmap(matrix_test,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Test Confusion Matrix')
```

```
Test confusion matrix
[[ 2051  1306]
 [ 5469 10974]]
```

Out[194]:

```
Text(0.5,1,'Test Confusion Matrix')
```



2.4.3 Applying Random Forests on AVG W2V, SET 3

In [199]:

```
X_train_features_avgw2v = X_train_features_avgw2v[0:20000]
y_train = y_train[0:20000]
```

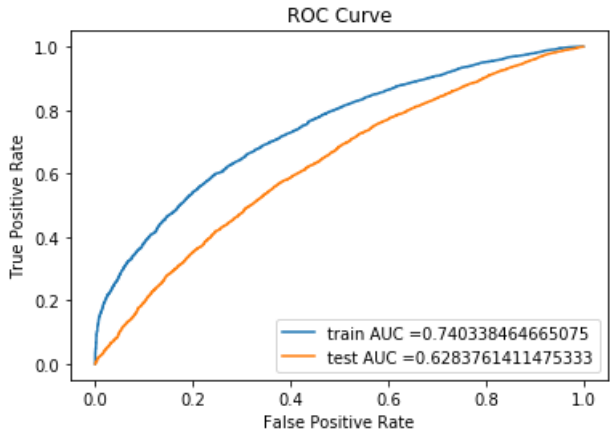
In [200]:

```
# https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt

n_estimators = [10, 50, 100, 150, 200, 300, 500, 1000]
max_depth = [2, 3, 4, 5, 6, 7, 8, 9, 10]
parameters = {'n_estimators': n_estimators, 'max_depth':max_depth}
rf = RandomForestClassifier(class_weight = 'balanced')
clf = GridSearchCV(rf, parameters, cv=3, scoring='roc_auc')
clf.fit(X_train_features_avgw2v, y_train)
best_est = clf.best_estimator_.get_params()['n_estimators']
best_depth = clf.best_estimator_.get_params()['max_depth']
print('Best estimator:', best_est)
print('Best Depth:', best_depth)


train_auc = []
cv_auc = []
for i in range(0,72,8):
    train_auc.append(clf.cv_results_['mean_train_score'][i])
    cv_auc.append(clf.cv_results_['mean_test_score'][i])

trace1 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=train_auc, name = 'train')
trace2 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=cv_auc, name = 'Cross validation
')
data = [trace1, trace2]
enable_plotly_in_cell()

layout = go.Layout(scene = dict(
        xaxis = dict(title='n_estimators'),
        yaxis = dict(title='max_depth'),
        zaxis = dict(title='AUC'),))

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='3d-scatter-colorscale')
```

```
Best estimator: 1000
Best Depth: 7
```

In [201]:

```
from sklearn.metrics import roc_curve, auc


rf = RandomForestClassifier(class_weight = 'balanced',n_estimators= best_est, max_depth =
best_depth)
rf.fit(X_train_features_avgw2v, y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs
y_train_pred_avgw2v =  rf.predict_proba(X_train_features_avgw2v)[:,1]
y_test_pred_avgw2v = rf.predict_proba(X_test_features_avgw2v)[:,1]
train_fpr, train_tpr, thresholds = roc_curve(y_train, y_train_pred_avgw2v)
test_fpr, test_tpr, thresholds = roc_curve(y_test, y_test_pred_avgw2v)

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("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve")
plt.show()

print("="*100)
```



==========

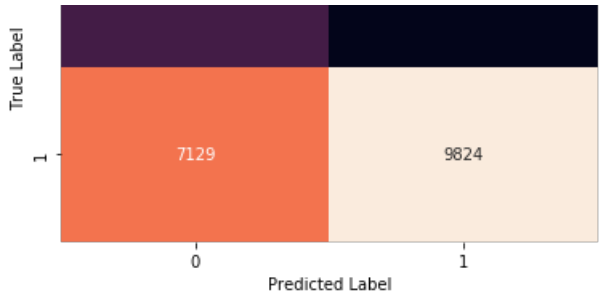
In [202]:

```
best_t = find_best_threshold(thresholds, train_fpr, train_tpr)

print("Train confusion matrix")
matrix_train= confusion_matrix(y_train, predict_with_best_t(y_train_pred_avgw2v, best_t))
print(matrix_train)
sns.heatmap(matrix_train,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Train Confusion Matrix')
```

```
the maximum value of tpr*(1-fpr) 0.4556190805402632 for threshold 0.523
Train confusion matrix
[[2340  707]
 [7129 9824]]
```

Out[202]:

```
Text(0.5,1,'Train Confusion Matrix')
```



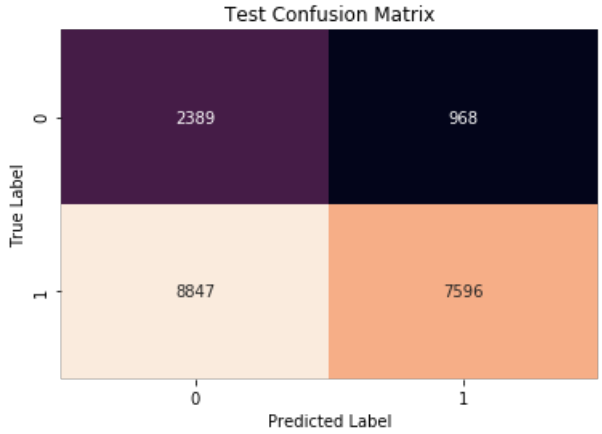
In [203]:

```
print("Test confusion matrix")
matrix_test= confusion_matrix(y_test, predict_with_best_t(y_test_pred_avgw2v, best_t))
print(matrix_test)
sns.heatmap(matrix_test,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Test Confusion Matrix')
```

```
Test confusion matrix
[[2389  968]
 [8847 7596]]
```

Out[203]:

```
Text(0.5,1,'Test Confusion Matrix')
```



2.4.4 Applying Random Forests on TFIDF W2V, SET 4

In [211]:

```
X_train_features_tfidfw2v = X_train_features_tfidfw2v[0:20000]
```

In [212]:

```
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt

n_estimators = [10, 50, 100, 150, 200, 300, 500, 1000]
max_depth = [2, 3, 4, 5, 6, 7, 8, 9, 10]
parameters = {'n_estimators': n_estimators, 'max_depth':max_depth}
rf = RandomForestClassifier(class_weight = 'balanced')
clf = GridSearchCV(rf, parameters, cv=3, scoring='roc_auc')
clf.fit(X_train_features_tfidfw2v, y_train)
best_est = clf.best_estimator_.get_params()['n_estimators']
best_depth = clf.best_estimator_.get_params()['max_depth']
print('Best estimator:', best_est)
print('Best Depth:', best_depth)
```

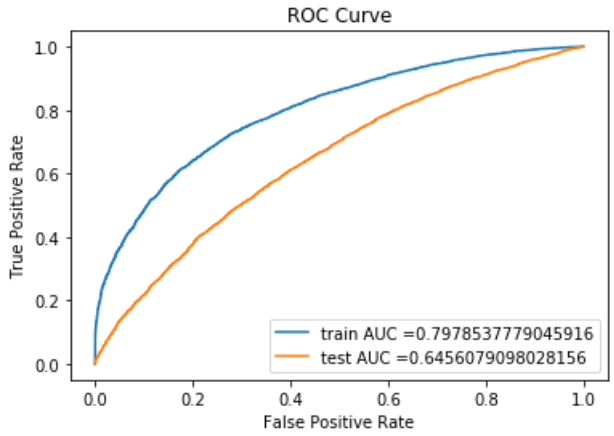
```
train_auc = []
cv_auc = []
for i in range(0,72,8):
    train_auc.append(clf.cv_results_['mean_train_score'][i])
    cv_auc.append(clf.cv_results_['mean_test_score'][i])

trace1 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=train_auc, name = 'train')
trace2 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=cv_auc, name = 'Cross validation
')
data = [trace1, trace2]
enable_plotly_in_cell()

layout = go.Layout(scene = dict(
        xaxis = dict(title='n_estimators'),
        yaxis = dict(title='max_depth'),
        zaxis = dict(title='AUC'),))

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='3d-scatter-colorscale')
```

```
Best estimator: 1000
Best Depth: 5
```

In [213]:

```
from sklearn.metrics import roc_curve, auc


rf = RandomForestClassifier(class_weight = 'balanced',n_estimators= best_est, max_depth =
best_depth)
rf.fit(X_train_features_tfidfw2v, y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs
y_train_pred_tfidfw2v =  rf.predict_proba(X_train_features_tfidfw2v)[:,1]
y_test_pred_tfidfw2v = rf.predict_proba(X_test_features_tfidfw2v)[:,1]
train_fpr, train_tpr, thresholds = roc_curve(y_train, y_train_pred_tfidfw2v)
test_fpr, test_tpr, thresholds = roc_curve(y_test, y_test_pred_tfidfw2v)

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("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve")
plt.show()

print("="*100)
```



==========

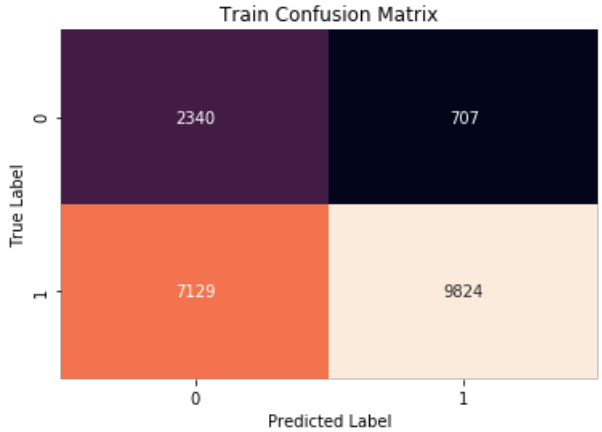
In [202]:

```
best_t = find_best_threshold(thresholds, train_fpr, train_tpr)

print("Train confusion matrix")
matrix_train= confusion_matrix(y_train, predict_with_best_t(y_train_pred_tfidfw2v, best_t))
print(matrix_train)
sns.heatmap(matrix_train,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Train Confusion Matrix')
```

```
the maximum value of tpr*(1-fpr) 0.4556190805402632 for threshold 0.523
Train confusion matrix
[[2340  707]
 [7129 9824]]
```

Out[202]:

```
Text(0.5,1,'Train Confusion Matrix')
```



In [203]:

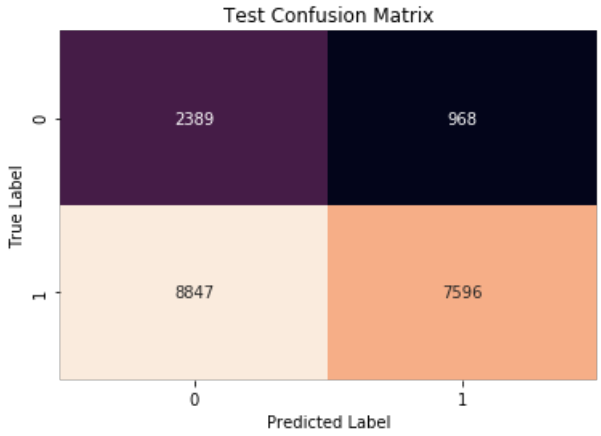
```
print("Test confusion matrix")
matrix_test= confusion_matrix(y_test, predict_with_best_t(y_test_pred_tfidfw2v, best_t))
print(matrix_test)
sns.heatmap(matrix_test,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
```

```
plt.xlabel('Predicted Label')
plt.title('Test Confusion Matrix')
```

```
Test confusion matrix
[[2389  968]
 [8847 7596]]
```

Out[203]:

```
Text(0.5,1,'Test Confusion Matrix')
```



2.5 Applying GBDT

Apply GBDT on different kind of featurization as mentioned in the instructions
For Every model that you work on make sure you do the step 2 and step 3 of instrucations

2.5.1 Applying XGBOOST on BOW, SET 1

In [218]:

```
w = X_train_features_bow

y_train= z
```

In [250]:

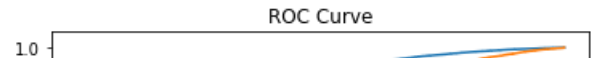
```
# https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
from sklearn.model_selection import GridSearchCV
from xgboost import XGBClassifier as xgb
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt

n_estimators = [10, 50, 100, 200, 300, 500, 1000]
max_depth = [2, 3, 5, 6, 7, 8, 10]
parameters = {'n_estimators': n_estimators, 'max_depth':max_depth}
xgb = xgb()
clf = GridSearchCV(xgb, parameters, cv=3, scoring='roc_auc')
clf.fit(X_train_features_bow, y_train)
best_est = clf.best_estimator_.get_params()['n_estimators']
best_depth = clf.best_estimator_.get_params()['max_depth']
print('Best estimator:', best_est)
print('Best Depth:', best_depth)


train_auc = []
cv_auc = []
for i in range(0,49,7):
    train_auc.append(clf.cv_results_['mean_train_score'][i])
    cv_auc.append(clf.cv_results_['mean_test_score'][i])

trace1 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=train_auc, name = 'train')
trace2 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=cv_auc, name = 'Cross validation
```

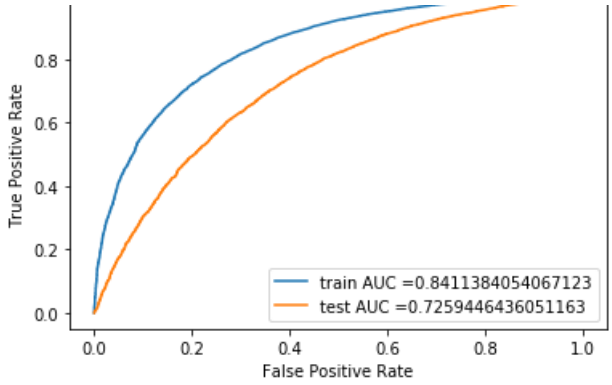
```
')
data = [trace1, trace2]
enable_plotly_in_cell()

layout = go.Layout(scene = dict(
        xaxis = dict(title='n_estimators'),
        yaxis = dict(title='max_depth'),
        zaxis = dict(title='AUC'),))

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='3d-scatter-colorscale')
```

```
Best estimator: 300
Best Depth: 2
```

In [251]:

```
from sklearn.metrics import roc_curve, auc
from xgboost import XGBClassifier

xgb = XGBClassifier(class_weight = 'balanced',n_estimators= best_est, max_depth = best_depth)
xgb.fit(X_train_features_bow, y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs
y_train_pred_bow =  xgb.predict_proba(X_train_features_bow)[:,1]
y_test_pred_bow = xgb.predict_proba(X_test_features_bow)[:,1]
train_fpr, train_tpr, thresholds = roc_curve(y_train, y_train_pred_bow)
test_fpr, test_tpr, thresholds = roc_curve(y_test, y_test_pred_bow)

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("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve")
plt.show()

print("="*100)
```





==========

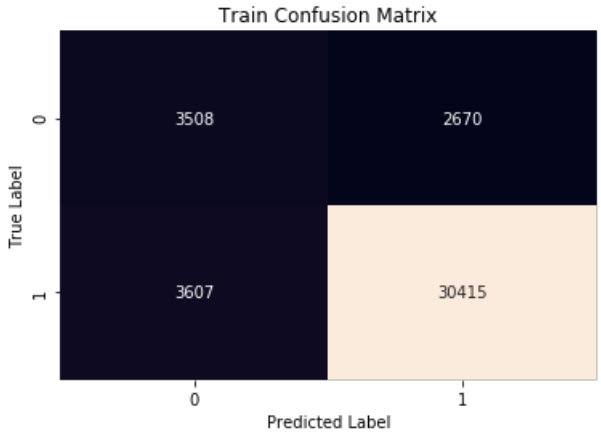
In [252]:

```
from sklearn.metrics import confusion_matrix
best_t = find_best_threshold(thresholds, train_fpr, train_tpr)

print("Train confusion matrix")
matrix_train= confusion_matrix(y_train, predict_with_best_t(y_train_pred_bow, best_t))
print(matrix_train)
sns.heatmap(matrix_train,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Train Confusion Matrix')
```

```
the maximum value of tpr*(1-fpr) 0.580264148011249 for threshold 0.742
Train confusion matrix
[[ 3508  2670]
 [ 3607 30415]]
```

Out[252]:

```
Text(0.5,1,'Train Confusion Matrix')
```



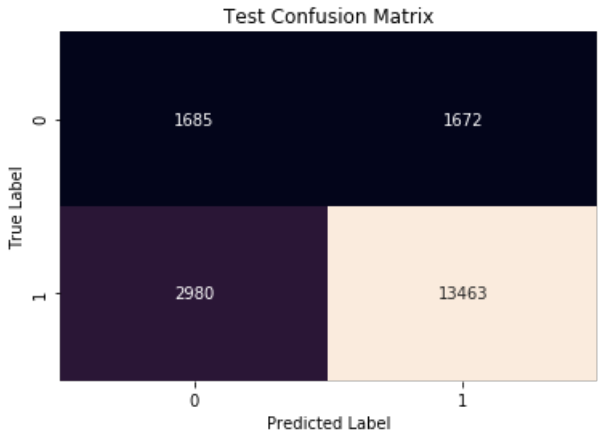
In [253]:

```
print("Test confusion matrix")
matrix_test= confusion_matrix(y_test, predict_with_best_t(y_test_pred_bow, best_t))
print(matrix_test)
sns.heatmap(matrix_test,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Test Confusion Matrix')
```

```
Test confusion matrix
[[ 1685  1672]
 [ 2980 13463]]
```

Out[253]:

```
Text(0.5,1,'Test Confusion Matrix')
```



2.5.2 Applying XGBOOST on TFIDF, SET 2

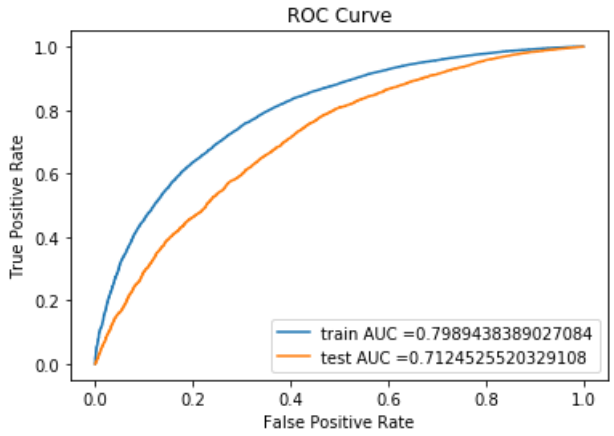
In [264]:

```
# Please write all the code with proper documentation

# https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
from sklearn.model_selection import GridSearchCV
from xgboost import XGBRegressor as xgb
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt

n_estimators = [10, 50, 100, 200, 300, 500, 1000]
max_depth = [2, 3, 5, 6, 7, 8, 10]
parameters = {'n_estimators': n_estimators, 'max_depth':max_depth}
xgb = xgb()
clf = GridSearchCV(xgb, parameters, cv=3, scoring='roc_auc')
clf.fit(X_train_features_tfidf, y_train)
best_est = clf.best_estimator_.get_params()['n_estimators']
best_depth = clf.best_estimator_.get_params()['max_depth']
print('Best estimator:', best_est)
print('Best Depth:', best_depth)


train_auc = []
cv_auc = []
for i in range(0,49,7):
    train_auc.append(clf.cv_results_['mean_train_score'][i])
    cv_auc.append(clf.cv_results_['mean_test_score'][i])

trace1 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=train_auc, name = 'train')
trace2 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=cv_auc, name = 'Cross validation
')
data = [trace1, trace2]
enable_plotly_in_cell()

layout = go.Layout(scene = dict(
        xaxis = dict(title='n_estimators'),
        yaxis = dict(title='max_depth'),
        zaxis = dict(title='AUC'),))

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='3d-scatter-colorscale')
```

```
Best estimator: 100
Best Depth: 2
```

In [265]:

```
from sklearn.metrics import roc_curve, auc


xgb = XGBClassifier(class_weight = 'balanced',n_estimators= best_est, max_depth = best_depth)
xgb.fit(X_train_features_tfidf, y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs
y_train_pred_tfidf =  xgb.predict_proba(X_train_features_tfidf)[:,1]
y_test_pred_tfidf = xgb.predict_proba(X_test_features_tfidf)[:,1]
train_fpr, train_tpr, thresholds = roc_curve(y_train, y_train_pred_tfidf)
test_fpr, test_tpr, thresholds = roc_curve(y_test, y_test_pred_tfidf)

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("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve")
plt.show()

print("="*100)
```



==========

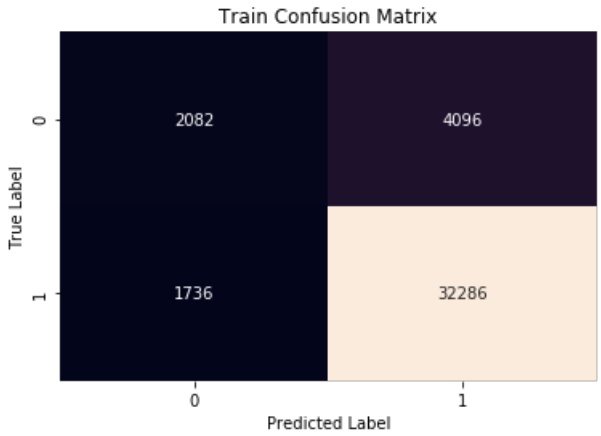
In [266]:

```
best_t = find_best_threshold(thresholds, train_fpr, train_tpr)

print("Train confusion matrix")
matrix_train= confusion_matrix(y_train, predict_with_best_t(y_train_pred_tfidf, best_t))
print(matrix_train)
sns.heatmap(matrix_train,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Train Confusion Matrix')
```

```
the maximum value of tpr*(1-fpr) 0.5255839874258043 for threshold 0.687
Train confusion matrix
[[ 2082  4096]
 [ 1736 32286]]
```

Out[266]:

```
Text(0.5,1,'Train Confusion Matrix')
```



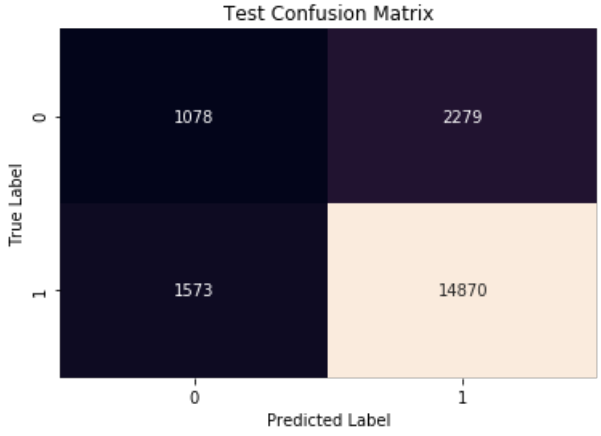
In [267]:

```
print("Test confusion matrix")
matrix_test= confusion_matrix(y_test, predict_with_best_t(y_test_pred_tfidf, best_t))
print(matrix_test)
sns.heatmap(matrix_test,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Test Confusion Matrix')
```

```
Test confusion matrix
[[ 1078  2279]
 [ 1573 14870]]
```

Out[267]:

```
Text(0.5,1,'Test Confusion Matrix')
```



2.5.3 Applying XGBOOST on AVG W2V, SET 3

In [ ]:

```
# Please write all the code with proper documentation
```

In [278]:

```
X_train_features_avgw2v = X_train_features_avgw2v[0:20000]
y_train = y_train[0:20000]
```

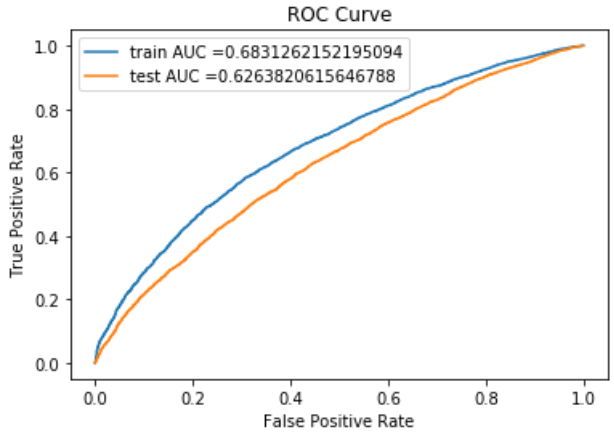
In [307]:

```
from sklearn.model_selection import GridSearchCV
from xgboost import XGBRegressor as xgb
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt

n_estimators = [10, 50, 100, 200, 300, 500, 1000]
max_depth = [2, 3, 5, 6, 7, 8, 10]
parameters = {'n_estimators': n_estimators, 'max_depth':max_depth}
xgb = xgb()
#clf = GridSearchCV(xgb, parameters, cv=3, scoring='roc_auc')
#clf.fit(X_train_features_avgw2v, y_train)
best_est = clf.best_estimator_.get_params()['n_estimators']
best_depth = clf.best_estimator_.get_params()['max_depth']
print('Best estimator:', best_est)
print('Best Depth:', best_depth)


train_auc = []
cv_auc = []
for i in range(0,49,7):
    train_auc.append(clf.cv_results_['mean_train_score'][i])
    cv_auc.append(clf.cv_results_['mean_test_score'][i])

trace1 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=train_auc, name = 'train')
trace2 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=cv_auc, name = 'Cross validation
')
data = [trace1, trace2]
enable_plotly_in_cell()

layout = go.Layout(scene = dict(
        xaxis = dict(title='n_estimators'),
        yaxis = dict(title='max_depth'),
        zaxis = dict(title='AUC'),))

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='3d-scatter-colorscale')
```

```
Best estimator: 300
Best Depth: 4
```

In [306]:

```
from sklearn.metrics import roc_curve, auc


xgb = RandomForestClassifier(class_weight = 'balanced',n_estimators= best_est, max_depth = best_dep
th)
xgb.fit(X_train_features_avgw2v, y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs
y_train_pred_avgw2v =  xgb.predict_proba(X_train_features_avgw2v)[:,1]
y_test_pred_avgw2v = xgb.predict_proba(X_test_features_avgw2v)[:,1]
train_fpr, train_tpr, thresholds = roc_curve(y_train, y_train_pred_avgw2v)
test_fpr, test_tpr, thresholds = roc_curve(y_test, y_test_pred_avgw2v)

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("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve")
plt.show()

print("="*100)
```



==========

In [308]:

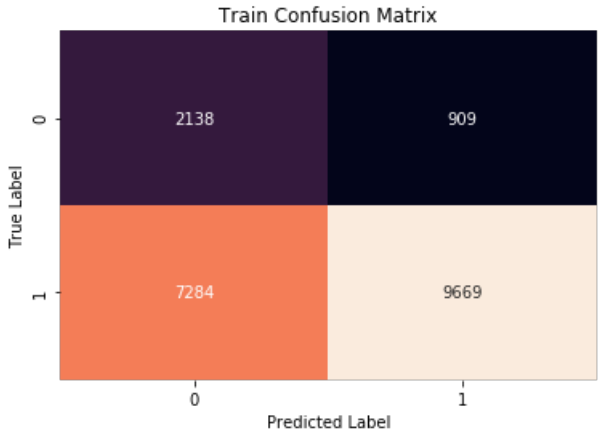
```
best_t = find_best_threshold(thresholds, train_fpr, train_tpr)

print("Train confusion matrix")
matrix_train= confusion_matrix(y_train, predict_with_best_t(y_train_pred_avgw2v, best_t))
print(matrix_train)
sns.heatmap(matrix_train,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Train Confusion Matrix')
```

```
the maximum value of tpr*(1-fpr) 0.40455870668982685 for threshold 0.508
Train confusion matrix
[[2138  909]
 [7284 96691]
```

[7284 9669]]

Out[308]:

Text(0.5,1,'Train Confusion Matrix')



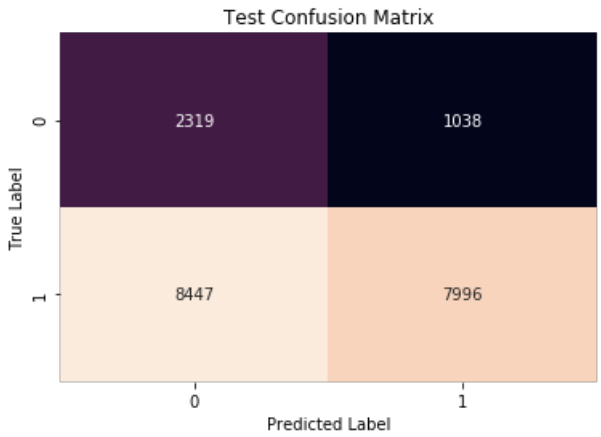
In [309]:

```
print("Test confusion matrix")
matrix_test= confusion_matrix(y_test, predict_with_best_t(y_test_pred_avgw2v, best_t))
print(matrix_test)
sns.heatmap(matrix_test,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Test Confusion Matrix')
```

```
Test confusion matrix
[[2319 1038]
 [8447 7996]]
```

Out[309]:

Text(0.5,1,'Test Confusion Matrix')



2.5.4 Applying XGBOOST on TFIDF W2V, SET 4

In [310]:

```
X_train_features_tfidfw2v = X_train_features_tfidfw2v[0:20000]
y_train = y_train[0:20000]
```

In [311]:

```
from sklearn.model_selection import GridSearchCV
from xgboost import XGBRegressor as xgb
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt
```

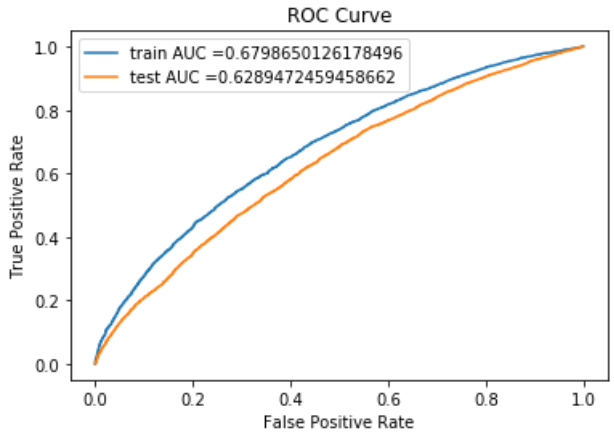
```
import matplotlib.pyplot as plt

n_estimators = [10, 50, 100, 200, 300, 500, 1000]
max_depth = [2, 3, 5, 6, 7, 8, 10]
parameters = {'n_estimators': n_estimators, 'max_depth':max_depth}
xgb = xgb()
clf = GridSearchCV(xgb, parameters, cv=3, scoring='roc_auc')
clf.fit(X_train_features_tfidfw2v, y_train)
best_est = clf.best_estimator_.get_params()['n_estimators']
best_depth = clf.best_estimator_.get_params()['max_depth']
print('Best estimator:', best_est)
print('Best Depth:', best_depth)


train_auc = []
cv_auc = []
for i in range(0,49,7):
    train_auc.append(clf.cv_results_['mean_train_score'][i])
    cv_auc.append(clf.cv_results_['mean_test_score'][i])

trace1 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=train_auc, name = 'train')
trace2 = go.Scatter3d(x=np.log(n_estimators),y=np.log(max_depth),z=cv_auc, name = 'Cross validation
')
data = [trace1, trace2]
enable_plotly_in_cell()

layout = go.Layout(scene = dict(
        xaxis = dict(title='n_estimators'),
        yaxis = dict(title='max_depth'),
        zaxis = dict(title='AUC'),))

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='3d-scatter-colorscale')
```

```
Best estimator: 50
Best Depth: 2
```

In [312]:

```
from sklearn.metrics import roc_curve, auc


xgb = RandomForestClassifier(class_weight = 'balanced',n_estimators= best_est, max_depth = best_dep
th)
```

```
tn)
xgb.fit(X_train_features_tfidfw2v, y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs
y_train_pred_tfidfw2v =  xgb.predict_proba(X_train_features_tfidfw2v)[:,1]
y_test_pred_tfidfw2v = xgb.predict_proba(X_test_features_tfidfw2v)[:,1]
train_fpr, train_tpr, thresholds = roc_curve(y_train, y_train_pred_tfidfw2v)
test_fpr, test_tpr, thresholds = roc_curve(y_test, y_test_pred_tfidfw2v)

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("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve")
plt.show()

print("="*100)
```




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

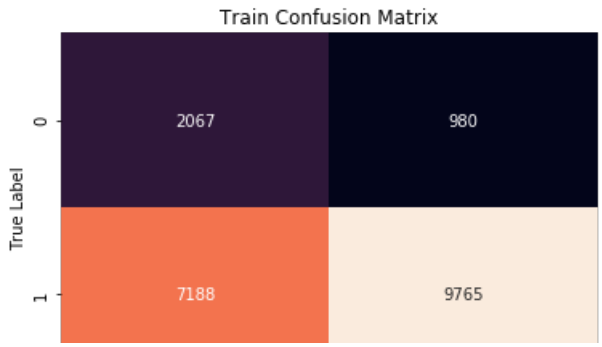
In [313]:

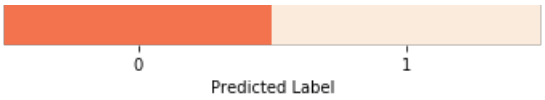
```
best_t = find_best_threshold(thresholds, train_fpr, train_tpr)

print("Train confusion matrix")
matrix_train= confusion_matrix(y_train, predict_with_best_t(y_train_pred_tfidfw2v, best_t))
print(matrix_train)
sns.heatmap(matrix_train,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Train Confusion Matrix')
```

```
the maximum value of tpr*(1-fpr) 0.3939630505319336 for threshold 0.504
Train confusion matrix
[[2067  980]
 [7188 9765]]
```

Out[313]:

```
Text(0.5,1,'Train Confusion Matrix')
```





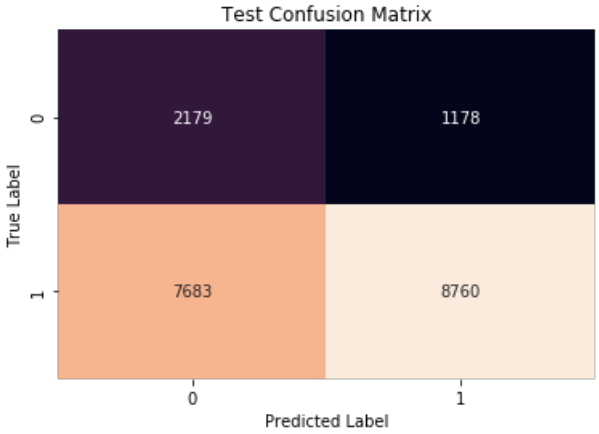
In [314]:

```
print("Test confusion matrix")
matrix_test= confusion_matrix(y_test, predict_with_best_t(y_test_pred_tfidfw2v, best_t))
print(matrix_test)
sns.heatmap(matrix_test,annot=True,cbar=False,fmt='d')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.title('Test Confusion Matrix')
```

```
Test confusion matrix
[[2179 1178]
 [7683 8760]]
```

Out[314]:

```
Text(0.5,1,'Test Confusion Matrix')
```



3. Conclusion

In [316]:

```
from prettytable import PrettyTable

x = PrettyTable()

x.field_names = ["Vectorizer", "Model", "n_estimators","Max_depth", "AUC"]


x.add_row(["BOW", "Random Forest", 1000,10, 0.695])
x.add_row(["TFIDF", "Random Forest", 1000,10, 0.692])
x.add_row(["AVGW2V","Random Forest", 1000,7, 0.628])
x.add_row(["TFIDFW2V", "Random Forest", 1000,5, 0.645])
x.add_row(["BOW", "XGBoost", 300,2, 0.725])
x.add_row(["TFIDF","XGBoost", 100,2, 0.712])
x.add_row(["AVGW2V", "XGBoost", 300,4, 0.628])
x.add_row(["TFIDFW2V", "XGBoost", 50 ,2, 0.628])


print(x)
```

```
+------------+---------------+--------------+-----------+-------+
| Vectorizer |     Model     | n_estimators | Max_depth |  AUC  |
+------------+---------------+--------------+-----------+-------+
|    BOW     | Random Forest |     1000     |     10    | 0.695 |
|   TFIDF    | Random Forest |     1000     |     10    | 0.692 |
|   AVGW2V   | Random Forest |     1000     |     7     | 0.628 |
|  TFIDFW2V  | Random Forest |     1000     |     5     | 0.645 |
|    BOW     |    XGBoost    |     300      |     2     | 0.725 |
```

TFIDF	XGBoost	100	2	0.712
AVGW2V	XGBoost	300	4	0.628
TFIDFW2V	XGBoost	50	2	0.628

From the above table, we can conclude that XGBoost with BOW and TFIDF vectorizations perform better for this case.