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`

- Literature & Writing, Social Sciences

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

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

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

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

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

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

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

Notes on the Essay Data

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

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

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

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

your neighborhood, and your school are all helpful.

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

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

In [1]:

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

1.1 Reading Data

In [2]:

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

In [3]:

```
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 [4]:

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


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


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


project_data.head(2)
```

Out[4]:

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

In [5]:

```
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[5]:

	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 [6]:

```
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 & Hunger"]
        if 'The' in j.split(): # this will split each of the catogory based on space "Math & Science"=> "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 [7]:

```
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 [8]:

```
# 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 [9]:

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

Out[9]:

Out[9]:

	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 [10]:

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

In [11]:

```
# 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 is 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 students 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 for 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 about changes over time. Students will be studying photos to learn about how their community has changed 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. Through 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 biggest 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. \r\nStudents 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 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 knowledge of where the ingredients came from as well as how it's 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 classroom garden in the spring. We will also create our own cookbooks to be printed and shared with families. \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-working 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 with them is limited, I want to ensure they get the most of this time and enjoy it to the best of their abilities.Currently, we have twenty-two desks of differing sizes, yet the desks are similar to the ones the students will use in middle school. We also have a kidney table with crates for seating. 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 constricting desks and move toward more “fun” seating options. I am requesting various seating so my students have more options to sit. Currently, I have a stool and a papasan chair I inherited from the previous sixth-grade teacher as well as five milk crate seats I made, but I would like to give them more options and reduce the competition for the “good seats”. I am also requesting two rugs as not only more seating options but to make the classroom more welcoming and appealing. In order 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 tables that we can fold up when we are not using them to leave more room for our flexible seating options.\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 [12]:

```python
# 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 [13]:

```python
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 [14]:

```python
# \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
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 [15]:

```
#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 [16]:

```
# 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', "dc
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 [17]:

```
# Combining all the above stundents
from tqdm import tqdm_notebook as 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.lower() not in stopwords)
    preprocessed_essays.append(sent.lower().strip())
```

In [18]:

```
# after preprocesing
preprocessed_essays[20000]
```

Out[18]:

```
'person person no matter small dr seuss teach smallest students biggest enthusiasm learning
students learn many different ways using senses multiple intelligences use wide range techniques h
elp students succeed students class come variety different backgrounds makes wonderful sharing exp
eriences cultures including native americans school caring community successful learners seen coll
aborative student project based learning classroom kindergarteners class love work hands materials
many different opportunities practice skill mastered social skills work cooperatively friends cruc
ial aspect kindergarten curriculum montana perfect place learn agriculture nutrition students love
role play pretend kitchen early childhood classroom several kids ask try cooking real food take id
ea create common core cooking lessons learn important math writing concepts cooking delicious heal
thy food snack time students grounded appreciation work went making food knowledge ingredients cam
e well healthy bodies project would expand learning nutrition agricultural cooking recipes us peel
apples make homemade applesauce make bread mix healthy plants classroom garden spring also create
cookbooks printed shared families students gain math literature skills well life long enjoyment he
althy cooking nannan'
```

1.4 Preprocessing of `project_title`

In [19]:

```
# similarly you can preprocess the titles also
from tqdm import tqdm_notebook as tqdm
preprocessed_title = []
# tqdm is for printing the status bar
for sentance in tqdm(project_data['project_title'].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.lower() not in stopwords)
    preprocessed_title.append(sent.lower().strip())
```

1.5 Preparing data for models

In [20]:

```
project_data.columns
```

Out[20]:

```
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'],
      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
```

1.5.1 Vectorizing Categorical data

- https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/handling-categorical-and-numerical-features/

In [211]:

```
# we use count vectorizer to convert the values into one
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(vocabulary=list(sorted_cat_dict.keys()), lowercase=False, binary=True)
categories_one_hot = vectorizer.fit_transform(project_data['clean_categories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (109248, 9)
```

In [0]:

```
# we use count vectorizer to convert the values into one
vectorizer = CountVectorizer(vocabulary=list(sorted_sub_cat_dict.keys()), lowercase=False, binary=True)
sub_categories_one_hot = vectorizer.fit_transform(project_data['clean_subcategories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot.shape)
```

In [0]:

```
# you can do the similar thing with state, teacher_prefix and project_grade_category also
```

1.5.2 Vectorizing Text data

1.5.2.1 Bag of words

In [0]:

```
# We are considering only the words which appeared in at least 10 documents(rows or projects).
vectorizer = CountVectorizer(min_df=10)
text_bow = vectorizer.fit_transform(preprocessed_essays)
print("Shape of matrix after one hot encodig ",text_bow.shape)
```

In [0]:

```
# you can vectorize the title also
# before you vectorize the title make sure you preprocess it
```

1.5.2.2 TFIDF vectorizer

In [0]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(min_df=10)
text_tfidf = vectorizer.fit_transform(preprocessed_essays)
print("Shape of matrix after one hot encodig ",text_tfidf.shape)
```

1.5.2.3 Using Pretrained Models: Avg W2V

In [0]:

```
'''
# Reading glove vectors in python: https://stackoverflow.com/a/38230349/4084039
def loadGloveModel(gloveFile):
    print ("Loading Glove Model")
    f = open(gloveFile,'r', encoding="utf8")
    model = {}
    for line in tqdm(f):
        splitLine = line.split()
        word = splitLine[0]
        embedding = np.array([float(val) for val in splitLine[1:]])
        model[word] = embedding
    print ("Done.",len(model)," words loaded!")
    return model
model = loadGloveModel('glove.42B.300d.txt')

# ============================
Output:

Loading Glove Model
1917495it [06:32, 4879.69it/s]
Done. 1917495  words loaded!

# ============================

words = []
for i in preproced_texts:
    words.extend(i.split(' '))

for i in preproced_titles:
    words.extend(i.split(' '))
print("all the words in the coupus", len(words))
words = set(words)
print("the unique words in the coupus", len(words))

inter_words = set(model.keys()).intersection(words)
print("The number of words that are present in both glove vectors and our coupus", \
      len(inter_words),"(",np.round(len(inter_words)/len(words)*100,3),"%)")

words_courpus = {}
words_glove = set(model.keys())
for i in words:
    if i in words_glove:
        words_courpus[i] = model[i]
print("word 2 vec length", len(words_courpus))


# stronging variables into pickle files python: http://www.jessicayung.com/how-to-use-pickle-to-sa
ve-and-load-variables-in-python/

import pickle
with open('glove_vectors', 'wb') as f:
    pickle.dump(words_courpus, f)


'''
```

In [0]:

```
# stronging variables into pickle files python: http://www.jessicayung.com/how-to-use-pickle-to-sa
ve-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())
```

In [0]:

```
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(preprocessed_essays): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    cnt_words =0; # num of words with a valid vector in the sentence/review
    for word in sentence.split(): # for each word in a review/sentence
        if word in glove_words:
            vector += model[word]
            cnt_words += 1
    if cnt_words != 0:
        vector /= cnt_words
    avg_w2v_vectors.append(vector)

print(len(avg_w2v_vectors))
print(len(avg_w2v_vectors[0]))
```

1.5.2.3 Using Pretrained Models: TFIDF weighted W2V

In [0]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(preprocessed_essays)
# we are converting a dictionary with word as a key, and the idf as a value
dictionary = dict(zip(tfidf_model.get_feature_names(), list(tfidf_model.idf_)))
tfidf_words = set(tfidf_model.get_feature_names())
```

In [0]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(preprocessed_essays): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    tf_idf_weight =0; # num of words with a valid vector in the sentence/review
    for word in sentence.split(): # for each word in a review/sentence
        if (word in glove_words) and (word in tfidf_words):
            vec = model[word] # getting the vector for each word
            # here we are multiplying idf value(dictionary[word]) and the tf
value((sentence.count(word)/len(sentence.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tf
idf value for each word
            vector += (vec * tf_idf) # calculating tfidf weighted w2v
            tf_idf_weight += tf_idf
    if tf_idf_weight != 0:
        vector /= tf_idf_weight
    tfidf_w2v_vectors.append(vector)

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

In [0]:

```
# Similarly you can vectorize for title also
```

1.5.3 Vectorizing Numerical features

In [0]:

```
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 [0]:

```
# check this one: https://www.youtube.com/watch?v=0HOqOcln3Z4&t=530s
# standardization sklearn: https://scikit-
learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
from sklearn.preprocessing import StandardScaler

# price_standardized = standardScalar.fit(project_data['price'].values)
# this will rise the error
# ValueError: Expected 2D array, got 1D array instead: array=[725.05 213.03 329.   ... 399.   287.
73   5.5 ].
# Reshape your data either using array.reshape(-1, 1)

price_scalar = StandardScaler()
price_scalar.fit(project_data['price'].values.reshape(-1,1)) # finding the mean and standard
deviation of this data
print(f"Mean : {price_scalar.mean_[0]}, Standard deviation : {np.sqrt(price_scalar.var_[0])}")

# Now standardize the data with above maen and variance.
price_standardized = price_scalar.transform(project_data['price'].values.reshape(-1, 1))
```

In [0]:

```
price_standardized
```

1.5.4 Merging all the above features

- we need to merge all the numerical vectors i.e catogorical, text, numerical vectors

In [0]:

```
print(categories_one_hot.shape)
print(sub_categories_one_hot.shape)
print(text_bow.shape)
print(price_standardized.shape)
```

In [0]:

```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
# with the same hstack function we are concatinating a sparse matrix and a dense matirx :)
X = hstack((categories_one_hot, sub_categories_one_hot, text_bow, price_standardized))
X.shape
```

Assignment 3: Apply KNN

1. **[Task-1] Apply KNN(brute force version) on these feature sets**
   - Set 1: categorical, numerical features + project_title(BOW) + preprocessed_essay (BOW)
   - Set 2: categorical, numerical features + project_title(TFIDF)+ preprocessed_essay (TFIDF)
   - Set 3: categorical, numerical features + project_title(AVG W2V)+ preprocessed_essay (AVG W2V)
   - Set 4: categorical, numerical features + project_title(TFIDF W2V)+ preprocessed_essay (TFIDF W2V)

2. **Hyper paramter tuning to find best K**
   - Find the best hyper parameter which results in the maximum AUC value
   - Find the best hyper paramter using k-fold cross validation (or) simple cross validation data
   - Use gridsearch-cv or randomsearch-cv or 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, as shown in the figure
   - Once you find the best hyper parameter, you need to train your model-M using the best hyper-param. Now, find the AUC on test data and plot the ROC curve on both train and test using model-M.

- Along with plotting ROC curve, you need to print the confusion matrix with predicted and original labels of test data points

4. **[Task-2]**

   - Select top 2000 features from feature Set 2 using `SelectKBest` and then apply KNN on top of these features

     -
       ```
       from sklearn.datasets import load_digits
       from sklearn.feature_selection import SelectKBest, chi2
       X, y = load_digits(return_X_y=True)
       X.shape
       X_new = SelectKBest(chi2, k=20).fit_transform(X, y)
       X_new.shape
       ========
       output:
       (1797, 64)
       (1797, 20)
       ```

   - Repeat the steps 2 and 3 on the data matrix after feature selection

5. **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. K Nearest Neighbor

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 [21]:

```
#merging price into project data.
price_data = resource_data.groupby('id').agg({'price':'sum', 'quantity':'sum'}).reset_index()
project_data = pd.merge(project_data, price_data, on='id', how='left')

data = project_data[:50000]#taking only 75k points
data.shape
```

Out[21]:

```
(50000, 20)
```

Resampling the data

Resampling the data

In [22]:

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

Out[22]:

```
(50000, 19)
```

In [31]:

```
# train test split
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, stratify=y)
X_train, X_cv, y_train, y_cv = train_test_split(X_train, y_train, test_size=0.2, stratify=y_train)

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

```
(32000, 19) (32000,)
(8000, 19) (8000,)
(10000, 19) (10000,)
```

2.2 Make Data Model Ready: encoding numerical, categorical features

In [25]:

```
# 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
```

One hot encoding: Clean categories

In [32]:

```
from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer(vocabulary=list(sorted_cat_dict.keys()), lowercase=False, binary=True
)
vectorizer.fit(X_train['clean_categories'].values) # fit has to happen only on train data

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

print("After vectorizations")
print(X_train_cat_ohe.shape, y_train.shape)
print(X_cv_cat_ohe.shape, y_cv.shape)
print(X_test_cat_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(32000, 9) (32000,)
(8000, 9) (8000,)
(10000, 9) (10000,)
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
===================================================================================================
```

one hot encoding: school state

In [33]:

```
vectorizer = CountVectorizer(vocabulary=list(X_train['school_state'].unique()), lowercase=False,
binary=True)
vectorizer.fit(X_train['school_state'].values) # fit has to happen only on train data

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

print("After vectorizations")
print(X_train_ss_ohe.shape, y_train.shape)
print(X_cv_ss_ohe.shape, y_cv.shape)
print(X_test_ss_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(32000, 51) (32000,)
(8000, 51) (8000,)
(10000, 51) (10000,)
['SC', 'WI', 'PA', 'NY', 'IL', 'NC', 'CA', 'VA', 'NJ', 'DC', 'MI', 'OR', 'AZ', 'TX', 'CT', 'WA', 'M
O', 'IN', 'FL', 'GA', 'LA', 'DE', 'UT', 'MA', 'TN', 'AR', 'OK', 'WY', 'AK', 'WV', 'MN', 'NV', 'KY',
'ND', 'HI', 'OH', 'ID', 'NM', 'AL', 'MS', 'MD', 'KS', 'NE', 'IA', 'CO', 'ME', 'NH', 'MT', 'RI', 'SD
', 'VT']
===================================================================================================
```

One hot encoding: Subcatecories

In [34]:

```
vectorizer = CountVectorizer(vocabulary=list(sorted_sub_cat_dict.keys()), lowercase=False, binary=
True)
vectorizer.fit(X_train['clean_subcategories'].values) # fit has to happen only on train data

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

print("After vectorizations")
print(X_train_sub_cat_ohe.shape, y_train.shape)
print(X_cv_sub_cat_ohe.shape, y_cv.shape)
print(X_test_sub_cat_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(32000, 30) (32000,)
(8000, 30) (8000,)
(10000, 30) (10000,)
['Economics', 'CommunityService', 'FinancialLiteracy', 'ParentInvolvement', 'Extracurricular',
'Civics_Government', 'ForeignLanguages', 'NutritionEducation', 'Warmth', 'Care_Hunger',
'SocialSciences', 'PerformingArts', 'CharacterEducation', 'TeamSports', 'Other',
'College_CareerPrep', 'Music', 'History_Geography', 'Health_LifeScience', 'EarlyDevelopment', 'ESL
', 'Gym_Fitness', 'EnvironmentalScience', 'VisualArts', 'Health_Wellness', 'AppliedSciences',
'SpecialNeeds', 'Literature_Writing', 'Mathematics', 'Literacy']
===================================================================================================
```

In [35]:

```
# similarly you can preprocess the titles also
```

```
preprocessed_pgc_train = []
# tqdm is for printing the status bar
for sentance in X_train['project_grade_category']:
    sent = decontracted(sentance)
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/55428
    preprocessed_pgc_train.append(sent.lower().strip())

preprocessed_pgc_cv = []
# tqdm is for printing the status bar
for sentance in X_cv['project_grade_category']:
    sent = decontracted(sentance)
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e.lower() not in stopwords)
    preprocessed_pgc_cv.append(sent.lower().strip())

from tqdm import tqdm_notebook as tqdm
preprocessed_pgc_test = []
for sentance in X_test['project_grade_category']:
    sent = decontracted(sentance)
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e.lower() not in stopwords)
    preprocessed_pgc_test.append(sent.lower().strip())
```

One hot encoding: Project grade category

In [36]:

```
vectorizer = CountVectorizer(vocabulary = np.unique(preprocessed_pgc_train),lowercase=False,binary
=True)

vectorizer.fit(preprocessed_pgc_train) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_pgc_ohe = vectorizer.transform(preprocessed_pgc_train)
X_cv_pgc_ohe = vectorizer.transform(preprocessed_pgc_cv)
X_test_pgc_ohe = vectorizer.transform(preprocessed_pgc_test)

print("After vectorizations")
print(X_train_pgc_ohe.shape, y_train.shape)
print(X_cv_pgc_ohe.shape, y_cv.shape)
print(X_test_pgc_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(32000, 4) (32000,)
(8000, 4) (8000,)
(10000, 4) (10000,)
['grades 3 5', 'grades 6 8', 'grades 9 12', 'grades prek 2']
====================================================================================================
```

One hot encoding: Teacher prefix

In [37]:

```
#https://stackoverflow.com/questions/14162723/replacing-pandas-or-numpy-nan-with-a-none-to-use-wit
h-mysqldb
X_train['teacher_prefix'] = project_data['teacher_prefix'].replace(np.nan,'empty',regex = True)
X_test['teacher_prefix'] = project_data['teacher_prefix'].replace(np.nan,'empty',regex = True)
X_cv['teacher_prefix'] = project_data['teacher_prefix'].replace(np.nan,'empty',regex = True)




vectorizer = CountVectorizer(vocabulary=list(X_train['teacher_prefix'].unique()),lowercase=False,b
inary=True)
vectorizer.fit(X_train['teacher_prefix'].values) # fit has to happen only on train data
```

```
# we use the fitted CountVectorizer to convert the text to vector
X_train_tp_ohe = vectorizer.transform(X_train['teacher_prefix'].values)
X_cv_tp_ohe = vectorizer.transform(X_cv['teacher_prefix'].values)
X_test_tp_ohe = vectorizer.transform(X_test['teacher_prefix'].values)

print("After vectorizations")
print(X_train_tp_ohe.shape, y_train.shape)
print(X_cv_tp_ohe.shape, y_cv.shape)
print(X_test_tp_ohe.shape, y_test.shape)
print(vectorizer.get_feature_names())
print("="*100)
```

```
After vectorizations
(32000, 6) (32000,)
(8000, 6) (8000,)
(10000, 6) (10000,)
['Ms.', 'Mrs.', 'Mr.', 'Teacher', 'Dr.', 'empty']
====================================================================================================
```

Vectorizing numerical data: Price

In [38]:

```
# check this one: https://www.youtube.com/watch?v=0HOqOcln3Z4&t=530s
# standardization sklearn: https://scikit-
learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
from sklearn.preprocessing import StandardScaler

# price_standardized = standardScalar.fit(project_data['price'].values)
# this will rise the error
# ValueError: Expected 2D array, got 1D array instead: array=[725.05 213.03 329.   ... 399.   287.
73   5.5 ].
# Reshape your data either using array.reshape(-1, 1)

price_scalar = StandardScaler()
price_scalar.fit(X_train['price'].values.reshape(-1,1)) # finding the mean and standard deviation
of this data
print(f"Mean : {price_scalar.mean_[0]}, Standard deviation : {np.sqrt(price_scalar.var_[0])}")

# Now standardize the data with above mean and variance.
X_train_price_standardized = price_scalar.transform(X_train['price'].values.reshape(-1, 1))
X_cv_price_standardized = price_scalar.transform(X_cv['price'].values.reshape(-1, 1))
X_test_price_standardized = price_scalar.transform(X_test['price'].values.reshape(-1, 1))
```

```
Mean : 314.767265625, Standard deviation : 375.20117925611515
```

Vectorizing numerical data: teacher_number_of_previously_posted_projects

In [39]:

```
tppp_scalar = StandardScaler()
tppp_scalar.fit(X_train['teacher_number_of_previously_posted_projects'].values.reshape(-1,1)) # fin
ding the mean and standard deviation of this data
print(f"Mean : {tppp_scalar.mean_[0]}, Standard deviation : {np.sqrt(price_scalar.var_[0])}")

# Now standardize the data with above maen and variance.
X_train_tppp_standardized =
tppp_scalar.transform(X_train['teacher_number_of_previously_posted_projects'].values.reshape(-1, 1)
)
X_cv_tppp_standardized = tppp_scalar.transform(X_cv['teacher_number_of_previously_posted_projects'
].values.reshape(-1, 1))
X_test_tppp_standardized =
tppp_scalar.transform(X_test['teacher_number_of_previously_posted_projects'].values.reshape(-1, 1))
```

```
Mean : 9.38603125, Standard deviation : 375.20117925611515
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:595: DataConversionWarning:

Data with input dtype int64 was converted to float64 by StandardScaler.
Data with input dtype int64 was converted to float64 by StandardScaler.

/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:595: DataConversionWarning:

Data with input dtype int64 was converted to float64 by StandardScaler.

/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:595: DataConversionWarning:

Data with input dtype int64 was converted to float64 by StandardScaler.

/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:595: DataConversionWarning:

Data with input dtype int64 was converted to float64 by StandardScaler.
```

2.3 Make Data Model Ready: encoding eassay, and project_title

In [40]:

```
# 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 [41]:

```
#https://pythonprogramming.net/lemmatizing-nltk-tutorial/
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
nltk.download("wordnet")
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/varadamurthiacharya/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[41]:

```
True
```

Preprocessing train, test and cv data seperately

In [42]:

```
# similarly you can preprocess the titles also
ppt_train = []
# tqdm is for printing the status bar
for sentance in X_train['project_title'].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(lemmatizer.lemmatize(e) for e in sent.split())
    sent = ' '.join(e for e in sent.split() if e.lower() not in stopwords)
    ppt_train.append(sent.lower().strip())

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

ppt_test = []
# tqdm is for printing the status bar
for sentance in X_test['project_title'].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(lemmatizer.lemmatize(e) for e in sent.split())
    sent = ' '.join(e for e in sent.split() if e.lower() not in stopwords)
    ppt_test.append(sent.lower().strip())
```

In [43]:

```
# similarly you can preprocess the titles also
ppe_train = []
# tqdm is for printing the status bar
for sentance in X_train['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(lemmatizer.lemmatize(e) for e in sent.split())
    sent = ' '.join(e for e in sent.split() if e.lower() not in stopwords)
    ppe_train.append(sent.lower().strip())

ppe_cv = []
# tqdm is for printing the status bar
for sentance in X_cv['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(lemmatizer.lemmatize(e) for e in sent.split())
    sent = ' '.join(e for e in sent.split() if e.lower() not in stopwords)
    ppe_cv.append(sent.lower().strip())

ppe_test = []
# tqdm is for printing the status bar
for sentance in X_test['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(lemmatizer.lemmatize(e) for e in sent.split())
    sent = ' '.join(e for e in sent.split() if e.lower() not in stopwords)
    ppe_test.append(sent.lower().strip())
```

BAG OF WORDS: Project essay

In [44]:

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

# we use the fitted CountVectorizer to convert the text to vector
X_train_essay_bow = vectorizer.transform(ppe_train)
X_cv_essay_bow = vectorizer.transform(ppe_cv)
X_test_essay_bow = vectorizer.transform(ppe_test)

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

```
After vectorizations
(32000, 8874) (32000,)
(8000, 8874) (8000,)
(10000, 8874) (10000,)
====================================================================================================
```

BAG OF WORDS: Project title

In [45]:

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

# we use the fitted CountVectorizer to convert the text to vector
X_train_title_bow = vectorizer.transform(ppt_train)
X_cv_title_bow = vectorizer.transform(ppt_cv)
X_test_title_bow = vectorizer.transform(ppt_test)

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

```
After vectorizations
(32000, 1464) (32000,)
(8000, 1464) (8000,)
(10000, 1464) (10000,)
====================================================================================================
```

2.4 Appling KNN on different kind of featurization as mentioned in the instructions

Apply KNN 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 instructions

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
```

2.4.1 Applying KNN brute force on BOW, SET 1

In [0]:

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

In [176]:

```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
from scipy.sparse import hstack
X_tr1 = hstack((X_train_cat_ohe,X_train_ss_ohe,X_train_sub_cat_ohe,X_train_pgc_ohe,X_train_tp_ohe,X
_train_price_standardized,X_train_tppp_standardized,X_train_essay_bow,X_train_title_bow)).tocsr()
X_cv1 =
hstack((X_cv_cat_ohe,X_cv_ss_ohe,X_cv_sub_cat_ohe,X_cv_pgc_ohe,X_cv_tp_ohe,X_cv_price_standardized
,X_cv_tppp_standardized,X_cv_essay_bow,X_cv_title_bow)).tocsr()
X_te1 = hstack((X_test_cat_ohe,X_test_ss_ohe,X_test_sub_cat_ohe,X_test_pgc_ohe,X_test_tp_ohe,X_test
_price_standardized,X_test_tppp_standardized,X_test_essay_bow,X_test_title_bow)).tocsr()

print("Final Data matrix")
print(X_tr1.shape, y_train.shape)
print(X_cv1.shape, y_cv.shape)
print(X_te1.shape, y_test.shape)
print("="*100)
```

```
Final Data matrix
(32000, 10440) (32000,)
(8000, 10440) (8000,)
(10000, 10440) (10000,)
====================================================================================================
```

In [177]:

```
#https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html
from sklearn.decomposition import TruncatedSVD
from sklearn.random_projection import sparse_random_matrix
svd = TruncatedSVD(n_components=1000, n_iter=7, random_state=42)
X_tr1 = svd.fit_transform(X_tr1)
X_cv1=svd.transform(X_cv1)
X_te1=svd.transform(X_te1)

print(X_tr1.shape, y_train.shape)
print(X_cv1.shape, y_cv.shape)
print(X_te1.shape, y_test.shape)
print("="*100)
```

```
(32000, 1000) (32000,)
(8000, 1000) (8000,)
(10000, 1000) (10000,)
====================================================================================================
```

In [178]:

```
#https://imbalanced-
learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.RandomOverSampler.html
from imblearn.over_sampling import RandomOverSampler
ros = RandomOverSampler(random_state=0,sampling_strategy=0.75)
X_tr1_res, y_train_res = ros.fit_resample(X_tr1, y_train)
X_cv1_res,y_cv_res = ros.fit_resample(X_cv1, y_cv)
X_te1_res,y_test_res = ros.fit_resample(X_te1,y_test)

print("Final Data matrix")
print(X_tr1_res.shape, y_train_res.shape)
print(X_cv1_res.shape, y_cv_res.shape)
print(X_te1_res.shape, y_test_res.shape)
print("="*100)
```

```
Final Data matrix
(47031, 1000) (47031,)
(11758, 1000) (11758,)
(14698, 1000) (14698,)
====================================================================================================
```

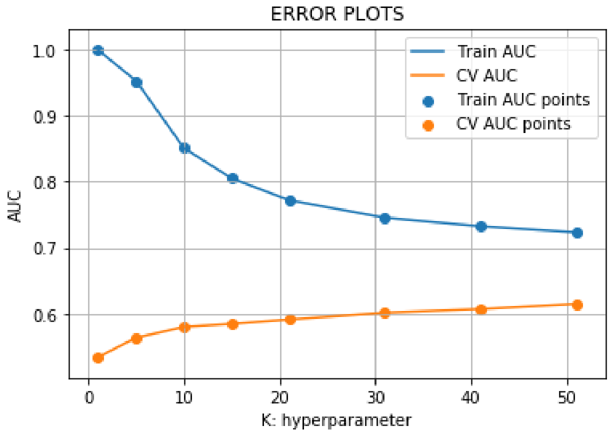
In [179]:

```
#the below set of code is taken from the sample solution ipython notebook

def batch_predict(clf, data):
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the posi
tive class
```

```
    # not the predicted outputs

    y_data_pred = []
    tr_loop = data.shape[0] - data.shape[0]%1000
    # consider you X_tr shape is 49041, then your cr_loop will be 49041 - 49041%1000 = 49000
    # in this for loop we will iterate unti the last 1000 multiplier
    for i in range(0, tr_loop, 1000):
        y_data_pred.extend(clf.predict_proba(data[i:i+1000])[:,1])
    # we will be predicting for the last data points
    y_data_pred.extend(clf.predict_proba(data[tr_loop:])[:,1])
    return y_data_pred

import matplotlib.pyplot as plt
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import roc_auc_score
"""
y_true : array, shape = [n_samples] or [n_samples, n_classes]
True binary labels or binary label indicators.

y_score : array, shape = [n_samples] or [n_samples, n_classes]
Target scores, can either be probability estimates of the positive class, confidence values, or no
n-thresholded measure of
decisions (as returned by “decision_function” on some classifiers).
For binary y_true, y_score is supposed to be the score of the class with greater label.

"""

train_auc = []
cv_auc = []
K = [1, 5, 10, 15, 21, 31, 41, 51]
for i in tqdm(K):
    neigh = KNeighborsClassifier(n_neighbors=i,n_jobs=-1)
    neigh.fit(X_tr1_res, y_train_res)

    y_train_pred = batch_predict(neigh, X_tr1_res)
    y_cv_pred = batch_predict(neigh, X_cv1_res)

    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the posi
tive class
    # not the predicted outputs
    train_auc.append(roc_auc_score(y_train_res,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv_res, y_cv_pred))
```

In [180]:

```
plt.plot(K, train_auc, label='Train AUC')
plt.plot(K, cv_auc, label='CV AUC')

plt.scatter(K, train_auc, label='Train AUC points')
plt.scatter(K, cv_auc, label='CV AUC points')

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

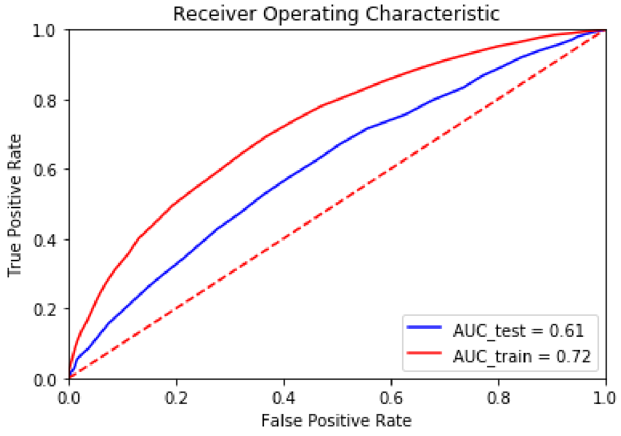


In [181]:

```
# from the above graph, taking value of K=51
#https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
from sklearn.neighbors import KNeighborsClassifier
neigh = KNeighborsClassifier(n_neighbors=51,algorithm='brute',n_jobs=-1)
neigh.fit(X_tr1_res, y_train_res)


pred = neigh.predict(X_te1_res)
pred1 = neigh.predict(X_tr1_res)
#https://stackoverflow.com/questions/25009284/how-to-plot-roc-curve-in-python
import sklearn.metrics as metrics
# calculate the fpr and tpr for all thresholds of the classification
probs = neigh.predict_proba(X_te1_res)
probs1 = neigh.predict_proba(X_tr1_res)
preds = probs[:,1]
preds1 = probs1[:,1]
fpr, tpr, threshold = metrics.roc_curve(y_test_res, preds)
fpr1, tpr1, threshold = metrics.roc_curve(y_train_res, preds1)
roc_auc = metrics.auc(fpr, tpr)
roc_auc1 = metrics.auc(fpr1, tpr1)
```

In [182]:

```
# method I: plt
import matplotlib.pyplot as plt
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b', label = 'AUC_test = %0.2f' % roc_auc)
plt.plot(fpr1, tpr1, 'r', label = 'AUC_train = %0.2f' % roc_auc1)
plt.legend(loc = 'lower right')
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
plt.ylim([0, 1])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```



In [183]:

```
#the below set of code is taken from the sample solution ipython notebook

# we are writing our own function for predict, with defined thresould
# we will pick a threshold that will give the least fpr
def predict(proba, 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))
    predictions = []
    for i in proba:
        if i>=t:
            predictions.append(1)
        else:
            predictions.append(0)
    return predictions
```

In [184]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
cm_train = confusion_matrix(y_train_res, predict(y_train_pred, threshold,tpr, fpr))
print(cm_train)

print("Test confusion matrix")
cm_test= confusion_matrix(y_test_res, predict(pred, threshold, tpr1,fpr1))
print(cm_test)
```

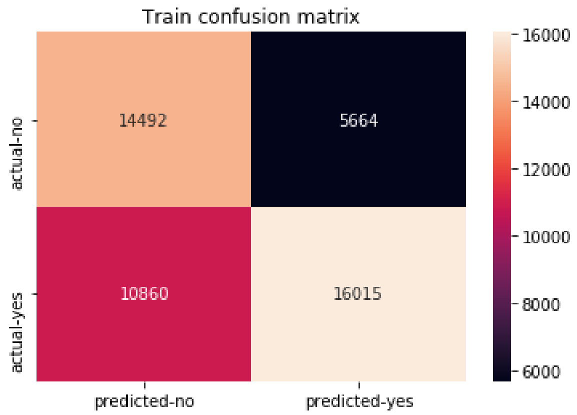
```
====================================================================================================

Train confusion matrix
the maximum value of tpr*(1-fpr) 0.1741599390957061 for threshold 0.529
[[14492  5664]
 [10860 16015]]
Test confusion matrix
the maximum value of tpr*(1-fpr) 0.11470513990871203 for threshold 0.51
[[3400 2899]
 [3164 5235]]
```

In [185]:

```
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
ylabel = ["actual-no","actual-yes"]
xlabel = ["predicted-no","predicted-yes"]
plt.title("Train confusion matrix")
sns.heatmap(cm_train, annot = True,yticklabels=ylabel, xticklabels=xlabel,fmt="d")
```

Out[185]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a448ba588>
```

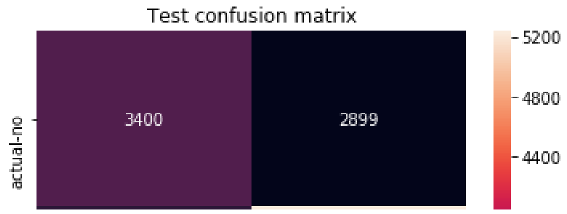


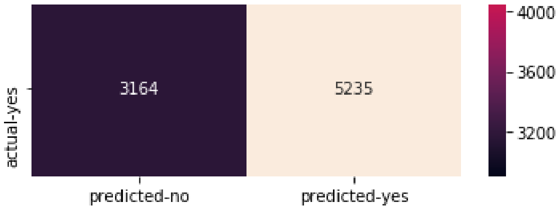
In [186]:

```
plt.title("Test confusion matrix")
sns.heatmap(cm_test, annot = True,yticklabels=ylabel, xticklabels=xlabel,fmt="d")
```

Out[186]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a4473f470>
```





2.4.2 Applying KNN brute force on TFIDF, SET 2

In [46]:

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

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

In [47]:

```
X_train_essay_tfidf = vectorizer.transform(ppe_train)
X_cv_essay_tfidf = vectorizer.transform(ppe_cv)
X_test_essay_tfidf = vectorizer.transform(ppe_test)

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

```
After vectorizations
(32000, 8874) (32000,)
(8000, 8874) (8000,)
(10000, 8874) (10000,)
====================================================================================================
```

In [48]:

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

In [49]:

```
X_train_title_tfidf = vectorizer.transform(ppt_train)
X_cv_title_tfidf = vectorizer.transform(ppt_cv)
X_test_title_tfidf = vectorizer.transform(ppt_test)

print("After vectorizations")
print(X_train_title_tfidf.shape, y_train.shape)
print(X_cv_title_tfidf.shape, y_cv.shape)
print(X_test_title_tfidf.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(32000, 1464) (32000,)
(8000, 1464) (8000,)
(10000, 1464) (10000,)
====================================================================================================
```

In [130]:

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

# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
X_tr2 = hstack((X_train_cat_ohe,X_train_ss_ohe,X_train_sub_cat_ohe,X_train_pgc_ohe,X_train_tp_ohe,X
_train_price_standardized,X_train_tppp_standardized,X_train_essay_tfidf,X_train_title_tfidf)).tocs
r()
X_cv2 =
hstack((X_cv_cat_ohe,X_cv_ss_ohe,X_cv_sub_cat_ohe,X_cv_pgc_ohe,X_cv_tp_ohe,X_cv_price_standardized
,X_cv_tppp_standardized,X_cv_essay_tfidf,X_cv_title_tfidf)).tocsr()
X_te2 = hstack((X_test_cat_ohe,X_test_ss_ohe,X_test_sub_cat_ohe,X_test_pgc_ohe,X_test_tp_ohe,X_test
_price_standardized,X_test_tppp_standardized,X_test_essay_tfidf,X_test_title_tfidf)).tocsr()

print("Final Data matrix")
print(X_tr2.shape, y_train.shape)
print(X_cv2.shape, y_cv.shape)
print(X_te2.shape, y_test.shape)
print("="*100)
```

```
Final Data matrix
(32000, 10440) (32000,)
(8000, 10440) (8000,)
(10000, 10440) (10000,)
====================================================================================================
```

In [131]:

```
#https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html
from sklearn.decomposition import TruncatedSVD
from sklearn.random_projection import sparse_random_matrix
svd = TruncatedSVD(n_components=1000, n_iter=7, random_state=42)
X_tr2 = svd.fit_transform(X_tr2)
X_cv2=svd.transform(X_cv2)
X_te2=svd.transform(X_te2)

print(X_tr2.shape, y_train.shape)
print(X_cv2.shape, y_cv.shape)
print(X_te2.shape, y_test.shape)
print("="*100)
```

```
(32000, 1000) (32000,)
(8000, 1000) (8000,)
(10000, 1000) (10000,)
====================================================================================================
```

In [132]:

```
#https://imbalanced-
learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.RandomOverSampler.html
from imblearn.over_sampling import RandomOverSampler
ros = RandomOverSampler(random_state=0,sampling_strategy='minority')
X_tr2_res, y_train_res = ros.fit_resample(X_tr2, y_train)
X_cv2_res,y_cv_res = ros.fit_resample(X_cv2, y_cv)
X_te2_res,y_test_res = ros.fit_resample(X_te2,y_test)
```

In [133]:

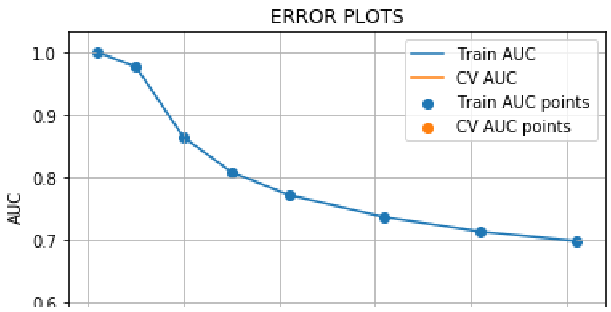
```
print("Final Data matrix")
print(X_tr2_res.shape, y_train_res.shape)
print(X_cv2_res.shape, y_cv_res.shape)
print(X_te2_res.shape, y_test_res.shape)
print("="*100)
```

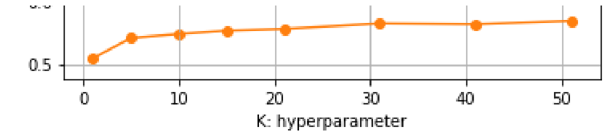
```
Final Data matrix
(53750, 1000) (53750,)
(13438, 1000) (13438,)
(16798, 1000) (16798,)
====================================================================================================
```

In [57]:

```
#the below set of code is taken from the sample solution ipython notebook
```

```
def batch_predict(clf, data):
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the posi
tive class
    # not the predicted outputs

    y_data_pred = []
    tr_loop = data.shape[0] - data.shape[0]%1000
    # consider you X_tr shape is 49041, then your cr_loop will be 49041 - 49041%1000 = 49000
    # in this for loop we will iterate unti the last 1000 multiplier
    for i in range(0, tr_loop, 1000):
        y_data_pred.extend(clf.predict_proba(data[i:i+1000])[:,1])
    # we will be predicting for the last data points
    y_data_pred.extend(clf.predict_proba(data[tr_loop:])[:,1])

    return y_data_pred


import matplotlib.pyplot as plt
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import roc_auc_score
"""
y_true : array, shape = [n_samples] or [n_samples, n_classes]
True binary labels or binary label indicators.

y_score : array, shape = [n_samples] or [n_samples, n_classes]
Target scores, can either be probability estimates of the positive class, confidence values, or no
n-thresholded measure of
decisions (as returned by "decision_function" on some classifiers).
For binary y_true, y_score is supposed to be the score of the class with greater label.

"""

train_auc = []
cv_auc = []

K = [1, 5, 10, 15, 21, 31, 41, 51]
for i in tqdm(K):
    neigh = KNeighborsClassifier(n_neighbors=i,n_jobs=-1)
    neigh.fit(X_tr2_res, y_train_res)

    y_train_pred = batch_predict(neigh, X_tr2_res)
    y_cv_pred = batch_predict(neigh, X_cv2_res)

    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the posi
tive class
    # not the predicted outputs
    train_auc.append(roc_auc_score(y_train_res,y_train_pred))

    cv_auc.append(roc_auc_score(y_cv_res, y_cv_pred))


plt.plot(K, train_auc, label='Train AUC')
plt.plot(K, cv_auc, label='CV AUC')

plt.scatter(K, train_auc, label='Train AUC points')
plt.scatter(K, cv_auc, label='CV AUC points')

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



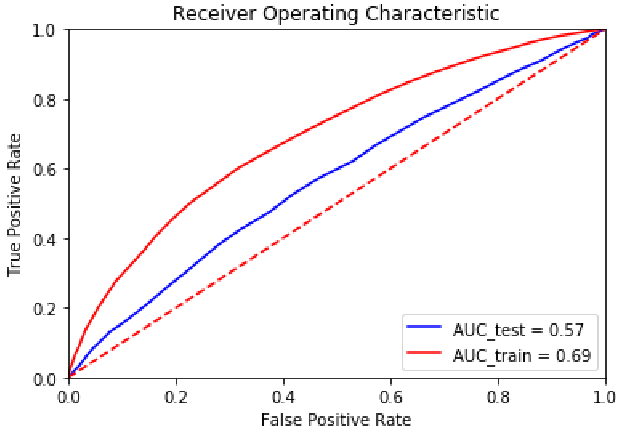


In [134]:

```
# from the above graph, taking value of K=51
#https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
from sklearn.neighbors import KNeighborsClassifier
neigh = KNeighborsClassifier(n_neighbors=51,algorithm='brute')
neigh.fit(X_tr2_res, y_train_res)

pred = neigh.predict(X_te2_res)

#https://stackoverflow.com/questions/25009284/how-to-plot-roc-curve-in-python
import sklearn.metrics as metrics
# calculate the fpr and tpr for all thresholds of the classification
probs = neigh.predict_proba(X_te2_res)
probs1 = neigh.predict_proba(X_tr2_res)
preds = probs[:,1]
preds1 = probs1[:,1]
fpr, tpr, threshold = metrics.roc_curve(y_test_res, preds)
fpr1, tpr1, threshold = metrics.roc_curve(y_train_res, preds1)
roc_auc = metrics.auc(fpr, tpr)
roc_auc1 = metrics.auc(fpr1, tpr1)
```

In [135]:

```
# method I: plt
import matplotlib.pyplot as plt
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b', label = 'AUC_test = %0.2f' % roc_auc)
plt.plot(fpr1, tpr1, 'r', label = 'AUC_train = %0.2f' % roc_auc1)
plt.legend(loc = 'lower right')
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
plt.ylim([0, 1])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```



In [136]:

```
#the below set of code is taken from the sample solution ipython notebook

# we are writing our own function for predict, with defined thresould
# we will pick a threshold that will give the least fpr
def predict(proba, 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))
    predictions = []
    for i in proba:
        if i>=t:
            predictions.append(1)
        else:
            predictions.append(0)
    return predictions
```

In [137]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
cm_train = confusion_matrix(y_train_res, predict(y_train_pred, threshold,tpr,fpr))
print(cm_train)

print("Test confusion matrix")
cm_test= confusion_matrix(y_test_res, predict(pred, threshold, tpr1,fpr1))
print(cm_test)
```

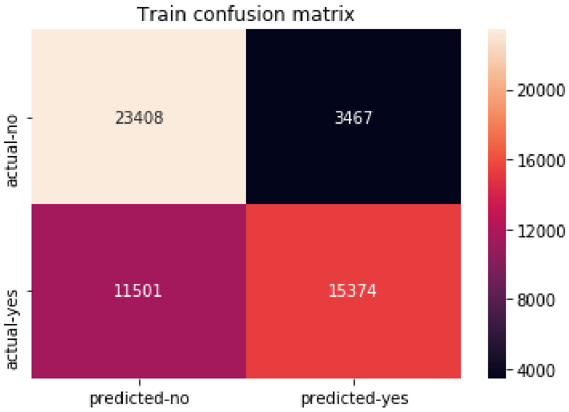
```
====================================================================================================

Train confusion matrix
the maximum value of tpr*(1-fpr) 0.20081616370087885 for threshold 0.49
[[23408  3467]
 [11501 15374]]
Test confusion matrix
the maximum value of tpr*(1-fpr) 0.13052599450513794 for threshold 0.471
[[4467 3932]
 [3580 4819]]
```

In [138]:

```
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
ylabel = ["actual-no","actual-yes"]
xlabel = ["predicted-no","predicted-yes"]
plt.title("Train confusion matrix")
sns.heatmap(cm_train, annot = True,yticklabels=ylabel, xticklabels=xlabel,fmt="d")
```

Out[138]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a448916a0>
```

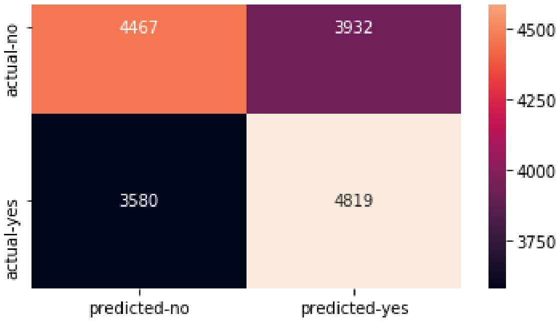


In [139]:

```
plt.title("Test confusion matrix")
sns.heatmap(cm_test, annot = True,yticklabels=ylabel, xticklabels=xlabel,fmt="d")
```

Out[139]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a448bacc0>
```



2.4.3 Applying KNN brute force on AVG W2V, SET 3

In [64]:

```
# stronging variables into pickle files python: http://www.jessicayung.com/how-to-use-pickle-to-sa
ve-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())
```

In [65]:

```
#Splitting into train test and CV dtata
pe_train, pe_test = train_test_split(preprocessed_essays,test_size=0.33)
pe_train, pe_cv = train_test_split(pe_train,test_size=0.33)
```

In [66]:

```
# average Word2Vec
# compute average word2vec for each review.
from tqdm import tqdm_notebook as tqdm
awv_pe_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(ppe_train): # 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
    awv_pe_train.append(vector)

print(len(awv_pe_train))
print(len(awv_pe_train[0]))

# average Word2Vec
# compute average word2vec for each review.
awv_pe_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(ppe_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
    awv_pe_test.append(vector)

print(len(awv_pe_test))
print(len(awv_pe_test[0]))


# average Word2Vec
# compute average word2vec for each review.
awv_pe_cv = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(ppe_cv): # 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
    awv_pe_cv.append(vector)

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

```
32000
300


10000
300


8000
300
```

In [67]:

```
# average Word2Vec
# compute average word2vec for each review.
awv_pt_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(ppt_train): # 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
    awv_pt_train.append(vector)

print(len(awv_pt_train))
print(len(awv_pt_train[0]))

# average Word2Vec
# compute average word2vec for each review.
awv_pt_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(ppt_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
    awv_pt_test.append(vector)

print(len(awv_pt_test))
print(len(awv_pt_test[0]))

# average Word2Vec
# compute average word2vec for each review.
awv_pt_cv = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(ppt_cv): # 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
    awv_pt_cv.append(vector)

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

```
32000
300


10000
300


8000
300
```

In [68]:

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

# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
X_tr3 = hstack((X_train_cat_ohe,X_train_ss_ohe,X_train_sub_cat_ohe,X_train_pgc_ohe,X_train_tp_ohe,X
_train_price_standardized,X_train_tppp_standardized,awv_pe_train,awv_pt_train)).tocsr()
X_cv3 =
hstack((X_cv_cat_ohe,X_cv_ss_ohe,X_cv_sub_cat_ohe,X_cv_pgc_ohe,X_cv_tp_ohe,X_cv_price_standardized
,X_cv_tppp_standardized,awv_pe_cv,awv_pt_cv)).tocsr()
X_te3 = hstack((X_test_cat_ohe,X_test_ss_ohe,X_test_sub_cat_ohe,X_test_pgc_ohe,X_test_tp_ohe,X_test
_price_standardized,X_test_tppp_standardized,awv_pe_test,awv_pt_test)).tocsr()

print("Final Data matrix")
print(X_tr3.shape, y_train.shape)
print(X_cv3.shape, y_cv.shape)
print(X_te3.shape, y_test.shape)
print("="*100)
```

```
Final Data matrix
(32000, 702) (32000,)
(8000, 702) (8000,)
(10000, 702) (10000,)
====================================================================================================
```

In [69]:

```
from sklearn.decomposition import TruncatedSVD
from sklearn.random_projection import sparse_random_matrix
svd = TruncatedSVD(n_components=100, n_iter=7, random_state=42)
X_tr3 = svd.fit_transform(X_tr3)
X_cv3=svd.transform(X_cv3)
X_te3=svd.transform(X_te3)

print(X_tr3.shape, y_train.shape)
print(X_cv3.shape, y_cv.shape)
print(X_te3.shape, y_test.shape)
print("="*100)
```

```
(32000, 100) (32000,)
(8000, 100) (8000,)
(10000, 100) (10000,)
====================================================================================================
```

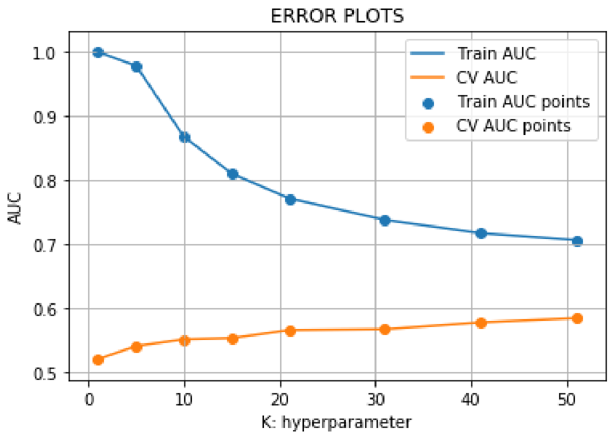
In [70]:

```
#https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html
from imblearn.over_sampling import RandomOverSampler
ros = RandomOverSampler(random_state=0,sampling_strategy='minority')
X_tr3_res, y_train_res = ros.fit_resample(X_tr3, y_train)
X_cv3_res,y_cv_res = ros.fit_resample(X_cv3, y_cv)
X_te3_res,y_test_res = ros.fit_resample(X_te3,y_test)

print("Final Data matrix")
print(X_tr3_res.shape, y_train_res.shape)
print(X_cv3_res.shape, y_cv_res.shape)
print(X_te3_res.shape, y_test_res.shape)
print("="*100)
```

```
Final Data matrix
(53750, 100) (53750,)
(13438, 100) (13438,)
(16798, 100) (16798,)
====================================================================================================
```

In [71]:

```
#the below set of code is taken from the sample solution ipython notebook

def batch_predict(clf, data):
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the posi
tive class
    # not the predicted outputs

    y_data_pred = []
    tr_loop = data.shape[0] - data.shape[0]%1000
    # consider you X_tr shape is 49041, then your cr_loop will be 49041 - 49041%1000 = 49000
    # in this for loop we will iterate unti the last 1000 multiplier
    for i in range(0, tr_loop, 1000):
        y_data_pred.extend(clf.predict_proba(data[i:i+1000])[:,1])
    # we will be predicting for the last data points
    y_data_pred.extend(clf.predict_proba(data[tr_loop:])[:,1])

    return y_data_pred


import matplotlib.pyplot as plt
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import roc_auc_score
"""
y_true : array, shape = [n_samples] or [n_samples, n_classes]
True binary labels or binary label indicators.

y_score : array, shape = [n_samples] or [n_samples, n_classes]
Target scores, can either be probability estimates of the positive class, confidence values, or no
n-thresholded measure of
decisions (as returned by "decision_function" on some classifiers).
For binary y_true, y_score is supposed to be the score of the class with greater label.

"""

train_auc = []
cv_auc = []
K = [1, 5, 10, 15, 21, 31, 41, 51]
for i in tqdm(K):
    neigh = KNeighborsClassifier(n_neighbors=i,n_jobs=-1)
    neigh.fit(X_tr3_res, y_train_res)

    y_train_pred = batch_predict(neigh, X_tr3_res)
    y_cv_pred = batch_predict(neigh, X_cv3_res)

    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the posi
tive class
    # not the predicted outputs
    train_auc.append(roc_auc_score(y_train_res,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv_res, y_cv_pred))

plt.plot(K, train_auc, label='Train AUC')
plt.plot(K, cv_auc, label='CV AUC')

plt.scatter(K, train_auc, label='Train AUC points')
plt.scatter(K, cv_auc, label='CV AUC points')

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

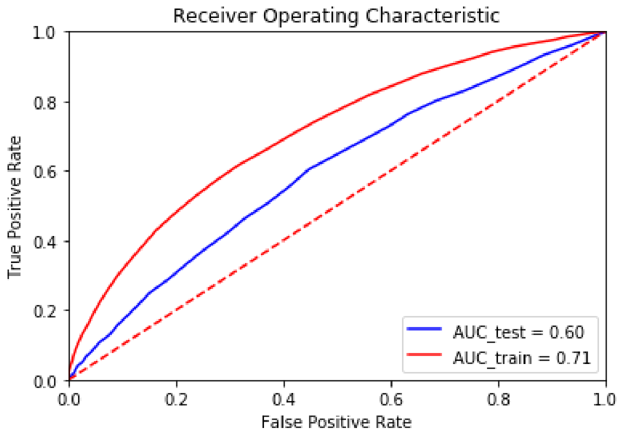


In [73]:

```
# from the above graph, taking value of K=51
#https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
from sklearn.neighbors import KNeighborsClassifier
neigh = KNeighborsClassifier(n_neighbors=51,algorithm='brute',n_jobs=-1)
neigh.fit(X_tr3_res, y_train_res)

pred = neigh.predict(X_te3_res)
pred1 = neigh.predict(X_tr3_res)

#https://stackoverflow.com/questions/25009284/how-to-plot-roc-curve-in-python
import sklearn.metrics as metrics
# calculate the fpr and tpr for all thresholds of the classification
probs = neigh.predict_proba(X_te3_res)
probs1 = neigh.predict_proba(X_tr3_res)
preds = probs[:,1]
preds1 = probs1[:,1]
fpr, tpr, threshold = metrics.roc_curve(y_test_res, preds)
fpr1, tpr1, threshold = metrics.roc_curve(y_train_res, preds1)
roc_auc = metrics.auc(fpr, tpr)
roc_auc1 = metrics.auc(fpr1, tpr1)
```

In [74]:

```
# method I: plt
import matplotlib.pyplot as plt
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b', label = 'AUC_test = %0.2f' % roc_auc)
plt.plot(fpr1, tpr1, 'r', label = 'AUC_train = %0.2f' % roc_auc1)
plt.legend(loc = 'lower right')
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
plt.ylim([0, 1])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```



In [75]:

```
#the below set of code is taken from the sample solution ipython notebook

# we are writing our own function for predict, with defined thresould
# we will pick a threshold that will give the least fpr
def predict(proba, 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))
    predictions = []
    for i in proba:
        if i>=t:
            predictions.append(1)
        else:
            predictions.append(0)
    return predictions
```

In [76]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
cm_train = confusion_matrix(y_train_res, predict(y_train_pred, threshold,tpr,fpr))
print(cm_train)

print("Test confusion matrix")
cm_test= confusion_matrix(y_test_res, predict(pred, threshold, tpr1,fpr1))
print(cm_test)
```

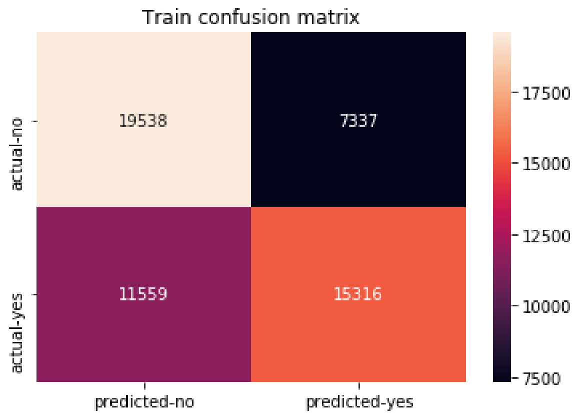
```
====================================================================================================

Train confusion matrix
the maximum value of tpr*(1-fpr) 0.18316614240400006 for threshold 0.529
[[19538  7337]
 [11559 15316]]
Test confusion matrix
the maximum value of tpr*(1-fpr) 0.12497640324499733 for threshold 0.49
[[4643 3756]
 [3329 5070]]
```

In [77]:

```
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
ylabel = ["actual-no","actual-yes"]
xlabel = ["predicted-no","predicted-yes"]
plt.title("Train confusion matrix")
sns.heatmap(cm_train, annot = True,yticklabels=ylabel, xticklabels=xlabel,fmt="d")
```

Out[77]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a3eff6b00>
```

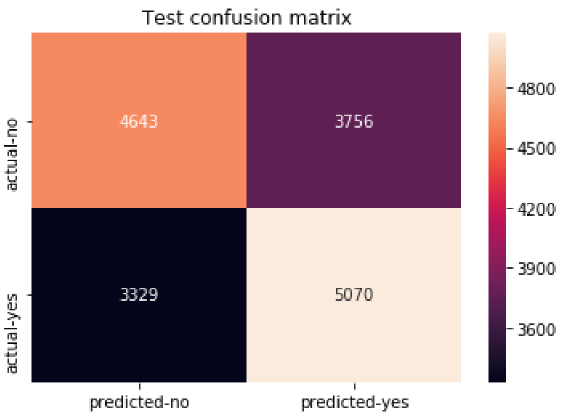


In [78]:

```
plt.title("Test confusion matrix")
sns.heatmap(cm_test, annot = True,yticklabels=ylabel, xticklabels=xlabel,fmt="d")
```

Out[78]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a427ed828>
```



2.4.4 Applying KNN brute force on TFIDF W2V, SET 4

In [ ]:

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

In [81]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model1 = TfidfVectorizer()
tfidf_model1.fit(ppe_train)
# we are converting a dictionary with word as a key, and the idf as a value
dictionary = dict(zip(tfidf_model1.get_feature_names(), list(tfidf_model1.idf_)))
tfidf_words1 = set(tfidf_model1.get_feature_names())
```

In [82]:

```
# average Word2Vec
# compute average word2vec for each review.
from tqdm import tqdm_notebook as tqdm
tfidf_w2v_pe_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(ppe_train): # 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_words1):
            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_pe_train.append(vector)

print(len(tfidf_w2v_pe_train))
print(len(tfidf_w2v_pe_train[0]))
```

```
32000
300
```

In [83]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_pe_cv = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(ppe_cv): # 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_words1):
            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_pe_cv.append(vector)

print(len(tfidf_w2v_pe_cv))
print(len(tfidf_w2v_pe_cv[0]))
```

```
8000
300
```

In [84]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_pe_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(ppe_test): # 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_words1):
            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_pe_test.append(vector)

print(len(tfidf_w2v_pe_test))
print(len(tfidf_w2v_pe_test[0]))
```

```
10000
300
```

In [85]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model2 = TfidfVectorizer()
tfidf_model2.fit(ppt_train)
# we are converting a dictionary with word as a key, and the idf as a value
dictionary = dict(zip(tfidf_model2.get_feature_names(), list(tfidf_model2.idf_)))
tfidf_words2 = set(tfidf_model2.get_feature_names())

# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_pt_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(ppt_train): # 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_words2):
            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_pt_train.append(vector)

print(len(tfidf_w2v_pt_train))
print(len(tfidf_w2v_pt_train[0]))
```

```
32000
300
```

In [86]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_pt_cv = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(ppt_cv): # 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_words2):
            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_pt_cv.append(vector)

print(len(tfidf_w2v_pt_cv))
print(len(tfidf_w2v_pt_cv[0]))
```

```
8000
300
```

In [87]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_pt_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(ppt_test): # 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_words2):
            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_pt_test.append(vector)

print(len(tfidf_w2v_pt_test))
print(len(tfidf_w2v_pt_test[0]))
```

```
10000
300
```

In [88]:

```
# Please write all the code with proper documentation
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
X_tr4 = hstack((X_train_cat_ohe,X_train_ss_ohe,X_train_sub_cat_ohe,X_train_pgc_ohe,X_train_tp_ohe,X
_train_price_standardized,X_train_tppp_standardized,tfidf_w2v_pe_train,tfidf_w2v_pt_train)).tocsr(
)
X_cv4 =
hstack((X_cv_cat_ohe,X_cv_ss_ohe,X_cv_sub_cat_ohe,X_cv_pgc_ohe,X_cv_tp_ohe,X_cv_price_standardized
,X_cv_tppp_standardized,tfidf_w2v_pe_cv,tfidf_w2v_pt_cv)).tocsr()
X_te4 = hstack((X_test_cat_ohe,X_test_ss_ohe,X_test_sub_cat_ohe,X_test_pgc_ohe,X_test_tp_ohe,X_test
_price_standardized,X_test_tppp_standardized,tfidf_w2v_pe_test,tfidf_w2v_pt_test)).tocsr()

print("Final Data matrix")
print(X_tr4.shape, y_train.shape)
print(X_cv4.shape, y_cv.shape)
print(X_te4.shape, y_test.shape)
print("="*100)
```

```
Final Data matrix
(32000, 702) (32000,)
(8000, 702) (8000,)
(10000, 702) (10000,)
====================================================================================================
```

In [89]:

```
from sklearn.decomposition import TruncatedSVD
from sklearn.random_projection import sparse_random_matrix
svd = TruncatedSVD(n_components=100, n_iter=7, random_state=42)
X_tr4 = svd.fit_transform(X_tr4)
X_cv4=svd.transform(X_cv4)
X_te4=svd.transform(X_te4)

print(X_tr4.shape, y_train.shape)
print(X_cv4.shape, y_cv.shape)
print(X_te4.shape, y_test.shape)
print("="*100)
```

```
(32000, 100) (32000,)
(8000, 100) (8000,)
(10000, 100) (10000,)
====================================================================================================
```

In [92]:

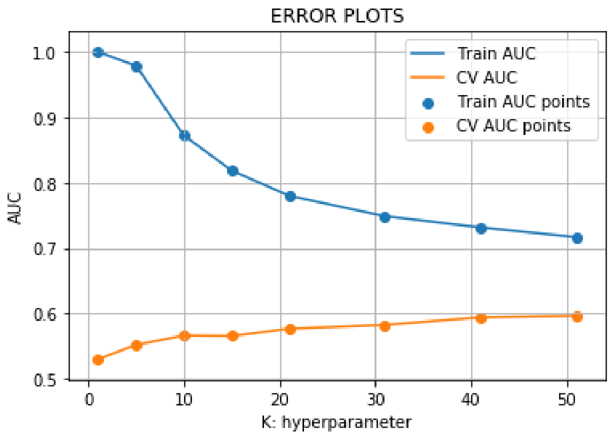
```
#https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html
from imblearn.over_sampling import RandomOverSampler
ros = RandomOverSampler(random_state=0,sampling_strategy='minority')
X_tr4_res, y_train_res = ros.fit_resample(X_tr4, y_train)
X_cv4_res,y_cv_res = ros.fit_resample(X_cv4, y_cv)
X_te4_res,y_test_res = ros.fit_resample(X_te4,y_test)

print("Final Data matrix")
print(X_tr4_res.shape, y_train_res.shape)
print(X_cv4_res.shape, y_cv_res.shape)
print(X_te4_res.shape, y_test_res.shape)
print("="*100)
```

```
Final Data matrix
(53750, 100) (53750,)
(13438, 100) (13438,)
(16798, 100) (16798,)
====================================================================================================
```

In [93]:

```
def batch_predict(clf, data):
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the posi
tive class
```

```
    # not the predicted outputs

    y_data_pred = []
    tr_loop = data.shape[0] - data.shape[0]%1000
    # consider you X_tr shape is 49041, then your cr_loop will be 49041 - 49041%1000 = 49000
    # in this for loop we will iterate unti the last 1000 multiplier
    for i in range(0, tr_loop, 1000):
        y_data_pred.extend(clf.predict_proba(data[i:i+1000])[:,1])
    # we will be predicting for the last data points
    y_data_pred.extend(clf.predict_proba(data[tr_loop:])[:,1])

    return y_data_pred


import matplotlib.pyplot as plt
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import roc_auc_score
"""
y_true : array, shape = [n_samples] or [n_samples, n_classes]
True binary labels or binary label indicators.

y_score : array, shape = [n_samples] or [n_samples, n_classes]
Target scores, can either be probability estimates of the positive class, confidence values, or non-thresholded measure of
decisions (as returned by “decision_function” on some classifiers).
For binary y_true, y_score is supposed to be the score of the class with greater label.

"""

train_auc = []
cv_auc = []
K = [1, 5, 10, 15, 21, 31, 41, 51]
for i in tqdm(K):
    neigh = KNeighborsClassifier(n_neighbors=i,n_jobs=-1)
    neigh.fit(X_tr4_res, y_train_res)

    y_train_pred = batch_predict(neigh, X_tr4_res)
    y_cv_pred = batch_predict(neigh, X_cv4_res)

    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive class
    # not the predicted outputs
    train_auc.append(roc_auc_score(y_train_res,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv_res, y_cv_pred))

plt.plot(K, train_auc, label='Train AUC')
plt.plot(K, cv_auc, label='CV AUC')

plt.scatter(K, train_auc, label='Train AUC points')
plt.scatter(K, cv_auc, label='CV AUC points')

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

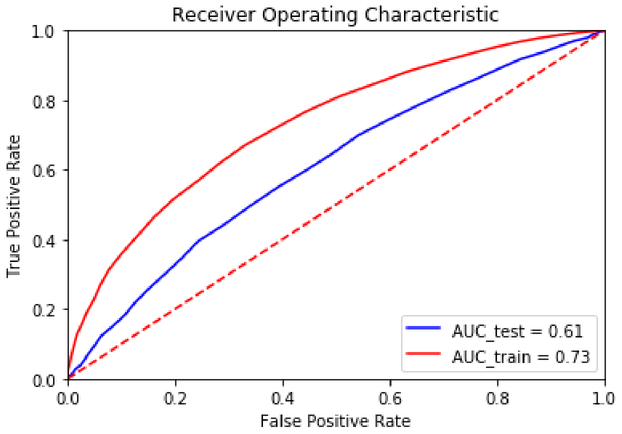


In [94]:

```
# from the above graph, taking value of K=51
#https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
from sklearn.neighbors import KNeighborsClassifier
neigh = KNeighborsClassifier(n_neighbors=41,algorithm='brute')
neigh.fit(X_tr4_res, y_train_res)

pred = neigh.predict(X_te4_res)
pred1 = neigh.predict(X_tr4_res)
#https://stackoverflow.com/questions/25009284/how-to-plot-roc-curve-in-python
import sklearn.metrics as metrics
# calculate the fpr and tpr for all thresholds of the classification
probs = neigh.predict_proba(X_te4_res)
probs1 = neigh.predict_proba(X_tr4_res)
preds = probs[:,1]
preds1 = probs1[:,1]
fpr, tpr, threshold = metrics.roc_curve(y_test_res, preds)
fpr1, tpr1, threshold = metrics.roc_curve(y_train_res, preds1)
roc_auc = metrics.auc(fpr, tpr)
roc_auc1 = metrics.auc(fpr1, tpr1)
```

In [95]:

```
# method I: plt
import matplotlib.pyplot as plt
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b', label = 'AUC_test = %0.2f' % roc_auc)
plt.plot(fpr1, tpr1, 'r', label = 'AUC_train = %0.2f' % roc_auc1)
plt.legend(loc = 'lower right')
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
plt.ylim([0, 1])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```



In [96]:

```
# we are writing our own function for predict, with defined thresould
# we will pick a threshold that will give the least fpr
def predict(proba, 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))
    predictions = []
    for i in proba:
        if i>=t:
            predictions.append(1)
        else:
            predictions.append(0)
    return predictions
```

In [97]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
cm_train = confusion_matrix(y_train_res, predict(y_train_pred, threshold,tpr,fpr))
print(cm_train)

print("Test confusion matrix")
cm_test= confusion_matrix(y_test_res, predict(pred, threshold, tpr1,fpr1))
print(cm_test)
```

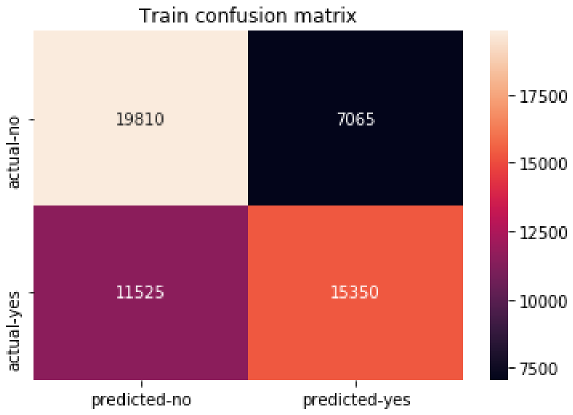
```
====================================================================================================

Train confusion matrix
the maximum value of tpr*(1-fpr) 0.1785701785775216 for threshold 0.512
[[19810  7065]
 [11525 15350]]
Test confusion matrix
the maximum value of tpr*(1-fpr) 0.10899147249323958 for threshold 0.488
[[4671 3728]
 [3379 5020]]
```

In [98]:

```
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
ylabel = ["actual-no","actual-yes"]
xlabel = ["predicted-no","predicted-yes"]
plt.title("Train confusion matrix")
sns.heatmap(cm_train, annot = True,yticklabels=ylabel, xticklabels=xlabel,fmt="d")
```

Out[98]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a42ff2f28>
```

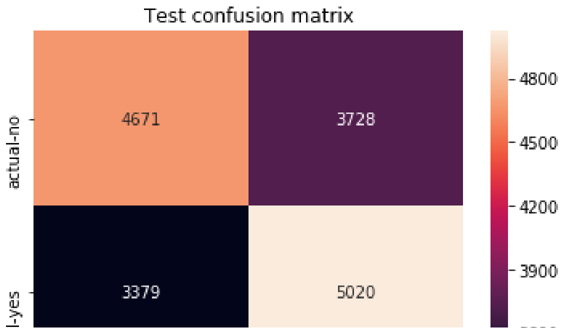


In [99]:

```
plt.title("Test confusion matrix")
sns.heatmap(cm_test, annot = True,yticklabels=ylabel, xticklabels=xlabel,fmt="d")
```

Out[99]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a44790da0>
```



2.5 Feature selection with `SelectKBest`

In [73]:

```
# 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 [187]:

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

# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
X_tr2 = hstack((X_train_cat_ohe,X_train_ss_ohe,X_train_sub_cat_ohe,X_train_pgc_ohe,X_train_tp_ohe,X
_train_price_standardized.astype(int),X_train_tppp_standardized.astype(int),X_train_essay_tfidf,X_
train_title_tfidf)).tocsr()
X_cv2 =
hstack((X_cv_cat_ohe,X_cv_ss_ohe,X_cv_sub_cat_ohe,X_cv_pgc_ohe,X_cv_tp_ohe,X_cv_price_standardized
.astype(int),X_cv_tppp_standardized.astype(int),X_cv_essay_tfidf,X_cv_title_tfidf)).tocsr()
X_te2 = hstack((X_test_cat_ohe,X_test_ss_ohe,X_test_sub_cat_ohe,X_test_pgc_ohe,X_test_tp_ohe,X_test
_price_standardized.astype(int),X_test_tppp_standardized.astype(int),X_test_essay_tfidf,X_test_titl
e_tfidf)).tocsr()

print("Final Data matrix")
print(X_tr2.shape, y_train.shape)
print(X_cv2.shape, y_cv.shape)
print(X_te2.shape, y_test.shape)
print("="*100)
```

```
Final Data matrix
(32000, 10440) (32000,)
(8000, 10440) (8000,)
(10000, 10440) (10000,)
====================================================================================================
```

In [188]:

```
from sklearn.datasets import load_digits
from sklearn.feature_selection import SelectKBest, chi2

X=SelectKBest(k=2000)

Kbest_Xtr = X.fit(X_tr2, y_train)

Kbest_Xtr = X.transform(X_tr2)
Kbest_Xcv = X.transform(X_cv2)
Kbest_Xte = X.transform(X_te2)
print(Kbest_Xtr.shape)
print(Kbest_Xcv.shape)
print(Kbest_Xte.shape)
```

```
(32000, 2000)
(8000, 2000)
(10000, 2000)
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/feature_selection/univariate_selection.py:114: User
Warning:
```

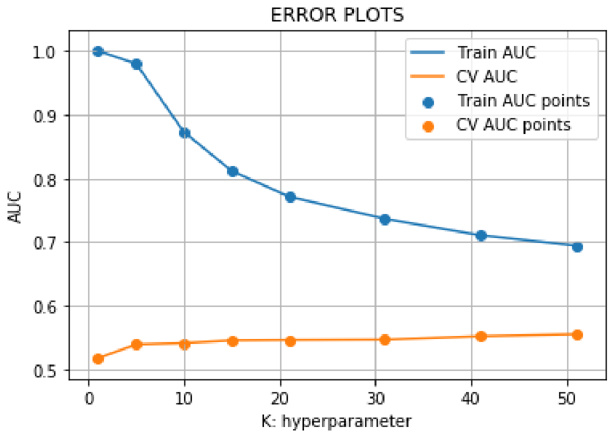
```
Features [0 0 0 0 0 0 0 0] are constant.
```

In [189]:

```
from imblearn.over_sampling import RandomOverSampler
ros = RandomOverSampler(random_state=0,sampling_strategy='minority')
Kbest_Xtr_res, y_train_res = ros.fit_resample(Kbest_Xtr, y_train)
Kbest_Xcv_res,y_cv_res = ros.fit_resample(Kbest_Xcv, y_cv)
Kbest_Xte_res,y_test_res = ros.fit_resample(Kbest_Xte,y_test)
```

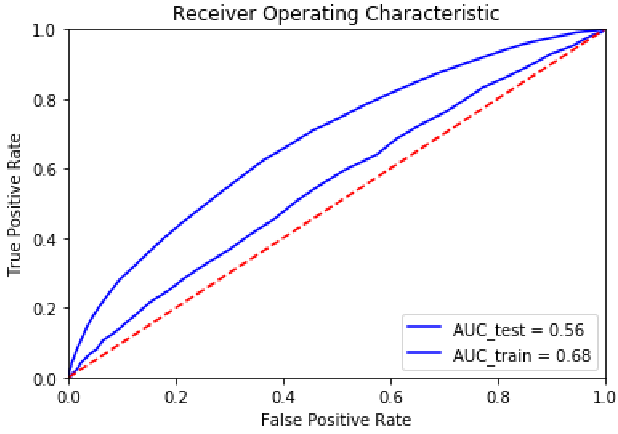
In [190]:

```
#the below set of code is taken from the sample solution ipython notebook

def batch_predict(clf, data):
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive class
    # not the predicted outputs

    y_data_pred = []
    tr_loop = data.shape[0] - data.shape[0]%1000
    # consider you X_tr shape is 49041, then your cr_loop will be 49041 - 49041%1000 = 49000
    # in this for loop we will iterate unti the last 1000 multiplier
    for i in range(0, tr_loop, 1000):
        y_data_pred.extend(clf.predict_proba(data[i:i+1000])[:,1])
    # we will be predicting for the last data points
    y_data_pred.extend(clf.predict_proba(data[tr_loop:])[:,1])

    return y_data_pred


import matplotlib.pyplot as plt
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import roc_auc_score
"""3
y_true : array, shape = [n_samples] or [n_samples, n_classes]
True binary labels or binary label indicators.

y_score : array, shape = [n_samples] or [n_samples, n_classes]
Target scores, can either be probability estimates of the positive class, confidence values, or non-thresholded measure of
decisions (as returned by "decision_function" on some classifiers).
For binary y_true, y_score is supposed to be the score of the class with greater label.

"""

train_auc = []
cv_auc = []
K = [1, 5, 10, 15, 21, 31, 41, 51]
for i in tqdm(K):
    neigh = KNeighborsClassifier(n_neighbors=i,n_jobs=-1)
    neigh.fit(Kbest_Xtr_res, y_train_res)

    y_train_pred = batch_predict(neigh, Kbest_Xtr_res)
    y_cv_pred = batch_predict(neigh, Kbest_Xcv_res)

    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive class
    # not the predicted outputs
    train_auc.append(roc_auc_score(y_train_res,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv_res, y_cv_pred))

plt.plot(K, train_auc, label='Train AUC')
plt.plot(K, cv_auc, label='CV AUC')

plt.scatter(K, train_auc, label='Train AUC points')
plt.scatter(K, cv_auc, label='CV AUC points')

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



In [201]:

```
#https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
from sklearn.neighbors import KNeighborsClassifier
neigh = KNeighborsClassifier(n_neighbors=61,algorithm='brute')
neigh.fit(Kbest_Xtr_res, y_train_res)


pred = neigh.predict(Kbest_Xte_res)


#https://stackoverflow.com/questions/25009284/how-to-plot-roc-curve-in-python
import sklearn.metrics as metrics
# calculate the fpr and tpr for all thresholds of the classification
probs = neigh.predict_proba(Kbest_Xte_res)
probs1 = neigh.predict_proba(Kbest_Xtr_res)
preds = probs[:,1]
preds1 = probs1[:,1]
fpr, tpr, threshold = metrics.roc_curve(y_test_res, preds)
fpr1, tpr1, threshold = metrics.roc_curve(y_train_res, preds1)
roc_auc = metrics.auc(fpr, tpr)
roc_auc1 = metrics.auc(fpr1, tpr1)

# method I: plt
import matplotlib.pyplot as plt
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b', label = 'AUC_test = %0.2f' % roc_auc)
plt.plot(fpr1, tpr1, 'b', label = 'AUC_train = %0.2f' % roc_auc1)
plt.legend(loc = 'lower right')
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
plt.ylim([0, 1])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```



In [206]:

In [206]:

```
# we are writing our own function for predict, with defined thresould
# we will pick a threshold that will give the least fpr
def predict(proba, 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))
    predictions = []
    for i in proba:
        if i>=t:
            predictions.append(1)
        else:
            predictions.append(0)
    return predictions
```

In [207]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
cm_train = confusion_matrix(y_train_res, predict(y_train_pred, threshold,tpr,fpr),labels=[0,1])
print(cm_train)

print("Test confusion matrix")
cm_test= confusion_matrix(y_test_res, predict(pred, threshold, tpr1,fpr1),labels=[0,1])
print(cm_test)
```

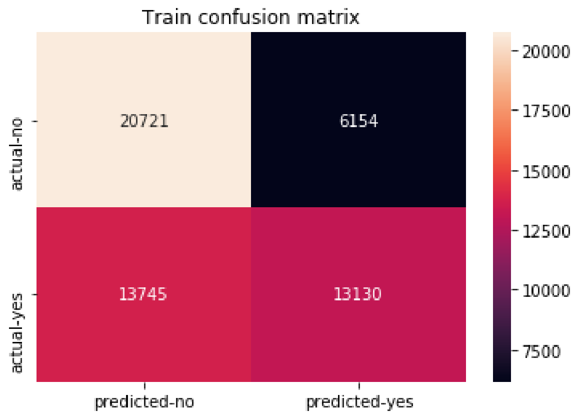
```
====================================================================================================

Train confusion matrix
the maximum value of tpr*(1-fpr) 0.20987077124555203 for threshold 0.557
[[20721  6154]
 [13745 13130]]
Test confusion matrix
the maximum value of tpr*(1-fpr) 0.1371315601514332 for threshold 0.492
[[4069 4330]
 [3390 5009]]
```

In [208]:

```
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
ylabel = ["actual-no","actual-yes"]
xlabel = ["predicted-no","predicted-yes"]
plt.title("Train confusion matrix")
sns.heatmap(cm_train, annot = True,yticklabels=ylabel, xticklabels=xlabel,fmt="d")
```

Out[208]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a4c0b9470>
```



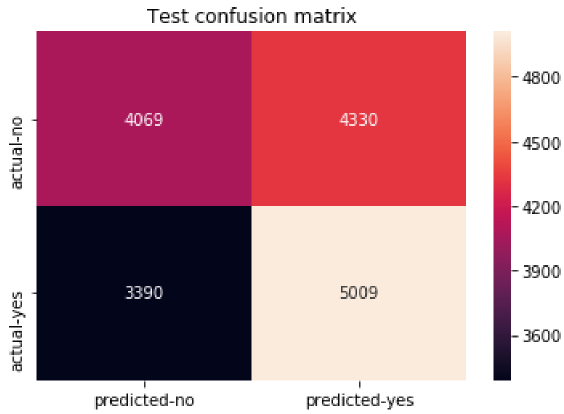
In [209]:

```
plt.title("Test confusion matrix")
```

```
sns.heatmap(cm_test, annot = True,yticklabels=ylabel, xticklabels=xlabel,fmt="d")
```

Out[209]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a4c0f1b70>
```



3. Conclusions

In [210]:

```
# Please compare all your models using Prettytable library
from prettytable import PrettyTable

x = PrettyTable()

x.field_names = ["Vectorizer", "Model", "Hyper parameter", "AUC"]


x.add_row(["BOW", "Brute", 51, 0.61])
x.add_row(["TFIDF", "Brute",51, 0.57])
x.add_row(["W2V", "Brute",51,0.60])
x.add_row(["TFIDF-W2V", "Brute",41,0.61])

print(x)
```

```
+------------+-------+-----------------+------+
| Vectorizer | Model | Hyper parameter | AUC  |
+------------+-------+-----------------+------+
|    BOW     | Brute |        51       | 0.61 |
|   TFIDF    | Brute |        51       | 0.57 |
|    W2V     | Brute |        51       | 0.6  |
| TFIDF-W2V  | Brute |        41       | 0.61 |
+------------+-------+-----------------+------+
```

Observations

1. Dataset is imbalanced. Earlier the AUC scores for any K values were around 0.5.
2. For this I have oversampled the data and AUC scores have improved.
3. For any K value taken, the TP and TN values were lesser and later I have tried doing SVD truncated dimensionality reduction which helped to get better confusion matrix.