# 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:**<br>• `Art Will Make You Happy!`<br>• `First Grade Fun` |
| **`project_grade_category`** | Grade level of students for which the project is targeted. One of the following enumerated values:<br>• `Grades PreK-2`<br>• `Grades 3-5`<br>• `Grades 6-8`<br>• `Grades 9-12` |
| **`project_subject_categories`** | One or more (comma-separated) subject categories for the project from the following enumerated list of values:<br>• `Applied Learning`<br>• `Care & Hunger`<br>• `Health & Sports`<br>• `History & Civics`<br>• `Literacy & Language`<br>• `Math & Science`<br>• `Music & The Arts`<br>• `Special Needs`<br>• `Warmth`<br>**Examples:**<br>• `Music & The Arts`<br>• `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:**<br>• `Literacy` |

| Feature | Description |
| --- | --- |
| | • `Literature & Writing, Social Sciences` |
| **`project_resource_summary`** | An explanation of the resources needed for the project. **Example:**<br>• `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:<br>• `nan`<br>• `Dr.`<br>• `Mr.`<br>• `Mrs.`<br>• `Ms.`<br>• `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
import os

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

```
C:\Users\Shashank\Anaconda3\lib\site-packages\gensim\utils.py:1209: UserWarning: detected Windows;
aliasing chunkize to chunkize_serial
  warnings.warn("detected Windows; aliasing chunkize to chunkize_serial")
```

## 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"
```

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

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

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

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

## 1.3 preprocessing of `project_subject_subcategories`

In [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]:

| | 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]:

```
y=project_data['project_is_approved']
```

In [11]:

```
y.head()
```

Out[11]:

```
55660    1
76127    1
51140    1
473      1
41558    1
Name: project_is_approved, dtype: int64
```

In [12]:

```
project_data.drop(['project_is_approved'], axis=1, inplace=True)
```

In [13]:

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

In [14]:

```
# printing some random reviews
print(project_data['essay'].values[0])
print("="*50)
print(project_data['essay'].values[150])
print("="*50)
print(project_data['essay'].values[1000])
print("="*50)
print(project_data['essay'].values[20000])
print("="*50)
print(project_data['essay'].values[99999])
print("="*50)
```

```
I have been fortunate enough to use the Fairy Tale STEM kits in my classroom as well as the STEM j
ournals, which my students really enjoyed.  I would love to implement more of the Lakeshore STEM k
its in my classroom for the next school year as they provide excellent and engaging STEM
```

lessons.My students come from a variety of backgrounds, including language and socioeconomic statu
s.  Many of them don't have a lot of experience in science and engineering and these kits give me
the materials to provide these exciting opportunities for my students.Each month I try to do
several science or STEM/STEAM projects.  I would use the kits and robot to help guide my science i
nstruction in engaging and meaningful ways.  I can adapt the kits to my current language arts paci
ng guide where we already teach some of the material in the kits like tall tales (Paul Bunyan) or
Johnny Appleseed.  The following units will be taught in the next school year where I will
implement these kits: magnets, motion, sink vs. float, robots.  I often get to these units and don
't know If I am teaching the right way or using the right materials.  The kits will give me
additional ideas, strategies, and lessons to prepare my students in science.It is challenging to d
evelop high quality science activities.  These kits give me the materials I need to provide my
students with science activities that will go along with the curriculum in my classroom.  Although
I have some things (like magnets) in my classroom, I don't know how to use them effectively.  The
kits will provide me with the right amount of materials and show me how to use them in an
appropriate way.
==========
I teach high school English to students with learning and behavioral disabilities. My students all
vary in their ability level. However, the ultimate goal is to increase all students literacy level
s. This includes their reading, writing, and communication levels.I teach a really dynamic group o
f students. However, my students face a lot of challenges. My students all live in poverty and in
a dangerous neighborhood. Despite these challenges, I have students who have the the desire to def
eat these challenges. My students all have learning disabilities and currently all are performing
below grade level. My students are visual learners and will benefit from a classroom that fulfills
their preferred learning style.The materials I am requesting will allow my students to be prepared
for the classroom with the necessary supplies.  Too often I am challenged with students who come t
o school unprepared for class due to economic challenges.  I want my students to be able to focus
on learning and not how they will be able to get school supplies.  The supplies will last all year
.  Students will be able to complete written assignments and maintain a classroom journal.  The ch
art paper will be used to make learning more visual in class and to create posters to aid students
in their learning.  The students have access to a classroom printer.  The toner will be used to pr
int student work that is completed on the classroom Chromebooks.I want to try and remove all barri
ers for the students learning and create opportunities for learning. One of the biggest barriers i
s the students not having the resources to get pens, paper, and folders. My students will be able
to increase their literacy skills because of this project.
==========
\"Life moves pretty fast. If you don't stop and look around once in awhile, you could miss it.\"
from the movie, Ferris Bueller's Day Off.  Think back...what do you remember about your
grandparents?  How amazing would it be to be able to flip through a book to see a day in their
lives?My second graders are voracious readers! They love to read both fiction and nonfiction books
.  Their favorite characters include Pete the Cat, Fly Guy, Piggie and Elephant, and Mercy Watson.
They also love to read about insects, space and plants. My students are hungry bookworms! My stude
nts are eager to learn and read about the world around them. My kids love to be at school and are
like little sponges absorbing everything around them. Their parents work long hours and usually do
not see their children. My students are usually cared for by their grandparents or a family
friend. Most of my students do not have someone who speaks English at home. Thus it is difficult f
or my students to acquire language.Now think forward... wouldn't it mean a lot to your kids,
nieces or nephews or grandchildren, to be able to see a day in your life today 30 years from now?
Memories are so precious to us and being able to share these memories with future generations will
be a rewarding experience.  As part of our social studies curriculum, students will be learning ab
out changes over time.  Students will be studying photos to learn about how their community has ch
anged over time.  In particular, we will look at photos to study how the land, buildings,
clothing, and schools have changed over time.  As a culminating activity, my students will capture
a slice of their history and preserve it through scrap booking. Key important events in their
young lives will be documented with the date, location, and names.  Students will be using photos
from home and from school to create their second grade memories.  Their scrap books will preserve
their unique stories for future generations to enjoy.Your donation to this project will provide my
second graders with an opportunity to learn about social studies in a fun and creative manner.  Th
rough their scrapbooks, children will share their story with others and have a historical document
for the rest of their lives.
==========
\"A person's a person, no matter how small.\" (Dr.Seuss) I teach the smallest students with the bi
ggest enthusiasm for learning. My students learn in many different ways using all of our senses an
d multiple intelligences. I use a wide range of techniques to help all my students succeed. \r\nSt
udents in my class come from a variety of different backgrounds which makes for wonderful sharing
of experiences and cultures, including Native Americans.\r\nOur school is a caring community of su
ccessful learners which can be seen through collaborative student project based learning in and ou
t of the classroom. Kindergarteners in my class love to work with hands-on materials and have many
different opportunities to practice a skill before it is mastered. Having the social skills to wor
k cooperatively with friends is a crucial aspect of the kindergarten curriculum.Montana is the
perfect place to learn about agriculture and nutrition. My students love to role play in our
pretend kitchen in the early childhood classroom. I have had several kids ask me, \"Can we try coo
king with REAL food?\" I will take their idea and create \"Common Core Cooking Lessons\" where we
learn important math and writing concepts while cooking delicious healthy food for snack time. My
students will have a grounded appreciation for the work that went into making the food and knowled
ge of where the ingredients came from as well as how it's healthy for their bodies. This project w
ould expand our learning of nutrition and agricultural cooking recipes by having us peel our own a
pples to make homemade applesauce, make our own bread, and mix up healthy plants from our classroo

m garden in the spring. We will also create our own cookbooks to be printed and shared with famili
es. \r\nStudents will gain math and literature skills as well as a life long enjoyment for healthy
cooking.nannan
==================================================
My classroom consists of twenty-two amazing sixth graders from different cultures and backgrounds.
They are a social bunch who enjoy working in partners and working with groups. They are hard-worki
ng and eager to head to middle school next year. My job is to get them ready to make this
transition and make it as smooth as possible. In order to do this, my students need to come to
school every day and feel safe and ready to learn. Because they are getting ready to head to
middle school, I give them lots of choice- choice on where to sit and work, the order to complete
assignments, choice of projects, etc. Part of the students feeling safe is the ability for them to
come into a welcoming, encouraging environment. My room is colorful and the atmosphere is casual.
I want them to take ownership of the classroom because we ALL share it together. Because my time w
ith them is limited, I want to ensure they get the most of this time and enjoy it to the best of t
heir abilities.Currently, we have twenty-two desks of differing sizes, yet the desks are similar t
o the ones the students will use in middle school. We also have a kidney table with crates for sea
ting. I allow my students to choose their own spots while they are working independently or in
groups. More often than not, most of them move out of their desks and onto the crates. Believe it
or not, this has proven to be more successful than making them stay at their desks! It is because
of this that I am looking toward the “Flexible Seating” option for my classroom.\r\n The students
look forward to their work time so they can move around the room. I would like to get rid of the c
onstricting desks and move toward more “fun” seating options. I am requesting various seating so m
y students have more options to sit. Currently, I have a stool and a papasan chair I inherited fro
m the previous sixth-grade teacher as well as five milk crate seats I made, but I would like to gi
ve them more options and reduce the competition for the “good seats”. I am also requesting two rug
s as not only more seating options but to make the classroom more welcoming and appealing. In orde
r for my students to be able to write and complete work without desks, I am requesting a class set
of clipboards. Finally, due to curriculum that requires groups to work together, I am requesting t
ables that we can fold up when we are not using them to leave more room for our flexible seating o
ptions.\r\nI know that with more seating options, they will be that much more excited about coming
to school! Thank you for your support in making my classroom one students will remember
forever!nannan
==================================================

In [15]:

```
# 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 [16]:

```
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

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 [17]:

```
# \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 [18]:

```
#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 [19]:

```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'not'
stopwords= ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've",
\
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his',
'himself', \
```

```
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them',
'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll",
'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having',
'do', 'does', \
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', '
while', 'of', \
            'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during',
'before', 'after',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under'
, 'again', 'further',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'e
ach', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'too', 'very', \
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll'
, 'm', 'o', 're', \
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "d
esn't", 'hadn',\
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn',
"mightn't", 'mustn',\
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn',
"wasn't", 'weren', "weren't", \
            'won', "won't", 'wouldn', "wouldn't"]
```

In [20]:

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

```
100%|
|
109248/109248 [02:10<00:00, 835.17it/s]
```

In [21]:

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

Out[21]:

```
'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'
```

In [22]:

```
project_data['preprocessed_essays'] = preprocessed_essays
```

In [23]:

```
project_data.drop(['essay'], axis=1, inplace=True)
```

## 1.4 Preprocessing of `project_title`

In [24]:

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

In [25]:

```
# Combining all the above statemennts
from tqdm import tqdm
project_title_list = []
# 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)
    project_title_list.append(sent.lower().strip())
```

```
100%|
| 109248/109248
[00:07<00:00, 14974.47it/s]
```

In [26]:

```
project_data['project_title_list'] = project_title_list
```

In [27]:

```
project_data.drop(['project_title'], axis=1, inplace=True)
```

## 1.5 Preparing data for models

In [28]:

```
project_data.columns
```

Out[28]:

```
Index(['Unnamed: 0', 'id', 'teacher_id', 'teacher_prefix', 'school_state',
       'Date', 'project_grade_category', 'project_essay_1', 'project_essay_2',
       'project_essay_3', 'project_essay_4', 'project_resource_summary',
       'teacher_number_of_previously_posted_projects', 'clean_categories',
       'clean_subcategories', 'preprocessed_essays', 'project_title_list'],
      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
- quantity : numerical
- teacher_number_of_previously_posted_projects : numerical

- 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 [29]:

```
# 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 [30]:

```
# 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)
```

```
['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']
Shape of matrix after one hot encodig  (109248, 30)
```

In [31]:

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

In [32]:

```
#onehotencoding for school_state
one_hot_encoding_school_state=pd.get_dummies(project_data.school_state)
print("Shape of dataframe for school_state",  one_hot_encoding_school_state.shape)
```

```
Shape of dataframe for school_state (109248, 51)
```

In [33]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix=pd.get_dummies(project_data.teacher_prefix)

print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix.shape)
```

```
Shape of dataframe for teacher_prefix (109248, 5)
```

In [34]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category=pd.get_dummies(project_data.project_grade_category)

print("Shape of dataframe for project_grade_category", one_hot_encoding_project_grade_category.sha
pe)
```

```
Shape of dataframe for project_grade_category (109248, 4)
```

Shape of dataframe for project_grade_category (109248, 4)

## 1.5.2 Vectorizing Text data

### 1.5.2.1 Bag of words

In [35]:

```
# 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)
```

```
Shape of matrix after one hot encodig  (109248, 16512)
```

In [36]:

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

In [37]:

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

```
Shape of matrix after one hot encodig  (109248, 3222)
```

### 1.5.2.2 TFIDF vectorizer

In [38]:

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

```
Shape of matrix after one hot encodig  (109248, 16512)
```

### 1.5.2.3 Using Pretrained Models: Avg W2V

In [39]:

```
'''
# 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)


'''
```

Out[39]:

```
'\n# Reading glove vectors in python: https://stackoverflow.com/a/38230349/4084039\ndef
loadGloveModel(gloveFile):\n    print ("Loading Glove Model")\n    f = open(gloveFile,\'r\',
encoding="utf8")\n    model = {}\n    for line in tqdm(f):\n        splitLine = line.split()\n
word = splitLine[0]\n        embedding = np.array([float(val) for val in splitLine[1:]])\n        m
odel[word] = embedding\n    print ("Done.",len(model)," words loaded!")\n    return model\nmodel =
loadGloveModel(\'glove.42B.300d.txt\')\n\n# ============================\nOutput:\n    \nLoading G
love Model\n1917495it [06:32, 4879.69it/s]\nDone. 1917495  words loaded!\n\n#
============================\n\nwords = []\nfor i in preproced_texts:\n    words.extend(i.split(\'
\'))\n\nfor i in preproced_titles:\n    words.extend(i.split(\' \'))\nprint("all the words in the
coupus", len(words))\nwords = set(words)\nprint("the unique words in the coupus",
len(words))\n\ninter_words = set(model.keys()).intersection(words)\nprint("The number of words tha
t are present in both glove vectors and our coupus",       len(inter_words),"
(",np.round(len(inter_words)/len(words)*100,3),"%)")\n\nwords_courpus = {}\nwords_glove =
set(model.keys())\nfor i in words:\n    if i in words_glove:\n        words_courpus[i] = model[i]\n
print("word 2 vec length", len(words_courpus))\n\n\n# stronging variables into pickle files python
: http://www.jessicayung.com/how-to-use-pickle-to-save-and-load-variables-in-python/\n\nimport pic
kle\nwith open(\'glove_vectors\', \'wb\') as f:\n    pickle.dump(words_courpus, f)\n\n\n'
```

In [40]:

```
# 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 [41]:

```
# 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]))
```

```
100%|
| 109248/109248
[01:06<00:00, 1639.31it/s]
```

```
109248
300
```

#### 1.5.2.3 Using Pretrained Models: TFIDF weighted W2V

In [42]:

```
# 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 [43]:

```
# 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]))
```

```
100%|
|
109248/109248 [06:33<00:00, 277.63it/s]
```

```
109248
300
```

In [44]:

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

In [45]:

```
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_project_title = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(project_title_list): # 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
```

```
    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_project_title.append(vector)

print(len(avg_w2v_vectors_project_title))
print(len(avg_w2v_vectors_project_title[0]))
```

```
100%|
| 109248/109248
[00:04<00:00, 24711.92it/s]
```

```
109248
300
```

In [46]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(project_title_list)
# 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 [47]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_project_title = []; # the avg-w2v for each sentence/review is stored in this lis
t
for sentence in tqdm(project_title_list): # 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_project_title.append(vector)

print(len(tfidf_w2v_vectors_project_title))
print(len(tfidf_w2v_vectors_project_title[0]))
```

```
100%|
| 109248/109248
[00:09<00:00, 11803.81it/s]
```

```
109248
300
```

### 1.5.3 Vectorizing Numerical features

In [48]:

```
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 [49]:

In [49]:

```
project_data.columns
```

Out[49]:

```
Index(['Unnamed: 0', 'id', 'teacher_id', 'teacher_prefix', 'school_state',
       'Date', 'project_grade_category', 'project_essay_1', 'project_essay_2',
       'project_essay_3', 'project_essay_4', 'project_resource_summary',
       'teacher_number_of_previously_posted_projects', 'clean_categories',
       'clean_subcategories', 'preprocessed_essays', 'project_title_list',
       'price', 'quantity'],
      dtype='object')
```

In [50]:

```
# 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))
```

```
Mean : 298.1193425966608, Standard deviation : 367.49634838483496
```

In [51]:

```
price_standardized
```

Out[51]:

```
array([[ 1.16172762],
       [-0.23153793],
       [ 0.08402983],
       ...,
       [ 0.27450792],
       [-0.0282706 ],
       [-0.79625102]])
```

### 1.5.4 Merging all the above features

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

In [52]:

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

```
(109248, 9)
(109248, 30)
(109248, 16512)
(109248, 1)
```

In [53]:

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

```
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
```

Out[53]:

```
(109248, 16552)
```

# 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 [103]:

```
from sklearn.model_selection import train_test_split
X1_train, X_test, y1_train, y_test = train_test_split(
    project_data, y, test_size=0.2,stratify=y, random_state=42)
X_cv,X_train,y_cv,y_train=train_test_split(X1_train,y1_train,test_size=0.7,stratify=y1_train,random
_state=42)
```

In [104]:

```
X_train.shape
```

Out[104]:

```
(61179, 19)
```

In [105]:

```
y_train.shape
```

Out[105]:

```
(61179,)
```

In [106]:

```
X_cv.shape
```

Out[106]:

```
(26219, 19)
```

In [107]:

```
y_cv.shape
```

Out[107]:

```
(26219,)
```

In [108]:

```
X_test.shape
```

Out[108]:

```
(21850, 19)
```

In [109]:

```
y_test.shape
```

Out[109]:

```
(21850,)
```

In [110]:

```
X_train.head(2)
```

Out[110]:

| | Unnamed: 0 | id | teacher_id | teacher_prefix | school_state | Date | project_grade_cate |
|---|---|---|---|---|---|---|---|
| **70311** | 75821 | p002574 | 72ec3ef1a9c08a7d5bcc80406c21a02d | Teacher | DC | 2016-11-23 18:58:56 | Grades PreK-2 |
| **7977** | 4140 | p217775 | 1e0795765edbda0a10149b66bfca6a2e | Mrs. | CA | 2016-06-08 22:59:54 | Grades 3-5 |

## 2.2 Make Data Model Ready: encoding numerical, categorical features

In [111]:

```
# 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 [112]:

```
# 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 maen and variance.
price_standardized_train = price_scalar.transform(X_train['price'].values.reshape(-1, 1))
```

```
Mean : 297.4337254613511, Standard deviation : 356.94930693022343
```

In [113]:

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

In [114]:

```
#onehotencoding for school_state
one_hot_encoding_school_state_train=pd.get_dummies(X_train.school_state)
print("Shape of dataframe for school_state", one_hot_encoding_school_state_train.shape)
```

```
Shape of dataframe for school_state (61179, 51)
```

In [115]:

```
#onehotencoding for school_state
one_hot_encoding_school_state_cv=pd.get_dummies(X_cv.school_state)
print("Shape of dataframe for school_state",  one_hot_encoding_school_state_cv.shape)
```

```
Shape of dataframe for school_state (26219, 51)
```

In [116]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix_train=pd.get_dummies(X_train.teacher_prefix)

print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix_train.shape)
```

```
Shape of dataframe for teacher_prefix (61179, 5)
```

In [117]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix_cv=pd.get_dummies(X_cv.teacher_prefix)

print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix_cv.shape)
```

```
Shape of dataframe for teacher_prefix (26219, 5)
```

In [118]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category_train=pd.get_dummies(X_train.project_grade_category)

print("Shape of dataframe for project_grade_category",
one_hot_encoding_project_grade_category_train.shape)
```

```
Shape of dataframe for project_grade_category (61179, 4)
```

In [119]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category_cv=pd.get_dummies(X_cv.project_grade_category)

print("Shape of dataframe for project_grade_category", one_hot_encoding_project_grade_category_cv.
shape)
```

```
Shape of dataframe for project_grade_category (26219, 4)
```

In [120]:

```
# 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_train = vectorizer.fit_transform(X_train['clean_categories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot_train.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (61179, 9)
```

In [121]:

```
# we use count vectorizer to convert the values into one
vectorizer = CountVectorizer(vocabulary=list(sorted_cat_dict.keys()), lowercase=False, binary=True
)
categories_one_hot_cv = vectorizer.transform(X_cv['clean_categories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot_cv.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (26219, 9)
```

In [122]:

```
categories_one_hot_test = vectorizer.transform(X_test['clean_categories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot_test.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (21850, 9)
```

In [123]:

```
# 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_train = vectorizer.fit_transform(X_train['clean_subcategories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot_train.shape)
```

```
['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']
Shape of matrix after one hot encodig  (61179, 30)
```

In [124]:

```
# 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_cv = vectorizer.transform(X_cv['clean_subcategories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot_cv.shape)
```

```
['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']
Shape of matrix after one hot encodig  (26219, 30)
```

In [125]:

```
# we use count vectorizer to convert the values into one

sub_categories_one_hot_test = vectorizer.transform(X_test['clean_subcategories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot_test.shape)
```

```
['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']
Shape of matrix after one hot encodig  (21850, 30)
```

In [126]:

```
# 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)


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

In [127]:

```
#onehotencoding for school_state
one_hot_encoding_school_state_test=pd.get_dummies(X_test.school_state)
print("Shape of dataframe for school_state",  one_hot_encoding_school_state_test.shape)
```

```
Shape of dataframe for school_state (21850, 51)
```

In [128]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix_test=pd.get_dummies(X_test.teacher_prefix)

print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix_test.shape)
```

```
Shape of dataframe for teacher_prefix (21850, 5)
```

In [129]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category_test=pd.get_dummies(X_test.project_grade_category)

print("Shape of dataframe for project_grade_category",
one_hot_encoding_project_grade_category_test.shape)
```

```
Shape of dataframe for project_grade_category (21850, 4)
```

## 2.3 Make Data Model Ready: encoding eassay, and project_title

In [130]:

```
# 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 [131]:

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

```
Shape of matrix after one hot encodig  (61179, 13153)
```

In [132]:

```
# We are considering only the words which appeared in at least 10 documents(rows or projects).
```

```
text_essay_cv = vectorizer.transform(X_cv['preprocessed_essays'])
print("Shape of matrix after one hot encodig ",text_essay_cv.shape)
```

```
Shape of matrix after one hot encodig  (26219, 13153)
```

In [133]:

```
# We are considering only the words which appeared in at least 10 documents(rows or projects).

text_bow_essay_test = vectorizer.transform(X_test['preprocessed_essays'])
print("Shape of matrix after one hot encodig ",text_bow_essay_test.shape)
```

```
Shape of matrix after one hot encodig  (21850, 13153)
```

In [134]:

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

```
Shape of matrix after one hot encodig  (61179, 2276)
```

In [135]:

```
# We are considering only the words which appeared in at least 10 documents(rows or projects).

text_bow_project_title_cv= vectorizer.transform(X_cv['project_title_list'])
print("Shape of matrix after one hot encodig ",text_bow_project_title_cv.shape)
```

```
Shape of matrix after one hot encodig  (26219, 2276)
```

In [136]:

```
# We are considering only the words which appeared in at least 10 documents(rows or projects).

text_bow_project_title_test = vectorizer.transform(X_test['project_title_list'])
print("Shape of matrix after one hot encodig ",text_bow_project_title_test.shape)
```

```
Shape of matrix after one hot encodig  (21850, 2276)
```

## 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 [137]:

```
# 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 [138]:

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

In [139]:

```
# 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 :)
bow_data_matrix_train=
hstack((one_hot_encoding_school_state_train,one_hot_encoding_teacher_prefix_train,one_hot_encoding_
project_grade_category_train,categories_one_hot_train,sub_categories_one_hot_train,price_standardiz
ed_train,text_essay_train,
text_bow_project_title_train))
bow_data_matrix_train.shape
```

Out[139]:

```
(61179, 15529)
```

In [140]:

```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
# with the same hstack function we are concatenating a sparse matrix and a dense matirx :)
bow_data_matrix_test=
hstack((one_hot_encoding_school_state_test,one_hot_encoding_teacher_prefix_test,one_hot_encoding_pr
oject_grade_category_test.shape,categories_one_hot_test,sub_categories_one_hot_test,price_standardi
zed_test,text_bow_essay_test,text_bow_project_title_test))
bow_data_matrix_test.shape
```

Out[140]:

```
(21850, 15529)
```

In [141]:

```
# 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 :)
bow_data_matrix_cv= hstack((one_hot_encoding_school_state_cv,one_hot_encoding_teacher_prefix_cv,on
e_hot_encoding_project_grade_category_cv,categories_one_hot_cv,sub_categories_one_hot_cv,price_star
dardized_cv,text_essay_cv,text_bow_project_title_cv))
bow_data_matrix_cv.shape
```

Out[141]:

```
(26219, 15529)
```

In [142]:

```
y_train.shape
```

Out[142]:

```
(61179,)
```

In [143]:

```
from scipy.sparse import coo_matrix
m = coo_matrix(bow_data_matrix_train)
m1 = m.tocsr()
```

In [150]:

```
new_bow_data_matrix_train=m1[:60001]
```

In [151]:

```
new_y_train=y_train[:60001]
```

In [162]:

```
from scipy.sparse import coo_matrix
m2 = coo_matrix(bow_data_matrix_test)
m3 = m2.tocsr()
```

In [154]:

```
new_bow_data_matrix_test=m3[:20001]
```

In [155]:

```
new_y_test=y_test[:20001]
```

In [156]:

```
from scipy.sparse import coo_matrix
m4 = coo_matrix(bow_data_matrix_cv)
m5 = m4.tocsr()
```

In [157]:

```
new_bow_data_matrix_cv=m5[:20001]
```

In [158]:

```
new_y_cv=y_cv[:20001]
```

In [159]:

```
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
```

In [160]:

```
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_bow = []
cv_auc_bow = []
K = [1, 5, 10, 15, 21, 31, 41, 51, 101]
for i in K:
    neigh = KNeighborsClassifier(n_neighbors=i)
    neigh.fit(new_bow_data_matrix_train, new_y_train)
```

```
    neigh.fit(new_bow_data_matrix_train, new_y_train)

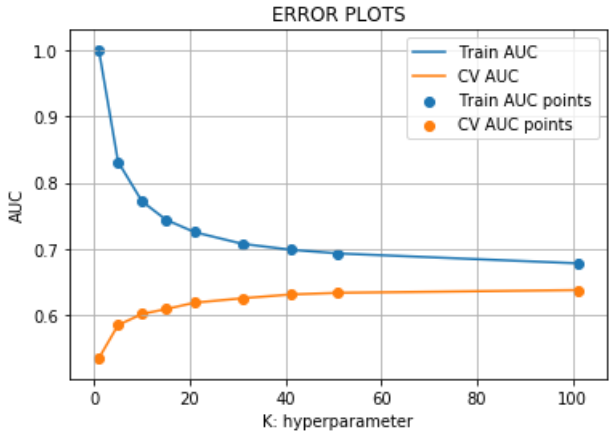
    y_train_pred_bow = batch_predict(neigh,new_bow_data_matrix_train )
    y_cv_pred_bow = batch_predict(neigh, new_bow_data_matrix_cv)

    # 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_bow.append(roc_auc_score(new_y_train,y_train_pred_bow))
    cv_auc_bow.append(roc_auc_score(new_y_cv, y_cv_pred_bow))

plt.plot(K, train_auc_bow, label='Train AUC')
plt.plot(K, cv_auc_bow, label='CV AUC')

plt.scatter(K, train_auc_bow, label='Train AUC points')
plt.scatter(K, cv_auc_bow, label='CV AUC points')

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




In [161]:

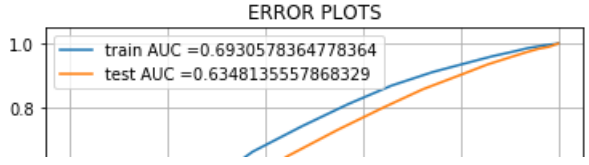
```
# https://scikit-
learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
from sklearn.metrics import roc_curve, auc

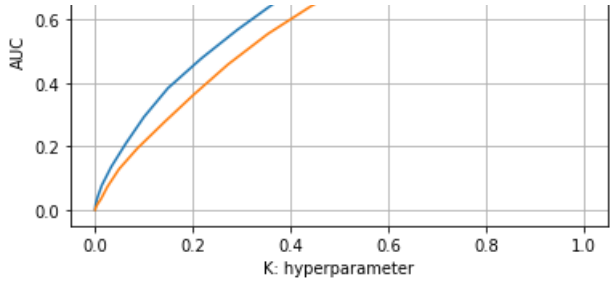

neigh = KNeighborsClassifier(n_neighbors=51)
neigh.fit(new_bow_data_matrix_train, new_y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs

y_train_pred_bow = batch_predict(neigh, new_bow_data_matrix_train)
y_test_pred_bow = batch_predict(neigh, new_bow_data_matrix_test)

train_fpr_bow, train_tpr_bow, tr_thresholds_bow = roc_curve(new_y_train, y_train_pred_bow)
test_fpr_bow, test_tpr_bow, te_thresholds_bow = roc_curve(new_y_test, y_test_pred_bow)

plt.plot(train_fpr_bow, train_tpr_bow, label="train AUC ="+str(auc(train_fpr_bow, train_tpr_bow)))
plt.plot(test_fpr_bow, test_tpr_bow, label="test AUC ="+str(auc(test_fpr_bow, test_tpr_bow)))
plt.legend()
plt.xlabel("K: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.grid()
plt.show()
```






In [163]:

```
# 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, threshold, fpr, tpr):

    t = threshold[np.argmax(fpr*(1-tpr))]

    # (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 [164]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
df_cm=confusion_matrix(new_y_train, predict(y_train_pred_bow, tr_thresholds_bow, train_fpr_bow, tra
in_fpr_bow))

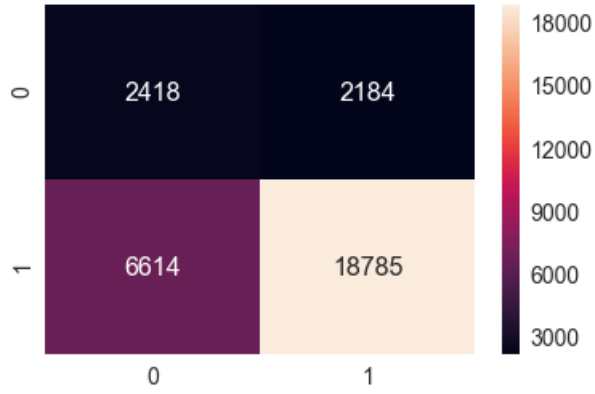

sns.set(font_scale=1.4)#for label size
sns.heatmap(df_cm, annot=True,annot_kws={"size": 16}, fmt='g')
```

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

Train confusion matrix
the maximum value of tpr*(1-fpr) 0.24935363401321456 for threshold 0.765
```

Out[164]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x256bfd69278>
```




In [165]:

```
print("Test confusion matrix")

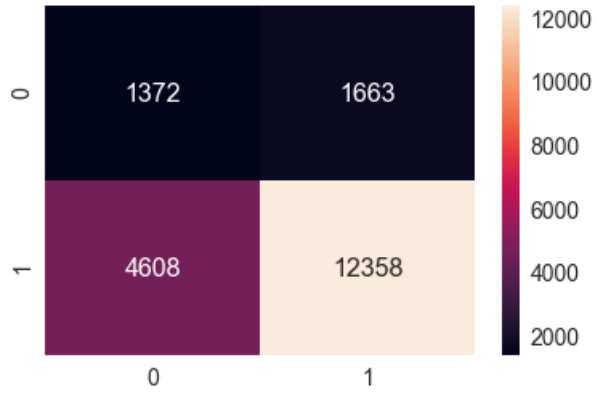
df_cm_test=confusion_matrix(new_y_test, predict(y_test_pred_bow, tr_thresholds_bow, test_fpr_bow,
test_fpr_bow))
```

```
sns.set(font_scale=1.4)#for label size
sns.heatmap(df_cm_test, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Test confusion matrix
the maximum value of tpr*(1-fpr) 0.24770169005751136 for threshold 0.765
```

Out[165]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x256b1427940>
```




### 2.4.2 Applying KNN brute force on TFIDF, SET 2

In [56]:

```
from sklearn.model_selection import train_test_split
X1_train, X_test_tfidf, y1_train, y_test_tfidf = train_test_split(
    project_data, y, test_size=0.20,stratify=y, random_state=42)
X_cv_tfidf,X_train_tfidf,y_cv_tfidf,y_train_tfidf=train_test_split(X1_train,y1_train,test_size=0.70
,stratify=y1_train,random_state=42)
```

In [57]:

```
# 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_tfidf['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_train_tfidf = price_scalar.transform(X_train_tfidf['price'].values.reshape(-1, 1
))
```

```
Mean : 297.4337254613511, Standard deviation : 356.94930693022343
```

In [58]:

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

In [59]:

```
#onehotencoding for school_state
one_hot_encoding_school_state_train_tfidf=pd.get_dummies(X_train_tfidf.school_state)
print("Shape of dataframe for school state",  one hot encoding school state train tfidf.shape)
```

```
Shape of dataframe for school_state (61179, 51)
```

In [60]:

```
#onehotencoding for school_state
one_hot_encoding_school_state_cv_tfidf=pd.get_dummies(X_cv_tfidf.school_state)
print("Shape of dataframe for school_state",  one_hot_encoding_school_state_cv_tfidf.shape)
```

```
Shape of dataframe for school_state (26219, 51)
```

In [61]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix_train_tfidf=pd.get_dummies(X_train_tfidf.teacher_prefix)

print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix_train_tfidf.shape)
```

```
Shape of dataframe for teacher_prefix (61179, 5)
```

In [62]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix_cv_tfidf=pd.get_dummies(X_cv_tfidf.teacher_prefix)

print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix_cv_tfidf.shape)
```

```
Shape of dataframe for teacher_prefix (26219, 5)
```

In [63]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category_train_tfidf=pd.get_dummies(X_train_tfidf.project_grade_cate
gory)

print("Shape of dataframe for project_grade_category",
one_hot_encoding_project_grade_category_train_tfidf.shape)
```

```
Shape of dataframe for project_grade_category (61179, 4)
```

In [64]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category_cv_tfidf=pd.get_dummies(X_cv_tfidf.project_grade_category)

print("Shape of dataframe for project_grade_category",
one_hot_encoding_project_grade_category_cv_tfidf.shape)
```

```
Shape of dataframe for project_grade_category (26219, 4)
```

In [65]:

```
# 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_train_tfidf = vectorizer.fit_transform(X_train_tfidf['clean_categories'].values
)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot_train_tfidf.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (61179, 9)
```

In [66]:

```
# we use count vectorizer to convert the values into one
vectorizer = CountVectorizer(vocabulary=list(sorted_cat_dict.keys()), lowercase=False, binary=True
)
categories_one_hot_cv_tfidf = vectorizer.transform(X_cv_tfidf['clean_categories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot_cv_tfidf.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (26219, 9)
```

In [67]:

```
categories_one_hot_test_tfidf = vectorizer.transform(X_test_tfidf['clean_categories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot_test_tfidf.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (21850, 9)
```

In [68]:

```
# 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_train_tfidf = vectorizer.fit_transform(X_train_tfidf['clean_subcategories']
.values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot_train_tfidf.shape)
```

```
['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']
Shape of matrix after one hot encodig  (61179, 30)
```

In [69]:

```
# 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_cv_tfidf = vectorizer.transform(X_cv_tfidf['clean_subcategories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot_cv_tfidf.shape)
```

```
['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']
Shape of matrix after one hot encodig  (26219, 30)
```

In [70]:

```
# we use count vectorizer to convert the values into one

sub_categories_one_hot_test_tfidf =
vectorizer.transform(X_test_tfidf['clean_subcategories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot_test_tfidf.shape)
```

```
['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']
Shape of matrix after one hot encodig  (21850, 30)
```

In [71]:

```
# 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)


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

In [72]:

```
#onehotencoding for school_state
one_hot_encoding_school_state_test_tfidf=pd.get_dummies(X_test_tfidf.school_state)
print("Shape of dataframe for school_state",  one_hot_encoding_school_state_test_tfidf.shape)
```

```
Shape of dataframe for school_state (21850, 51)
```

In [73]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix_test_tfidf=pd.get_dummies(X_test_tfidf.teacher_prefix)

print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix_test_tfidf.shape)
```

```
Shape of dataframe for teacher_prefix (21850, 5)
```

In [74]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category_test_tfidf=pd.get_dummies(X_test_tfidf.project_grade_categ
ry)

print("Shape of dataframe for project_grade_category",
one_hot_encoding_project_grade_category_test_tfidf.shape)
```

```
Shape of dataframe for project_grade_category (21850, 4)
```

In [75]:

```
vectorizer = TfidfVectorizer(min_df=10)
tfidf_essay_train = vectorizer.fit_transform(X_train_tfidf['preprocessed_essays'])
print("Shape of matrix after one hot encodig ",tfidf_essay_train.shape)
```

```
Shape of matrix after one hot encodig  (61179, 13153)
```

In [76]:

```
tfidf_essay_cv = vectorizer.transform(X_cv_tfidf['preprocessed_essays'])
print("Shape of matrix after one hot encodig ",tfidf_essay_cv.shape)
```

```
Shape of matrix after one hot encodig  (26219, 13153)
```

In [77]:

```
tfidf_essay_test = vectorizer.transform(X_test_tfidf['preprocessed_essays'])
```

```
print("Shape of matrix after one hot encodig ",tfidf_essay_test.shape)
```

```
Shape of matrix after one hot encodig  (21850, 13153)
```

In [78]:

```
vectorizer = TfidfVectorizer(min_df=10)
tfidf_project_title_train = vectorizer.fit_transform(X_train_tfidf['project_title_list'])
print("Shape of matrix after one hot encodig ",tfidf_project_title_train.shape)
```

```
Shape of matrix after one hot encodig  (61179, 2276)
```

In [79]:

```
tfidf_project_title_cv = vectorizer.transform(X_cv_tfidf['project_title_list'])
print("Shape of matrix after one hot encodig ",tfidf_project_title_cv.shape)
```

```
Shape of matrix after one hot encodig  (26219, 2276)
```

In [80]:

```
tfidf_project_title_test = vectorizer.transform(X_test_tfidf['project_title_list'])
print("Shape of matrix after one hot encodig ",tfidf_project_title_test.shape)
```

```
Shape of matrix after one hot encodig  (21850, 2276)
```

In [81]:

```
# 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 :)
tfidf_data_matrix_train=
hstack((one_hot_encoding_school_state_train_tfidf,one_hot_encoding_teacher_prefix_train_tfidf,one_
hot_encoding_project_grade_category_train_tfidf,categories_one_hot_train_tfidf,sub_categories_one_h
ot_train_tfidf,price_standardized_train_tfidf,tfidf_essay_train,tfidf_project_title_train))
tfidf_data_matrix_train.shape
```

Out[81]:

```
(61179, 15529)
```

In [82]:

```
# 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 :)
tfidf_data_matrix_test=
hstack((one_hot_encoding_school_state_test_tfidf,one_hot_encoding_teacher_prefix_test_tfidf,one_ho
t_encoding_project_grade_category_test_tfidf,categories_one_hot_test_tfidf,sub_categories_one_hot_t
est_tfidf,price_standardized_test_tfidf,tfidf_essay_test,tfidf_project_title_test))
tfidf_data_matrix_test.shape
```

Out[82]:

```
(21850, 15529)
```

In [83]:

```
# 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 :)
tfidf_data_matrix_cv=
hstack((one_hot_encoding_school_state_cv_tfidf,one_hot_encoding_teacher_prefix_cv_tfidf,one_hot_enc
oding_project_grade_category_cv_tfidf,categories_one_hot_cv_tfidf,sub_categories_one_hot_cv_tfidf,
price_standardized_cv_tfidf,tfidf_essay_cv,tfidf_project_title_cv))
tfidf_data_matrix_cv.shape
```

Out[83]:

Out[83]:

(26219, 15529)

In [84]:

```
from scipy.sparse import coo_matrix
m = coo_matrix(tfidf_data_matrix_train)
m1 = m.tocsr()
```

In [204]:

```
new_tfidf_data_matrix_train=m1[:60001]
```

In [205]:

```
new_y_train_tfidf=y_train_tfidf[:60001]
```

In [87]:

```
from scipy.sparse import coo_matrix
m2 = coo_matrix(tfidf_data_matrix_test)
m3 = m2.tocsr()
```

In [88]:

```
new_tfidf_data_matrix_test=m3[:20001]
```

In [89]:

```
new_y_test_tfidf=y_test_tfidf[:20001]
```

In [90]:

```
new_y_test_tfidf.shape
```

Out[90]:

(20001,)

In [91]:

```
from scipy.sparse import coo_matrix
m4 = coo_matrix(tfidf_data_matrix_cv)
m5 = m4.tocsr()
```

In [92]:

```
new_tfidf_data_matrix_cv=m5[:20001]
```

In [93]:

```
new_y_cv_tfidf=y_cv_tfidf[:20001]
```

In [211]:

```
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
```

In [212]:

```
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_tfidf = []
cv_auc_tfidf = []
K = [1, 5, 10, 15, 21, 31, 41, 51, 101]
for i in K:
    neigh = KNeighborsClassifier(n_neighbors=i)
    neigh.fit(new_tfidf_data_matrix_train, new_y_train_tfidf)

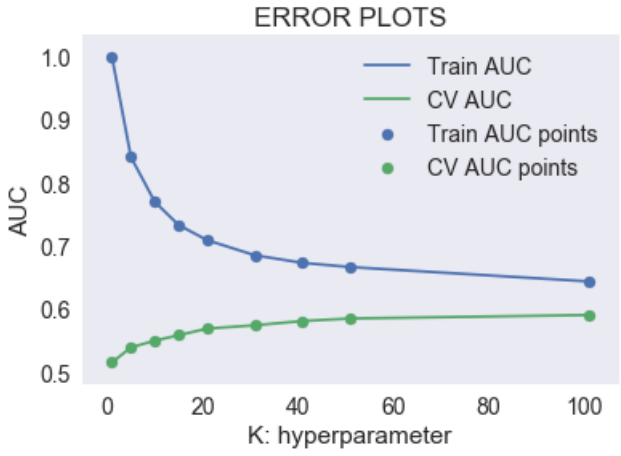
    y_train_pred_tfidf = batch_predict(neigh,new_tfidf_data_matrix_train )
    y_cv_pred_tfidf = batch_predict(neigh, new_tfidf_data_matrix_cv)

    # 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_tfidf.append(roc_auc_score(new_y_train_tfidf,y_train_pred_tfidf))
    cv_auc_tfidf.append(roc_auc_score(new_y_cv_tfidf, y_cv_pred_tfidf))

plt.plot(K, train_auc_tfidf, label='Train AUC')
plt.plot(K, cv_auc_tfidf, label='CV AUC')

plt.scatter(K, train_auc_tfidf, label='Train AUC points')
plt.scatter(K, cv_auc_tfidf, label='CV AUC points')

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




In [213]:

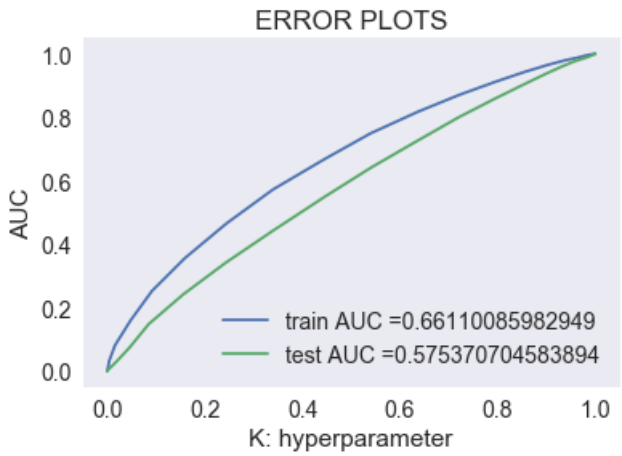
```
# https://scikit-
learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
from sklearn.metrics import roc_curve, auc
```

```
neigh = KNeighborsClassifier(n_neighbors=57)
neigh.fit(new_tfidf_data_matrix_train, new_y_train_tfidf)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs

y_train_pred_tfidf = batch_predict(neigh, new_tfidf_data_matrix_train)
y_test_pred_tfidf = batch_predict(neigh, new_tfidf_data_matrix_test)

train_fpr_tfidf, train_tpr_tfidf, tr_threshold_tfidf = roc_curve(new_y_train_tfidf,
y_train_pred_tfidf)
test_fpr_tfidf, test_tpr_tfidf, te_thresholds_tfidf = roc_curve(new_y_test_tfidf, y_test_pred_tfidf
)

plt.plot(train_fpr_tfidf, train_tpr_tfidf, label="train AUC
="+str(auc(train_fpr_tfidf,train_tpr_tfidf)))
plt.plot(test_fpr_tfidf, test_tpr_tfidf, label="test AUC ="+str(auc(test_fpr_tfidf, test_tpr_tfidf)
))
plt.legend()
plt.xlabel("K: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.grid()
plt.show()
```




In [214]:

```
# 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, threshold, fpr, tpr):

    t = threshold[np.argmax(fpr*(1-tpr))]

    # (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 [215]:

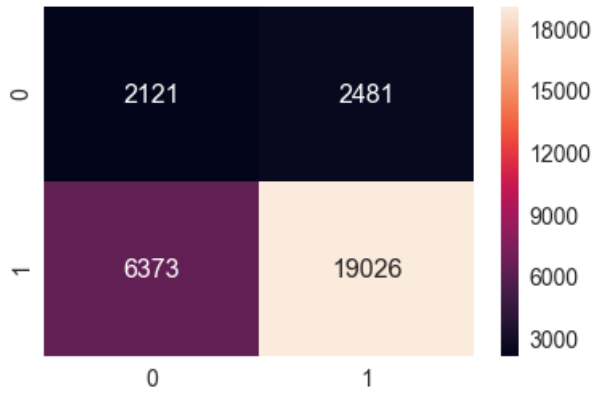
```
from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
df_cm_train=confusion_matrix(new_y_train_tfidf, predict(y_train_pred_tfidf, tr_threshold_tfidf,
train_fpr_tfidf, train_fpr_tfidf))


sns.set(font_scale=1.4)#for label size
sns.heatmap(df_cm_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Train confusion matrix
the maximum value of tpr*(1-fpr) 0.24847013967624756 for threshold 0.825
```

Out[215]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x256b845a080>
```




In [216]:

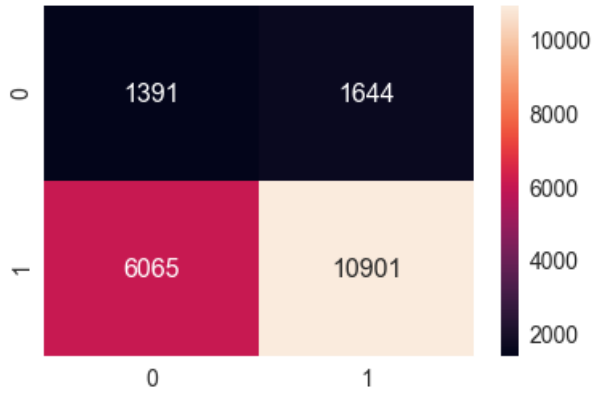
```
print("Test confusion matrix")

df_cm_test=confusion_matrix(new_y_test_tfidf, predict(y_test_pred_tfidf, tr_threshold_tfidf,
test_fpr_tfidf, test_fpr_tfidf))
sns.set(font_scale=1.4)#for label size
sns.heatmap(df_cm_test, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Test confusion matrix
the maximum value of tpr*(1-fpr) 0.24826274464037085 for threshold 0.842
```

Out[216]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x256b7b0fda0>
```




## 2.4.3 Applying KNN brute force on AVG W2V, SET 3

In [96]:

```
from sklearn.model_selection import train_test_split
X1_train, X_test_avg, y1_train, y_test_avg = train_test_split(
    project_data, y, test_size=0.20,stratify=y, random_state=42)

X_train_avg, X_cv_avg, y_train_avg, y_cv_avg = train_test_split(
   X1_train ,  y1_train, test_size=0.70,stratify=y1_train, random_state=42)
```

In [97]:

```
# average Word2Vec
# compute average word2vec for each review.
```

```
avg_w2v_vectors_essay_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_train_avg['preprocessed_essays'].values): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    cnt_words =0; # num of words with a valid vector in the sentence/review
    for word in sentence.split(): # for each word in a review/sentence
        if word in glove_words:
            vector += model[word]
            cnt_words += 1
    if cnt_words != 0:
        vector /= cnt_words
    avg_w2v_vectors_essay_train.append(vector)

print(len(avg_w2v_vectors_essay_train))
print(len(avg_w2v_vectors_essay_train[0]))
```

```
100%|
| 26219/26219
[00:18<00:00, 1422.23it/s]
```

```
26219
300
```

In [98]:

```
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_essay_cv = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_cv_avg['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_essay_cv.append(vector)

print(len(avg_w2v_vectors_essay_cv))
print(len(avg_w2v_vectors_essay_cv[0]))
```

```
100%|
| 61179/61179
[00:42<00:00, 1425.44it/s]
```

```
61179
300
```

In [99]:

```
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_essay_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_test_avg['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_essay_test.append(vector)

print(len(avg_w2v_vectors_essay_test))
print(len(avg_w2v_vectors_essay_test[0]))
```

```
100%|
| 21850/21850
```

```
[00:16<00:00, 1353.28it/s]
```

```
21850
300
```

In [100]:

```
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_project_title_train = []; # the avg-w2v for each sentence/review is stored in this
list
for sentence in tqdm(X_train_avg['project_title_list']): # 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_project_title_train.append(vector)

print(len(avg_w2v_vectors_project_title_train))
print(len(avg_w2v_vectors_project_title_train[0]))
```

```
100%|
| 26219/26219
[00:01<00:00, 23971.28it/s]
```

```
26219
300
```

In [101]:

```
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_project_title_cv = []; # the avg-w2v for each sentence/review is stored in this li
st
for sentence in tqdm(X_cv_avg['project_title_list']): # 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_project_title_cv.append(vector)

print(len(avg_w2v_vectors_project_title_cv))
print(len(avg_w2v_vectors_project_title_cv[0]))
```

```
100%|
| 61179/61179
[00:02<00:00, 21578.35it/s]
```

```
61179
300
```

In [102]:

```
# average Word2Vec
# compute average word2vec for each review.
avg_w2v_vectors_project_title_test = []; # the avg-w2v for each sentence/review is stored in this
list
for sentence in tqdm(X_test_avg['project_title_list']): # 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:
```

```
            if word in glove_words:
                vector += model[word]
                cnt_words += 1
    if cnt_words != 0:
        vector /= cnt_words
    avg_w2v_vectors_project_title_test.append(vector)

print(len(avg_w2v_vectors_project_title_test))
print(len(avg_w2v_vectors_project_title_test[0]))
```

```
100%|
| 21850/21850
[00:00<00:00, 25216.27it/s]
```

```
21850
300
```

In [103]:

```
# 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_avg['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_train_avg = price_scalar.transform(X_train_avg['price'].values.reshape(-1, 1))
```

```
Mean : 300.0371013387238, Standard deviation : 382.31744805973403
```

In [104]:

```
price_standardized_cv_avg = price_scalar.transform(X_cv_avg['price'].values.reshape(-1, 1))
```

In [105]:

```
price_standardized_test_avg = price_scalar.transform(X_test_avg['price'].values.reshape(-1, 1))
```

In [106]:

```
#onehotencoding for school_state
one_hot_encoding_school_state_train_avg=pd.get_dummies(X_train_avg.school_state)
print("Shape of dataframe for school_state",  one_hot_encoding_school_state_train_avg.shape)
```

```
Shape of dataframe for school_state (26219, 51)
```

In [107]:

```
#onehotencoding for school_state
one_hot_encoding_school_state_cv_avg=pd.get_dummies(X_cv_avg.school_state)
print("Shape of dataframe for school_state",  one_hot_encoding_school_state_cv_avg.shape)
```

```
Shape of dataframe for school_state (61179, 51)
```

In [108]:

```
#onehotencoding for school_state
one_hot_encoding_school_state_test_avg=pd.get_dummies(X_test_avg.school_state)
```

```
print("Shape of dataframe for school_state",  one_hot_encoding_school_state_test_avg.shape)
```

```
Shape of dataframe for school_state (21850, 51)
```

In [109]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix_train_avg=pd.get_dummies(X_train_avg.teacher_prefix)

print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix_train_avg.shape)
```

```
Shape of dataframe for teacher_prefix (26219, 5)
```

In [110]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix_cv_avg=pd.get_dummies(X_cv_avg.teacher_prefix)

print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix_cv_avg.shape)
```

```
Shape of dataframe for teacher_prefix (61179, 5)
```

In [111]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix_test_avg=pd.get_dummies(X_test_avg.teacher_prefix)

print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix_test_avg.shape)
```

```
Shape of dataframe for teacher_prefix (21850, 5)
```

In [112]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category_train_avg=pd.get_dummies(X_train_avg.project_grade_category
)

print("Shape of dataframe for project_grade_category",
one_hot_encoding_project_grade_category_train_avg.shape)
```

```
Shape of dataframe for project_grade_category (26219, 4)
```

In [113]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category_cv_avg=pd.get_dummies(X_cv_avg.project_grade_category)

print("Shape of dataframe for project_grade_category",
one_hot_encoding_project_grade_category_cv_avg.shape)
```

```
Shape of dataframe for project_grade_category (61179, 4)
```

In [114]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category_test_avg=pd.get_dummies(X_test_avg.project_grade_category)

print("Shape of dataframe for project_grade_category",
one_hot_encoding_project_grade_category_test_avg.shape)
```

```
Shape of dataframe for project_grade_category (21850, 4)
```

In [115]:

```
# 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_train_avg = vectorizer.fit_transform(X_train_avg['clean_categories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot_train_avg.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (26219, 9)
```

In [116]:

```
vectorizer = CountVectorizer(vocabulary=list(sorted_cat_dict.keys()), lowercase=False, binary=True
)
categories_one_hot_cv_avg = vectorizer.transform(X_cv_avg['clean_categories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot_cv_avg.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (61179, 9)
```

In [117]:

```
categories_one_hot_test_avg = vectorizer.transform(X_test_avg['clean_categories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot_test_avg.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (21850, 9)
```

In [118]:

```
# 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_train_avg =
vectorizer.fit_transform(X_train_avg['clean_subcategories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot_train_avg.shape)
```

```
['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']
Shape of matrix after one hot encodig  (26219, 30)
```

In [119]:

```
# 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_cv_avg = vectorizer.transform(X_cv_avg['clean_subcategories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot_cv_avg.shape)
```

```
['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']
Shape of matrix after one hot encodig  (61179, 30)
```

In [120]:

```
# we use count vectorizer to convert the values into one

sub_categories_one_hot_test_avg =
vectorizer.fit_transform(X_test_avg['clean_subcategories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot_test_avg.shape)
```

```
['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']
Shape of matrix after one hot encodig  (21850, 30)
```

In [121]:

```
# 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 :)
avgw2v_data_matrix_train=
hstack((one_hot_encoding_school_state_train_avg,one_hot_encoding_teacher_prefix_train_avg,one_hot_e
ncoding_project_grade_category_train_avg,categories_one_hot_train_avg,sub_categories_one_hot_train_
avg,price_standardized_train_avg,avg_w2v_vectors_essay_train,avg_w2v_vectors_project_title_train))
avgw2v_data_matrix_train.shape
```

Out[121]:

```
(26219, 700)
```

In [122]:

```
# 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 :)
avgw2v_data_matrix_cv=
hstack((one_hot_encoding_school_state_cv_avg,one_hot_encoding_teacher_prefix_cv_avg,one_hot_encodir
g_project_grade_category_cv_avg,categories_one_hot_cv_avg,sub_categories_one_hot_cv_avg,price_stanc
ardized_cv_avg,avg_w2v_vectors_essay_cv,avg_w2v_vectors_project_title_cv))
avgw2v_data_matrix_cv.shape
```

Out[122]:

```
(61179, 700)
```

In [123]:

```
# 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 :)
avgw2v_data_matrix_test=
hstack((one_hot_encoding_school_state_test_avg,one_hot_encoding_teacher_prefix_test_avg,one_hot_enc
oding_project_grade_category_test_avg,categories_one_hot_test_avg,sub_categories_one_hot_test_avg,
price_standardized_test_avg,avg_w2v_vectors_essay_test,avg_w2v_vectors_project_title_test))
avgw2v_data_matrix_test.shape
```

Out[123]:

```
(21850, 700)
```

In [133]:

```
from scipy.sparse import coo_matrix
n = coo_matrix(avgw2v_data_matrix_train)
n1 = n.tocsr()
```

In [134]:

```
new_avgw2v_data_matrix_train=n1[:10001]
```

In [135]:

```
new_y_train_avgw2v=y_train_avg[:10001]
```

In [136]:

```
from scipy.sparse import coo_matrix
n4 = coo_matrix(avgw2v_data_matrix_cv)
n5 = n4.tocsr()
```

In [137]:

```
new_avgw2v_data_matrix_cv=n5[:10001]
```

In [138]:

```
new_y_cv_avgw2v=y_cv_avg[:10001]
```

In [139]:

```
from scipy.sparse import coo_matrix
n2 = coo_matrix(avgw2v_data_matrix_test)
n3 = n2.tocsr()
```

In [140]:

```
new_avgw2v_data_matrix_test=n3[:10001]
```

In [141]:

```
new_y_test_avgw2v=y_test_avg[:10001]
```

In [142]:

```
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
```

In [143]:

```
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_avgw2v = []
cv_auc_avgw2v = []
```

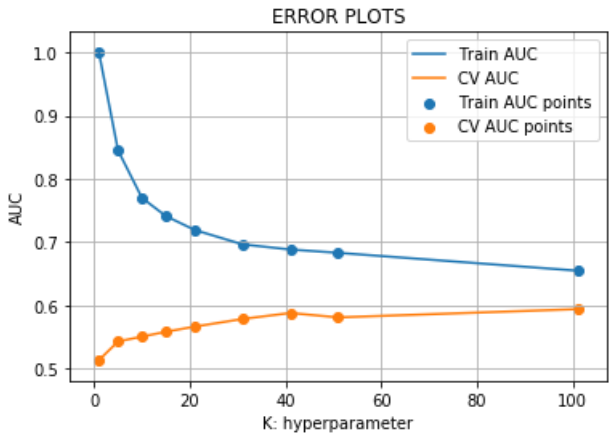
```
K = [1, 5, 10, 15, 21, 31, 41, 51, 101]
for i in K:
    neigh = KNeighborsClassifier(n_neighbors=i)
    neigh.fit(new_avgw2v_data_matrix_train, new_y_train_avgw2v)
    y_train_pred_avgw2v = batch_predict(neigh,new_avgw2v_data_matrix_train )
    y_cv_pred_avgw2v = batch_predict(neigh, new_avgw2v_data_matrix_cv)

    # 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_avgw2v.append(roc_auc_score(new_y_train_avgw2v,y_train_pred_avgw2v))
    cv_auc_avgw2v.append(roc_auc_score(new_y_cv_avgw2v, y_cv_pred_avgw2v))

plt.plot(K, train_auc_avgw2v, label='Train AUC')
plt.plot(K, cv_auc_avgw2v, label='CV AUC')

plt.scatter(K, train_auc_avgw2v, label='Train AUC points')
plt.scatter(K, cv_auc_avgw2v, label='CV AUC points')

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




In [144]:

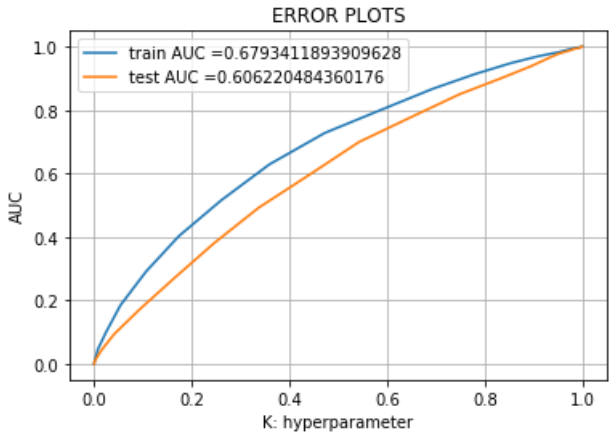
```
# https://scikit-
learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
from sklearn.metrics import roc_curve, auc


neigh = KNeighborsClassifier(n_neighbors=55)
neigh.fit(new_avgw2v_data_matrix_train, new_y_train_avgw2v)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs

y_train_pred_avgw2v = batch_predict(neigh, new_avgw2v_data_matrix_train)
y_test_pred_avgw2v = batch_predict(neigh, new_avgw2v_data_matrix_test)

train_fpr_avgw2v, train_tpr_avgw2v, tr_threshold_avgw2v = roc_curve(new_y_train_avgw2v, y_train_pre
d_avgw2v)
test_fpr_avgw2v, test_tpr_avgw2v, te_thresholds_avgw2v = roc_curve(new_y_test_avgw2v,
y_test_pred_avgw2v)

plt.plot(train_fpr_avgw2v, train_tpr_avgw2v, label="train AUC ="+str(auc(train_fpr_avgw2v,train_tpr
_avgw2v)))
plt.plot(test_fpr_avgw2v, test_tpr_avgw2v, label="test AUC ="+str(auc(test_fpr_avgw2v,
test_tpr_avgw2v)))
plt.legend()
plt.xlabel("K: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.grid()
plt.show()
```




In [145]:

```
# 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, threshold, fpr, tpr):

    t = threshold[np.argmax(fpr*(1-tpr))]

    # (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 [146]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
df_cm_train=confusion_matrix(new_y_train_avgw2v, predict(y_train_pred_avgw2v, tr_threshold_avgw2v,
train_fpr_avgw2v, train_fpr_avgw2v))

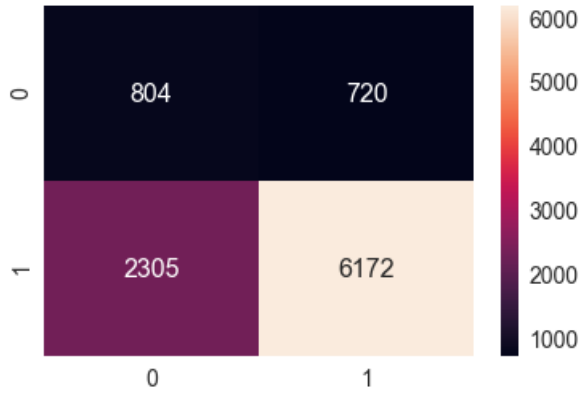

sns.set(font_scale=1.4)#for label size
sns.heatmap(df_cm_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

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

Train confusion matrix
the maximum value of tpr*(1-fpr) 0.24924049848099694 for threshold 0.836
```

Out[146]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1b316279160>
```



In [147]:

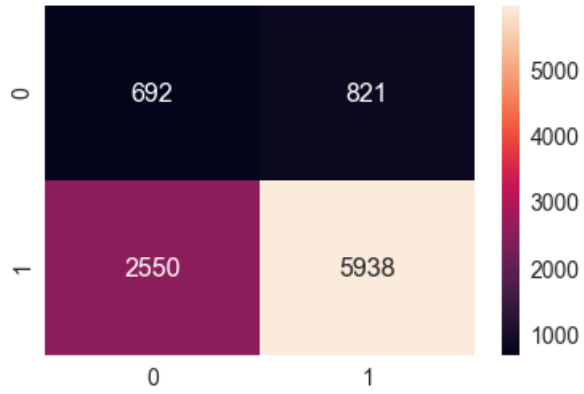
```
print("Test confusion matrix")

df_cm_test=confusion_matrix(new_y_test_avgw2v, predict(y_test_pred_avgw2v, tr_threshold_avgw2v,
test_fpr_avgw2v, test_fpr_avgw2v))
sns.set(font_scale=1.4)#for label size
sns.heatmap(df_cm_test, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Test confusion matrix
the maximum value of tpr*(1-fpr) 0.24818263745490174 for threshold 0.836
```

Out[147]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1b3161f7160>
```




### 2.4.4 Applying KNN brute force on TFIDF W2V, SET 4

In [149]:

```
from sklearn.model_selection import train_test_split
X1_train, X_test_tfidf_w2v, y1_train, y_test_tfidf_w2v = train_test_split(
    project_data, y, test_size=0.20,stratify=y, random_state=42)
X_cv_tfidf_w2v,X_train_tfidf_w2v,y_cv_tfidf_w2v,y_train_tfidf_w2v=train_test_split(X1_train,y1_trai
n,test_size=0.70,stratify=y1_train,random_state=42)
```

In [150]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train_tfidf_w2v['project_title_list'].values)
# 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 [151]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_project_title_train = []; # the avg-w2v for each sentence/review is stored in th
is list
for sentence in tqdm(X_train_tfidf_w2v['project_title_list'].values): # 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_project_title_train.append(vector)

print(len(tfidf_w2v_vectors_project_title_train))
print(len(tfidf_w2v_vectors_project_title_train[0]))
```

```
100%|
61179/61179 [00:04<00:00, 13583.19it/s]
```

```
61179
300
```

In [152]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_cv_tfidf_w2v['project_title_list'].values)
# 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 [153]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_project_title_cv = []; # the avg-w2v for each sentence/review is stored in this
list
for sentence in tqdm(X_cv_tfidf_w2v['project_title_list'].values): # 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_project_title_cv.append(vector)

print(len(tfidf_w2v_vectors_project_title_cv))
print(len(tfidf_w2v_vectors_project_title_cv[0]))
```

```
100%|
26219/26219 [00:02<00:00, 11496.08it/s]
```

```
26219
300
```

In [154]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_test_tfidf_w2v['project_title_list'].values)
# 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 [155]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_project_title_test = []; # the avg-w2v for each sentence/review is stored in thi
s list
for sentence in tqdm(X_test_tfidf_w2v['project_title_list'].values): # 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_project_title_test.append(vector)

print(len(tfidf_w2v_vectors_project_title_test))
print(len(tfidf_w2v_vectors_project_title_test[0]))
```

```
100%|
21850/21850 [00:01<00:00, 14095.76it/s]
```

```
21850
300
```

In [156]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train_tfidf_w2v['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 [157]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_preprocessed_essays_train = []; # the avg-w2v for each sentence/review is stored
in this list
for sentence in tqdm(X_train_tfidf_w2v['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_preprocessed_essays_train.append(vector)
print(len(tfidf_w2v_vectors_preprocessed_essays_train))
print(len(tfidf_w2v_vectors_project_title_train[0]))
```

```
100%|
61179/61179 [03:56<00:00, 258.21it/s]
```

```
61179
300
```

In [158]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_cv_tfidf_w2v['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 [159]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_preprocessed_essays_cv = []; # the avg-w2v for each sentence/review is stored in
this list
for sentence in tqdm(X_cv_tfidf_w2v['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_preprocessed_essays_cv.append(vector)
print(len(tfidf_w2v_vectors_preprocessed_essays_cv))
print(len(tfidf_w2v_vectors_preprocessed_essays_cv[0]))
```

```
100%|
26219/26219 [01:29<00:00, 291.94it/s]
```

```
26219
300
```

In [160]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_test_tfidf_w2v['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 [161]:

```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_preprocessed_essays_test = []; # the avg-w2v for each sentence/review is stored
in this list
for sentence in tqdm(X_test_tfidf_w2v['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_preprocessed_essays_test.append(vector)
print(len(tfidf_w2v_vectors_preprocessed_essays_test))
print(len(tfidf_w2v_vectors_preprocessed_essays_test[0]))
```

```
100%|
```

```
21850/21850 [01:08<00:00, 321.29it/s]
```

```
21850
300
```

In [162]:

```
# 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_tfidf_w2v['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_train_tfidf_w2v =
price_scalar.transform(X_train_tfidf_w2v['price'].values.reshape(-1, 1))
```

```
Mean : 297.4337254613511, Standard deviation : 356.94930693022343
```

In [163]:

```
price_standardized_cv_tfidf_w2v = price_scalar.transform(X_cv_tfidf_w2v['price'].values.reshape(-1,
1))
```

In [164]:

```
price_standardized_test_tfidf_w2v =
price_scalar.transform(X_test_tfidf_w2v['price'].values.reshape(-1, 1))
```

In [165]:

```
#onehotencoding for school_state
one_hot_encoding_school_state_train_tfidf_w2v=pd.get_dummies(X_train_tfidf_w2v.school_state)
print("Shape of dataframe for school_state",  one_hot_encoding_school_state_train_tfidf_w2v.shape)
```

```
Shape of dataframe for school_state (61179, 51)
```

In [166]:

```
#onehotencoding for school_state
one_hot_encoding_school_state_cv_tfidf_w2v=pd.get_dummies(X_cv_tfidf_w2v.school_state)
print("Shape of dataframe for school_state",  one_hot_encoding_school_state_cv_tfidf_w2v.shape)
```

```
Shape of dataframe for school_state (26219, 51)
```

In [167]:

```
#onehotencoding for school_state
one_hot_encoding_school_state_test_tfidf_w2v=pd.get_dummies(X_test_tfidf_w2v.school_state)
print("Shape of dataframe for school_state",  one_hot_encoding_school_state_test_tfidf_w2v.shape)
```

```
Shape of dataframe for school_state (21850, 51)
```

In [168]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix_train_tfidf_w2v=pd.get_dummies(X_train_tfidf_w2v.teacher_prefix)
```

```
print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix_train_tfidf_w2v.sha
pe)
```

```
Shape of dataframe for teacher_prefix (61179, 5)
```

In [169]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix_cv_tfidf_w2v=pd.get_dummies(X_cv_tfidf_w2v.teacher_prefix)

print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix_cv_tfidf_w2v.shape)
```

```
Shape of dataframe for teacher_prefix (26219, 5)
```

In [170]:

```
#onehotencoding for teacher_prefix
one_hot_encoding_teacher_prefix_test_tfidf_w2v=pd.get_dummies(X_test_tfidf_w2v.teacher_prefix)

print("Shape of dataframe for teacher_prefix", one_hot_encoding_teacher_prefix_test_tfidf_w2v.shap
e)
```

```
Shape of dataframe for teacher_prefix (21850, 5)
```

In [171]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category_train_tfidf_w2v=pd.get_dummies(X_train_tfidf_w2v.project_gr
ade_category)

print("Shape of dataframe for project_grade_category",
one_hot_encoding_project_grade_category_train_tfidf_w2v.shape)
```

```
Shape of dataframe for project_grade_category (61179, 4)
```

In [172]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category_cv_tfidf_w2v=pd.get_dummies(X_cv_tfidf_w2v.project_grade_ca
tegory)

print("Shape of dataframe for project_grade_category",
one_hot_encoding_project_grade_category_cv_tfidf_w2v.shape)
```

```
Shape of dataframe for project_grade_category (26219, 4)
```

In [173]:

```
#onehotencoding for project_grade_category
one_hot_encoding_project_grade_category_test_tfidf_w2v=pd.get_dummies(X_test_tfidf_w2v.project_grad
e_category)

print("Shape of dataframe for project_grade_category",
one_hot_encoding_project_grade_category_test_tfidf_w2v.shape)
```

```
Shape of dataframe for project_grade_category (21850, 4)
```

In [174]:

```
# 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_train_tfidf_w2v = vectorizer.fit_transform(X_train_tfidf_w2v['clean_categories'
].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot_train_tfidf_w2v.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (61179, 9)
```

In [175]:

```
vectorizer = CountVectorizer(vocabulary=list(sorted_cat_dict.keys()), lowercase=False, binary=True
)
categories_one_hot_cv_tfidf_w2v = vectorizer.transform(X_cv_tfidf_w2v['clean_categories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot_cv_tfidf_w2v.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (26219, 9)
```

In [176]:

```
categories_one_hot_test_tfidf_w2v=
vectorizer.transform(X_test_tfidf_w2v['clean_categories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",categories_one_hot_test_tfidf_w2v.shape)
```

```
['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds',
'Health_Sports', 'Math_Science', 'Literacy_Language']
Shape of matrix after one hot encodig  (21850, 9)
```

In [177]:

```
# 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_train_tfidf_w2v =
vectorizer.fit_transform(X_train_tfidf_w2v['clean_subcategories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot_train_tfidf_w2v.shape)
```

```
['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']
Shape of matrix after one hot encodig  (61179, 30)
```

In [178]:

```
# 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_cv_tfidf_w2v= vectorizer.transform(X_cv_tfidf_w2v['clean_subcategories'].va
lues)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot_cv_tfidf_w2v.shape)
```

```
['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']
Shape of matrix after one hot encodig  (26219, 30)
```

In [179]:

```
# we use count vectorizer to convert the values into one

sub_categories_one_hot_test_tfidf_w2v =
```

```
vectorizer.fit_transform(X_test_tfidf_w2v['clean_subcategories'].values)
print(vectorizer.get_feature_names())
print("Shape of matrix after one hot encodig ",sub_categories_one_hot_test_tfidf_w2v.shape)
```

```
['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']
Shape of matrix after one hot encodig  (21850, 30)
```

In [180]:

```
# 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 :)
tfidf_w2v_data_matrix_train=
hstack((one_hot_encoding_school_state_train_tfidf_w2v,one_hot_encoding_teacher_prefix_train_tfidf_w
v,one_hot_encoding_project_grade_category_train_tfidf_w2v,categories_one_hot_train_tfidf_w2v,sub_ca
tegories_one_hot_train_tfidf_w2v,price_standardized_train_tfidf_w2v,tfidf_w2v_vectors_preprocessed_
ssays_train,tfidf_w2v_vectors_project_title_train))
tfidf_w2v_data_matrix_train.shape
```

Out[180]:

```
(61179, 700)
```

In [181]:

```
# 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 :)
tfidf_w2v_data_matrix_cv=
hstack((one_hot_encoding_school_state_cv_tfidf_w2v,one_hot_encoding_teacher_prefix_cv_tfidf_w2v,on
e_hot_encoding_project_grade_category_cv_tfidf_w2v,categories_one_hot_cv_tfidf_w2v,sub_categories_o
ne_hot_cv_tfidf_w2v,price_standardized_cv_tfidf_w2v,tfidf_w2v_vectors_preprocessed_essays_cv,tfidf
_w2v_vectors_project_title_cv))
tfidf_w2v_data_matrix_cv.shape
```

Out[181]:

```
(26219, 700)
```

In [182]:

```
# 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 :)
tfidf_w2v_data_matrix_test=
hstack((one_hot_encoding_school_state_test_tfidf_w2v,one_hot_encoding_teacher_prefix_test_tfidf_w2v
,one_hot_encoding_project_grade_category_test_tfidf_w2v,categories_one_hot_test_tfidf_w2v,sub_categ
ories_one_hot_test_tfidf_w2v,price_standardized_test_tfidf_w2v,tfidf_w2v_vectors_preprocessed_essay
_test,tfidf_w2v_vectors_project_title_test))
tfidf_w2v_data_matrix_test.shape
```

Out[182]:

```
(21850, 700)
```

In [183]:

```
from scipy.sparse import coo_matrix
k= coo_matrix(tfidf_w2v_data_matrix_train)
k1 = k.tocsr()
```

In [186]:

```
new_tfidf_w2v_data_matrix_train=k1[:10001]
```

In [187]:

```
new_y_train_tfidf_w2v=y_train_tfidf_w2v[:10001]
```

In [188]:

```
from scipy.sparse import coo_matrix
k4 = coo_matrix(tfidf_w2v_data_matrix_cv)
k5 = k4.tocsr()
```

In [189]:

```
new_tfidf_w2v_data_matrix_cv=k5[:10001]
```

In [190]:

```
new_y_cv_tfidf_w2v=y_cv_tfidf_w2v[:10001]
```

In [191]:

```
from scipy.sparse import coo_matrix
k2 = coo_matrix(tfidf_w2v_data_matrix_test)
k3 = k2.tocsr()
```

In [192]:

```
new_tfidf_w2v_data_matrix_test=k3[:10001]
```

In [193]:

```
new_y_test_tfidf_w2v=y_test_tfidf_w2v[:10001]
```

In [194]:

```
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
```

In [195]:

```
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_tfidf_w2v = []
cv_auc_tfidf_w2v = []
K = [1, 5, 10, 15, 21, 31, 41, 51, 101]
```

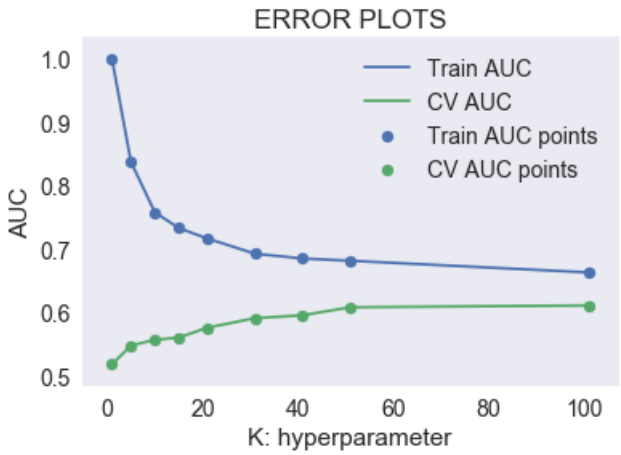
```
K = [1, 5, 10, 15, 21, 31, 41, 51, 101]
for i in K:
    neigh = KNeighborsClassifier(n_neighbors=i)
    neigh.fit(new_tfidf_w2v_data_matrix_train, new_y_train_tfidf_w2v)
    y_train_pred_tfidf_w2v = batch_predict(neigh,new_tfidf_w2v_data_matrix_train )
    y_cv_pred_tfidf_w2v = batch_predict(neigh, new_tfidf_w2v_data_matrix_cv)

    # 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_tfidf_w2v.append(roc_auc_score(new_y_train_tfidf_w2v,y_train_pred_tfidf_w2v))
    cv_auc_tfidf_w2v.append(roc_auc_score(new_y_cv_tfidf_w2v, y_cv_pred_tfidf_w2v))

plt.plot(K, train_auc_tfidf_w2v, label='Train AUC')
plt.plot(K, cv_auc_tfidf_w2v, label='CV AUC')

plt.scatter(K, train_auc_tfidf_w2v, label='Train AUC points')
plt.scatter(K, cv_auc_tfidf_w2v, label='CV AUC points')

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




In [196]:

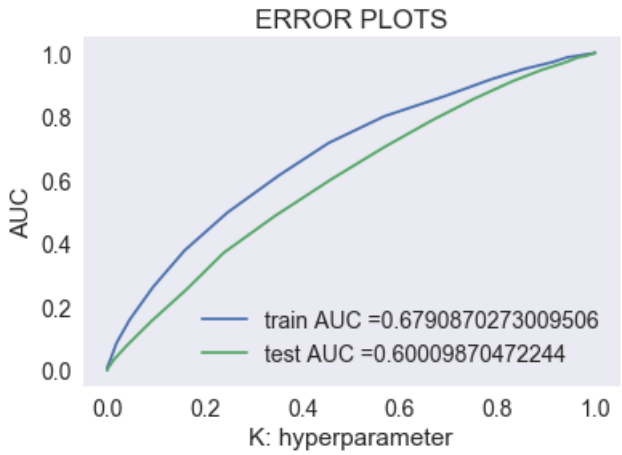
```
# https://scikit-
learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
from sklearn.metrics import roc_curve, auc


neigh = KNeighborsClassifier(n_neighbors=53)
neigh.fit(new_tfidf_w2v_data_matrix_train, new_y_train_tfidf_w2v)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs

y_train_pred_tfidf_w2v = batch_predict(neigh, new_tfidf_w2v_data_matrix_train)
y_test_pred_tfidf_w2v = batch_predict(neigh, new_tfidf_w2v_data_matrix_test)

train_fpr_tfidf_w2v, train_tpr_tfidf_w2v, tr_threshold_tfidf_w2v = roc_curve(new_y_train_tfidf_w2v
, y_train_pred_tfidf_w2v)
test_fpr_tfidf_w2v, test_tpr_tfidf_w2v, te_thresholds_tfidf_w2v = roc_curve(new_y_test_tfidf_w2v,
y_test_pred_tfidf_w2v)

plt.plot(train_fpr_tfidf_w2v, train_tpr_tfidf_w2v, label="train AUC ="+str(auc(train_fpr_tfidf_w2v,
train_tpr_tfidf_w2v)))
plt.plot(test_fpr_tfidf_w2v, test_tpr_tfidf_w2v, label="test AUC ="+str(auc(test_fpr_tfidf_w2v, tes
t_tpr_tfidf_w2v)))
plt.legend()
plt.xlabel("K: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.grid()
plt.show()
```




In [197]:

```
# 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, threshold, fpr, tpr):

    t = threshold[np.argmax(fpr*(1-tpr))]

    # (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 [198]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
df_cm_train=confusion_matrix(new_y_train_tfidf_w2v, predict(y_train_pred_tfidf_w2v,
tr_threshold_tfidf_w2v, train_fpr_tfidf_w2v, train_fpr_tfidf_w2v))

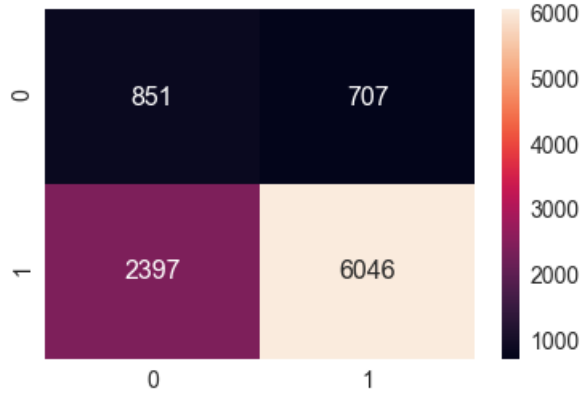

sns.set(font_scale=1.4)#for label size
sns.heatmap(df_cm_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

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

Train confusion matrix
the maximum value of tpr*(1-fpr) 0.24786434996976145 for threshold 0.83
```

Out[198]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1b3163cc8d0>
```



In [199]:

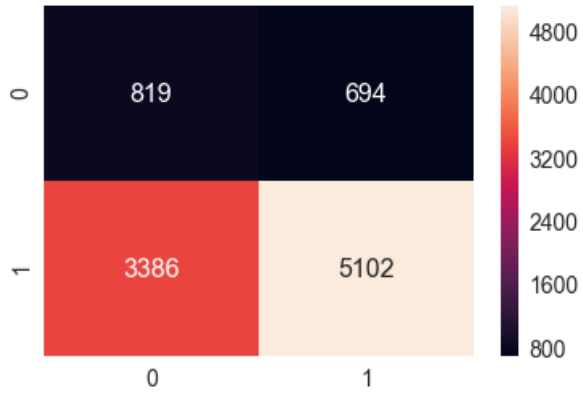
```
print("Test confusion matrix")

df_cm_test=confusion_matrix(new_y_test_tfidf_w2v, predict(y_test_pred_tfidf_w2v,
tr_threshold_tfidf_w2v, test_fpr_tfidf_w2v, test_fpr_tfidf_w2v))
sns.set(font_scale=1.4)#for label size
sns.heatmap(df_cm_test, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Test confusion matrix
the maximum value of tpr*(1-fpr) 0.24829359474988522 for threshold 0.849
```

Out[199]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1b3159c9518>
```




## 2.5 Feature selection with `SelectKBest`

In [97]:

```
# 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 [261]:

```
from sklearn.feature_selection import SelectKBest, f_classif
X_new = SelectKBest(f_classif, k=2000).fit_transform(new_tfidf_data_matrix_train,
new_y_train_tfidf)
X_new.shape
```

Out[261]:

```
(60001, 2000)
```

In [263]:

```
from sklearn.model_selection import train_test_split
X1_train, X_test_2k, y1_train, y_test_2k = train_test_split(
   X_new , new_y_train_tfidf, test_size=0.20,stratify=new_y_train_tfidf, random_state=42)

X_train_2k, X_cv_2k, y_train_2k, y_cv_2k = train_test_split(
   X1_train ,  y1_train, test_size=0.70,stratify=y1_train, random_state=42)
```

In [264]:

```
from scipy.sparse import coo_matrix
n = coo_matrix(X_train_2k)
n1 = n.tocsr()
```

In [265]:

```
new_X_train_2k=n1[:30001]
```

In [266]:

```
new_y_train_2k=y_train_2k[:30001]
```

In [277]:

```
new_y_train_2k_reshape=new_y_train_2k.values.reshape(-1,1)
```

In [267]:

```
from scipy.sparse import coo_matrix
c = coo_matrix(X_cv_2k)
c1 = c.tocsr()
```

In [268]:

```
new_X_cv_2k=X_cv_2k[:10001]
```

In [286]:

```
new_y_cv_2k=y_cv_2k[:10001]
```

In [280]:

```
new_y_cv_2k_reshape=new_y_cv_2k.values.reshape(-1,1)
```

In [270]:

```
from scipy.sparse import coo_matrix
f = coo_matrix(X_test_2k)
f1 = f.tocsr()
```

In [254]:

```
new_X_test_2k=X_test_2k[:10001]
```

In [271]:

```
new_y_test_2k= y_test_2k[:10001]
```

In [282]:

```
new_y_test_2k_reshape=new_y_test_2k.values.reshape(-1,1)
```

In [283]:

```
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
```

In [287]:

```
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.

"""

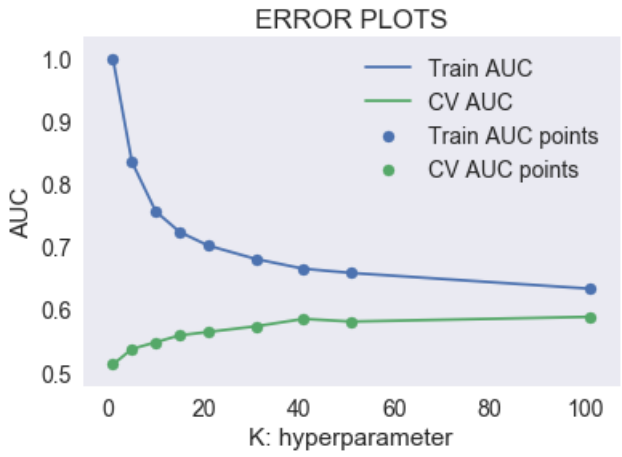
train_auc_2k = []
cv_auc_2k = []
K = [1, 5, 10, 15, 21, 31, 41, 51, 101]
for i in K:
    neigh = KNeighborsClassifier(n_neighbors=i)
    neigh.fit(new_X_train_2k,new_y_train_2k )
    y_train_2k = batch_predict(neigh,new_X_train_2k)
    y_cv_pred_2k = batch_predict(neigh,new_X_cv_2k)

    # 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_2k.append(roc_auc_score(new_y_train_2k,y_train_2k))
    cv_auc_2k.append(roc_auc_score(new_y_cv_2k, y_cv_pred_2k))

plt.plot(K, train_auc_2k, label='Train AUC')
plt.plot(K, cv_auc_2k, label='CV AUC')

plt.scatter(K, train_auc_2k, label='Train AUC points')
plt.scatter(K, cv_auc_2k, label='CV AUC points')

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




In [290]:

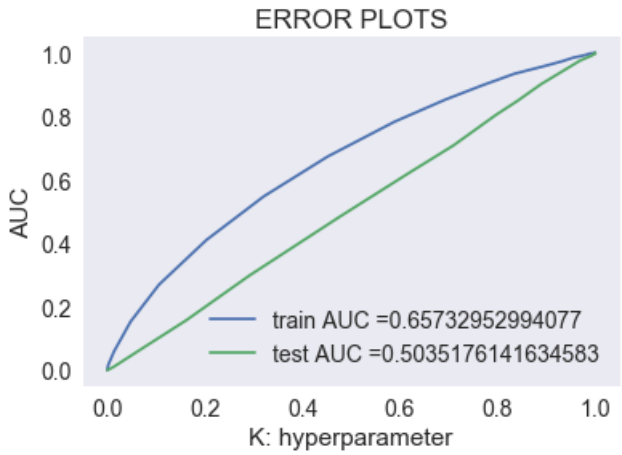
```
# https://scikit-
learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
from sklearn.metrics import roc_curve, auc
```

```
neigh = KNeighborsClassifier(n_neighbors=53)
neigh.fit(new_X_train_2k, new_y_train_2k)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs

y_train_pred_2k = batch_predict(neigh, new_X_train_2k)
y_test_pred_2k = batch_predict(neigh,new_X_test_2k )

train_fpr_2k, train_tpr_2k, tr_threshold_2k = roc_curve(new_y_train_2k, y_train_pred_2k)
test_fpr_2k, test_tpr_2k, te_thresholds_2k = roc_curve(new_y_test_2k, y_test_pred_2k)

plt.plot(train_fpr_2k, train_tpr_2k, label="train AUC ="+str(auc(train_fpr_2k,train_tpr_2k)))
plt.plot(test_fpr_2k, test_tpr_2k, label="test AUC ="+str(auc(test_fpr_2k, test_tpr_2k)))
plt.legend()
plt.xlabel("K: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.grid()
plt.show()
```




In [291]:

```
# 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, threshold, fpr, tpr):

    t = threshold[np.argmax(fpr*(1-tpr))]

    # (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 [293]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
df_cm_train=confusion_matrix(new_y_train_2k, predict(y_train_pred_2k, tr_threshold_2k,
train_fpr_2k, train_fpr_2k))


sns.set(font_scale=1.4)#for label size
sns.heatmap(df_cm_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

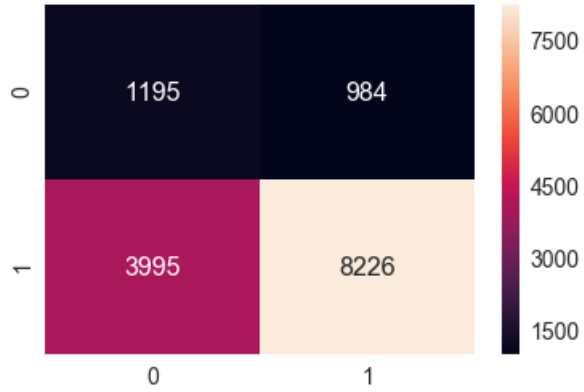
```
====================================================================================================

Train confusion matrix
```

```
the maximum value of tpr*(1-fpr) 0.24765582268560865 for threshold 0.849
```

Out[293]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1b3142f09e8>
```




In [294]:

```
print("="*100)
from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
df_cm_train=confusion_matrix(new_y_train_2k, predict(y_train_pred_2k, tr_threshold_2k,
train_fpr_2k, train_fpr_2k))

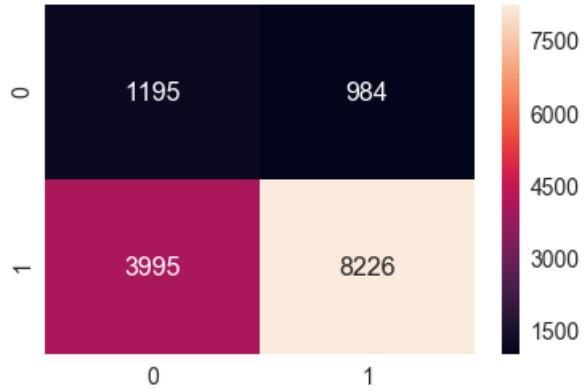

sns.set(font_scale=1.4)#for label size
sns.heatmap(df_cm_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

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

Train confusion matrix
the maximum value of tpr*(1-fpr) 0.24765582268560865 for threshold 0.849
```

Out[294]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1b31653cd30>
```




In [297]:

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

#If you get a ModuleNotFoundError error , install prettytable using: pip3 install prettytable

x = PrettyTable()
x.field_names = ["Featurization","train_auc","test_auc","threshold_for train","tpr*(1-fpr) for
train","threshold_for_test","tpr*(1-fpr) for test"  ,]
x.add_row(["BOW",0.6930,0.6348,0.765,0.24935,0.765,0.24770 ])
x.add_row(["TFIDF",0.6611,0.5753,0.825,0.24847,0.842,0.2482 ])
x.add_row(["AVG_W2V", 0.6793,0.6062,0.83,0.24920,0.84,0.24818])
x.add_row(["TFIDF_W2v", 0.6790 , 0.6000,0.83,0.24786,0.84,0.2482])
```

```
x.add_row(["Best _2k_features", 0.6573 , 0.5035,0.849,0.24765,0.849,0.24765])
print(x)
```

```
+------------------+-----------+----------+--------------------+----------------------+---------
--------+----------------------+
|   Featurization    | train_auc | test_auc | threshold_for train | tpr*(1-fpr) for train |
threshold_for_test | tpr*(1-fpr) for test |
+------------------+-----------+----------+--------------------+----------------------+---------
--------+----------------------+
|        BOW         |   0.693   |  0.6348  |        0.765        |        0.24935        |       0.
5        |        0.2477        |
|       TFIDF        |   0.6611  |  0.5753  |        0.825        |        0.24847        |       0.
2        |        0.2482        |
|      AVG_W2V       |   0.6793  |  0.6062  |         0.83        |         0.2492        |
0.84        |       0.24818        |
|     TFIDF_W2v      |   0.679   |   0.6    |         0.83        |        0.24786        |        0
4        |        0.2482        |
| Best _2k_features |   0.6573  |  0.5035  |        0.849        |        0.24765        |       0.
49        |       0.24765        |
+------------------+-----------+----------+--------------------+----------------------+---------
--------+----------------------+
```

From the above table one can infer that Bow is performing better in this kind of data set.