DonorsChoose

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

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

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

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

About the DonorsChoose Data Set

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

Feature	Description
`project_id`	A unique identifier for the proposed project. **Example:** `p036502`
`project_title`	Title of the project. **Examples:** • `Art Will Make You Happy!` • `First Grade Fun`
`project_grade_category`	Grade level of students for which the project is targeted. One of the following enumerated values: • `Grades PreK-2` • `Grades 3-5` • `Grades 6-8` • `Grades 9-12`
`project_subject_categories`	One or more (comma-separated) subject categories for the project from the following enumerated list of values: • `Applied Learning` • `Care & Hunger` • `Health & Sports` • `History & Civics` • `Literacy & Language` • `Math & Science` • `Music & The Arts` • `Special Needs` • `Warmth` **Examples:** • `Music & The Arts` • `Literacy & Language, Math & Science`
`school_state`	State where school is located (Two-letter U.S. postal code). **Example:** `WY`
`project_subject_subcategories`	One or more (comma-separated) subject subcategories for the project. **Examples:** • `Literacy` • `Literature & Writing, Social Sciences`
`project_resource_summary`	An explanation of the resources needed for the project. **Example:** • `My students need hands on literacy materials to manage sensory needs!`
`project_essay_1`	First application essay*
`project_essay_2`	Second application essay*
`project_essay_3`	Third application essay*

Feature	Description
`project_essay_4`	Fourth application essay
`project_submitted_datetime`	Datetime when project application was submitted. **Example:** `2016-04-28 12:43:56.245`
`teacher_id`	A unique identifier for the teacher of the proposed project. **Example:** `bdf8baa8fedef6bfeec7ae4ff1c15c56`
`teacher_prefix`	Teacher's title. One of the following enumerated values: • `nan` • `Dr.` • `Mr.` • `Mrs.` • `Ms.` • `Teacher.`
`teacher_number_of_previously_posted_projects`	Number of project applications previously submitted by the same teacher. **Example:** `2`

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

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

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

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

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

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

Notes on the Essay Data

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

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

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

- __project_essay_1:__ "Describe your students: What makes your students special? Specific details about their background, your neighborhood, and your school are all helpful."
- __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 [206]:

```
%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
```

1.1 Reading Data

In [207]:

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

In [208]:

```
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 [209]:

```
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[209]:

	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

In [210]:

```
#https://stackoverflow.com/questions/29530232/how-to-check-if-any-value-is-nan-in-a-pandas-datafra
me

project_data[project_data['teacher_prefix'].isnull()]
#Handle null values in pandas https://www.geeksforgeeks.org/python-pandas-dataframe-fillna-to-repl
ace-null-values-in-dataframe/

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

1.2 preprocessing of `project_subject_categories`

In [211]:

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

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

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

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

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

1.3 preprocessing of `project_subject_subcategories`

In [212]:

```
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 preprocessing of `project_grade_category`

In [213]:

```
grades = list(project_data['project_grade_category'].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
grade_list = []
for i in grades:
    if 'Grades' in i.split(): # this will split each of the catogory based on space "Math & Science
"=> "Math","&", "Science"
            i=i.replace(' ','_') # if we have the words "The" we are going to replace it with ''(i.
e removing 'The')
            i = i.replace('-','_') # we are placeing all the ' '(space) with ''(empty) ex:"Math & S
cience"=>"Math&Science"
    grade_list.append(i.strip())

project_data['project_grade_category'] = grade_list

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

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

1.3 Text preprocessing

In [214]:

```
# 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 [215]:

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

Out[215]:

Unnamed: 0	id	teacher_id	teacher_prefix	school_state	project_submitted_datetime	project_grade_cate
0	160221	p253737	c90749f5d961ff158d4b4d1e7dc665fc	Mrs.	IN	2016-12-05 13:43:57	Grades_Pr

Unnamed: 0	id	teacher_id	teacher_prefix	school_state	project_submitted_datetime	project_grade_cate
1	140945	p258326	897464ce9ddc600bced1151f324dd63a	Mr.	FL	2016-10-25 09:22:10	Grades

In [216]:

```
# 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)
```

```
My students are English learners that are working on English as their second or third languages. W
e are a melting pot of refugees, immigrants, and native-born Americans bringing the gift of langua
ge to our school. \r\n\r\n We have over 24 languages represented in our English Learner program wi
th students at every level of mastery.  We also have over 40 countries represented with the
families within our school.  Each student brings a wealth of knowledge and experiences to us that
open our eyes to new cultures, beliefs, and respect.\"The limits of your language are the limits o
f your world.\"-Ludwig Wittgenstein  Our English learner's have a strong support system at home th
at begs for more resources.  Many times our parents are learning to read and speak English along s
ide of their children.  Sometimes this creates barriers for parents to be able to help their child
learn phonetics, letter recognition, and other reading skills.\r\n\r\nBy providing these dvd's and
players, students are able to continue their mastery of the English language even if no one at hom
e is able to assist.  All families with students within the Level 1 proficiency status, will be a
offered to be a part of this program.  These educational videos will be specially chosen by the En
glish Learner Teacher and will be sent home regularly to watch.  The videos are to help the child
develop early reading skills.\r\n\r\nParents that do not have access to a dvd player will have the
opportunity to check out a dvd player to use for the year.  The plan is to use these videos and ed
ucational dvd's for the years to come for other EL students.\r\nnannan
==================================================
The 51 fifth grade students that will cycle through my classroom this year all love learning, at l
east most of the time. At our school, 97.3% of the students receive free or reduced price lunch. O
f the 560 students, 97.3% are minority students. \r\nThe school has a vibrant community that loves
to get together and celebrate. Around Halloween there is a whole school parade to show off the bea
utiful costumes that students wear. On Cinco de Mayo we put on a big festival with crafts made by
the students, dances, and games. At the end of the year the school hosts a carnival to celebrate t
he hard work put in during the school year, with a dunk tank being the most popular activity.My st
udents will use these five brightly colored Hokki stools in place of regular, stationary, 4-legged
chairs. As I will only have a total of ten in the classroom and not enough for each student to hav
e an individual one, they will be used in a variety of ways. During independent reading time they
will be used as special chairs students will each use on occasion. I will utilize them in place of
chairs at my small group tables during math and reading times. The rest of the day they will be us
ed by the students who need the highest amount of movement in their life in order to stay focused
on school.\r\n\r\nWhenever asked what the classroom is missing, my students always say more Hokki
Stools. They can't get their fill of the 5 stools we already have. When the students are sitting i
n group with me on the Hokki Stools, they are always moving, but at the same time doing their
work. Anytime the students get to pick where they can sit, the Hokki Stools are the first to be ta
ken. There are always students who head over to the kidney table to get one of the stools who are
disappointed as there are not enough of them. \r\n\r\nWe ask a lot of students to sit for 7 hours
a day. The Hokki stools will be a compromise that allow my students to do desk work and move at th
e same time. These stools will help students to meet their 60 minutes a day of movement by
allowing them to activate their core muscles for balance while they sit. For many of my students,
these chairs will take away the barrier that exists in schools for a child who can't sit
still.nannan
==================================================
How do you remember your days of school? Was it in a sterile environment with plain walls, rows of
desks, and a teacher in front of the room? A typical day in our room is nothing like that. I work
hard to create a warm inviting themed room for my students look forward to coming to each day.\r\n
\r\nMy class is made up of 28 wonderfully unique boys and girls of mixed races in
Arkansas.\r\nThey attend a Title I school, which means there is a high enough percentage of free a
nd reduced-price lunch to qualify. Our school is an \"open classroom\" concept, which is very uniq
ue as there are no walls separating the classrooms. These 9 and 10 year-old students are very eage
r learners; they are like sponges, absorbing all the information and experiences and keep on wanti
```

ng more.With these resources such as the comfy red throw pillows and the whimsical nautical hangin g decor and the blue fish nets, I will be able to help create the mood in our classroom setting to be one of a themed nautical environment. Creating a classroom environment is very important in the success in each and every child's education. The nautical photo props will be used with each child as they step foot into our classroom for the first time on Meet the Teacher evening. I'll take pic tures of each child with them, have them developed, and then hung in our classroom ready for their first day of 4th grade. This kind gesture will set the tone before even the first day of school! The nautical thank you cards will be used throughout the year by the students as they create thank you cards to their team groups.\r\n\r\nYour generous donations will help me to help make our classroom a fun, inviting, learning environment from day one.\r\n\r\nIt costs lost of money out of my own pocket on resources to get our classroom ready. Please consider helping with this project t o make our new school year a very successful one. Thank you!nannan
==
My kindergarten students have varied disabilities ranging from speech and language delays, cogniti ve delays, gross/fine motor delays, to autism. They are eager beavers and always strive to work th eir hardest working past their limitations. \r\n\r\nThe materials we have are the ones I seek out for my students. I teach in a Title I school where most of the students receive free or reduced pr ice lunch. Despite their disabilities and limitations, my students love coming to school and come eager to learn and explore.Have you ever felt like you had ants in your pants and you needed to gr oove and move as you were in a meeting? This is how my kids feel all the time. The want to be able to move as they learn or so they say.Wobble chairs are the answer and I love then because they dev elop their core, which enhances gross motor and in Turn fine motor skills. \r\nThey also want to l earn through games, my kids don't want to sit and do worksheets. They want to learn to count by ju mping and playing. Physical engagement is the key to our success. The number toss and color and sh ape mats can make that happen. My students will forget they are doing work and just have the fun a 6 year old deserves.nannan
==
The mediocre teacher tells. The good teacher explains. The superior teacher demonstrates. The grea t teacher inspires. -William A. Ward\r\n\r\nMy school has 803 students which is makeup is 97.6% Af rican-American, making up the largest segment of the student body. A typical school in Dallas is m ade up of 23.2% African-American students. Most of the students are on free or reduced lunch. We a ren't receiving doctors, lawyers, or engineers children from rich backgrounds or neighborhoods. As an educator I am inspiring minds of young children and we focus not only on academics but one smar t, effective, efficient, and disciplined students with good character.In our classroom we can util ize the Bluetooth for swift transitions during class. I use a speaker which doesn't amplify the so und enough to receive the message. Due to the volume of my speaker my students can't hear videos o r books clearly and it isn't making the lessons as meaningful. But with the bluetooth speaker my s tudents will be able to hear and I can stop, pause and replay it at any time.\r\nThe cart will all ow me to have more room for storage of things that are needed for the day and has an extra part to it I can use. The table top chart has all of the letter, words and pictures for students to learn about different letters and it is more accessible.nannan
==

In [217]:

```
# 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 [218]:

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

My kindergarten students have varied disabilities ranging from speech and language delays, cogniti ve delays, gross/fine motor delays, to autism. They are eager beavers and always strive to work th eir hardest working past their limitations. \r\n\r\nThe materials we have are the ones I seek out for my students. I teach in a Title I school where most of the students receive free or reduced pr

```
ice lunch.  Despite their disabilities and limitations, my students love coming to school and come
eager to learn and explore.Have you ever felt like you had ants in your pants and you needed to gr
oove and move as you were in a meeting? This is how my kids feel all the time. The want to be able
to move as they learn or so they say.Wobble chairs are the answer and I love then because they dev
elop their core, which enhances gross motor and in Turn fine motor skills. \r\nThey also want to l
earn through games, my kids do not want to sit and do worksheets. They want to learn to count by j
umping and playing. Physical engagement is the key to our success. The number toss and color and s
hape mats can make that happen. My students will forget they are doing work and just have the fun
a 6 year old deserves.nannan
==================================================
```

In [219]:

```
# \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)
```

```
My kindergarten students have varied disabilities ranging from speech and language delays, cogniti
ve delays, gross/fine motor delays, to autism. They are eager beavers and always strive to work th
eir hardest working past their limitations.     The materials we have are the ones I seek out for
my students. I teach in a Title I school where most of the students receive free or reduced price
lunch.  Despite their disabilities and limitations, my students love coming to school and come eag
er to learn and explore.Have you ever felt like you had ants in your pants and you needed to groov
e and move as you were in a meeting? This is how my kids feel all the time. The want to be able to
move as they learn or so they say.Wobble chairs are the answer and I love then because they develo
p their core, which enhances gross motor and in Turn fine motor skills.   They also want to learn t
hrough games, my kids do not want to sit and do worksheets. They want to learn to count by jumping
and playing. Physical engagement is the key to our success. The number toss and color and shape ma
ts can make that happen. My students will forget they are doing work and just have the fun a 6 yea
r old deserves.nannan
```

In [220]:

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

```
My kindergarten students have varied disabilities ranging from speech and language delays cognitiv
e delays gross fine motor delays to autism They are eager beavers and always strive to work their
hardest working past their limitations The materials we have are the ones I seek out for my studen
ts I teach in a Title I school where most of the students receive free or reduced price lunch
Despite their disabilities and limitations my students love coming to school and come eager to lea
rn and explore Have you ever felt like you had ants in your pants and you needed to groove and mov
e as you were in a meeting This is how my kids feel all the time The want to be able to move as th
ey learn or so they say Wobble chairs are the answer and I love then because they develop their co
re which enhances gross motor and in Turn fine motor skills They also want to learn through games
my kids do not want to sit and do worksheets They want to learn to count by jumping and playing Ph
ysical engagement is the key to our success The number toss and color and shape mats can make that
happen My students will forget they are doing work and just have the fun a 6 year old deserves nan
nan
```

In [221]:

```
# 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 [222]:

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

```
100%|██████████████████████████████████████████████████████████████████████████████| 109248/109248
[02:07<00:00, 776.83it/s]
```

In [223]:

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

1.4 Preprocessing of `project_title`

In [224]:

```
# similarly you can preprocess the titles also

from tqdm import tqdm
preprocessed_titles = []
# 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)
    sent = sent.lower()
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e not in stopwords)
    preprocessed_titles.append(sent.lower().strip())
```

```
100%|██████████████████████████████████████████████████████████████████████████████| 109248/109248
[00:06<00:00, 17937.43it/s]
```

In [225]:

```
project_data['project_title'] = preprocessed_titles
```

Removing Unnecessary Columns

In [226]:

```
#Removing unnecessary columns
# drop columns fron pandas dataframe https://stackoverflow.com/questions/13411544/delete-column-from-pandas-dataframe

project_data.drop(['project_essay_1','project_essay_2', 'project_essay_3', 'project_essay_4'], axis=1, inplace=True)
```

1.5 Preparing data for models

In [227]:

```
project_data.columns
```

Out[227]:

```
Index(['Unnamed: 0', 'id', 'teacher_id', 'teacher_prefix', 'school_state',
       'project_submitted_datetime', 'project_grade_category', 'project_title',
       'project_resource_summary',
       'teacher_number_of_previously_posted_projects', 'project_is_approved',
       'clean_categories', 'clean_subcategories', 'essay'],
      dtype='object')
```

we are going to consider

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

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

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

In [228]:

```
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')
project_data.columns
```

Out[228]:

```
Index(['Unnamed: 0', 'id', 'teacher_id', 'teacher_prefix', 'school_state',
       'project_submitted_datetime', 'project_grade_category', 'project_title',
       'project_resource_summary',
       'teacher_number_of_previously_posted_projects', 'project_is_approved',
       'clean_categories', 'clean_subcategories', 'essay', 'price',
       'quantity'],
      dtype='object')
```

In [229]:

```
# move columns in pandas dataframe https://stackoverflow.com/questions/35321812/move-column-in-pandas-dataframe/35321983
project_data = project_data[[c for c in project_data if c not in ['project_is_approved']]
       + ['project_is_approved']]
project_data.columns
```

Out[229]:

```
Index(['Unnamed: 0', 'id', 'teacher_id', 'teacher_prefix', 'school_state',
       'project_submitted_datetime', 'project_grade_category', 'project_title',
       'project_resource_summary',
```

```
       'teacher_number_of_previously_posted_projects', 'clean_categories',
       'clean_subcategories', 'essay', 'price', 'quantity',
       'project_is_approved'],
      dtype='object')
```

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

In [230]:

```
#importing necessary modules

from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score
from collections import Counter
from sklearn.metrics import accuracy_score

# Splitting the data into X , Y labels

# create design matrix X and target vector y
X = np.array(project_data.iloc[:, :-1]) # end index is exclusive
y = np.array(project_data['project_is_approved']) # showing you two ways of indexing a pandas df
# split the data set into train and test
X_1, X_test, y_1, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

# split the train data set into cross validation train and cross validation test
X_tr, X_cv, y_tr, y_cv = train_test_split(X_1, y_1, test_size=0.3)
print(len(X_tr))
print(len(X_cv))
print(len(X_test))
```

```
53531
22942
32775
```

In [231]:

```
X_tr = pd.DataFrame(data=X_tr[0:,0:], columns=project_data.columns[0:-1])
X_cv = pd.DataFrame(data=X_cv[0:,0:], columns=project_data.columns[0:-1])
X_test = pd.DataFrame(data=X_test[0:,0:], columns=project_data.columns[0:-1])
```

2.2 Make Data Model Ready: encoding numerical, categorical features

In [232]:

```
# 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 neededtHuWDX6yizwIhai
    # c. X-axis label
    # d. Y-axis label
print("="*25+"encoding categorical features"+"="*25)
#Vectorizing categorical data :
# 1 Clean_Categories

# we use count vectorizer to convert the values into one
from sklearn.feature_extraction.text import CountVectorizer
vectorizer1 = CountVectorizer(lowercase=False, binary=True)
vectorizer1.fit(X_tr['clean_categories'].values)

categories_one_hot_train = vectorizer1.transform(X_tr['clean_categories'].values)
categories_one_hot_test = vectorizer1.transform(X_test['clean_categories'].values)
categories_one_hot_cv = vectorizer1.transform(X_cv['clean_categories'].values)

print(vectorizer1.get_feature_names())
```

```
print(vectorizer1.get_feature_names())
print("Shape of train matrix after one hot encodig ",categories_one_hot_train.shape)
print("Shape of test matrix after one hot encodig ",categories_one_hot_test.shape)
print("Shape of cv matrix after one hot encodig ",categories_one_hot_cv.shape)
print("="*100)
```

```
==========================encoding categorical features==========================
['AppliedLearning', 'Care_Hunger', 'Health_Sports', 'History_Civics', 'Literacy_Language',
'Math_Science', 'Music_Arts', 'SpecialNeeds', 'Warmth']
Shape of train matrix after one hot encodig  (53531, 9)
Shape of test matrix after one hot encodig  (32775, 9)
Shape of cv matrix after one hot encodig  (22942, 9)
==================================================================================================
```

In [233]:

```
# 2 clean_subcategories

vectorizer2 = CountVectorizer(lowercase=False, binary=True)
vectorizer2.fit(X_tr['clean_subcategories'].values)
print(vectorizer2.get_feature_names())

sub_categories_one_hot_train = vectorizer2.transform(X_tr['clean_subcategories'].values)
sub_categories_one_hot_test = vectorizer2.transform(X_test['clean_subcategories'].values)
sub_categories_one_hot_cv = vectorizer2.transform(X_cv['clean_subcategories'].values)

print("Shape of train matrix after one hot encodig ",sub_categories_one_hot_train.shape)
print("Shape of test matrix after one hot encodig ",sub_categories_one_hot_test.shape)
print("Shape of cv matrix after one hot encodig ",sub_categories_one_hot_cv.shape)
print("="*100)
```

```
['AppliedSciences', 'Care_Hunger', 'CharacterEducation', 'Civics_Government',
'College_CareerPrep', 'CommunityService', 'ESL', 'EarlyDevelopment', 'Economics',
'EnvironmentalScience', 'Extracurricular', 'FinancialLiteracy', 'ForeignLanguages', 'Gym_Fitness',
'Health_LifeScience', 'Health_Wellness', 'History_Geography', 'Literacy', 'Literature_Writing', 'M
athematics', 'Music', 'NutritionEducation', 'Other', 'ParentInvolvement', 'PerformingArts', 'Socia
lSciences', 'SpecialNeeds', 'TeamSports', 'VisualArts', 'Warmth']
Shape of train matrix after one hot encodig  (53531, 30)
Shape of test matrix after one hot encodig  (32775, 30)
Shape of cv matrix after one hot encodig  (22942, 30)
==================================================================================================
```

In [234]:

```
my_counter = Counter()
for state in project_data['school_state'].values:
    my_counter.update(state.split())
school_state_cat_dict = dict(my_counter)
sorted_school_state_cat_dict = dict(sorted(school_state_cat_dict.items(), key=lambda kv: kv[1]))
```

In [235]:

```
# 3 school_state

vectorizer3 = CountVectorizer(lowercase=False, binary=True)
vectorizer3.fit(X_tr['school_state'].values)
print(vectorizer3.get_feature_names())
school_state_one_hot_train = vectorizer3.transform(X_tr['school_state'].values)
school_state_one_hot_test = vectorizer3.transform(X_test['school_state'].values)
school_state_one_hot_cv = vectorizer3.transform(X_cv['school_state'].values)

print("Shape of train matrix after one hot encodig ",school_state_one_hot_train.shape)
print("Shape of test matrix after one hot encodig ",school_state_one_hot_test.shape)
print("Shape of cv matrix after one hot encodig ",school_state_one_hot_cv.shape)
print("="*100)
```

```
['AK', 'AL', 'AR', 'AZ', 'CA', 'CO', 'CT', 'DC', 'DE', 'FL', 'GA', 'HI', 'IA', 'ID', 'IL', 'IN', 'K
S', 'KY', 'LA', 'MA', 'MD', 'ME', 'MI', 'MN', 'MO', 'MS', 'MT', 'NC', 'ND', 'NE', 'NH', 'NJ', 'NM',
'NV', 'NY', 'OH', 'OK', 'OR', 'PA', 'RI', 'SC', 'SD', 'TN', 'TX', 'UT', 'VA', 'VT', 'WA', 'WI', 'WV
', 'WY']
Shape of train matrix after one hot encodig  (53531, 51)
```

```
Shape of train matrix after one hot encodig  (53531, 51)
Shape of test matrix after one hot encodig  (32775, 51)
Shape of cv matrix after one hot encodig  (22942, 51)
==================================================================================================
```

In [236]:

```
my_counter = Counter()
for teacher in project_data['teacher_prefix'].values:
    my_counter.update(teacher.split())
teacher_prefix_cat_dict = dict(my_counter)
sorted_teacher_prefix_cat_dict = dict(sorted(teacher_prefix_cat_dict.items(), key=lambda kv: kv[1])
)
```

In [237]:

```
# 4 teacher_prefix

#one hot encoding for teacher_prefix feature

vectorizer4 = CountVectorizer(lowercase=False, binary=True)
vectorizer4.fit(X_tr['teacher_prefix'].values)
print(vectorizer4.get_feature_names())
teacher_prefix_one_hot_train = vectorizer4.transform(X_tr['teacher_prefix'].values)
teacher_prefix_one_hot_test = vectorizer4.transform(X_test['teacher_prefix'].values)
teacher_prefix_one_hot_cv = vectorizer4.transform(X_cv['teacher_prefix'].values)

print("Shape of train matrix after one hot encodig ",teacher_prefix_one_hot_train.shape)
print("Shape of test matrix after one hot encodig ",teacher_prefix_one_hot_test.shape)
print("Shape of cv matrix after one hot encodig ",teacher_prefix_one_hot_cv.shape)
print("="*100)
```

```
['Dr', 'Mr', 'Mrs', 'Ms', 'Teacher']
Shape of train matrix after one hot encodig  (53531, 5)
Shape of test matrix after one hot encodig  (32775, 5)
Shape of cv matrix after one hot encodig  (22942, 5)
==================================================================================================
```

In [238]:

```
# 5 project_grade_category

vectorizer5 = CountVectorizer(lowercase=False, binary=True)
vectorizer5.fit(X_tr['project_grade_category'].values)
print(vectorizer5.get_feature_names())
project_grade_one_hot_train = vectorizer5.transform(X_tr['project_grade_category'].values)
project_grade_one_hot_test = vectorizer5.transform(X_test['project_grade_category'].values)
project_grade_one_hot_cv = vectorizer5.transform(X_cv['project_grade_category'].values)

print("Shape of train matrix after one hot encodig ",project_grade_one_hot_train.shape)
print("Shape of test matrix after one hot encodig ",project_grade_one_hot_test.shape)
print("Shape of cv matrix after one hot encodig ",project_grade_one_hot_cv.shape)
print("="*100)
```

```
['Grades_3_5', 'Grades_6_8', 'Grades_9_12', 'Grades_PreK_2']
Shape of train matrix after one hot encodig  (53531, 4)
Shape of test matrix after one hot encodig  (32775, 4)
Shape of cv matrix after one hot encodig  (22942, 4)
==================================================================================================
```

Encoding numerical features

In [239]:

```
print("_"*25+"encoding numerical features"+"_"*25)
# 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)

#1 price
price_scalar = StandardScaler()
price_scalar.fit(X_tr['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_tr['price'].values.reshape(-1, 1))
price_standardized_test = price_scalar.transform(X_test['price'].values.reshape(-1, 1))
price_standardized_cv = price_scalar.transform(X_cv['price'].values.reshape(-1, 1))


print("Shape of train matrix after one hot encodig ",price_standardized_train.shape)
print("Shape of test matrix after one hot encodig ",price_standardized_test.shape)
print("Shape of cv matrix after one hot encodig ",price_standardized_cv.shape)
print("="*100)
#2 teacher_number_of_previously_posted_projects

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

# Now standardize the data with above maen and variance.
previous_project_standardized_train =
previous_project_scalar.transform(X_tr['teacher_number_of_previously_posted_projects'].values.resh
ape(-1, 1))
previous_project_standardized_test =
previous_project_scalar.transform(X_test['teacher_number_of_previously_posted_projects'].values.re
shape(-1, 1))
previous_project_standardized_cv =
previous_project_scalar.transform(X_cv['teacher_number_of_previously_posted_projects'].values.resh
ape(-1, 1))

print("Shape of train matrix after one hot encodig ",previous_project_standardized_train.shape)
print("Shape of test matrix after one hot encodig ",previous_project_standardized_test.shape)
print("Shape of cv matrix after one hot encodig ",previous_project_standardized_cv.shape)
```

```
_____________________encoding numerical features_____________________
Mean : 296.2095355961966, Standard deviation : 369.3449608014876
Shape of train matrix after one hot encodig  (53531, 1)
Shape of test matrix after one hot encodig  (32775, 1)
Shape of cv matrix after one hot encodig  (22942, 1)
====================================================================================================

Mean : 11.03179466103753, Standard deviation : 27.312462937258964
Shape of train matrix after one hot encodig  (53531, 1)
Shape of test matrix after one hot encodig  (32775, 1)
Shape of cv matrix after one hot encodig  (22942, 1)
```

In [240]:

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

# Now standardize the data with above maen and variance.
quantity_standardized_train = quantity_scalar.transform(X_tr['quantity'].values.reshape(-1, 1))
quantity_standardized_test = quantity_scalar.transform(X_test['quantity'].values.reshape(-1, 1))
quantity_standardized_cv = quantity_scalar.transform(X_cv['quantity'].values.reshape(-1, 1))

print("Shape of train matrix ",quantity_standardized_train.shape)
print("Shape of test matrix  ",quantity_standardized_test.shape)
```

```
print("Shape of cv matrix ",quantity_standardized_cv.shape)
```

```
Mean : 17.03256057237862, Standard deviation : 26.068962077293794
Shape of train matrix  (53531, 1)
Shape of test matrix   (32775, 1)
Shape of cv matrix  (22942, 1)
```

In [241]:

```
quantity_standardized_train = np.where(quantity_standardized_train<0, 0,
quantity_standardized_train)
quantity_standardized_cv = np.where(quantity_standardized_cv<0, 0, quantity_standardized_cv)
quantity_standardized_test = np.where(quantity_standardized_test<0, 0, quantity_standardized_test)
```

In [242]:

```
price_standardized_train = np.where(price_standardized_train<0, 0, price_standardized_train)
price_standardized_cv = np.where(price_standardized_cv<0, 0, price_standardized_cv)
price_standardized_test = np.where(price_standardized_test<0, 0, price_standardized_test)
```

In [243]:

```
previous_project_standardized_train = np.where(previous_project_standardized_train<0, 0,
previous_project_standardized_train)
previous_project_standardized_cv = np.where(previous_project_standardized_cv<0, 0,
previous_project_standardized_cv)
previous_project_standardized_test = np.where(previous_project_standardized_test<0, 0, previous_pro
ject_standardized_test)
```

2.3 Make Data Model Ready: encoding eassay, and project_title

Encoding Essay and Project_title columns using Bag Of Words

In [244]:

```
# 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
print("_"*25+"Essay BOW"+"_"*25)
vectorizer6 = CountVectorizer(min_df=10)
vectorizer6.fit(X_tr['essay'])
essay_bow_train = vectorizer6.transform(X_tr['essay'])
essay_bow_test = vectorizer6.transform(X_test['essay'])
essay_bow_cv = vectorizer6.transform(X_cv['essay'])
print("Shape of train matrix after one hot encodig ",essay_bow_train.shape)
print("Shape of test matrix after one hot encodig ",essay_bow_test.shape)
print("Shape of cv matrix after one hot encodig ",essay_bow_cv.shape)
print("="*100)
print("_"*25+"Project_Title BOW"+"_"*25)
vectorizer7 = CountVectorizer(min_df=10)
vectorizer7.fit(X_tr['project_title'])
title_bow_train = vectorizer7.transform(X_tr['project_title'])
title_bow_test = vectorizer7.transform(X_test['project_title'])
title_bow_cv = vectorizer7.transform(X_cv['project_title'])
print("Shape of train matrix after one hot encodig ",title_bow_train.shape)
print("Shape of test matrix after one hot encodig ",title_bow_test.shape)
print("Shape of cv matrix after one hot encodig ",title_bow_cv.shape)
```

```
_________________________Essay BOW_________________________
Shape of train matrix after one hot encodig  (53531, 12502)
```

```
Shape of test matrix after one hot encodig  (32775, 12502)
Shape of cv matrix after one hot encodig  (22942, 12502)
==========================================================================================================

____________________________Project_Title BOW_____________________________
Shape of train matrix after one hot encodig  (53531, 2118)
Shape of test matrix after one hot encodig  (32775, 2118)
Shape of cv matrix after one hot encodig  (22942, 2118)
```

Encoding Essay and Project_title columns using TFIDF

In [245]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
print("_"*25+"Essay TFIDF"+"_"*25)
vectorizer8 = TfidfVectorizer(min_df=10)
vectorizer8.fit(X_tr['essay'])
essay_tfidf_train = vectorizer8.transform(X_tr['essay'])
essay_tfidf_test = vectorizer8.transform(X_test['essay'])
essay_tfidf_cv = vectorizer8.transform(X_cv['essay'])
print("Shape of train matrix after one hot encodig ",essay_tfidf_train.shape)
print("Shape of test matrix after one hot encodig ",essay_tfidf_test.shape)
print("Shape of cv matrix after one hot encodig ",essay_tfidf_cv.shape)
print("="*100)
print("_"*25+"Project_Title TFIDF"+"_"*25)
vectorizer9 = TfidfVectorizer(min_df=10)
vectorizer9.fit_transform(X_tr['project_title'])
title_tfidf_train = vectorizer9.transform(X_tr['project_title'])
title_tfidf_test = vectorizer9.transform(X_test['project_title'])
title_tfidf_cv = vectorizer9.transform(X_cv['project_title'])
print("Shape of train matrix after one hot encodig ",title_tfidf_train.shape)
print("Shape of test matrix after one hot encodig ",title_tfidf_test.shape)
print("Shape of cv matrix after one hot encodig ",title_tfidf_cv.shape)
```

```
____________________________Essay TFIDF_____________________________
Shape of train matrix after one hot encodig  (53531, 12502)
Shape of test matrix after one hot encodig  (32775, 12502)
Shape of cv matrix after one hot encodig  (22942, 12502)
==========================================================================================================

____________________________Project_Title TFIDF_____________________________
Shape of train matrix after one hot encodig  (53531, 2118)
Shape of test matrix after one hot encodig  (32775, 2118)
Shape of cv matrix after one hot encodig  (22942, 2118)
```

1.5.4 Merging all the above features

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

Set 1: categorical, numerical features + project_title(BOW) + preprocessed_essay (BOW)

Assignment 4: Naive Bayes

1. **Apply Multinomial NaiveBayes on these feature sets**
   - **Set 1: categorical, numerical features + project_title(BOW) + preprocessed_eassay (BOW)**
   - **Set 2: categorical, numerical features + project_title(TFIDF)+ preprocessed_eassay (TFIDF)**
2. **The hyper paramter tuning(find best Alpha)**
   - **Find the best hyper parameter which will give the maximum AUC value**
   - **Consider a wide range of alpha values for hyperparameter tuning, start as low as 0.00001**
   - **Find the best hyper paramter using k-fold cross validation or simple cross validation data**
   - **Use gridsearch cv or randomsearch cv or you can also write your own for loops to do this task of hyperparameter**

tuning

3. Feature importance

- Find the top 10 features of positive class and top 10 features of negative class for both feature sets Set 1 and Set 2 using values of `feature_log_prob_` parameter of MultinomialNB and print their corresponding feature names

4. Representation of results

- You need to plot the performance of model both on train data and cross validation data for each hyper parameter, like shown in the figure. Here on X-axis you will have alpha values, since they have a wide range, just to represent those alpha values on the graph, apply log function on those alpha values.
- Once after you found the best hyper parameter, you need to train your model with it, and find the AUC on test data and plot the ROC curve on both train and test.
- Along with plotting ROC curve, you need to print the confusion matrix with predicted and original labels of test data points. Please visualize your confusion matrices using seaborn heatmaps.

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

2. Naive Bayes

2.4 Appling NB() on different kind of featurization as mentioned in the instructions

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

2.4.1 Applying Naive Bayes on BOW, SET 1

In [281]:

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

# Set 1: categorical, numerical features + project_title(BOW) + preprocessed_essay (BOW)

from scipy.sparse import hstack
# with the same hstack function we are concatinating a sparse matrix and a dense matirx :)
X_tr = hstack((school_state_one_hot_train,categories_one_hot_train, sub_categories_one_hot_train,
teacher_prefix_one_hot_train, project_grade_one_hot_train,essay_bow_train,title_bow_train))
X_cv = hstack((school_state_one_hot_cv,categories_one_hot_cv, sub_categories_one_hot_cv,
teacher_prefix_one_hot_cv, project_grade_one_hot_cv,essay_bow_cv,title_bow_cv))
X_test = hstack((school_state_one_hot_test,categories_one_hot_test, sub_categories_one_hot_test, t
eacher_prefix_one_hot_test, project_grade_one_hot_test,essay_bow_test,title_bow_test))
X_tr = X_tr.tocsr()
X_cv = X_cv.tocsr()
X_test = X_test.tocsr()
print(X_tr.shape , y_tr.shape)
print(X_cv.shape , y_cv.shape)
print(X_test.shape , y_test.shape)
```

```
(53531, 14719) (53531,)
(22942, 14719) (22942,)
(32775, 14719) (32775,)
```

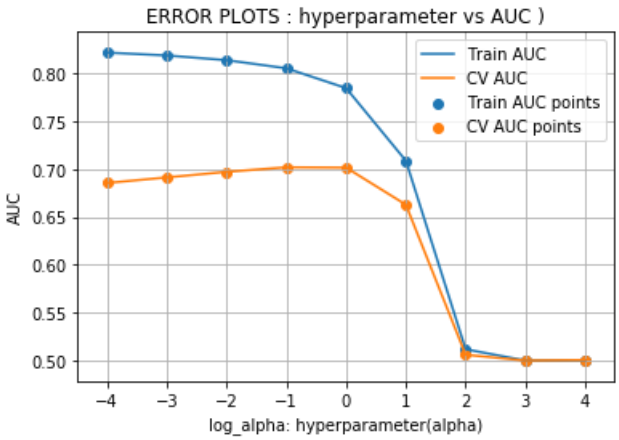
In [283]:

```
import math

alpha = [0.0001, 0.001, 0.01, 0.1, 1.0, 10,100, 1000, 10000]
log_alpha = list(map(lambda x : math.log10(x), alpha))
```

In [284]:

```
from sklearn.naive_bayes import MultinomialNB
import matplotlib.pyplot as plt
from sklearn.metrics import roc_auc_score

train_auc = []
cv_auc = []

for i in tqdm(alpha):
    neigh = MultinomialNB(alpha = i, class_prior = [0.5,0.5])
    neigh.fit(X_tr, y_tr)

    y_train_pred = neigh.predict_proba( X_tr)[:, 1]
    y_cv_pred = neigh.predict_proba(X_cv)[:, 1]

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

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

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

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

```
100%|██████████████████████████████████████████████████████████████████████████████████| 9/9 [00
:02<00:00,  3.33it/s]
```



In the above case the best alpha is choosen to be alpha=1 because at alpha=1 the cv_auc is maximum. If the alpha is further increased both the cv_auc and train_auc decreases drastically.

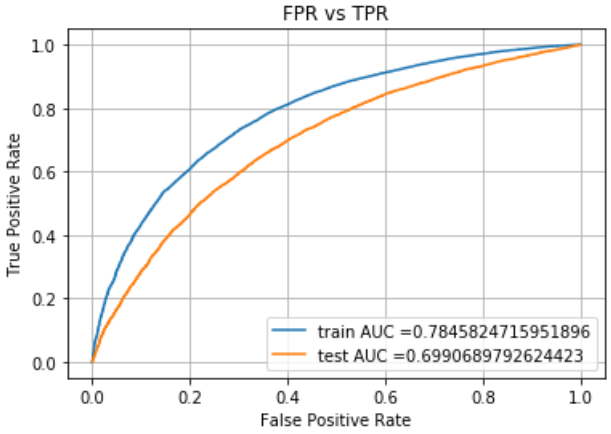
In [285]:

```
best_alpha1 = 1
```

In [286]:

```
from sklearn.metrics import roc_curve, auc


neigh = MultinomialNB(alpha = best_alpha1, class_prior = [0.5,0.5])
neigh.fit(X_tr, y_tr)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of the positive
class
# not the predicted outputs
```

```
# not the predicted outputs

y_train_pred = neigh.predict_proba( X_tr)[:, 1]
y_test_pred = neigh.predict_proba(X_test)[:, 1]

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

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
plt.legend()
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("FPR vs TPR")
plt.grid()
plt.show()
```



In [287]:

```
test_auc1 = auc(test_fpr, test_tpr)
```

In [288]:

```
# we are writing our own function for predict, with defined thresould
# we will pick a threshold that will give the least fpr
def find_best_threshold(threshould, fpr, tpr):
    t = threshould[np.argmax(tpr*(1-fpr))]
    # (tpr*(1-fpr)) will be maximum if your fpr is very low and tpr is very high
    print("the maximum value of tpr*(1-fpr)", max(tpr*(1-fpr)), "for threshold", np.round(t,3))
    return t

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

In [289]:

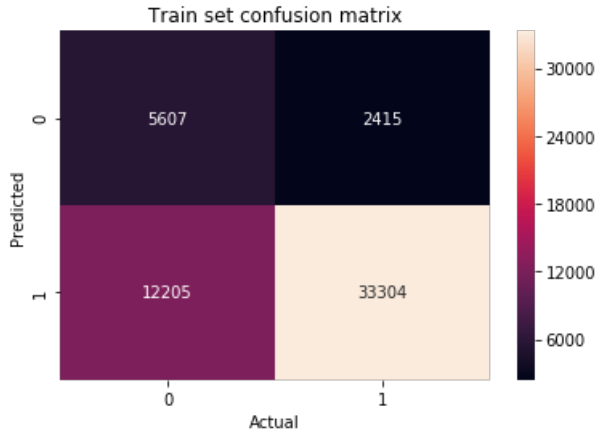
```
print("="*100)
from sklearn.metrics import confusion_matrix
best_t1 = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")
print(confusion_matrix(y_tr, predict_with_best_t(y_train_pred, best_t1)))
print("Test confusion matrix")
print(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t1)))
```

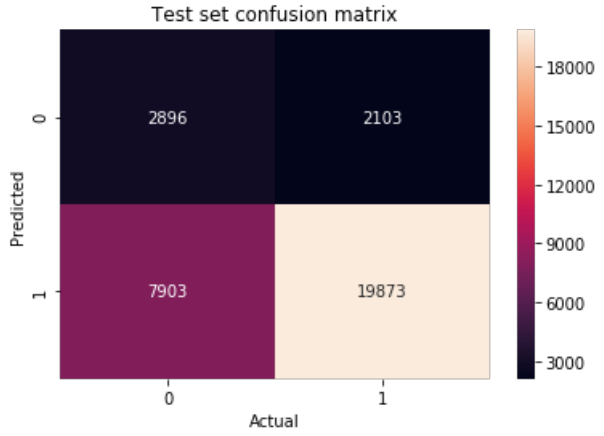
```
====================================================================================================

the maximum value of tpr*(1-fpr) 0.5115016085075629 for threshold 0.426
Train confusion matrix
[[ 5607  2415]
 [12205 33304]]
```

```
Test confusion matrix
[[ 2896  2103]
 [ 7903 19873]]
```

In [290]:

```
# Heatmap for train set confusion matrix(Select K best)
heatmap_train = sns.heatmap(confusion_matrix(y_tr, predict_with_best_t(y_train_pred,best_t1)),
annot=True, fmt="d")
plt.title("Train set confusion matrix")
plt.xlabel("Actual")
plt.ylabel("Predicted")
plt.show()

# Heatmap for test set confusion matrix(Select K best)
heatmap_train = sns.heatmap(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t1)), an
not=True, fmt="d")
plt.title("Test set confusion matrix")
plt.xlabel("Actual")
plt.ylabel("Predicted")
plt.show()
```





2.4.1.1 Top 10 important features of positive class from SET 1

In [291]:

```
# To find the top features of positive/negative class
https://github.com/shashimanyam/NaiveBayes/blob/master/NAVIEBAYES.pdf
bow_features_probs = []
for a in range(14719):
    bow_features_probs.append(neigh.feature_log_prob_[1,a] )
print(len(bow_features_probs))
```

```
14719
```

In [292]:

```
bow_features_names = []
for a in vectorizer1.get_feature_names(): # clean_categories
    bow_features_names.append(a)
for a in vectorizer2.get_feature_names(): # clean_sub_categories
    bow_features_names.append(a)
for a in vectorizer3.get_feature_names(): # school state
    bow_features_names.append(a)
for a in vectorizer4.get_feature_names(): # teacher_prefix
    bow_features_names.append(a)
for a in vectorizer5.get_feature_names(): # Grades
    bow_features_names.append(a)
for a in vectorizer6.get_feature_names(): # bow_essay
    bow_features_names.append(a)
for a in vectorizer7.get_feature_names(): # bow_title
    bow_features_names.append(a)
print(len(bow_features_names))
```

```
14719
```

In [293]:

```
final_bow_features = pd.DataFrame({'feature_prob_estimates' : bow_features_probs, 'feature_names':
bow_features_names})
a =final_bow_features.sort_values(by = ['feature_prob_estimates'], ascending = False)
#print(final_bow_features.head(6))
a.head(10)
```

Out[293]:

	feature_prob_estimates	feature_names
10903	-2.999336	students
9891	-4.145343	school
6546	-4.506431	learning
2202	-4.529571	classroom
7614	-4.801790	not
6542	-4.844556	learn
5414	-4.877866	help
6928	-5.017352	many
7441	-5.034355	nannan
7490	-5.146248	need

2.4.1.2 Top 10 important features of negative class from SET 1

In [294]:

```
# To find the top features of positive/negative class
https://github.com/shashimanyam/NaiveBayes/blob/master/NAVIEBAYES.pdf
bow_features_probs = []
for a in range(14719):
    bow_features_probs.append(neigh.feature_log_prob_[0,a] )
print(len(bow_features_probs))
```

```
14719
```

In [295]:

```
bow_features_names = []
for a in vectorizer1.get_feature_names(): # clean_categories
    bow_features_names.append(a)
for a in vectorizer2.get_feature_names(): # clean_sub_categories
    bow_features_names.append(a)
for a in vectorizer3.get_feature_names(): # school state
    bow_features_names.append(a)
for a in vectorizer4.get_feature_names(): # teacher_prefix
```

```
    bow_features_names.append(a)
for a in vectorizer5.get_feature_names(): # Grades
    bow_features_names.append(a)
for a in vectorizer6.get_feature_names(): # bow_essay
    bow_features_names.append(a)
for a in vectorizer7.get_feature_names(): # bow_title
    bow_features_names.append(a)
print(len(bow_features_names))
```

```
14719
```

In [296]:

```
final_bow_features = pd.DataFrame({'feature_prob_estimates' : bow_features_probs, 'feature_names':
bow_features_names})
a =final_bow_features.sort_values(by = ['feature_prob_estimates'], ascending = False)
#print(final_bow_features.head(6))
a.head(10)
```

Out[296]:

	feature_prob_estimates	feature_names
10903	-3.021591	students
9891	-4.120659	school
6546	-4.442721	learning
2202	-4.580629	classroom
7614	-4.780674	not
6542	-4.793774	learn
5414	-4.824030	help
7441	-4.992089	nannan
6928	-5.034273	many
7490	-5.105470	need

2.4.2 Applying Naive Bayes on TFIDF, SET 2

In [297]:

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

from scipy.sparse import hstack
# with the same hstack function we are concatinating a sparse matrix and a dense matirx :)
X_tr = hstack((school_state_one_hot_train,categories_one_hot_train, sub_categories_one_hot_train,
teacher_prefix_one_hot_train, project_grade_one_hot_train,essay_tfidf_train,title_tfidf_train))
X_cv = hstack((school_state_one_hot_cv,categories_one_hot_cv, sub_categories_one_hot_cv,
teacher_prefix_one_hot_cv, project_grade_one_hot_cv,essay_tfidf_cv,title_tfidf_cv))
X_test = hstack((school_state_one_hot_test,categories_one_hot_test, sub_categories_one_hot_test, t
eacher_prefix_one_hot_test, project_grade_one_hot_test,essay_tfidf_test,title_tfidf_test))
X_tr = X_tr.tocsr()
X_cv = X_cv.tocsr()
X_test = X_test.tocsr()
print(X_tr.shape , y_tr.shape)
print(X_cv.shape , y_cv.shape)
print(X_test.shape , y_test.shape)
```

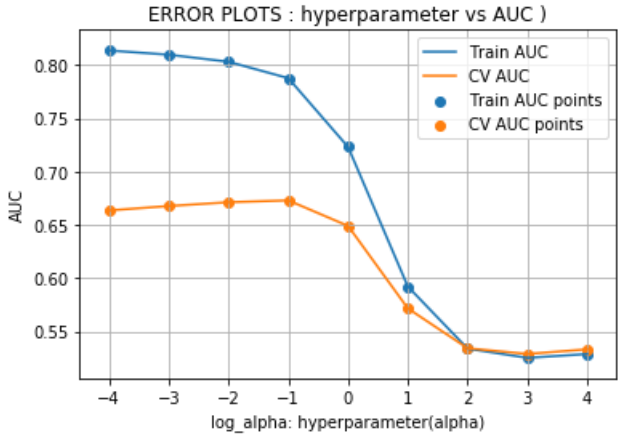
```
(53531, 14719) (53531,)
(22942, 14719) (22942,)
(32775, 14719) (32775,)
```

In [298]:

```
from sklearn.naive_bayes import MultinomialNB
import matplotlib.pyplot as plt
from sklearn.metrics import roc_auc_score

train_auc = []
```

```
cv_auc = []

for i in tqdm(alpha):
    neigh = MultinomialNB(alpha = i, class_prior = [0.5,0.5])
    neigh.fit(X_tr, y_tr)

    y_train_pred = neigh.predict_proba( X_tr)[:, 1]
    y_cv_pred = neigh.predict_proba(X_cv)[:, 1]

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

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

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

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

```
100%|██████████████████████████████████████████████████████████████████████████████| 9/9 [00
:02<00:00,  3.64it/s]
```



In the above case the best alpha is choosen to be alpha=0.1 because at alpha=0.1 the cv_auc is maximum. If the alpha is further increased both the cv_auc and train_auc decreases drastically.

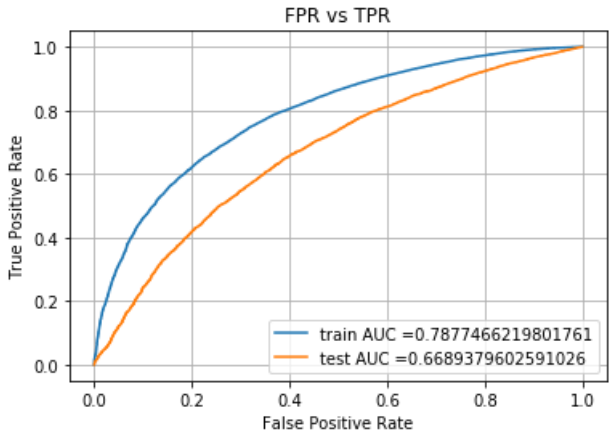
In [299]:

```
best_alpha2 = 0.1
```

In [300]:

```
from sklearn.metrics import roc_curve, auc


neigh = MultinomialNB(alpha = best_alpha2, class_prior = [0.5,0.5])
neigh.fit(X_tr, y_tr)
# 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 = neigh.predict_proba( X_tr)[:, 1]
y_test_pred = neigh.predict_proba(X_test)[:, 1]

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

```
plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
plt.legend()
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("FPR vs TPR")
plt.grid()
plt.show()
```



In [301]:

```
test_auc2 = auc(test_fpr, test_tpr)
```

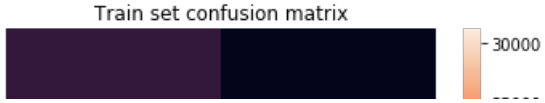
In [302]:

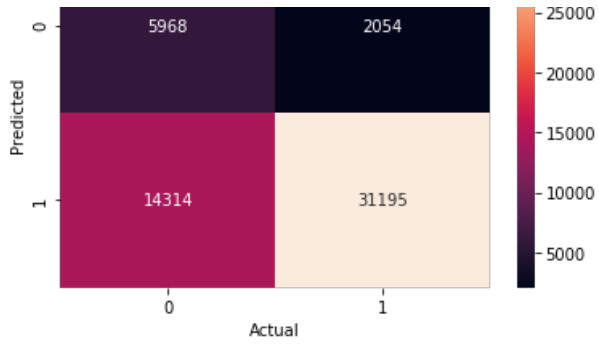
```
print("="*100)
from sklearn.metrics import confusion_matrix
best_t2 = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")
print(confusion_matrix(y_tr, predict_with_best_t(y_train_pred, best_t1)))
print("Test confusion matrix")
print(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t1)))
```

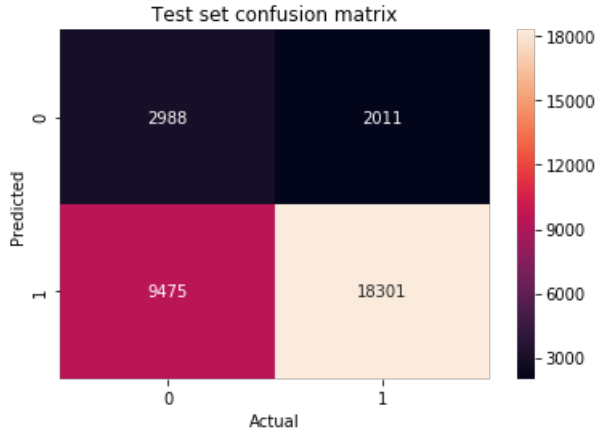
```
====================================================================================================

the maximum value of tpr*(1-fpr) 0.5099573483342922 for threshold 0.519
Train confusion matrix
[[ 4926  3096]
 [ 9219 36290]]
Test confusion matrix
[[ 2289  2710]
 [ 6238 21538]]
```

In [303]:

```
# Heatmap for train set confusion matrix(Select K best)
heatmap_train = sns.heatmap(confusion_matrix(y_tr, predict_with_best_t(y_train_pred,best_t2)),
annot=True, fmt="d")
plt.title("Train set confusion matrix")
plt.xlabel("Actual")
plt.ylabel("Predicted")
plt.show()

# Heatmap for test set confusion matrix(Select K best)
heatmap_train = sns.heatmap(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t2)), an
not=True, fmt="d")
plt.title("Test set confusion matrix")
plt.xlabel("Actual")
plt.ylabel("Predicted")
plt.show()
```







2.4.2.1 Top 10 important features of positive class from SET 2

In [304]:

```
# Please write all the code with proper documentation
# To find the top features of positive/negative class
https://github.com/shashimanyam/NaiveBayes/blob/master/NAVIEBAYES.pdf
tfidf_features_probs = []
for a in range(14719):
    tfidf_features_probs.append(neigh.feature_log_prob_[1,a] )
print(len(tfidf_features_probs))
```

```
14719
```

In [305]:

```
tfidf_features_names = []
for a in vectorizer1.get_feature_names(): # clean_categories
    tfidf_features_names.append(a)
for a in vectorizer2.get_feature_names(): # clean_sub_categories
    tfidf_features_names.append(a)
for a in vectorizer3.get_feature_names(): # school state
    tfidf_features_names.append(a)
for a in vectorizer4.get_feature_names(): # teacher_prefix
    tfidf_features_names.append(a)
for a in vectorizer5.get_feature_names(): # Grades
    tfidf_features_names.append(a)
for a in vectorizer6.get_feature_names(): # bow_essay
    tfidf_features_names.append(a)
for a in vectorizer7.get_feature_names(): # bow_title
    tfidf_features_names.append(a)
print(len(tfidf_features_names))
```

```
14719
```

In [306]:

```
final_tfidf_features = pd.DataFrame({'feature_prob_estimates' : tfidf_features_probs,
'feature_names': tfidf_features_names})
a =final_tfidf_features.sort_values(by = ['feature_prob_estimates'], ascending = False)
```

```
a.head(10)
```

Out[306]:

	feature_prob_estimates	feature_names
92	-3.429790	Mrs
55	-3.510144	KS
98	-3.697332	Grades_PreK_2
56	-3.776092	KY
93	-3.830576	Ms
95	-3.864789	Grades_3_5
77	-3.944340	PA
79	-4.160341	SC
78	-4.360513	RI
96	-4.670501	Grades_6_8

2.4.2.2 Top 10 important features of negative class from SET 2

In [307]:

```
# To find the top features of positive/negative class
https://github.com/shashimanyam/NaiveBayes/blob/master/NAVIEBAYES.pdf
tfidf_features_probs = []
for a in range(14719):
    tfidf_features_probs.append(neigh.feature_log_prob_[0,a] )
print(len(tfidf_features_probs))
```

```
14719
```

In [308]:

```
tfidf_features_names = []
for a in vectorizer1.get_feature_names(): # clean_categories
    tfidf_features_names.append(a)
for a in vectorizer2.get_feature_names(): # clean_sub_categories
    tfidf_features_names.append(a)
for a in vectorizer3.get_feature_names(): # school state
    tfidf_features_names.append(a)
for a in vectorizer4.get_feature_names(): # teacher_prefix
    tfidf_features_names.append(a)
for a in vectorizer5.get_feature_names(): # Grades
    tfidf_features_names.append(a)
for a in vectorizer6.get_feature_names(): # bow_essay
    tfidf_features_names.append(a)
for a in vectorizer7.get_feature_names(): # bow_title
    tfidf_features_names.append(a)
print(len(tfidf_features_names))
```

```
14719
```

In [309]:

```
final_tfidf_features = pd.DataFrame({'feature_prob_estimates' : tfidf_features_probs,
'feature_names': tfidf_features_names})
a =final_tfidf_features.sort_values(by = ['feature_prob_estimates'], ascending = False)
#print(final_bow_features.head(6))
a.head(10)
```

Out[309]:

	feature_prob_estimates	feature_names
92	-3.472887	Mrs
55	-3.638505	KS

	feature_prob_estimates	feature_names
98	-3.689460	Grades_PreK_2
56	-3.715789	KY
93	-3.813988	Ms
95	-3.930642	Grades_3_5
77	-4.129749	PA
79	-4.147473	SC
78	-4.442739	RI
96	-4.633132	Grades_6_8

3. Conclusions

In [312]:

```
from prettytable import PrettyTable

model_compare = PrettyTable()
model_compare.field_names = ["Feature_sets", "Best_alpha_value","Test_AUC", "Best_threshold"]
model_compare.add_row(["Bag of words", best_alpha1,np.round(test_auc1,4), np.round(best_t1,3)])
model_compare.add_row(["TF-IDF", best_alpha2, np.round(test_auc2,4),np.round(best_t2,3)])

print(model_compare)
```

```
+--------------+------------------+----------+----------------+
| Feature_sets | Best_alpha_value | Test_AUC | Best_threshold |
+--------------+------------------+----------+----------------+
| Bag of words |         1        |  0.6991  |     0.426      |
|    TF-IDF    |        0.1       |  0.6689  |     0.519      |
+--------------+------------------+----------+----------------+
```

1) The Best Hyperparameter K is found to be different in all the cases based on the features. 2) The Best threshold value is found to be different in all the cases based on the features that are used to train the model.3)We get maximum auc when Bag of Words Encoding is used to transform the text features than using TFIDF.