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 (https://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations#Postal_codes)). **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*
`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 [1]:
```python
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")

import sqlite3
import pandas as pd
import numpy as np
import nltk
import string
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import roc_curve, auc
from nltk.stem.porter import PorterStemmer

import re
# Tutorial about Python regular expressions: https://pymotw.com/2/re/
import string
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer

from gensim.models import Word2Vec
from gensim.models import KeyedVectors
import pickle

from tqdm import tqdm
import os

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

1.1 Reading Data

In [2]:
```python
project_data = pd.read_csv('../train_data.csv')
resource_data = pd.read_csv('../resources.csv')
```

In [3]:
```python
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]:
```
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[4]:

	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 Categorical Data

1.2.1 preprocessing `project_subject_categories`

In [5]:
```
catogories = list(project_data['project_subject_categories'].values)
# remove special characters from list of strings python: https://stackoverflow.com/a/47301924/4084039

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

In [6]:
```
sorted_cat_dict.keys()
```

Out[6]:
```
dict_keys(['Warmth', 'Care_Hunger', 'History_Civics', 'Music_Arts', 'AppliedLearning', 'SpecialNeeds', 'Health_Spo
rts', 'Math_Science', 'Literacy_Language'])
```

1.2.2 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 & Hunger"]
        if 'The' in j.split(): # this will split each of the catogory based on space "Math & Science"=> "Math","&",
"Science"
            j=j.replace('The','') # if we have the words "The" we are going to replace it with ''(i.e removing 'Th
e')
        j = j.replace(' ','') # we are placeing all the ' '(space) with ''(empty) ex:"Math & Science"=>"Math&Scienc
e"
        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]))
```

In [8]:
```
sorted_sub_cat_dict.keys()
```

Out[8]:
```
dict_keys(['Economics', 'CommunityService', 'FinancialLiteracy', 'ParentInvolvement', 'Extracurricular', 'Civics_G
overnment', 'ForeignLanguages', 'NutritionEducation', 'Warmth', 'Care_Hunger', 'SocialSciences', 'PerformingArts',
'CharacterEducation', 'TeamSports', 'Other', 'College_CareerPrep', 'Music', 'History_Geography', 'Health_LifeScien
ce', 'EarlyDevelopment', 'ESL', 'Gym_Fitness', 'EnvironmentalScience', 'VisualArts', 'Health_Wellness', 'AppliedSc
iences', 'SpecialNeeds', 'Literature_Writing', 'Mathematics', 'Literacy'])
```

1.2.3 preprocessing of `School State`

In [9]:
```
project_data['school_state'].unique()
```

Out[9]:
```
array(['IN', 'FL', 'AZ', 'KY', 'TX', 'CT', 'GA', 'SC', 'NC', 'CA', 'NY',
       'OK', 'MA', 'NV', 'OH', 'PA', 'AL', 'LA', 'VA', 'AR', 'WA', 'WV',
       'ID', 'TN', 'MS', 'CO', 'UT', 'IL', 'MI', 'HI', 'IA', 'RI', 'NJ',
       'MO', 'DE', 'MN', 'ME', 'WY', 'ND', 'OR', 'AK', 'MD', 'WI', 'SD',
       'NE', 'NM', 'DC', 'KS', 'MT', 'NH', 'VT'], dtype=object)
```

In [10]:
```
project_data['school_state'][project_data['school_state'].isnull()==True]
```

Out[10]:
```
Series([], Name: school_state, dtype: object)
```

In [11]:
```
# count of all the words in corpus python: https://stackoverflow.com/a/22898595/4084039
my_counter = Counter()
for word in project_data['school_state'].values:
    my_counter.update(word.split())

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

In [12]:
```
sorted_school_state_dict.keys()
```

Out[12]:
```
dict_keys(['VT', 'WY', 'ND', 'MT', 'RI', 'SD', 'NE', 'DE', 'AK', 'NH', 'WV', 'ME', 'HI', 'DC', 'NM', 'KS', 'IA',
'ID', 'AR', 'CO', 'MN', 'OR', 'KY', 'MS', 'NV', 'MD', 'CT', 'TN', 'UT', 'AL', 'WI', 'VA', 'AZ', 'NJ', 'OK', 'WA',
'MA', 'LA', 'OH', 'MO', 'IN', 'PA', 'MI', 'SC', 'GA', 'IL', 'NC', 'FL', 'NY', 'TX', 'CA'])
```

1.2.4 preprocessing of `Teacher Prefix`

In [13]:
```
project_data.groupby(['teacher_prefix'])['teacher_prefix'].count()
```
Out[13]:
```
teacher_prefix
Dr.           13
Mr.        10648
Mrs.       57269
Ms.        38955
Teacher     2360
Name: teacher_prefix, dtype: int64
```

In [14]:
```
project_data['teacher_prefix'][project_data['teacher_prefix'].isnull()==True]
```
Out[14]:
```
7820     NaN
30368    NaN
57654    NaN
Name: teacher_prefix, dtype: object
```

In [15]:
```
project_data['teacher_prefix'].fillna(project_data['teacher_prefix'].mode()[0],inplace=True)
```

In [16]:
```
project_data['teacher_prefix'][project_data['teacher_prefix'].isnull()==True]
```
Out[16]:
```
Series([], Name: teacher_prefix, dtype: object)
```

In [17]:
```
project_data['teacher_prefix'].unique()
```
Out[17]:
```
array(['Mrs.', 'Mr.', 'Ms.', 'Teacher', 'Dr.'], dtype=object)
```

In [18]:
```
teacher_prefix = list(project_data['teacher_prefix'].values)

teacher_prefix_list = []
for i in teacher_prefix:
    temp = ""
    temp = i.split('.')
    temp = i.replace('.','')
    teacher_prefix_list.append(temp)

project_data['clean_teacher_prefix'] = teacher_prefix_list
project_data.drop(['teacher_prefix'], 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_teacher_prefix'].values:
    my_counter.update(word.split())

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

In [19]:
```
sorted_teacher_prefix_dict.keys()
```
Out[19]:
```
dict_keys(['Dr', 'Teacher', 'Mr', 'Ms', 'Mrs'])
```

In [20]:
```
project_data.groupby(['clean_teacher_prefix'])['clean_teacher_prefix'].count()
```
Out[20]:
```
clean_teacher_prefix
Dr            13
Mr         10648
Mrs        57272
Ms         38955
Teacher     2360
Name: clean_teacher_prefix, dtype: int64
```

1.2.5 preprocessing of `Project Grade Category`

In [21]:
```
project_data.groupby(['project_grade_category'])['project_grade_category'].count()
```
Out[21]:
```
project_grade_category
Grades 3-5       37137
Grades 6-8       16923
Grades 9-12      10963
Grades PreK-2    44225
Name: project_grade_category, dtype: int64
```

In [22]:
```
project_data['project_grade_category'][project_data['project_grade_category'].isnull()==True]
```
Out[22]:
```
Series([], Name: project_grade_category, dtype: object)
```

In [23]:
```
project_grade_category = list(project_data['project_grade_category'].values)

project_grade_category_list = []
for i in project_grade_category:
    temp = ""
    temp = i.split(' ')
    temp = i.replace('Grades ','')
    project_grade_category_list.append(temp)

project_data['clean_project_grade_category'] = project_grade_category_list
project_data.drop(['project_grade_category'], 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_project_grade_category'].values:
    my_counter.update(word.split())

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

In [24]:
```
sorted_project_grade_category_dict.keys()
```

Out[24]:
```
dict_keys(['9-12', '6-8', '3-5', 'PreK-2'])
```

In [25]:
```
project_data.groupby(['clean_project_grade_category'])['clean_project_grade_category'].count()
```

Out[25]:
```
clean_project_grade_category
3-5       37137
6-8       16923
9-12      10963
PreK-2    44225
Name: clean_project_grade_category, dtype: int64
```

In [ ]:

1.3 Text preprocessing

In [26]:
```
# 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 [27]:
```
project_data.head(2)
```

Out[27]:

	Unnamed: 0	id	teacher_id	school_state	project_submitted_datetime	project_title	project_essay_1	project_essa
0	160221	p253737	c90749f5d961ff158d4b4d1e7dc665fc	IN	2016-12-05 13:43:57	Educational Support for English Learners at Home	My students are English learners that are work...	\"The lim your langu are the limit
1	140945	p258326	897464ce9ddc600bced1151f324dd63a	FL	2016-10-25 09:22:10	Wanted: Projector for Hungry Learners	Our students arrive to our school eager to lea...	The projecto need fo school is ver

In [28]:
```
#### 1.4.2.3 Using Pretrained Models: TFIDF weighted W2V
```

```
In [29]: # 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. We are a melting
pot of refugees, immigrants, and native-born Americans bringing the gift of language to our school. \r\n\r\n We ha
ve over 24 languages represented in our English Learner program with 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 t
he limits of your world.\"-Ludwig Wittgenstein  Our English learner's have a strong support system at home that be
gs for more resources.  Many times our parents are learning to read and speak English along side of their childre
n.  Sometimes this creates barriers for parents to be able to help their child learn phonetics, letter recognitio
n, and other reading skills.\r\n\r\nBy providing these dvd's and players, students are able to continue their mast
ery of the English language even if no one at home 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 English 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 educational dvd's for the years t
o come for other EL students.\r\nnannan
==================================================
The 51 fifth grade students that will cycle through my classroom this year all love learning, at least most of the
time. At our school, 97.3% of the students receive free or reduced price lunch. Of the 560 students, 97.3% are min
ority 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 beautiful 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 carniv
al to celebrate the hard work put in during the school year, with a dunk tank being the most popular activity.My s
tudents will use these five brightly colored Hokki stools in place of regular, stationary, 4-legged chairs. As I w
ill only have a total of ten in the classroom and not enough for each student to have 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 ea
ch 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 used by the students who need the highest amount of movement in their life in ord
er to stay focused on school.\r\n\r\nWhenever asked what the classroom is missing, my students always say more Hok
ki Stools. They can't get their fill of the 5 stools we already have. When the students are sitting in 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 taken. There are always students who head over to th
e 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 the 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 aw
ay 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 te
acher 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 uniqu
e boys and girls of mixed races in Arkansas.\r\nThey attend a Title I school, which means there is a high enough p
ercentage of free and reduced-price lunch to qualify. Our school is an \"open classroom\" concept, which is very u
nique as there are no walls separating the classrooms. These 9 and 10 year-old students are very eager learners; t
hey are like sponges, absorbing all the information and experiences and keep on wanting more.With these resources
such as the comfy red throw pillows and the whimsical nautical hanging decor and the blue fish nets, I will be abl
e to help create the mood in our classroom setting to be one of a themed nautical environment. Creating a classroo
m 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 pictures of each child with them, have them developed, and then hung in our classroom ready for their first d
ay 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 o
ne.\r\n\r\nIt costs lost of money out of my own pocket on resources to get our classroom ready. Please consider he
lping with this project to make our new school year a very successful one. Thank you!nannan
==================================================
My kindergarten students have varied disabilities ranging from speech and language delays, cognitive delays, gros
s/fine motor delays, to autism. They are eager beavers and always strive to work their 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 wh
ere most of the students receive free or reduced price lunch.  Despite their disabilities and limitations, my stud
ents 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 groove 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 develop their
core, which enhances gross motor and in Turn fine motor skills. \r\nThey also want to learn through games, my kids
don't want to sit and do worksheets. They want to learn to count by jumping and playing. Physical engagement is th
e 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.nannan
==================================================
The mediocre teacher tells. The good teacher explains. The superior teacher demonstrates. The great teacher inspir
es. -William A. Ward\r\n\r\nMy school has 803 students which is makeup is 97.6% African-American, making up the la
rgest segment of the student body. A typical school in Dallas is made up of 23.2% African-American students. Most
of the students are on free or reduced lunch. We aren't receiving doctors, lawyers, or engineers children from ric
h backgrounds or neighborhoods. As an educator I am inspiring minds of young children and we focus not only on aca
demics but one smart, effective, efficient, and disciplined students with good character.In our classroom we can u
tilize the Bluetooth for swift transitions during class. I use a speaker which doesn't amplify the sound enough to
receive the message. Due to the volume of my speaker my students can't hear videos or books clearly and it isn't m
aking the lessons as meaningful. But with the bluetooth speaker my students will be able to hear and I can stop, p
ause and replay it at any time.\r\nThe cart will allow 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 [30]:
```python
# https://stackoverflow.com/a/47091490/4084039
import re

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

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

In [31]:
```python
sent = decontracted(project_data['essay'].values[20000])
print(sent)
print("="*50)
```

```
My kindergarten students have varied disabilities ranging from speech and language delays, cognitive delays, gros
s/fine motor delays, to autism. They are eager beavers and always strive to work their 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 wh
ere most of the students receive free or reduced price lunch.  Despite their disabilities and limitations, my stud
ents 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 groove 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 develop their
core, which enhances gross motor and in Turn fine motor skills. \r\nThey 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. Physical engagement is t
he 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.nannan
==================================================
```

In [32]:
```python
# \r \n \t remove from string python: http://texthandler.com/info/remove-line-breaks-python/
sent = sent.replace('\\r', ' ')
sent = sent.replace('\\"', ' ')
sent = sent.replace('\\n', ' ')
print(sent)
```

```
My kindergarten students have varied disabilities ranging from speech and language delays, cognitive delays, gros
s/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 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 eager to learn and explore.Have you ever felt like you had ants in your pants and y
ou needed to groove 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 develop their core,
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. Physical 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 do
ing work and just have the fun a 6 year old deserves.nannan
```

In [33]:
```python
#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 cognitive delays gross f
ine motor delays to autism They are eager beavers and always strive to work their hardest working past their limit
ations 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 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 l
earn or so they say Wobble chairs are the answer and I love then because they develop their core which enhances gr
oss motor and in Turn fine motor skills They also want to learn through games my kids do not want to sit and do wo
rksheets They want to learn to count by jumping and playing Physical engagement is the key to our success The numb
er toss and color and shape mats can make that happen My students will forget they are doing work and just have th
e fun a 6 year old deserves nannan
```

In [34]:
```
# 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', 'afte
r',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'furt
her',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'm
ore',\
            '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', 'r
e', \
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn'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', "we
ren't", \
            'won', "won't", 'wouldn', "wouldn't"]
```

In [35]:
```
# Combining all the above stundents
from tqdm import tqdm
preprocessed_essays = []
# tqdm is for printing the status bar
for sentance in tqdm(project_data['essay'].values):
    sent = decontracted(sentance)
    sent = sent.replace('\\r', ' ')
    sent = sent.replace('\\"', ' ')
    sent = sent.replace('\\n', ' ')
    sent = re.sub('[^A-Za-z0-9]+', ' ', sent)
    # https://gist.github.com/sebleier/554280
    sent = ' '.join(e for e in sent.split() if e not in stopwords)
    preprocessed_essays.append(sent.lower().strip())
```

```
100%|████████████████████████████████████████████████████████████████████████| 109248/109248 [02:42<00:00, 673.80
it/s]
```

In [36]:
```
# after preprocesing
preprocessed_essays[20000]
```

Out[36]:
```
'my kindergarten students varied disabilities ranging speech language delays cognitive delays gross fine motor del
ays autism they eager beavers always strive work hardest working past limitations the materials ones i seek studen
ts i teach title i school students receive free reduced price lunch despite disabilities limitations students love
coming school come eager learn explore have ever felt like ants pants needed groove move meeting this kids feel ti
me the want able move learn say wobble chairs answer i love develop core enhances gross motor turn fine motor skil
ls they also want learn games kids not want sit worksheets they want learn count jumping playing physical engageme
nt key success the number toss color shape mats make happen my students forget work fun 6 year old deserves nanna
n'
```

In [37]:
```
project_data['preprocessed_essays'] = preprocessed_essays
project_data.drop(['essay'], axis=1, inplace=True)
```

1.4 Preprocessing of `project_title`

In [38]:
```
# similarly you can preprocess the titles also
```

In [39]:
```
project_data['project_title'][2000:2010]
```

Out[39]:
```
2000                    Steady Stools for Active Learning
2001                                   Classroom Supplies
2002    Kindergarten Students Deserve Quality  Books a...
2003                                Listen to Understand!
2004                             iPads to iGnite Learning
2005                                  Tablets For Learning
2006                                              Go P.E.!
2007                                  Making Learning Fun!
2008    Empowerment Through Silk Screen Designed Tee S...
2009                                   Let's Play Together!
Name: project_title, dtype: object
```

In [40]:
```
# Combining all the above statemennts
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)
    # 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 [00:07<00:00, 14601.46
it/s]
```

In [41]:
```
preprocessed_titles[2000:2010]
```

Out[41]:
```
['steady stools active learning',
 'classroom supplies',
 'kindergarten students deserve quality books vibrant rug',
 'listen understand',
 'ipads ignite learning',
 'tablets for learning',
 'go p e',
 'making learning fun',
 'empowerment through silk screen designed tee shirts',
 'let play together']
```

In [42]:
```
project_data['preprocessed_titles'] = preprocessed_titles
project_data.drop(['project_title'], axis=1, inplace=True)
```

1.5 Preparing data for models

In [43]:
```
project_data.columns
```

Out[43]:
```
Index(['Unnamed: 0', 'id', 'teacher_id', 'school_state',
       'project_submitted_datetime', 'project_essay_1', 'project_essay_2',
       'project_essay_3', 'project_essay_4', 'project_resource_summary',
       'teacher_number_of_previously_posted_projects', 'project_is_approved',
       'clean_categories', 'clean_subcategories', 'clean_teacher_prefix',
       'clean_project_grade_category', 'preprocessed_essays',
       'preprocessed_titles'],
      dtype='object')
```

we are going to consider

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

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

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

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

1.5.2 Vectorizing Text data

1.5.2.1 Bag of words

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

1.5.2.2 TFIDF vectorizer

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

1.5.2.3 Using Pretrained Models: Avg W2V

# 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-save-and-load-variables-in-python/ import pickle with open('glove_vectors', 'wb') as f: pickle.dump(words_courpus, f)# stronging variables into pickle files python: http://www.jessicayung.com/how-to-use-pickle-to-save-and-load-variables-in-python/ # make sure you have the glove_vectors file with open('glove_vectors', 'rb') as f: model = pickle.load(f) glove_words = set(model.keys()) # average Word2Vec # compute average word2vec for each review. avg_w2v_vectors = []; # the avg-w2v for each sentence/review is stored in this list for sentence in tqdm(preprocessed_essays): # for each review/sentence vector = np.zeros(300) # as word vectors are of zero length cnt_words =0; # num of words with a valid vector in the sentence/review for word in sentence.split(): # for each word in a review/sentence if word in glove_words: vector += model[word] cnt_words += 1 if cnt_words != 0: vector /= cnt_words avg_w2v_vectors.append(vector) print(len(avg_w2v_vectors)) print(len(avg_w2v_vectors[0]))

1.5.2.3 Using Pretrained Models: TFIDF weighted W2V

# 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())# 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 tfidf 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]))# Similarly you can vectorize for title also

1.5.3 Vectorizing Numerical features

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')# 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))price_standardized

1.5.4 Merging all the above features

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

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

```
In [ ]:
```

1.6 Merging Numerical data in Resources to project_data

In [44]:
```
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 [ ]:

In [ ]:

Computing Sentiment Scores

In [45]:
```
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer

# import nltk
# nltk.download('vader_lexicon')

sid = SentimentIntensityAnalyzer()

for_sentiment = 'a person is a person no matter how small dr seuss i teach the smallest students with the biggest e
nthusiasm \
for learning my students learn in many different ways using all of our senses and multiple intelligences i use a wi
de range\
of techniques to help all my students succeed students in my class come from a variety of different backgrounds whi
ch makes\
for wonderful sharing of experiences and cultures including native americans our school is a caring community of su
ccessful \
learners which can be seen through collaborative student project based learning in and out of the classroom kinderg
arteners \
in my class love to work with hands on materials and have many different opportunities to practice a skill before i
t is\
mastered having the social skills to work cooperatively with friends is a crucial aspect of the kindergarten curric
ulum\
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 h
ealthy \
food for snack time my students will have a grounded appreciation for the work that went into making the food and k
nowledge \
of where the ingredients came from as well as how it is healthy for their bodies this project would expand our lear
ning of \
nutrition and agricultural cooking recipes by having us peel our own apples to make homemade applesauce make our ow
n bread \
and mix up healthy plants from our classroom garden in the spring we will also create our own cookbooks to be print
ed and \
shared with families students will gain math and literature skills as well as a life long enjoyment for healthy coo
king \
nannan'
ss = sid.polarity_scores(for_sentiment)

for k in ss:
    print('{0}: {1}, '.format(k, ss[k]), end='')

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

```
neg: 0.01, neu: 0.745, pos: 0.245, compound: 0.9975,
```

In [ ]:

Assignment 9: RF and GBDT

Response Coding: Example

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

1. **Apply both Random Forrest and GBDT on these feature sets**

    - Set 1: categorical(instead of one hot encoding, try response coding (https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/handling-categorical-and-numerical-features/): use probability values), numerical features + project_title(BOW) + preprocessed_eassay (BOW)
    - Set 2: categorical(instead of one hot encoding, try response coding (https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/handling-categorical-and-numerical-features/): use probability values), numerical features + project_title(TFIDF)+ preprocessed_eassay (TFIDF)
    - Set 3: categorical(instead of one hot encoding, try response coding (https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/handling-categorical-and-numerical-features/): use probability values), numerical features + project_title(AVG W2V)+ preprocessed_eassay (AVG W2V). Here for this set take **20K** datapoints only.
    - Set 4: categorical(instead of one hot encoding, try response coding (https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/handling-categorical-and-numerical-features/): use probability values), numerical features + project_title(TFIDF W2V)+ preprocessed_eassay (TFIDF W2V). Here for this set take **20K** datapoints only.

2. **The hyper paramter tuning (Consider any two hyper parameters preferably n_estimators, max_depth)**

    - Consider the following range for hyperparameters **n_estimators** = [10, 50, 100, 150, 200, 300, 500, 1000], **max_depth** = [2, 3, 4, 5, 6, 7, 8, 9, 10]
    - Find the best hyper parameter which will give the maximum AUC (https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/receiver-operating-characteristic-curve-roc-curve-and-auc-1/) value
    - Find the best hyper paramter using simple cross validation data
    - You can write your own for loops to do this task

3. **Representation of results**

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

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

      or

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

      seaborn heat maps (https://seaborn.pydata.org/generated/seaborn.heatmap.html) with rows as **n_estimators**, columns as **max_depth**, and values inside the cell representing **AUC Score**
    - You can choose either of the plotting techniques: 3d plot or heat map
    - Once after you found the best hyper parameter, you need to train your model with it, and find the AUC on test data and plot the ROC curve on both train and test.
    - Along with plotting ROC curve, you need to print the confusion matrix (https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/confusion-matrix-tpr-fpr-fnr-tnr-1/) with predicted and original labels of test data points

4. **Conclusion**

    - You need to summarize the results at the end of the notebook, summarize it in the table format. To print out a table please refer to this prettytable library link (http://zetcode.com/python/prettytable/)

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. (https://soundcloud.com/applied-ai-course/leakage-bow-and-tfidf)

2. Random Forest and GBDT

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

```
In [46]: # 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
```

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
   - Essay : text data

   - quantity : numerical
   - teacher_number_of_previously_posted_projects : numerical
   - price : numerical
```

```
In [47]: data1 = project_data.drop(['Unnamed: 0', 'id','project_submitted_datetime','project_essay_1','project_essay_2','pro
         ject_essay_3','project_essay_4','project_resource_summary','teacher_id'], axis = 1)
```

```
In [48]: data1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 109248 entries, 0 to 109247
Data columns (total 11 columns):
school_state                                    109248 non-null object
teacher_number_of_previously_posted_projects    109248 non-null int64
project_is_approved                             109248 non-null int64
clean_categories                                109248 non-null object
clean_subcategories                             109248 non-null object
clean_teacher_prefix                            109248 non-null object
clean_project_grade_category                    109248 non-null object
preprocessed_essays                             109248 non-null object
preprocessed_titles                             109248 non-null object
price                                           109248 non-null float64
quantity                                        109248 non-null int64
dtypes: float64(1), int64(3), object(7)
memory usage: 10.0+ MB
```

```
In [49]: data1 = data1[:45000]
```

```
In [50]: y = data1['project_is_approved'].values
         X = data1.drop(['project_is_approved'], axis=1)
         X.head(1)
```

Out[50]:

	school_state	teacher_number_of_previously_posted_projects	clean_categories	clean_subcategories	clean_teacher_prefix	clean_project_grad
0	IN	0	Literacy_Language	ESL Literacy	Mrs	

```
In [51]: # train test split

         from sklearn.model_selection import train_test_split

         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, stratify = y)
```

```
In [52]: X.shape
Out[52]: (45000, 10)
```

```
In [53]: y.shape
Out[53]: (45000,)
```

2.2 Make Data Model Ready: encoding numerical, categorical features

In [54]:
```
# 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
```

2.2.1 Numerical features

1. teacher_number_of_previously_posted_projects
2. price
3. quantity

2.2.1.1 Teacher number of previously posted projects

In [55]:
```
from sklearn.preprocessing import Normalizer
normalizer = Normalizer()
# normalizer.fit(X_train['price'].values)
# this will rise an error Expected 2D array, got 1D array instead:
# array=[105.22 215.96  96.01 ... 368.98  80.53 709.67].
# Reshape your data either using
# array.reshape(-1, 1) if your data has a single feature
# array.reshape(1, -1)  if it contains a single sample.
normalizer.fit(X_train['teacher_number_of_previously_posted_projects'].values.reshape(1,-1))

X_train_TPPP_norm = normalizer.transform(X_train['teacher_number_of_previously_posted_projects'].values.reshape(1,-
1))
X_test_TPPP_norm = normalizer.transform(X_test['teacher_number_of_previously_posted_projects'].values.reshape(1,-1
))

print("After vectorizations")
print(X_train_TPPP_norm.shape, y_train.shape)
print(X_test_TPPP_norm.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(1, 30150) (30150,)
(1, 14850) (14850,)
====================================================================================================
```

In [56]:
```
print("Transpose of teacher number of previously posted projects")

X_train_TPPP_norm = X_train_TPPP_norm.transpose()
X_test_TPPP_norm = X_test_TPPP_norm.transpose()

print("After transpose")
print(X_train_TPPP_norm.shape, y_train.shape)
print(X_test_TPPP_norm.shape, y_test.shape)
print("="*100)
```

```
Transpose of teacher number of previously posted projects
After transpose
(30150, 1) (30150,)
(14850, 1) (14850,)
====================================================================================================
```

2.2.1.2 price

In [57]:
```
from sklearn.preprocessing import Normalizer
normalizer = Normalizer()
# normalizer.fit(X_train['price'].values)
# this will rise an error Expected 2D array, got 1D array instead:
# array=[105.22 215.96  96.01 ... 368.98  80.53 709.67].
# Reshape your data either using
# array.reshape(-1, 1) if your data has a single feature
# array.reshape(1, -1)  if it contains a single sample.
normalizer.fit(X_train['price'].values.reshape(1,-1))

X_train_price_norm = normalizer.transform(X_train['price'].values.reshape(1,-1))
X_test_price_norm = normalizer.transform(X_test['price'].values.reshape(1,-1))

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

```
After vectorizations
(1, 30150) (30150,)
(1, 14850) (14850,)
====================================================================================================
```

In [58]:
```
print("Transpose of price")

X_train_price_norm = X_train_price_norm.transpose()
X_test_price_norm = X_test_price_norm.transpose()


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

```
Transpose of price
After vectorizations
(30150, 1) (30150,)
(14850, 1) (14850,)
====================================================================================================
```

2.2.1.3 quantity

In [59]:
```
from sklearn.preprocessing import Normalizer
normalizer = Normalizer()
# normalizer.fit(X_train['price'].values)
# this will rise an error Expected 2D array, got 1D array instead:
# array=[105.22 215.96  96.01 ... 368.98  80.53 709.67].
# Reshape your data either using
# array.reshape(-1, 1) if your data has a single feature
# array.reshape(1, -1)  if it contains a single sample.
normalizer.fit(X_train['quantity'].values.reshape(1,-1))

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

print("After vectorizations")
print(X_train_quantity_norm.shape, y_train.shape)
print(X_test_quantity_norm.shape, y_test.shape)
print("="*100)
```

```
After vectorizations
(1, 30150) (30150,)
(1, 14850) (14850,)
====================================================================================================
```

In [60]:
```
print("Transpose of Quantity")

X_train_quantity_norm = X_train_quantity_norm.transpose()
X_test_quantity_norm = X_test_quantity_norm.transpose()


print("After vectorizations")
print(X_train_quantity_norm.shape, y_train.shape)
print(X_test_quantity_norm.shape, y_test.shape)
print("="*100)
```

```
Transpose of Quantity
After vectorizations
(30150, 1) (30150,)
(14850, 1) (14850,)
====================================================================================================
```

In [ ]:

2.2.2 Categorical Data

Categorical Features for vectorization

```
1. Clean Categories
2. Clean Sub Categories
3. School State
4. Teacher Prefix
5. Project grade category
```

In the below function, I have considered all the unique features and calculated the pos and neg pos of that particulat unique value

In [61]:
```
def response(features,label):
    df = pd.DataFrame({'A':features.values.tolist(),'label':label.tolist()})
    l = len(features.values.tolist())
    df1 = pd.DataFrame(np.nan, index=range(0,l),columns=['x','y'])
    abc = df['A'].unique()



    for i  in tqdm(range(0,len(abc))):


        count_neg = 0
        count_pos = 0

        for j in range(0,len(df)):

            if((abc[i] == df.A[j]) and df.label[j] == 0):
                count_neg = count_neg+1
                #print(i,df.A[j],df.label[j],count_neg)
            elif((abc[i] == df.A[j]) and df.label[j] == 1):
                count_pos = count_pos+1
                #print(i,df.A[j],df.label[j],count_pos)

        prob_neg = count_neg/(count_neg+count_pos)
        prob_pos = count_pos/(count_neg+count_pos)

        c = count_neg+count_pos

        for p in range(0,len(df)):
            if((abc[i] == df.A[p]) and (c>1)):
                df1.iloc[p,0]=prob_neg
                df1.iloc[p,1]=prob_pos
            elif((abc[i] == df.A[p]) and (c==1)):
                df1.iloc[p,0]= 0.5
                df1.iloc[p,1]= 0.5
                print('unique feature = ' + df.A[p])



        print('Feature = ' + abc[i])
        print('Prob neg = '+ str(prob_neg) + '      Prob pos = '+ str(prob_pos))
        print('count neg = '+ str(count_neg) + '     count pos = '+ str(count_pos) + '     sum = '+str(c))


    return df1
```

""" b=count_neg+count_pos print('Feature =' + i) print('Prob neg ='+ str(prob_neg) + 'Prob pos ='+ str(prob_pos)) print('count neg ='+ str(count_neg) + 'count pos ='+ str(count_pos) + 'sum = '+str(b)) """

In [ ]:

2.2.2.1 Teacher Prefix

In [62]:
```
X_train['clean_teacher_prefix'].value_counts()
```
Out[62]:
```
Mrs        15767
Ms         10853
Mr          2857
Teacher      671
Dr             2
Name: clean_teacher_prefix, dtype: int64
```
In [63]:
```
response_clean_teacher_prefix = response(X_train['clean_teacher_prefix'],y_train)
```
```
 20%|████████████████▌                                                                  | 1/5 [00:15<01:00, 15.15
s/it]

Feature = Mrs
Prob neg = 0.14955286357582293     Prob pos = 0.8504471364241771
count neg = 2358     count pos = 13409     sum = 15767

 40%|█████████████████████████████████▏                                                | 2/5 [00:26<00:42, 14.01
s/it]

Feature = Ms
Prob neg = 0.15562517276329127     Prob pos = 0.8443748272367088
count neg = 1689     count pos = 9164     sum = 10853

 60%|█████████████████████████████████████████████████▊                                 | 3/5 [00:32<00:23, 11.54
s/it]

Feature = Mr
Prob neg = 0.16100805040252011     Prob pos = 0.8389919495974799
count neg = 460     count pos = 2397     sum = 2857

 80%|██████████████████████████████████████████████████████████████████▍                | 4/5 [00:36<00:09,  9.29
s/it]

Feature = Teacher
Prob neg = 0.20268256333830104     Prob pos = 0.797317436661699
count neg = 136     count pos = 535     sum = 671

100%|██████████████████████████████████████████████████████████████████████████████████| 5/5 [00:39<00:00,  7.99
s/it]

Feature = Dr
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2
```
In [64]:
```
X_train['clean_teacher_prefix'][600:620]
```
Out[64]:
```
12675        Mrs
11670        Mrs
25900         Ms
8749          Mr
7880          Ms
27154        Mrs
18622         Ms
3458          Mr
23508         Ms
40992        Mrs
24134         Mr
35064         Ms
39518        Mrs
33881         Ms
3255         Mrs
35516        Mrs
6537         Mrs
37970    Teacher
39198         Ms
14886    Teacher
Name: clean_teacher_prefix, dtype: object
```

```
In [65]: response_clean_teacher_prefix[600:620]
```

Out[65]:

	x	y
600	0.149553	0.850447
601	0.149553	0.850447
602	0.155625	0.844375
603	0.161008	0.838992
604	0.155625	0.844375
605	0.149553	0.850447
606	0.155625	0.844375
607	0.161008	0.838992
608	0.155625	0.844375
609	0.149553	0.850447
610	0.161008	0.838992
611	0.155625	0.844375
612	0.149553	0.850447
613	0.155625	0.844375
614	0.149553	0.850447
615	0.149553	0.850447
616	0.149553	0.850447
617	0.202683	0.797317
618	0.155625	0.844375
619	0.202683	0.797317

```
In [ ]:
```

2.2.2.2 Clean categories

```
In [66]: X_train.clean_categories.value_counts()
```

```
Out[66]: Literacy_Language                      6554
         Math_Science                           4641
         Literacy_Language Math_Science         4009
         Health_Sports                          2808
         Music_Arts                             1397
         SpecialNeeds                           1204
         Literacy_Language SpecialNeeds         1113
         AppliedLearning                        1012
         Math_Science Literacy_Language          632
         AppliedLearning Literacy_Language       626
         Math_Science SpecialNeeds               554
         History_Civics                          518
         Math_Science Music_Arts                 476
         Literacy_Language Music_Arts            465
         History_Civics Literacy_Language        409
         AppliedLearning SpecialNeeds            402
         Health_Sports SpecialNeeds              385
         Warmth Care_Hunger                      367
         Math_Science AppliedLearning            347
         AppliedLearning Math_Science            269
         Literacy_Language History_Civics        223
         Health_Sports Literacy_Language         213
         AppliedLearning Music_Arts              211
         Math_Science History_Civics             173
         Literacy_Language AppliedLearning       171
         AppliedLearning Health_Sports           150
         Math_Science Health_Sports              121
         History_Civics Math_Science             107
         SpecialNeeds Music_Arts                  88
         History_Civics Music_Arts                87
         Health_Sports Math_Science               67
         History_Civics SpecialNeeds              66
         Health_Sports AppliedLearning            62
         AppliedLearning History_Civics           48
         Health_Sports Music_Arts                 42
         Music_Arts SpecialNeeds                  36
         Literacy_Language Health_Sports          21
         History_Civics AppliedLearning           12
         Health_Sports History_Civics             11
         SpecialNeeds Health_Sports               10
         Health_Sports Warmth Care_Hunger          8
         Music_Arts Health_Sports                  7
         AppliedLearning Warmth Care_Hunger        6
         Music_Arts History_Civics                 5
         Math_Science Warmth Care_Hunger           5
         History_Civics Health_Sports              4
         SpecialNeeds Warmth Care_Hunger           3
         Literacy_Language Warmth Care_Hunger      2
         Music_Arts AppliedLearning                2
         Music_Arts Warmth Care_Hunger             1
         Name: clean_categories, dtype: int64
```

In [67]:
```
response_clean_categories = response(X_train['clean_categories'],y_train)
```

```
  2%|█▌                                                                          | 1/50 [00:09<07:52,  9.65
s/it]

Feature = Literacy_Language
Prob neg = 0.13411657003356728     Prob pos = 0.8658834299664327
count neg = 879     count pos = 5675     sum = 6554

  4%|███▏                                                                        | 2/50 [00:14<06:34,  8.21
s/it]

Feature = Literacy_Language SpecialNeeds
Prob neg = 0.14734950584007186     Prob pos = 0.8526504941599281
count neg = 164     count pos = 949     sum = 1113

  6%|████▋                                                                       | 3/50 [00:19<05:37,  7.18
s/it]

Feature = SpecialNeeds
Prob neg = 0.18687707641196014     Prob pos = 0.8131229235880398
count neg = 225     count pos = 979     sum = 1204

  8%|██████▏                                                                     | 4/50 [00:26<05:33,  7.26
s/it]

Feature = Math_Science
Prob neg = 0.18336565395388923     Prob pos = 0.8166343460461107
count neg = 851     count pos = 3790     sum = 4641

 10%|███████▊                                                                    | 5/50 [00:32<05:11,  6.92
s/it]

Feature = Health_Sports
Prob neg = 0.150997150997151     Prob pos = 0.8490028490028491
count neg = 424     count pos = 2384     sum = 2808

 12%|█████████▎                                                                  | 6/50 [00:38<04:41,  6.40
s/it]

Feature = AppliedLearning
Prob neg = 0.1956521739130435     Prob pos = 0.8043478260869565
count neg = 198     count pos = 814     sum = 1012

 14%|███████████                                                                | 7/50 [00:43<04:16,  5.98
s/it]

Feature = Health_Sports SpecialNeeds
Prob neg = 0.14545454545454545     Prob pos = 0.8545454545454545
count neg = 56     count pos = 329     sum = 385

 16%|████████████▌                                                               | 8/50 [00:47<03:48,  5.44
s/it]

Feature = Literacy_Language Music_Arts
Prob neg = 0.18064516129032257     Prob pos = 0.8193548387096774
count neg = 84     count pos = 381     sum = 465

 18%|██████████████                                                              | 9/50 [00:51<03:23,  4.96
s/it]

Feature = Warmth Care_Hunger
Prob neg = 0.06267029972752043     Prob pos = 0.9373297002724795
count neg = 23     count pos = 344     sum = 367

 20%|███████████████▌                                                           | 10/50 [00:54<03:02,  4.56
s/it]

Feature = SpecialNeeds Music_Arts
Prob neg = 0.20454545454545456     Prob pos = 0.7954545454545454
count neg = 18     count pos = 70     sum = 88

 22%|█████████████████                                                         | 11/50 [01:01<03:24,  5.24
s/it]

Feature = Literacy_Language Math_Science
Prob neg = 0.13394861561486654     Prob pos = 0.8660513843851334
count neg = 537     count pos = 3472     sum = 4009

 24%|██████████████████▋                                                        | 12/50 [01:05<03:07,  4.94
s/it]

Feature = Math_Science Literacy_Language
Prob neg = 0.14082278481012658     Prob pos = 0.8591772151898734
count neg = 89     count pos = 543     sum = 632

 26%|████████████████████▏                                                      | 13/50 [01:10<02:59,  4.85
s/it]

Feature = Music_Arts
Prob neg = 0.1388690050107373     Prob pos = 0.8611309949892627
count neg = 194     count pos = 1203     sum = 1397
```

```
 28%|███████████████████████▋                                                              | 14/50 [01:14<02:41,  4.48
s/it]

Feature = AppliedLearning Health_Sports
Prob neg = 0.15333333333333332     Prob pos = 0.8466666666666667
count neg = 23     count pos = 127     sum = 150

 30%|█████████████████████████▌                                                           | 15/50 [01:17<02:28,  4.24
s/it]

Feature = AppliedLearning History_Civics
Prob neg = 0.1875     Prob pos = 0.8125
count neg = 9     count pos = 39     sum = 48

 32%|███████████████████████████▏                                                          | 16/50 [01:21<02:22,  4.18
s/it]

Feature = Math_Science SpecialNeeds
Prob neg = 0.18411552346570398     Prob pos = 0.8158844765342961
count neg = 102     count pos = 452     sum = 554

 34%|████████████████████████████▉                                                         | 17/50 [01:25<02:12,  4.02
s/it]

Feature = Health_Sports Literacy_Language
Prob neg = 0.18779342723004694     Prob pos = 0.812206572769953
count neg = 40     count pos = 173     sum = 213

 36%|██████████████████████████████▌                                                      | 18/50 [01:29<02:08,  4.01
s/it]

Feature = History_Civics
Prob neg = 0.16795366795366795     Prob pos = 0.832046332046332
count neg = 87     count pos = 431     sum = 518

 38%|████████████████████████████████▎                                                     | 19/50 [01:33<02:01,  3.93
s/it]

Feature = Literacy_Language History_Civics
Prob neg = 0.11210762331838565     Prob pos = 0.8878923766816144
count neg = 25     count pos = 198     sum = 223

 40%|██████████████████████████████████                                                    | 20/50 [01:37<01:58,  3.94
s/it]

Feature = Math_Science Health_Sports
Prob neg = 0.21487603305785125     Prob pos = 0.7851239669421488
count neg = 26     count pos = 95     sum = 121

 42%|███████████████████████████████████▋                                                  | 21/50 [01:41<01:55,  3.98
s/it]

Feature = AppliedLearning Literacy_Language
Prob neg = 0.1501597444089457     Prob pos = 0.8498402555910544
count neg = 94     count pos = 532     sum = 626

 44%|█████████████████████████████████████▍                                               | 22/50 [01:45<01:50,  3.95
s/it]

Feature = History_Civics Literacy_Language
Prob neg = 0.08557457212713937     Prob pos = 0.9144254278728606
count neg = 35     count pos = 374     sum = 409

 46%|███████████████████████████████████████                                               | 23/50 [01:48<01:43,  3.84
s/it]

Feature = History_Civics SpecialNeeds
Prob neg = 0.22727272727272727     Prob pos = 0.7727272727272727
count neg = 15     count pos = 51     sum = 66

 48%|████████████████████████████████████████▊                                             | 24/50 [01:52<01:39,  3.82
s/it]

Feature = Math_Science AppliedLearning
Prob neg = 0.1585014409221902     Prob pos = 0.8414985590778098
count neg = 55     count pos = 292     sum = 347

 50%|██████████████████████████████████████████▌                                          | 25/50 [01:56<01:35,  3.84
s/it]

Feature = Math_Science Music_Arts
Prob neg = 0.16596638655462184     Prob pos = 0.8340336134453782
count neg = 79     count pos = 397     sum = 476
unique feature = Music_Arts Warmth Care_Hunger

 52%|████████████████████████████████████████████▏                                         | 26/50 [01:59<01:29,  3.73
s/it]

Feature = Music_Arts Warmth Care_Hunger
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1
```

```
 54%|████████████████████████████████████████████▌                                     | 27/50 [02:03<01:26,  3.76
s/it]

Feature = AppliedLearning SpecialNeeds
Prob neg = 0.1791044776119403     Prob pos = 0.8208955223880597
count neg = 72     count pos = 330     sum = 402

 56%|██████████████████████████████████████████████▏                                   | 28/50 [02:07<01:25,  3.90
s/it]

Feature = AppliedLearning Music_Arts
Prob neg = 0.20853080568720378     Prob pos = 0.7914691943127962
count neg = 44     count pos = 167     sum = 211

 58%|███████████████████████████████████████████████▊                                  | 29/50 [02:11<01:19,  3.81
s/it]

Feature = Music_Arts SpecialNeeds
Prob neg = 0.027777777777777776     Prob pos = 0.9722222222222222
count neg = 1     count pos = 35     sum = 36

 60%|█████████████████████████████████████████████████▌                                | 30/50 [02:15<01:17,  3.87
s/it]

Feature = AppliedLearning Math_Science
Prob neg = 0.1970260223048327     Prob pos = 0.8029739776951673
count neg = 53     count pos = 216     sum = 269

 62%|███████████████████████████████████████████████████▏                              | 31/50 [02:18<01:11,  3.79
s/it]

Feature = Health_Sports History_Civics
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 11     sum = 11

 64%|████████████████████████████████████████████████████▉                             | 32/50 [02:22<01:08,  3.78
s/it]

Feature = Math_Science History_Civics
Prob neg = 0.15028901734104047     Prob pos = 0.8497109826589595
count neg = 26     count pos = 147     sum = 173

 66%|██████████████████████████████████████████████████████▍                           | 33/50 [02:26<01:02,  3.70
s/it]

Feature = Health_Sports Music_Arts
Prob neg = 0.30952380952380953     Prob pos = 0.6904761904761905
count neg = 13     count pos = 29     sum = 42

 68%|████████████████████████████████████████████████████████                          | 34/50 [02:29<00:58,  3.68
s/it]

Feature = Literacy_Language AppliedLearning
Prob neg = 0.16374269005847952     Prob pos = 0.8362573099415205
count neg = 28     count pos = 143     sum = 171

 70%|█████████████████████████████████████████████████████████▋                        | 35/50 [02:33<00:54,  3.66
s/it]

Feature = History_Civics Math_Science
Prob neg = 0.14953271028037382     Prob pos = 0.8504672897196262
count neg = 16     count pos = 91     sum = 107

 72%|███████████████████████████████████████████████████████████▎                      | 36/50 [02:37<00:50,  3.62
s/it]

Feature = SpecialNeeds Warmth Care_Hunger
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 74%|████████████████████████████████████████████████████████████▉                     | 37/50 [02:40<00:46,  3.61
s/it]

Feature = Health_Sports Math_Science
Prob neg = 0.208955223880597     Prob pos = 0.7910447761194029
count neg = 14     count pos = 53     sum = 67

 76%|██████████████████████████████████████████████████████████████▌                   | 38/50 [02:44<00:43,  3.60
s/it]

Feature = Music_Arts History_Civics
Prob neg = 0.4     Prob pos = 0.6
count neg = 2     count pos = 3     sum = 5

 78%|████████████████████████████████████████████████████████████████▏                 | 39/50 [02:47<00:39,  3.57
s/it]

Feature = Math_Science Warmth Care_Hunger
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5
```

```
 80%|█████████████████████████████████████████████████████████████████▌                | 40/50 [02:51<00:35,  3.55
s/it]

Feature = Health_Sports Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 8     sum = 8

 82%|███████████████████████████████████████████████████████████████████▎              | 41/50 [02:54<00:31,  3.54
s/it]

Feature = Health_Sports AppliedLearning
Prob neg = 0.16129032258064516     Prob pos = 0.8387096774193549
count neg = 10     count pos = 52     sum = 62

 84%|████████████████████████████████████████████████████████████████████▉             | 42/50 [02:58<00:28,  3.53
s/it]

Feature = History_Civics AppliedLearning
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 4     count pos = 8     sum = 12

 86%|██████████████████████████████████████████████████████████████████████▌           | 43/50 [03:02<00:26,  3.72
s/it]

Feature = History_Civics Music_Arts
Prob neg = 0.14942528735632185     Prob pos = 0.8505747126436781
count neg = 13     count pos = 74     sum = 87

 88%|████████████████████████████████████████████████████████████████████████▏         | 44/50 [03:05<00:22,  3.67
s/it]

Feature = Music_Arts AppliedLearning
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 90%|█████████████████████████████████████████████████████████████████████████▊        | 45/50 [03:09<00:18,  3.62
s/it]

Feature = Literacy_Language Health_Sports
Prob neg = 0.23809523809523808     Prob pos = 0.7619047619047619
count neg = 5     count pos = 16     sum = 21

 92%|███████████████████████████████████████████████████████████████████████████▍      | 46/50 [03:12<00:14,  3.60
s/it]

Feature = History_Civics Health_Sports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 94%|█████████████████████████████████████████████████████████████████████████████     | 47/50 [03:16<00:10,  3.60
s/it]

Feature = AppliedLearning Warmth Care_Hunger
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 2     count pos = 4     sum = 6

 96%|██████████████████████████████████████████████████████████████████████████████▋   | 48/50 [03:20<00:07,  3.56
s/it]

Feature = Literacy_Language Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 98%|████████████████████████████████████████████████████████████████████████████████▎ | 49/50 [03:23<00:03,  3.53
s/it]

Feature = SpecialNeeds Health_Sports
Prob neg = 0.4     Prob pos = 0.6
count neg = 4     count pos = 6     sum = 10

100%|██████████████████████████████████████████████████████████████████████████████████| 50/50 [03:28<00:00,  4.17
s/it]

Feature = Music_Arts Health_Sports
Prob neg = 0.2857142857142857     Prob pos = 0.7142857142857143
count neg = 2     count pos = 5     sum = 7
```

In [68]:
```
response_clean_categories.count()
```

Out[68]:
```
x    30150
y    30150
dtype: int64
```

2.2.2.3 Clean Sub categories

```
In [69]: X_train.clean_subcategories.value_counts()
```

```
Out[69]: Literacy                                   2671
         Literacy Mathematics                       2302
         Literature_Writing Mathematics             1607
         Literacy Literature_Writing                1516
         Mathematics                                1445
         SpecialNeeds                               1204
         Literature_Writing                         1198
         Health_Wellness                             968
         AppliedSciences Mathematics                 945
         Literacy SpecialNeeds                       686
         AppliedSciences                             674
         Gym_Fitness Health_Wellness                 654
         ESL Literacy                                634
         VisualArts                                  580
         Music                                       386
         Literature_Writing SpecialNeeds             373
         Warmth Care_Hunger                          367
         Gym_Fitness                                 340
         Mathematics SpecialNeeds                    337
         Health_Wellness SpecialNeeds                332
         EnvironmentalScience                        331
         TeamSports                                  280
         Music PerformingArts                        263
         AppliedSciences EnvironmentalScience        251
         EarlyDevelopment                            247
         EnvironmentalScience Health_LifeScience     247
         EnvironmentalScience Mathematics            241
         Health_Wellness NutritionEducation          234
         Other                                       225
         ESL Literature_Writing                      217
                                                    ... 
         TeamSports VisualArts                         1
         Other PerformingArts                          1
         Mathematics Warmth Care_Hunger                1
         Literature_Writing Warmth Care_Hunger         1
         CharacterEducation Economics                  1
         Economics Music                               1
         FinancialLiteracy History_Geography           1
         Civics_Government Extracurricular             1
         CommunityService Gym_Fitness                  1
         FinancialLiteracy Health_Wellness             1
         EnvironmentalScience ForeignLanguages         1
         CharacterEducation FinancialLiteracy          1
         EnvironmentalScience FinancialLiteracy        1
         Economics Literature_Writing                  1
         ForeignLanguages PerformingArts               1
         Other TeamSports                              1
         ForeignLanguages Gym_Fitness                  1
         Civics_Government TeamSports                  1
         Civics_Government PerformingArts              1
         EarlyDevelopment NutritionEducation           1
         CommunityService PerformingArts               1
         CommunityService History_Geography            1
         NutritionEducation VisualArts                 1
         Gym_Fitness Health_LifeScience                1
         Extracurricular Health_LifeScience            1
         History_Geography ParentInvolvement           1
         CommunityService Other                        1
         Extracurricular ForeignLanguages              1
         College_CareerPrep Gym_Fitness                1
         FinancialLiteracy Other                       1
         Name: clean_subcategories, Length: 362, dtype: int64
```

In [70]:
```
response_clean_subcategories = response(X_train['clean_subcategories'],y_train)
```

```
  0%|▏                                                                                | 1/362 [00:07<43:42,  7.26
s/it]

Feature = Literacy
Prob neg = 0.11718457506551853     Prob pos = 0.8828154249344815
count neg = 313     count pos = 2358     sum = 2671

  1%|▌                                                                               | 2/362 [00:11<37:48,  6.30
s/it]

Feature = Literacy SpecialNeeds
Prob neg = 0.13994169096209913     Prob pos = 0.8600583090379009
count neg = 96     count pos = 590     sum = 686

  1%|▊                                                                               | 3/362 [00:16<35:37,  5.95
s/it]

Feature = SpecialNeeds
Prob neg = 0.18687707641196014     Prob pos = 0.8131229235880398
count neg = 225     count pos = 979     sum = 1204

  1%|█                                                                               | 4/362 [00:20<32:05,  5.38
s/it]

Feature = AppliedSciences
Prob neg = 0.20474777448071216     Prob pos = 0.7952522255192879
count neg = 138     count pos = 536     sum = 674

  1%|▊                                                                               | 5/362 [00:24<29:51,  5.02
s/it]

Feature = Gym_Fitness Health_Wellness
Prob neg = 0.11314984709480122     Prob pos = 0.8868501529051988
count neg = 74     count pos = 580     sum = 654

  2%|█▏                                                                              | 6/362 [00:28<28:04,  4.73
s/it]

Feature = ESL Literacy
Prob neg = 0.14353312302839116     Prob pos = 0.8564668769716088
count neg = 91     count pos = 543     sum = 634

  2%|█                                                                              | 7/362 [00:32<25:49,  4.36
s/it]

Feature = Extracurricular
Prob neg = 0.17857142857142858     Prob pos = 0.8214285714285714
count neg = 5     count pos = 23     sum = 28

  2%|█▊                                                                              | 8/362 [00:36<24:47,  4.20
s/it]

Feature = TeamSports
Prob neg = 0.17857142857142858     Prob pos = 0.8214285714285714
count neg = 50     count pos = 230     sum = 280

  2%|█▌                                                                             | 9/362 [00:40<25:00,  4.25
s/it]

Feature = Health_Wellness
Prob neg = 0.13636363636363635     Prob pos = 0.8636363636363636
count neg = 132     count pos = 836     sum = 968

  3%|█▊                                                                             | 10/362 [00:44<24:10,  4.12
s/it]

Feature = Health_Wellness SpecialNeeds
Prob neg = 0.14156626506024098     Prob pos = 0.858433734939759
count neg = 47     count pos = 285     sum = 332

  3%|██▏                                                                            | 11/362 [00:47<23:16,  3.98
s/it]

Feature = Literacy PerformingArts
Prob neg = 0.14285714285714285     Prob pos = 0.8571428571428571
count neg = 4     count pos = 24     sum = 28

  3%|██▌                                                                            | 12/362 [00:51<22:58,  3.94
s/it]

Feature = Warmth Care_Hunger
Prob neg = 0.06267029972752043     Prob pos = 0.9373297002724795
count neg = 23     count pos = 344     sum = 367

  4%|██▊                                                                            | 13/362 [00:56<23:27,  4.03
s/it]

Feature = AppliedSciences Mathematics
Prob neg = 0.182010582010582     Prob pos = 0.817989417989418
count neg = 172     count pos = 773     sum = 945
```

```
  4%|███▏                                                                         | 14/362 [00:59<22:45,  3.92
s/it]

Feature = SpecialNeeds VisualArts
Prob neg = 0.20454545454545456     Prob pos = 0.7954545454545454
count neg = 18     count pos = 70     sum = 88

  4%|███▎                                                                         | 15/362 [01:03<22:20,  3.86
s/it]

Feature = EarlyDevelopment
Prob neg = 0.2145748987854251     Prob pos = 0.7854251012145749
count neg = 53     count pos = 194     sum = 247

  4%|███▌                                                                        | 16/362 [01:08<24:47,  4.30
s/it]

Feature = Literacy Mathematics
Prob neg = 0.1272806255430061     Prob pos = 0.8727193744569939
count neg = 293     count pos = 2009     sum = 2302

  5%|███▋                                                                         | 17/362 [01:13<25:48,  4.49
s/it]

Feature = Mathematics
Prob neg = 0.17301038062283736     Prob pos = 0.8269896193771626
count neg = 250     count pos = 1195     sum = 1445

  5%|████                                                                        | 18/362 [01:17<24:23,  4.25
s/it]

Feature = Health_LifeScience Literature_Writing
Prob neg = 0.10638297872340426     Prob pos = 0.8936170212765957
count neg = 5     count pos = 42     sum = 47

  5%|████▏                                                                        | 19/362 [01:22<25:25,  4.45
s/it]

Feature = Music
Prob neg = 0.09844559585492228     Prob pos = 0.9015544041450777
count neg = 38     count pos = 348     sum = 386

  6%|████▎                                                                        | 20/362 [01:26<24:29,  4.30
s/it]

Feature = Literature_Writing SpecialNeeds
Prob neg = 0.16890080428954424     Prob pos = 0.8310991957104558
count neg = 63     count pos = 310     sum = 373

  6%|████▌                                                                        | 21/362 [01:30<25:07,  4.42
s/it]

Feature = Literacy Literature_Writing
Prob neg = 0.13192612137203166     Prob pos = 0.8680738786279684
count neg = 200     count pos = 1316     sum = 1516

  6%|████▊                                                                        | 22/362 [01:36<26:15,  4.63
s/it]

Feature = Literature_Writing
Prob neg = 0.15025041736227046     Prob pos = 0.8497495826377296
count neg = 180     count pos = 1018     sum = 1198

  6%|████▉                                                                        | 23/362 [01:39<24:23,  4.32
s/it]

Feature = Health_Wellness TeamSports
Prob neg = 0.21428571428571427     Prob pos = 0.7857142857142857
count neg = 18     count pos = 66     sum = 84

  7%|█████▏                                                                       | 24/362 [01:43<23:30,  4.17
s/it]

Feature = Gym_Fitness
Prob neg = 0.16176470588235295     Prob pos = 0.8382352941176471
count neg = 55     count pos = 285     sum = 340

  7%|█████▍                                                                      | 25/362 [01:47<22:53,  4.07
s/it]

Feature = ESL Literature_Writing
Prob neg = 0.18433179723502305     Prob pos = 0.815668202764977
count neg = 40     count pos = 177     sum = 217

  7%|█████▋                                                                       | 26/362 [01:50<22:03,  3.94
s/it]

Feature = Health_LifeScience Mathematics
Prob neg = 0.19548872180451127     Prob pos = 0.8045112781954887
count neg = 26     count pos = 107     sum = 133
```

```
  7%|██████                                                                         | 27/362 [01:54<21:16,  3.81
s/it]

Feature = Literature_Writing Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 13     sum = 13

  8%|██████▍                                                                         | 28/362 [01:58<22:10,  3.98
s/it]

Feature = EarlyDevelopment Health_Wellness
Prob neg = 0.1323529411764706     Prob pos = 0.8676470588235294
count neg = 9     count pos = 59     sum = 68

  8%|██████▋                                                                         | 29/362 [02:03<22:27,  4.05
s/it]

Feature = EnvironmentalScience Health_LifeScience
Prob neg = 0.1862348178137652     Prob pos = 0.8137651821862348
count neg = 46     count pos = 201     sum = 247

  8%|███████                                                                         | 30/362 [02:06<22:09,  4.01
s/it]

Feature = College_CareerPrep SocialSciences
Prob neg = 0.2222222222222222     Prob pos = 0.7777777777777778
count neg = 2     count pos = 7     sum = 9

  9%|███████▌                                                                        | 31/362 [02:10<21:43,  3.94
s/it]

Feature = PerformingArts
Prob neg = 0.11538461538461539     Prob pos = 0.8846153846153846
count neg = 15     count pos = 115     sum = 130

  9%|███████▏                                                                        | 32/362 [02:14<21:11,  3.85
s/it]

Feature = EnvironmentalScience SpecialNeeds
Prob neg = 0.1346153846153846     Prob pos = 0.8653846153846154
count neg = 7     count pos = 45     sum = 52

  9%|███████▋                                                                        | 33/362 [02:18<21:21,  3.89
s/it]

Feature = AppliedSciences Literacy
Prob neg = 0.15337423312883436     Prob pos = 0.8466257668711656
count neg = 25     count pos = 138     sum = 163

  9%|███████▌                                                                       | 34/362 [02:21<20:52,  3.82
s/it]

Feature = Literacy VisualArts
Prob neg = 0.19205298013245034     Prob pos = 0.8079470198675497
count neg = 29     count pos = 122     sum = 151

 10%|████████▎                                                                       | 35/362 [02:25<20:34,  3.78
s/it]

Feature = Health_Wellness Literacy
Prob neg = 0.1984732824427481     Prob pos = 0.8015267175572519
count neg = 26     count pos = 105     sum = 131

 10%|████████                                                                       | 36/362 [02:29<20:39,  3.80
s/it]

Feature = SocialSciences
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 9     count pos = 45     sum = 54

 10%|████████▍                                                                       | 37/362 [02:33<20:24,  3.77
s/it]

Feature = Literature_Writing VisualArts
Prob neg = 0.20786516853932585     Prob pos = 0.7921348314606742
count neg = 37     count pos = 141     sum = 178

 10%|████████▌                                                                      | 38/362 [02:36<20:14,  3.75
s/it]

Feature = Other
Prob neg = 0.1511111111111111     Prob pos = 0.8488888888888889
count neg = 34     count pos = 191     sum = 225

 11%|█████████                                                                       | 39/362 [02:40<20:40,  3.84
s/it]

Feature = AppliedSciences Health_LifeScience
Prob neg = 0.175     Prob pos = 0.825
count neg = 28     count pos = 132     sum = 160
```

```
 11%|████████▍                                                                     | 40/362 [02:44<20:23,  3.80
s/it]

Feature = College_CareerPrep
Prob neg = 0.19387755102040816     Prob pos = 0.8061224489795918
count neg = 19     count pos = 79     sum = 98

 11%|█████████▏                                                                    | 41/362 [02:48<20:52,  3.90
s/it]

Feature = EnvironmentalScience
Prob neg = 0.18429003021148035     Prob pos = 0.8157099697885196
count neg = 61     count pos = 270     sum = 331

 12%|█████████▎                                                                    | 42/362 [02:55<25:25,  4.77
s/it]

Feature = Literature_Writing Mathematics
Prob neg = 0.1431238332296204     Prob pos = 0.8568761667703796
count neg = 230     count pos = 1377     sum = 1607

 12%|█████████▌                                                                   | 43/362 [02:59<24:27,  4.60
s/it]

Feature = Health_Wellness NutritionEducation
Prob neg = 0.1581196581196581     Prob pos = 0.8418803418803419
count neg = 37     count pos = 197     sum = 234

 12%|█████████▋                                                                    | 44/362 [03:04<24:04,  4.54
s/it]

Feature = Music PerformingArts
Prob neg = 0.11406844106463879     Prob pos = 0.8859315589353612
count neg = 30     count pos = 233     sum = 263

 12%|██████████                                                                   | 45/362 [03:08<22:57,  4.35
s/it]

Feature = Literature_Writing SocialSciences
Prob neg = 0.1     Prob pos = 0.9
count neg = 10     count pos = 90     sum = 100

 13%|██████████▏                                                                   | 46/362 [03:11<21:53,  4.16
s/it]

Feature = History_Geography
Prob neg = 0.2012987012987013     Prob pos = 0.7987012987012987
count neg = 31     count pos = 123     sum = 154

 13%|██████████▍                                                                  | 47/362 [03:15<21:19,  4.06
s/it]

Feature = Health_LifeScience
Prob neg = 0.1588785046728972     Prob pos = 0.8411214953271028
count neg = 34     count pos = 180     sum = 214

 13%|██████████▋                                                                   | 48/362 [03:19<20:42,  3.96
s/it]

Feature = AppliedSciences EnvironmentalScience
Prob neg = 0.2151394422310757     Prob pos = 0.7848605577689243
count neg = 54     count pos = 197     sum = 251

 14%|██████████▊                                                                   | 49/362 [03:22<19:55,  3.82
s/it]

Feature = AppliedSciences Health_Wellness
Prob neg = 0.15384615384615385     Prob pos = 0.8461538461538461
count neg = 2     count pos = 11     sum = 13

 14%|███████████                                                                   | 50/362 [03:26<19:28,  3.75
s/it]

Feature = CharacterEducation Literature_Writing
Prob neg = 0.23076923076923078     Prob pos = 0.7692307692307693
count neg = 12     count pos = 40     sum = 52

 14%|███████████▎                                                                  | 51/362 [03:30<19:14,  3.71
s/it]

Feature = History_Geography Literature_Writing
Prob neg = 0.10493827160493827     Prob pos = 0.8950617283950617
count neg = 17     count pos = 145     sum = 162

 14%|███████████▍                                                                  | 52/362 [03:34<19:50,  3.84
s/it]

Feature = SocialSciences SpecialNeeds
Prob neg = 0.21428571428571427     Prob pos = 0.7857142857142857
count neg = 3     count pos = 11     sum = 14
```

```
 15%|███████████▌                                                            | 53/362 [03:37<19:25,  3.77
s/it]

Feature = ForeignLanguages
Prob neg = 0.22330097087378642     Prob pos = 0.7766990291262136
count neg = 23     count pos = 80     sum = 103

 15%|████████████                                                           | 54/362 [03:41<19:01,  3.71
s/it]

Feature = CharacterEducation Literacy
Prob neg = 0.09278350515463918     Prob pos = 0.9072164948453608
count neg = 9     count pos = 88     sum = 97

 15%|████████████▏                                                           | 55/362 [03:45<18:51,  3.68
s/it]

Feature = EnvironmentalScience Mathematics
Prob neg = 0.17427385892116182     Prob pos = 0.8257261410788381
count neg = 42     count pos = 199     sum = 241

 15%|████████████▍                                                          | 56/362 [03:48<18:51,  3.70
s/it]

Feature = NutritionEducation
Prob neg = 0.28205128205128205     Prob pos = 0.717948717948718
count neg = 22     count pos = 56     sum = 78

 16%|████████████▌                                                           | 57/362 [03:52<18:28,  3.63
s/it]

Feature = Civics_Government FinancialLiteracy
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6

 16%|████████████▊                                                           | 58/362 [03:55<18:26,  3.64
s/it]

Feature = AppliedSciences College_CareerPrep
Prob neg = 0.15     Prob pos = 0.85
count neg = 18     count pos = 102     sum = 120

 16%|█████████████                                                           | 59/362 [03:59<18:16,  3.62
s/it]

Feature = Mathematics VisualArts
Prob neg = 0.16556291390728478     Prob pos = 0.8344370860927153
count neg = 25     count pos = 126     sum = 151

 17%|█████████████▏                                                          | 60/362 [04:03<18:11,  3.61
s/it]

Feature = CharacterEducation EarlyDevelopment
Prob neg = 0.34285714285714286     Prob pos = 0.6571428571428571
count neg = 12     count pos = 23     sum = 35

 17%|█████████████▍                                                          | 61/362 [04:06<18:34,  3.70
s/it]

Feature = EnvironmentalScience Literacy
Prob neg = 0.1440677966101695     Prob pos = 0.8559322033898306
count neg = 17     count pos = 101     sum = 118

 17%|█████████████▋                                                          | 62/362 [04:10<18:16,  3.65
s/it]

Feature = CommunityService Health_Wellness
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5

 17%|█████████████▉                                                         | 63/362 [04:14<18:13,  3.66
s/it]

Feature = CharacterEducation
Prob neg = 0.22340425531914893     Prob pos = 0.776595744680851
count neg = 21     count pos = 73     sum = 94

 18%|██████████████▏                                                         | 64/362 [04:17<18:05,  3.64
s/it]

Feature = Civics_Government History_Geography
Prob neg = 0.14705882352941177     Prob pos = 0.8529411764705882
count neg = 10     count pos = 58     sum = 68

 18%|██████████████▍                                                        | 65/362 [04:21<17:56,  3.63
s/it]

Feature = EarlyDevelopment Other
Prob neg = 0.06521739130434782     Prob pos = 0.9347826086956522
count neg = 3     count pos = 43     sum = 46
```

```
 18%|██████████████▌                                                              | 66/362 [04:24<17:52,  3.62
s/it]

Feature = EnvironmentalScience Literature_Writing
Prob neg = 0.14457831325301204     Prob pos = 0.8554216867469879
count neg = 12     count pos = 71     sum = 83

 19%|██████████████▋                                                              | 67/362 [04:28<17:33,  3.57
s/it]

Feature = Mathematics Other
Prob neg = 0.15384615384615385     Prob pos = 0.8461538461538461
count neg = 4     count pos = 22     sum = 26
unique feature = VisualArts Warmth Care_Hunger

 19%|███████████████                                                              | 68/362 [04:31<17:21,  3.54
s/it]

Feature = VisualArts Warmth Care_Hunger
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 19%|███████████████▏                                                             | 69/362 [04:35<17:32,  3.59
s/it]

Feature = EarlyDevelopment SpecialNeeds
Prob neg = 0.16097560975609757     Prob pos = 0.8390243902439024
count neg = 33     count pos = 172     sum = 205

 19%|███████████████▎                                                             | 70/362 [04:39<17:23,  3.57
s/it]

Feature = EnvironmentalScience VisualArts
Prob neg = 0.19117647058823528     Prob pos = 0.8088235294117647
count neg = 13     count pos = 55     sum = 68

 20%|███████████████▌                                                             | 71/362 [04:42<17:28,  3.60
s/it]

Feature = History_Geography Literacy
Prob neg = 0.05732484076433121     Prob pos = 0.9426751592356688
count neg = 9     count pos = 148     sum = 157

 20%|████████████████                                                            | 72/362 [04:46<17:19,  3.59
s/it]

Feature = EarlyDevelopment Music
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4

 20%|████████████████▏                                                            | 73/362 [04:49<17:15,  3.58
s/it]

Feature = AppliedSciences Literature_Writing
Prob neg = 0.15517241379310345     Prob pos = 0.8448275862068966
count neg = 18     count pos = 98     sum = 116

 20%|████████████████▌                                                           | 74/362 [04:53<17:10,  3.58
s/it]

Feature = College_CareerPrep Literature_Writing
Prob neg = 0.1111111111111111     Prob pos = 0.8888888888888888
count neg = 10     count pos = 80     sum = 90

 21%|████████████████▌                                                            | 75/362 [04:57<17:35,  3.68
s/it]

Feature = VisualArts
Prob neg = 0.1706896551724138     Prob pos = 0.8293103448275863
count neg = 99     count pos = 481     sum = 580

 21%|█████████████████                                                           | 76/362 [05:01<17:29,  3.67
s/it]

Feature = AppliedSciences SpecialNeeds
Prob neg = 0.15     Prob pos = 0.85
count neg = 18     count pos = 102     sum = 120

 21%|█████████████████▏                                                           | 77/362 [05:04<17:21,  3.65
s/it]

Feature = Music SpecialNeeds
Prob neg = 0.029411764705882353     Prob pos = 0.9705882352941176
count neg = 1     count pos = 33     sum = 34

 22%|█████████████████▎                                                           | 78/362 [05:09<19:23,  4.10
s/it]

Feature = College_CareerPrep Literacy
Prob neg = 0.12857142857142856     Prob pos = 0.8714285714285714
count neg = 9     count pos = 61     sum = 70
```

```
 22%|█████████████████▋                                                                | 79/362 [05:15<20:58,  4.45
s/it]

Feature = CharacterEducation VisualArts
Prob neg = 0.125     Prob pos = 0.875
count neg = 3     count pos = 21     sum = 24

 22%|█████████████████▊                                                                | 80/362 [05:18<19:51,  4.23
s/it]

Feature = College_CareerPrep PerformingArts
Prob neg = 0.4     Prob pos = 0.6
count neg = 4     count pos = 6     sum = 10

 22%|██████████████████                                                               | 81/362 [05:22<18:48,  4.02
s/it]

Feature = CommunityService
Prob neg = 0.2     Prob pos = 0.8
count neg = 3     count pos = 12     sum = 15

 23%|██████████████████▏                                                               | 82/362 [05:25<18:02,  3.87
s/it]

Feature = AppliedSciences ParentInvolvement
Prob neg = 0.15789473684210525     Prob pos = 0.8421052631578947
count neg = 3     count pos = 16     sum = 19

 23%|██████████████████                                                               | 83/362 [05:29<17:28,  3.76
s/it]

Feature = CharacterEducation CommunityService
Prob neg = 0.2857142857142857     Prob pos = 0.7142857142857143
count neg = 6     count pos = 15     sum = 21

 23%|██████████████████▊                                                               | 84/362 [05:33<18:04,  3.90
s/it]

Feature = CharacterEducation Mathematics
Prob neg = 0.10344827586206896     Prob pos = 0.896551724137931
count neg = 3     count pos = 26     sum = 29

 23%|███████████████████                                                              | 85/362 [05:37<17:36,  3.81
s/it]

Feature = History_Geography SocialSciences
Prob neg = 0.1511627906976744     Prob pos = 0.8488372093023255
count neg = 13     count pos = 73     sum = 86

 24%|███████████████████▏                                                              | 86/362 [05:40<17:09,  3.73
s/it]

Feature = EarlyDevelopment Literature_Writing
Prob neg = 0.20270270270270271     Prob pos = 0.7972972972972973
count neg = 15     count pos = 59     sum = 74

 24%|███████████████████▌                                                              | 87/362 [05:44<16:49,  3.67
s/it]

Feature = History_Geography SpecialNeeds
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 5     count pos = 25     sum = 30

 24%|███████████████████▉                                                              | 88/362 [05:48<16:58,  3.72
s/it]

Feature = Mathematics SpecialNeeds
Prob neg = 0.19881305637982197     Prob pos = 0.8011869436201781
count neg = 67     count pos = 270     sum = 337

 25%|████████████████████▌                                                             | 89/362 [05:51<16:28,  3.62
s/it]

Feature = AppliedSciences Music
Prob neg = 0.19047619047619047     Prob pos = 0.8095238095238095
count neg = 4     count pos = 17     sum = 21

 25%|████████████████████▎                                                             | 90/362 [05:55<16:26,  3.63
s/it]

Feature = EarlyDevelopment Literacy
Prob neg = 0.15463917525773196     Prob pos = 0.845360824742268
count neg = 30     count pos = 164     sum = 194

 25%|████████████████████▍                                                             | 91/362 [05:58<16:16,  3.60
s/it]

Feature = Civics_Government Literature_Writing
Prob neg = 0.16129032258064516     Prob pos = 0.8387096774193549
count neg = 5     count pos = 26     sum = 31
```

```
 25%|████████████████████                                                            | 92/362 [06:02<15:54,  3.54
s/it]

Feature = Health_Wellness SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 26%|████████████████████▋                                                            | 93/362 [06:05<16:08,  3.60
s/it]

Feature = Gym_Fitness TeamSports
Prob neg = 0.21568627450980393     Prob pos = 0.7843137254901961
count neg = 33     count pos = 120     sum = 153

 26%|████████████████████▉                                                           | 94/362 [06:09<15:56,  3.57
s/it]

Feature = Health_LifeScience VisualArts
Prob neg = 0.14285714285714285     Prob pos = 0.8571428571428571
count neg = 3     count pos = 18     sum = 21

 26%|█████████████████████▏                                                           | 95/362 [06:12<15:51,  3.57
s/it]

Feature = Literacy SocialSciences
Prob neg = 0.10891089108910891     Prob pos = 0.8910891089108911
count neg = 11     count pos = 90     sum = 101

 27%|█████████████████████▍                                                           | 96/362 [06:16<15:43,  3.55
s/it]

Feature = EnvironmentalScience History_Geography
Prob neg = 0.07317073170731707     Prob pos = 0.926829268292683
count neg = 3     count pos = 38     sum = 41

 27%|█████████████████████▋                                                           | 97/362 [06:19<15:42,  3.56
s/it]

Feature = ESL
Prob neg = 0.13274336283185842     Prob pos = 0.8672566371681416
count neg = 15     count pos = 98     sum = 113

 27%|█████████████████████▉                                                           | 98/362 [06:23<15:26,  3.51
s/it]

Feature = Health_Wellness VisualArts
Prob neg = 0.2727272727272727     Prob pos = 0.7272727272727273
count neg = 3     count pos = 8     sum = 11

 27%|██████████████████████▏                                                          | 99/362 [06:26<15:26,  3.52
s/it]

Feature = CharacterEducation Other
Prob neg = 0.17647058823529413     Prob pos = 0.8235294117647058
count neg = 6     count pos = 28     sum = 34

 28%|██████████████████████                                                          | 100/362 [06:30<15:16,  3.50
s/it]

Feature = College_CareerPrep SpecialNeeds
Prob neg = 0.2727272727272727     Prob pos = 0.7272727272727273
count neg = 9     count pos = 24     sum = 33

 28%|██████████████████████▎                                                          | 101/362 [06:33<15:20,  3.53
s/it]

Feature = ESL Mathematics
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 11     count pos = 55     sum = 66

 28%|██████████████████████▌                                                         | 102/362 [06:37<15:17,  3.53
s/it]

Feature = Health_Wellness Music
Prob neg = 0.23076923076923078     Prob pos = 0.7692307692307693
count neg = 3     count pos = 10     sum = 13

 28%|██████████████████████▊                                                          | 103/362 [06:40<15:15,  3.53
s/it]

Feature = Health_LifeScience History_Geography
Prob neg = 0.15     Prob pos = 0.85
count neg = 3     count pos = 17     sum = 20

 29%|███████████████████████                                                          | 104/362 [06:44<15:05,  3.51
s/it]

Feature = ESL EarlyDevelopment
Prob neg = 0.09523809523809523     Prob pos = 0.9047619047619048
count neg = 2     count pos = 19     sum = 21
```

```
 29%|███████████████████████▍                                                         | 105/362 [06:48<15:17,  3.57
s/it]

Feature = Health_LifeScience SpecialNeeds
Prob neg = 0.2222222222222222     Prob pos = 0.7777777777777778
count neg = 10     count pos = 35     sum = 45

 29%|███████████████████████▋                                                         | 106/362 [06:51<15:14,  3.57
s/it]

Feature = AppliedSciences EarlyDevelopment
Prob neg = 0.21153846153846154     Prob pos = 0.7884615384615384
count neg = 11     count pos = 41     sum = 52

 30%|███████████████████████▉                                                         | 107/362 [06:55<15:09,  3.57
s/it]

Feature = FinancialLiteracy Mathematics
Prob neg = 0.11627906976744186     Prob pos = 0.8837209302325582
count neg = 5     count pos = 38     sum = 43

 30%|████████████████████████▏                                                        | 108/362 [06:58<15:00,  3.55
s/it]

Feature = Economics Mathematics
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 8     sum = 8

 30%|████████████████████████                                                        | 109/362 [07:02<14:50,  3.52
s/it]

Feature = EnvironmentalScience SocialSciences
Prob neg = 0.22727272727272727     Prob pos = 0.7727272727272727
count neg = 5     count pos = 17     sum = 22

 30%|████████████████████████▍                                                        | 110/362 [07:05<14:38,  3.49
s/it]

Feature = Extracurricular SpecialNeeds
Prob neg = 0.2857142857142857     Prob pos = 0.7142857142857143
count neg = 2     count pos = 5     sum = 7
unique feature = EarlyDevelopment NutritionEducation

 31%|████████████████████████▌                                                       | 111/362 [07:09<14:34,  3.48
s/it]

Feature = EarlyDevelopment NutritionEducation
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 31%|████████████████████████▉                                                        | 112/362 [07:12<14:28,  3.47
s/it]

Feature = SpecialNeeds Warmth Care_Hunger
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 31%|█████████████████████████                                                        | 113/362 [07:16<14:31,  3.50
s/it]

Feature = Health_LifeScience Literacy
Prob neg = 0.075     Prob pos = 0.925
count neg = 6     count pos = 74     sum = 80

 31%|█████████████████████████▎                                                       | 114/362 [07:19<14:27,  3.50
s/it]

Feature = Mathematics ParentInvolvement
Prob neg = 0.2     Prob pos = 0.8
count neg = 4     count pos = 16     sum = 20

 32%|█████████████████████████▍                                                       | 115/362 [07:23<14:23,  3.50
s/it]

Feature = FinancialLiteracy
Prob neg = 0.20689655172413793     Prob pos = 0.7931034482758621
count neg = 12     count pos = 46     sum = 58

 32%|█████████████████████████▋                                                       | 116/362 [07:26<14:18,  3.49
s/it]

Feature = CharacterEducation College_CareerPrep
Prob neg = 0.14814814814814814     Prob pos = 0.8518518518518519
count neg = 4     count pos = 23     sum = 27

 32%|█████████████████████████▉                                                       | 117/362 [07:30<14:28,  3.55
s/it]

Feature = Other SpecialNeeds
Prob neg = 0.1797752808988764     Prob pos = 0.8202247191011236
count neg = 16     count pos = 73     sum = 89
```

```
 33%|██████████████████████████                                                     | 118/362 [07:34<15:22,  3.78
s/it]

Feature = NutritionEducation SpecialNeeds
Prob neg = 0.42857142857142855     Prob pos = 0.5714285714285714
count neg = 3     count pos = 4     sum = 7

 33%|██████████████████████████▍                                                    | 119/362 [07:38<15:00,  3.70
s/it]

Feature = ESL History_Geography
Prob neg = 0.1111111111111111     Prob pos = 0.8888888888888888
count neg = 1     count pos = 8     sum = 9

 33%|██████████████████████████▌                                                    | 120/362 [07:41<14:50,  3.68
s/it]

Feature = AppliedSciences VisualArts
Prob neg = 0.16477272727272727     Prob pos = 0.8352272727272727
count neg = 29     count pos = 147     sum = 176

 33%|██████████████████████████▋                                                    | 121/362 [07:45<14:38,  3.65
s/it]

Feature = Extracurricular VisualArts
Prob neg = 0.16     Prob pos = 0.84
count neg = 4     count pos = 21     sum = 25

 34%|██████████████████████████▊                                                    | 122/362 [07:48<14:34,  3.64
s/it]

Feature = Health_Wellness Literature_Writing
Prob neg = 0.11428571428571428     Prob pos = 0.8857142857142857
count neg = 8     count pos = 62     sum = 70

 34%|███████████████████████████▏                                                   | 123/362 [07:52<14:25,  3.62
s/it]

Feature = Literature_Writing PerformingArts
Prob neg = 0.09375     Prob pos = 0.90625
count neg = 3     count pos = 29     sum = 32

 34%|███████████████████████████▍                                                   | 124/362 [07:56<14:18,  3.61
s/it]

Feature = Gym_Fitness SpecialNeeds
Prob neg = 0.13043478260869565     Prob pos = 0.8695652173913043
count neg = 6     count pos = 40     sum = 46

 35%|███████████████████████████▌                                                   | 125/362 [07:59<14:10,  3.59
s/it]

Feature = CharacterEducation PerformingArts
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 35%|███████████████████████████▋                                                   | 126/362 [08:03<14:02,  3.57
s/it]

Feature = Literacy Other
Prob neg = 0.1509433962264151     Prob pos = 0.8490566037735849
count neg = 8     count pos = 45     sum = 53

 35%|████████████████████████████                                                   | 127/362 [08:06<14:02,  3.59
s/it]

Feature = EarlyDevelopment Mathematics
Prob neg = 0.24705882352941178     Prob pos = 0.7529411764705882
count neg = 21     count pos = 64     sum = 85

 35%|████████████████████████████▎                                                  | 128/362 [08:10<13:47,  3.54
s/it]

Feature = EarlyDevelopment ParentInvolvement
Prob neg = 0.2     Prob pos = 0.8
count neg = 3     count pos = 12     sum = 15

 36%|████████████████████████████▍                                                  | 129/362 [08:13<13:52,  3.57
s/it]

Feature = AppliedSciences Extracurricular
Prob neg = 0.043478260869565216     Prob pos = 0.9565217391304348
count neg = 2     count pos = 44     sum = 46

 36%|████████████████████████████▋                                                  | 130/362 [08:17<13:42,  3.54
s/it]

Feature = Health_LifeScience SocialSciences
Prob neg = 0.21739130434782608     Prob pos = 0.782608695652174
count neg = 5     count pos = 18     sum = 23
```

```
 36%|█████████████████████████████▌                                                  | 131/362 [08:20<13:37,  3.54
s/it]

Feature = AppliedSciences Other
Prob neg = 0.2     Prob pos = 0.8
count neg = 6     count pos = 24     sum = 30

 36%|█████████████████████████████▏                                                  | 132/362 [08:24<13:35,  3.54
s/it]

Feature = ESL ForeignLanguages
Prob neg = 0.15384615384615385     Prob pos = 0.8461538461538461
count neg = 2     count pos = 11     sum = 13

 37%|█████████████████████████████▍                                                  | 133/362 [08:27<13:26,  3.52
s/it]

Feature = Health_Wellness Mathematics
Prob neg = 0.22033898305084745     Prob pos = 0.7796610169491526
count neg = 13     count pos = 46     sum = 59

 37%|██████████████████████████████                                                 | 134/362 [08:31<13:24,  3.53
s/it]

Feature = History_Geography Mathematics
Prob neg = 0.21621621621621623     Prob pos = 0.7837837837837838
count neg = 8     count pos = 29     sum = 37

 37%|█████████████████████████████▋                                                  | 135/362 [08:34<13:13,  3.50
s/it]

Feature = College_CareerPrep EnvironmentalScience
Prob neg = 0.09090909090909091     Prob pos = 0.9090909090909091
count neg = 1     count pos = 10     sum = 11

 38%|██████████████████████████████                                                 | 136/362 [08:38<13:05,  3.48
s/it]

Feature = Health_Wellness History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 5     sum = 5

 38%|██████████████████████████████▎                                                 | 137/362 [08:41<12:59,  3.46
s/it]

Feature = CommunityService Literature_Writing
Prob neg = 0.25     Prob pos = 0.75
count neg = 2     count pos = 6     sum = 8

 38%|██████████████████████████████▍                                                 | 138/362 [08:45<12:54,  3.46
s/it]

Feature = PerformingArts SocialSciences
Prob neg = 0.5     Prob pos = 0.5
count neg = 2     count pos = 2     sum = 4

 38%|██████████████████████████████▋                                                 | 139/362 [08:48<13:06,  3.53
s/it]

Feature = Literacy ParentInvolvement
Prob neg = 0.175     Prob pos = 0.825
count neg = 7     count pos = 33     sum = 40

 39%|██████████████████████████████▉                                                 | 140/362 [08:52<13:00,  3.52
s/it]

Feature = Literature_Writing Other
Prob neg = 0.23076923076923078     Prob pos = 0.7692307692307693
count neg = 9     count pos = 30     sum = 39

 39%|███████████████████████████████▏                                                | 141/362 [08:55<12:58,  3.52
s/it]

Feature = Extracurricular Literacy
Prob neg = 0.36363636363636365     Prob pos = 0.6363636363636364
count neg = 4     count pos = 7     sum = 11

 39%|███████████████████████████████▍                                                | 142/362 [08:59<12:58,  3.54
s/it]

Feature = ForeignLanguages Literacy
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 11     count pos = 55     sum = 66

 40%|███████████████████████████████▌                                               | 143/362 [09:02<12:52,  3.53
s/it]

Feature = Health_LifeScience Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
```

```
 40%|███████████████████████████████▌                                               | 144/362 [09:06<12:43,  3.50
s/it]

Feature = PerformingArts VisualArts
Prob neg = 0.2692307692307692     Prob pos = 0.7307692307692307
count neg = 7     count pos = 19     sum = 26
unique feature = TeamSports VisualArts

 40%|████████████████████████████████                                              | 145/362 [09:09<12:44,  3.52
s/it]

Feature = TeamSports VisualArts
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 40%|████████████████████████████████▏                                              | 146/362 [09:13<12:42,  3.53
s/it]

Feature = Gym_Fitness Literature_Writing
Prob neg = 0.5     Prob pos = 0.5
count neg = 2     count pos = 2     sum = 4

 41%|████████████████████████████████▎                                              | 147/362 [09:17<12:37,  3.52
s/it]

Feature = College_CareerPrep Mathematics
Prob neg = 0.17142857142857143     Prob pos = 0.8285714285714286
count neg = 12     count pos = 58     sum = 70

 41%|████████████████████████████████▋                                              | 148/362 [09:20<12:35,  3.53
s/it]

Feature = AppliedSciences ESL
Prob neg = 0.19047619047619047     Prob pos = 0.8095238095238095
count neg = 4     count pos = 17     sum = 21

 41%|████████████████████████████████▉                                              | 149/362 [09:24<12:30,  3.53
s/it]

Feature = College_CareerPrep Music
Prob neg = 0.6     Prob pos = 0.4
count neg = 3     count pos = 2     sum = 5

 41%|█████████████████████████████████▏                                             | 150/362 [09:27<12:23,  3.51
s/it]

Feature = College_CareerPrep Extracurricular
Prob neg = 0.18181818181818182     Prob pos = 0.8181818181818182
count neg = 2     count pos = 9     sum = 11

 42%|█████████████████████████████████▍                                             | 151/362 [09:30<12:12,  3.47
s/it]

Feature = EnvironmentalScience Health_Wellness
Prob neg = 0.07142857142857142     Prob pos = 0.9285714285714286
count neg = 1     count pos = 13     sum = 14

 42%|█████████████████████████████████▌                                            | 152/362 [09:35<12:49,  3.66
s/it]

Feature = Health_Wellness Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 8     sum = 8

 42%|█████████████████████████████████▊                                             | 153/362 [09:38<12:45,  3.66
s/it]

Feature = AppliedSciences PerformingArts
Prob neg = 0.125     Prob pos = 0.875
count neg = 1     count pos = 7     sum = 8

 43%|██████████████████████████████████                                            | 154/362 [09:42<12:27,  3.59
s/it]

Feature = Civics_Government Literacy
Prob neg = 0.06521739130434782     Prob pos = 0.9347826086956522
count neg = 3     count pos = 43     sum = 46

 43%|██████████████████████████████████▏                                            | 155/362 [09:45<12:22,  3.59
s/it]

Feature = Health_Wellness Other
Prob neg = 0.16981132075471697     Prob pos = 0.8301886792452831
count neg = 9     count pos = 44     sum = 53

 43%|██████████████████████████████████▍                                            | 156/362 [09:49<12:15,  3.57
s/it]

Feature = Music VisualArts
Prob neg = 0.4166666666666667     Prob pos = 0.5833333333333334
count neg = 5     count pos = 7     sum = 12
```

```
 43%|█████████████████████████████████▉                                            | 157/362 [09:52<12:07,  3.55
s/it]

Feature = Health_LifeScience NutritionEducation
Prob neg = 0.4     Prob pos = 0.6
count neg = 8     count pos = 12     sum = 20

 44%|█████████████████████████████████▊                                            | 158/362 [09:56<12:02,  3.54
s/it]

Feature = ESL EnvironmentalScience
Prob neg = 0.1111111111111111     Prob pos = 0.8888888888888888
count neg = 1     count pos = 8     sum = 9
unique feature = FinancialLiteracy Other

 44%|██████████████████████████████████▏                                           | 159/362 [09:59<12:03,  3.56
s/it]

Feature = FinancialLiteracy Other
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = FinancialLiteracy SocialSciences

 44%|██████████████████████████████████▎                                           | 160/362 [10:03<11:56,  3.55
s/it]

Feature = FinancialLiteracy SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 44%|██████████████████████████████████▋                                          | 161/362 [10:06<11:48,  3.52
s/it]

Feature = Gym_Fitness NutritionEducation
Prob neg = 0.14285714285714285     Prob pos = 0.8571428571428571
count neg = 2     count pos = 12     sum = 14

 45%|██████████████████████████████████▉                                           | 162/362 [10:10<11:42,  3.51
s/it]

Feature = History_Geography VisualArts
Prob neg = 0.16     Prob pos = 0.84
count neg = 8     count pos = 42     sum = 50

 45%|███████████████████████████████████                                          | 163/362 [10:13<11:39,  3.51
s/it]

Feature = EarlyDevelopment VisualArts
Prob neg = 0.23684210526315788     Prob pos = 0.7631578947368421
count neg = 9     count pos = 29     sum = 38
unique feature = Civics_Government ESL

 45%|███████████████████████████████████▎                                          | 164/362 [10:18<12:51,  3.90
s/it]

Feature = Civics_Government ESL
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Civics_Government Extracurricular

 46%|███████████████████████████████████▌                                          | 165/362 [10:23<14:02,  4.27
s/it]

Feature = Civics_Government Extracurricular
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 46%|███████████████████████████████████▊                                          | 166/362 [10:27<13:42,  4.20
s/it]

Feature = AppliedSciences SocialSciences
Prob neg = 0.07142857142857142     Prob pos = 0.9285714285714286
count neg = 1     count pos = 13     sum = 14
unique feature = History_Geography ParentInvolvement

 46%|███████████████████████████████████▉                                          | 167/362 [10:31<12:57,  3.99
s/it]

Feature = History_Geography ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 46%|████████████████████████████████████▏                                         | 168/362 [10:34<12:23,  3.83
s/it]

Feature = Extracurricular PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 8     sum = 8

 47%|████████████████████████████████████▍                                         | 169/362 [10:38<11:58,  3.72
s/it]
```

```
Feature = Music ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 47%|█████████████████████████████████████▌                                          | 170/362 [10:41<11:42,  3.66
s/it]

Feature = AppliedSciences History_Geography
Prob neg = 0.22727272727272727     Prob pos = 0.7727272727272727
count neg = 5     count pos = 17     sum = 22

 47%|█████████████████████████████████████▊                                          | 171/362 [10:45<11:38,  3.66
s/it]

Feature = Gym_Fitness Literacy
Prob neg = 0.5     Prob pos = 0.5
count neg = 4     count pos = 4     sum = 8

 48%|██████████████████████████████████████                                          | 172/362 [10:49<11:32,  3.65
s/it]

Feature = SocialSciences VisualArts
Prob neg = 0.05263157894736842     Prob pos = 0.9473684210526315
count neg = 1     count pos = 18     sum = 19

 48%|██████████████████████████████████████▎                                         | 173/362 [10:52<11:22,  3.61
s/it]

Feature = Extracurricular Literature_Writing
Prob neg = 0.1     Prob pos = 0.9
count neg = 1     count pos = 9     sum = 10

 48%|██████████████████████████████████████▍                                         | 174/362 [10:56<11:19,  3.61
s/it]

Feature = Civics_Government SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 25     sum = 25

 48%|██████████████████████████████████████▋                                         | 175/362 [10:59<11:07,  3.57
s/it]

Feature = Economics
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 13     sum = 13

 49%|██████████████████████████████████████▉                                         | 176/362 [11:03<11:02,  3.56
s/it]

Feature = CharacterEducation SpecialNeeds
Prob neg = 0.13725490196078433     Prob pos = 0.8627450980392157
count neg = 7     count pos = 44     sum = 51

 49%|███████████████████████████████████████                                         | 177/362 [11:06<11:00,  3.57
s/it]

Feature = Literacy NutritionEducation
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 49%|███████████████████████████████████████▎                                        | 178/362 [11:10<10:58,  3.58
s/it]

Feature = ESL SpecialNeeds
Prob neg = 0.10204081632653061     Prob pos = 0.8979591836734694
count neg = 5     count pos = 44     sum = 49

 49%|███████████████████████████████████████▌                                        | 179/362 [11:13<10:49,  3.55
s/it]

Feature = College_CareerPrep Health_LifeScience
Prob neg = 0.1     Prob pos = 0.9
count neg = 1     count pos = 9     sum = 10

 50%|███████████████████████████████████████▊                                        | 180/362 [11:17<10:47,  3.56
s/it]

Feature = College_CareerPrep ForeignLanguages
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5

 50%|████████████████████████████████████████                                        | 181/362 [11:21<10:43,  3.56
s/it]

Feature = Civics_Government Health_LifeScience
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 50%|████████████████████████████████████████▏                                       | 182/362 [11:24<10:39,  3.55
s/it]
```

```
Feature = CharacterEducation Extracurricular
Prob neg = 0.3125     Prob pos = 0.6875
count neg = 5     count pos = 11     sum = 16

 51%|████████████████████████████████████████▍                                        | 183/362 [11:28<10:38,  3.56
s/it]

Feature = Economics FinancialLiteracy
Prob neg = 0.3125     Prob pos = 0.6875
count neg = 5     count pos = 11     sum = 16

 51%|████████████████████████████████████████▋                                        | 184/362 [11:31<10:41,  3.61
s/it]

Feature = CommunityService NutritionEducation
Prob neg = 1.0     Prob pos = 0.0
count neg = 2     count pos = 0     sum = 2

 51%|████████████████████████████████████████▊                                        | 185/362 [11:35<11:01,  3.74
s/it]

Feature = ForeignLanguages Literature_Writing
Prob neg = 0.17391304347826086     Prob pos = 0.8260869565217391
count neg = 4     count pos = 19     sum = 23

 51%|█████████████████████████████████████████                                        | 186/362 [11:39<10:44,  3.66
s/it]

Feature = ParentInvolvement SpecialNeeds
Prob neg = 0.25     Prob pos = 0.75
count neg = 2     count pos = 6     sum = 8

 52%|█████████████████████████████████████████▎                                       | 187/362 [11:42<10:35,  3.63
s/it]

Feature = Literacy Music
Prob neg = 0.125     Prob pos = 0.875
count neg = 5     count pos = 35     sum = 40

 52%|█████████████████████████████████████████▌                                       | 188/362 [11:46<10:28,  3.61
s/it]

Feature = Extracurricular Mathematics
Prob neg = 0.23529411764705882     Prob pos = 0.7647058823529411
count neg = 4     count pos = 13     sum = 17

 52%|█████████████████████████████████████████▊                                       | 189/362 [11:50<10:23,  3.61
s/it]

Feature = CharacterEducation Health_Wellness
Prob neg = 0.10344827586206896     Prob pos = 0.896551724137931
count neg = 3     count pos = 26     sum = 29

 52%|█████████████████████████████████████████▉                                       | 190/362 [11:53<10:17,  3.59
s/it]

Feature = EarlyDevelopment Extracurricular
Prob neg = 0.6666666666666666     Prob pos = 0.3333333333333333
count neg = 2     count pos = 1     sum = 3

 53%|██████████████████████████████████████████▎                                      | 191/362 [11:57<10:08,  3.56
s/it]

Feature = Civics_Government Economics
Prob neg = 0.09090909090909091     Prob pos = 0.9090909090909091
count neg = 1     count pos = 10     sum = 11

 53%|██████████████████████████████████████████▍                                      | 192/362 [12:00<10:03,  3.55
s/it]

Feature = NutritionEducation TeamSports
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 53%|██████████████████████████████████████████▋                                      | 193/362 [12:04<09:58,  3.54
s/it]

Feature = Mathematics TeamSports
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 54%|██████████████████████████████████████████▊                                      | 194/362 [12:07<09:49,  3.51
s/it]

Feature = CommunityService Extracurricular
Prob neg = 0.2222222222222222     Prob pos = 0.7777777777777778
count neg = 2     count pos = 7     sum = 9

 54%|███████████████████████████████████████████                                      | 195/362 [12:11<09:38,  3.47
s/it]
```

```
Feature = AppliedSciences Civics_Government
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 54%|███████████████████████████████████████████▍                                    | 196/362 [12:14<09:36,  3.47
s/it]

Feature = Extracurricular ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 54%|███████████████████████████████████████████▌                                    | 197/362 [12:18<09:34,  3.48
s/it]

Feature = Extracurricular Other
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 3     count pos = 6     sum = 9

 55%|███████████████████████████████████████████▊                                    | 198/362 [12:21<09:29,  3.47
s/it]

Feature = CommunityService Literacy
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6

 55%|███████████████████████████████████████████▉                                    | 199/362 [12:25<09:31,  3.50
s/it]

Feature = College_CareerPrep VisualArts
Prob neg = 0.1875     Prob pos = 0.8125
count neg = 6     count pos = 26     sum = 32

 55%|████████████████████████████████████████████▏                                   | 200/362 [12:28<09:29,  3.52
s/it]

Feature = College_CareerPrep FinancialLiteracy
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4
unique feature = Economics Other

 56%|████████████████████████████████████████████▍                                   | 201/362 [12:32<09:26,  3.52
s/it]

Feature = Economics Other
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 56%|████████████████████████████████████████████▋                                   | 202/362 [12:35<09:21,  3.51
s/it]

Feature = Mathematics Music
Prob neg = 0.1111111111111111     Prob pos = 0.8888888888888888
count neg = 2     count pos = 16     sum = 18

 56%|████████████████████████████████████████████▊                                   | 203/362 [12:39<09:17,  3.51
s/it]

Feature = Extracurricular Music
Prob neg = 0.125     Prob pos = 0.875
count neg = 1     count pos = 7     sum = 8

 56%|█████████████████████████████████████████████                                   | 204/362 [12:42<09:15,  3.52
s/it]

Feature = Literacy TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 6     sum = 6

 57%|█████████████████████████████████████████████▎                                  | 205/362 [12:46<09:12,  3.52
s/it]

Feature = CommunityService VisualArts
Prob neg = 0.38461538461538464     Prob pos = 0.6153846153846154
count neg = 5     count pos = 8     sum = 13

 57%|█████████████████████████████████████████████▌                                  | 206/362 [12:49<09:11,  3.53
s/it]

Feature = Health_LifeScience Health_Wellness
Prob neg = 0.15217391304347827     Prob pos = 0.8478260869565217
count neg = 7     count pos = 39     sum = 46
unique feature = EnvironmentalScience FinancialLiteracy

 57%|█████████████████████████████████████████████▋                                  | 207/362 [12:53<09:10,  3.55
s/it]

Feature = EnvironmentalScience FinancialLiteracy
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 57%|█████████████████████████████████████████████▉                                  | 208/362 [12:56<09:05,  3.54
s/it]
```

```
Feature = Mathematics SocialSciences
Prob neg = 0.1111111111111111     Prob pos = 0.8888888888888888
count neg = 3     count pos = 24     sum = 27

 58%|██████████████████████████████████████████████▏                                  | 209/362 [13:00<09:00,  3.53
s/it]

Feature = Gym_Fitness VisualArts
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4

 58%|██████████████████████████████████████████████▍                                  | 210/362 [13:04<09:02,  3.57
s/it]

Feature = Other VisualArts
Prob neg = 0.2222222222222222     Prob pos = 0.7777777777777778
count neg = 4     count pos = 14     sum = 18

 58%|██████████████████████████████████████████████▋                                  | 211/362 [13:07<08:54,  3.54
s/it]

Feature = Extracurricular TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 8     sum = 8

 59%|██████████████████████████████████████████████▊                                  | 212/362 [13:10<08:50,  3.54
s/it]

Feature = Gym_Fitness Music
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 2     count pos = 4     sum = 6

 59%|███████████████████████████████████████████████                                 | 213/362 [13:14<08:47,  3.54
s/it]

Feature = College_CareerPrep EarlyDevelopment
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 10     sum = 10

 59%|███████████████████████████████████████████████▎                                 | 214/362 [13:18<08:42,  3.53
s/it]

Feature = Literature_Writing ParentInvolvement
Prob neg = 0.08333333333333333     Prob pos = 0.9166666666666666
count neg = 1     count pos = 11     sum = 12

 59%|███████████████████████████████████████████████▌                                | 215/362 [13:21<08:36,  3.51
s/it]

Feature = CommunityService SpecialNeeds
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 3     count pos = 6     sum = 9

 60%|███████████████████████████████████████████████▋                                 | 216/362 [13:25<08:34,  3.52
s/it]

Feature = CharacterEducation ESL
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 60%|███████████████████████████████████████████████▉                                 | 217/362 [13:28<08:30,  3.52
s/it]

Feature = CommunityService EnvironmentalScience
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 2     count pos = 10     sum = 12

 60%|████████████████████████████████████████████████▏                                | 218/362 [13:32<08:37,  3.59
s/it]

Feature = History_Geography PerformingArts
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6
unique feature = Economics Literature_Writing

 60%|████████████████████████████████████████████████▍                                | 219/362 [13:36<08:44,  3.67
s/it]

Feature = Economics Literature_Writing
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 61%|████████████████████████████████████████████████▌                               | 220/362 [13:39<08:32,  3.61
s/it]

Feature = College_CareerPrep Economics
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 61%|████████████████████████████████████████████████▊                                | 221/362 [13:43<08:27,  3.60
s/it]
```

```
Feature = Economics EnvironmentalScience
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 61%|█████████████████████████████████████████████████                                        | 222/362 [13:46<08:28,  3.63
s/it]

Feature = College_CareerPrep Other
Prob neg = 0.23076923076923078     Prob pos = 0.7692307692307693
count neg = 6     count pos = 20     sum = 26

 62%|█████████████████████████████████████████████████▍                                       | 223/362 [13:50<08:22,  3.61
s/it]

Feature = CharacterEducation ParentInvolvement
Prob neg = 0.25     Prob pos = 0.75
count neg = 3     count pos = 9     sum = 12

 62%|█████████████████████████████████████████████████▌                                       | 224/362 [13:53<08:11,  3.56
s/it]

Feature = AppliedSciences TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 5     sum = 5

 62%|█████████████████████████████████████████████████▊                                       | 225/362 [13:57<08:03,  3.53
s/it]

Feature = History_Geography Other
Prob neg = 0.4     Prob pos = 0.6
count neg = 2     count pos = 3     sum = 5
unique feature = ForeignLanguages Gym_Fitness

 62%|██████████████████████████████████████████████████▊                                      | 226/362 [14:01<08:03,  3.56
s/it]

Feature = ForeignLanguages Gym_Fitness
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 63%|██████████████████████████████████████████████████▏                                      | 227/362 [14:04<07:57,  3.54
s/it]

Feature = CharacterEducation Music
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4
unique feature = Civics_Government NutritionEducation

 63%|██████████████████████████████████████████████████▌                                      | 228/362 [14:08<07:52,  3.52
s/it]

Feature = Civics_Government NutritionEducation
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 63%|██████████████████████████████████████████████████▊                                      | 229/362 [14:11<07:49,  3.53
s/it]

Feature = EarlyDevelopment SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 64%|███████████████████████████████████████████████████▌                                     | 230/362 [14:15<07:46,  3.54
s/it]

Feature = Health_Wellness PerformingArts
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4

 64%|███████████████████████████████████████████████████                                      | 231/362 [14:18<07:42,  3.53
s/it]

Feature = EarlyDevelopment Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 64%|███████████████████████████████████████████████████▍                                     | 232/362 [14:22<07:38,  3.53
s/it]

Feature = FinancialLiteracy Literature_Writing
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 64%|███████████████████████████████████████████████████▋                                     | 233/362 [14:25<07:32,  3.51
s/it]

Feature = CommunityService ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3
unique feature = Literature_Writing Warmth Care_Hunger
```

```
 65%|███████████████████████████████████████████████████▏                            | 234/362 [14:29<07:32,  3.53
s/it]

Feature = Literature_Writing Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 65%|███████████████████████████████████████████████████▋                            | 235/362 [14:32<07:30,  3.55
s/it]

Feature = CharacterEducation ForeignLanguages
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 65%|████████████████████████████████████████████████████▏                           | 236/362 [14:36<07:28,  3.56
s/it]

Feature = Civics_Government
Prob neg = 0.25     Prob pos = 0.75
count neg = 4     count pos = 12     sum = 16

 65%|████████████████████████████████████████████████████▌                           | 237/362 [14:39<07:21,  3.53
s/it]

Feature = ESL Other
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2
unique feature = College_CareerPrep Gym_Fitness

 66%|█████████████████████████████████████████████████████                          | 238/362 [14:43<07:17,  3.53
s/it]

Feature = College_CareerPrep Gym_Fitness
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 66%|████████████████████████████████████████████████████▊                           | 239/362 [14:46<07:15,  3.54
s/it]

Feature = Extracurricular Health_Wellness
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 66%|█████████████████████████████████████████████████████                          | 240/362 [14:50<07:10,  3.53
s/it]

Feature = ForeignLanguages Health_Wellness
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 67%|█████████████████████████████████████████████████████▏                          | 241/362 [14:53<07:04,  3.51
s/it]

Feature = ESL PerformingArts
Prob neg = 0.5     Prob pos = 0.5
count neg = 2     count pos = 2     sum = 4

 67%|█████████████████████████████████████████████████████▍                          | 242/362 [14:57<07:01,  3.51
s/it]

Feature = ForeignLanguages SocialSciences
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 67%|█████████████████████████████████████████████████████▊                          | 243/362 [15:00<06:58,  3.51
s/it]

Feature = EarlyDevelopment FinancialLiteracy
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 67%|██████████████████████████████████████████████████████▏                         | 244/362 [15:04<06:55,  3.52
s/it]

Feature = NutritionEducation Other
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 68%|██████████████████████████████████████████████████████▏                         | 245/362 [15:08<06:53,  3.54
s/it]

Feature = ParentInvolvement PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = Civics_Government TeamSports

 68%|██████████████████████████████████████████████████████▍                         | 246/362 [15:11<06:49,  3.53
s/it]

Feature = Civics_Government TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
```

```
 68%|██████████████████████████████████████████████████████                          | 247/362 [15:14<06:42,  3.50
s/it]

Feature = ESL Health_LifeScience
Prob neg = 0.18181818181818182     Prob pos = 0.8181818181818182
count neg = 2     count pos = 9     sum = 11

 69%|██████████████████████████████████████████████████████▊                         | 248/362 [15:18<06:40,  3.51
s/it]

Feature = EarlyDevelopment EnvironmentalScience
Prob neg = 0.21428571428571427     Prob pos = 0.7857142857142857
count neg = 3     count pos = 11     sum = 14

 69%|███████████████████████████████████████████████████████                         | 249/362 [15:21<06:33,  3.48
s/it]

Feature = Other ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 69%|███████████████████████████████████████████████████████▎                        | 250/362 [15:25<06:30,  3.48
s/it]

Feature = ForeignLanguages Mathematics
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 11     sum = 11

 69%|███████████████████████████████████████████████████████▍                        | 251/362 [15:30<07:24,  4.01
s/it]

Feature = College_CareerPrep Health_Wellness
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6

 70%|███████████████████████████████████████████████████████▋                        | 252/362 [15:36<08:13,  4.49
s/it]

Feature = CharacterEducation TeamSports
Prob neg = 0.75     Prob pos = 0.25
count neg = 3     count pos = 1     sum = 4

 70%|███████████████████████████████████████████████████████▉                        | 253/362 [15:39<07:40,  4.22
s/it]

Feature = Health_LifeScience Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 5     sum = 5

 70%|████████████████████████████████████████████████████████▏                       | 254/362 [15:43<07:12,  4.01
s/it]

Feature = ForeignLanguages SpecialNeeds
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 5     sum = 5

 70%|████████████████████████████████████████████████████████▎                       | 255/362 [15:46<06:55,  3.88
s/it]

Feature = EarlyDevelopment Gym_Fitness
Prob neg = 0.1111111111111111     Prob pos = 0.8888888888888888
count neg = 1     count pos = 8     sum = 9

 71%|████████████████████████████████████████████████████████                        | 256/362 [15:50<06:40,  3.78
s/it]

Feature = ParentInvolvement
Prob neg = 0.4444444444444444     Prob pos = 0.5555555555555556
count neg = 4     count pos = 5     sum = 9

 71%|████████████████████████████████████████████████████████▊                       | 257/362 [15:54<06:29,  3.71
s/it]

Feature = College_CareerPrep NutritionEducation
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = AppliedSciences Warmth Care_Hunger

 71%|█████████████████████████████████████████████████████████                       | 258/362 [15:57<06:19,  3.64
s/it]

Feature = AppliedSciences Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 72%|█████████████████████████████████████████████████████████▏                      | 259/362 [16:01<06:13,  3.62
s/it]

Feature = Economics History_Geography
Prob neg = 0.14285714285714285     Prob pos = 0.8571428571428571
count neg = 1     count pos = 6     sum = 7
```

```
 72%|████████████████████████████████████████████████████████████▌                        | 260/362 [16:04<06:10,  3.63
s/it]

Feature = EnvironmentalScience ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 5     sum = 5

 72%|████████████████████████████████████████████████████████████▊                        | 261/362 [16:08<06:03,  3.60
s/it]

Feature = AppliedSciences Gym_Fitness
Prob neg = 0.4     Prob pos = 0.6
count neg = 2     count pos = 3     sum = 5

 72%|████████████████████████████████████████████████████████████▉                        | 262/362 [16:12<06:10,  3.70
s/it]

Feature = Extracurricular Gym_Fitness
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 73%|█████████████████████████████████████████████████████████████                       | 263/362 [16:15<06:07,  3.72
s/it]

Feature = AppliedSciences CharacterEducation
Prob neg = 0.15384615384615385     Prob pos = 0.8461538461538461
count neg = 2     count pos = 11     sum = 13

 73%|█████████████████████████████████████████████████████████████▎                      | 264/362 [16:19<05:59,  3.66
s/it]

Feature = ParentInvolvement VisualArts
Prob neg = 0.1111111111111111     Prob pos = 0.8888888888888888
count neg = 1     count pos = 8     sum = 9
unique feature = Other Warmth Care_Hunger

 73%|█████████████████████████████████████████████████████████████▍                      | 265/362 [16:23<05:51,  3.62
s/it]

Feature = Other Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Extracurricular NutritionEducation

 73%|█████████████████████████████████████████████████████████████▋                      | 266/362 [16:26<05:48,  3.63
s/it]

Feature = Extracurricular NutritionEducation
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 74%|██████████████████████████████████████████████████████████████                      | 267/362 [16:30<05:43,  3.62
s/it]

Feature = ESL SocialSciences
Prob neg = 0.4     Prob pos = 0.6
count neg = 2     count pos = 3     sum = 5

 74%|██████████████████████████████████████████████████████████████▏                     | 268/362 [16:33<05:39,  3.61
s/it]

Feature = Gym_Fitness Mathematics
Prob neg = 0.14285714285714285     Prob pos = 0.8571428571428571
count neg = 1     count pos = 6     sum = 7
unique feature = CommunityService Music

 74%|██████████████████████████████████████████████████████████████▍                     | 269/362 [16:37<05:35,  3.60
s/it]

Feature = CommunityService Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 75%|██████████████████████████████████████████████████████████████▋                     | 270/362 [16:40<05:29,  3.58
s/it]

Feature = Economics Literacy
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 75%|██████████████████████████████████████████████████████████████▊                     | 271/362 [16:44<05:22,  3.54
s/it]

Feature = ESL Health_Wellness
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6

 75%|███████████████████████████████████████████████████████████████                     | 272/362 [16:48<05:21,  3.58
s/it]
```

```
Feature = Civics_Government EnvironmentalScience
Prob neg = 0.2857142857142857     Prob pos = 0.7142857142857143
count neg = 2     count pos = 5     sum = 7

 75%|████████████████████████████████████████████████████████████▎                   | 273/362 [16:51<05:16,  3.55
s/it]

Feature = ESL VisualArts
Prob neg = 0.3     Prob pos = 0.7
count neg = 3     count pos = 7     sum = 10

 76%|████████████████████████████████████████████████████████████▌                   | 274/362 [16:55<05:11,  3.54
s/it]

Feature = SpecialNeeds TeamSports
Prob neg = 0.4     Prob pos = 0.6
count neg = 4     count pos = 6     sum = 10

 76%|████████████████████████████████████████████████████████████▋                   | 275/362 [16:58<05:06,  3.52
s/it]

Feature = Health_LifeScience TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 76%|█████████████████████████████████████████████████████████████                   | 276/362 [17:02<05:20,  3.73
s/it]

Feature = ParentInvolvement SocialSciences
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 77%|█████████████████████████████████████████████████████████████▏                  | 277/362 [17:06<05:12,  3.67
s/it]

Feature = Civics_Government SpecialNeeds
Prob neg = 0.5     Prob pos = 0.5
count neg = 2     count pos = 2     sum = 4

 77%|█████████████████████████████████████████████████████████████▍                  | 278/362 [17:09<05:04,  3.63
s/it]

Feature = AppliedSciences CommunityService
Prob neg = 0.2857142857142857     Prob pos = 0.7142857142857143
count neg = 2     count pos = 5     sum = 7

 77%|█████████████████████████████████████████████████████████████▋                  | 279/362 [17:13<04:57,  3.59
s/it]

Feature = Extracurricular History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 77%|█████████████████████████████████████████████████████████████▉                  | 280/362 [17:16<04:52,  3.57
s/it]

Feature = CommunityService Mathematics
Prob neg = 0.4     Prob pos = 0.6
count neg = 2     count pos = 3     sum = 5

 78%|██████████████████████████████████████████████████████████████                  | 281/362 [17:20<04:48,  3.56
s/it]

Feature = CharacterEducation SocialSciences
Prob neg = 0.2857142857142857     Prob pos = 0.7142857142857143
count neg = 2     count pos = 5     sum = 7

 78%|██████████████████████████████████████████████████████████████▎                 | 282/362 [17:23<04:43,  3.54
s/it]

Feature = Literature_Writing TeamSports
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 78%|██████████████████████████████████████████████████████████████▌                 | 283/362 [17:27<04:37,  3.52
s/it]

Feature = History_Geography Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 78%|██████████████████████████████████████████████████████████████▊                 | 284/362 [17:30<04:33,  3.51
s/it]

Feature = Music TeamSports
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 79%|██████████████████████████████████████████████████████████████▉                 | 285/362 [17:35<04:49,  3.76
s/it]
```

```
Feature = ForeignLanguages History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 5     sum = 5
unique feature = Music SocialSciences

 79%|███████████████████████████████████████████████████████████████▏                | 286/362 [17:38<04:41,  3.70
s/it]

Feature = Music SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 79%|███████████████████████████████████████████████████████████████▍                | 287/362 [17:42<04:35,  3.67
s/it]

Feature = ForeignLanguages VisualArts
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5

 80%|███████████████████████████████████████████████████████████████▋                | 288/362 [17:46<04:32,  3.68
s/it]

Feature = EnvironmentalScience NutritionEducation
Prob neg = 0.5     Prob pos = 0.5
count neg = 5     count pos = 5     sum = 10

 80%|███████████████████████████████████████████████████████████████▊                | 289/362 [17:49<04:24,  3.62
s/it]

Feature = Mathematics PerformingArts
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4

 80%|████████████████████████████████████████████████████████████████                | 290/362 [17:53<04:16,  3.57
s/it]

Feature = EnvironmentalScience Other
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 80%|████████████████████████████████████████████████████████████████▎               | 291/362 [17:56<04:11,  3.54
s/it]

Feature = FinancialLiteracy SpecialNeeds
Prob neg = 0.23529411764705882     Prob pos = 0.7647058823529411
count neg = 4     count pos = 13     sum = 17

 81%|████████████████████████████████████████████████████████████████▌               | 292/362 [17:59<04:06,  3.52
s/it]

Feature = Gym_Fitness PerformingArts
Prob neg = 1.0     Prob pos = 0.0
count neg = 2     count pos = 0     sum = 2

 81%|████████████████████████████████████████████████████████████████▊               | 293/362 [18:03<04:00,  3.49
s/it]

Feature = EarlyDevelopment PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 5     sum = 5

 81%|████████████████████████████████████████████████████████████████▉               | 294/362 [18:06<03:58,  3.50
s/it]

Feature = Civics_Government CommunityService
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 81%|█████████████████████████████████████████████████████████████████▏              | 295/362 [18:10<03:56,  3.53
s/it]

Feature = EarlyDevelopment Health_LifeScience
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6

 82%|█████████████████████████████████████████████████████████████████▍              | 296/362 [18:13<03:50,  3.49
s/it]

Feature = EnvironmentalScience Extracurricular
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2
unique feature = Civics_Government PerformingArts

 82%|█████████████████████████████████████████████████████████████████▋              | 297/362 [18:17<03:46,  3.49
s/it]

Feature = Civics_Government PerformingArts
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 82%|█████████████████████████████████████████████████████████████████▊              | 298/362 [18:21<03:46,  3.53
s/it]
```

```
Feature = CommunityService SocialSciences
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 83%|██████████████████████████████████████████████████████████████████              | 299/362 [18:24<03:42,  3.54
s/it]

Feature = Civics_Government Mathematics
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5

 83%|██████████████████████████████████████████████████████████████████▎             | 300/362 [18:27<03:36,  3.49
s/it]

Feature = NutritionEducation SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 83%|██████████████████████████████████████████████████████████████████▌             | 301/362 [18:31<03:32,  3.48
s/it]

Feature = ForeignLanguages Health_LifeScience
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3
unique feature = Economics SpecialNeeds

 83%|██████████████████████████████████████████████████████████████████▋             | 302/362 [18:35<03:31,  3.52
s/it]

Feature = Economics SpecialNeeds
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 84%|██████████████████████████████████████████████████████████████████▉             | 303/362 [18:38<03:27,  3.52
s/it]

Feature = Gym_Fitness Other
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 84%|███████████████████████████████████████████████████████████████████▏            | 304/362 [18:42<03:24,  3.53
s/it]

Feature = College_CareerPrep ParentInvolvement
Prob neg = 0.2     Prob pos = 0.8
count neg = 2     count pos = 8     sum = 10
unique feature = ESL Gym_Fitness

 84%|███████████████████████████████████████████████████████████████████▍            | 305/362 [18:45<03:23,  3.56
s/it]

Feature = ESL Gym_Fitness
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 85%|███████████████████████████████████████████████████████████████████▌            | 306/362 [18:49<03:19,  3.56
s/it]

Feature = Economics SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 85%|███████████████████████████████████████████████████████████████████▊            | 307/362 [18:52<03:13,  3.51
s/it]

Feature = CommunityService Health_LifeScience
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 85%|████████████████████████████████████████████████████████████████████            | 308/362 [18:56<03:10,  3.53
s/it]

Feature = AppliedSciences ForeignLanguages
Prob neg = 0.6666666666666666     Prob pos = 0.3333333333333333
count neg = 2     count pos = 1     sum = 3

 85%|████████████████████████████████████████████████████████████████████▎           | 309/362 [18:59<03:07,  3.54
s/it]

Feature = CommunityService Economics
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = EnvironmentalScience ForeignLanguages

 86%|████████████████████████████████████████████████████████████████████▌           | 310/362 [19:03<03:03,  3.53
s/it]

Feature = EnvironmentalScience ForeignLanguages
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
```

```
 86%|████████████████████████████████████████████████████████████████████▌           | 311/362 [19:06<03:00,  3.53
s/it]

Feature = ForeignLanguages Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = Extracurricular Health_LifeScience

 86%|████████████████████████████████████████████████████████████████████▊           | 312/362 [19:10<02:55,  3.51
s/it]

Feature = Extracurricular Health_LifeScience
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 86%|█████████████████████████████████████████████████████████████████████▏          | 313/362 [19:13<02:52,  3.52
s/it]

Feature = CharacterEducation Gym_Fitness
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = ESL Extracurricular

 87%|█████████████████████████████████████████████████████████████████████▍          | 314/362 [19:17<02:49,  3.53
s/it]

Feature = ESL Extracurricular
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 87%|█████████████████████████████████████████████████████████████████████▌          | 315/362 [19:20<02:45,  3.53
s/it]

Feature = Mathematics NutritionEducation
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = Health_LifeScience ParentInvolvement

 87%|█████████████████████████████████████████████████████████████████████▊          | 316/362 [19:24<02:42,  3.53
s/it]

Feature = Health_LifeScience ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 88%|██████████████████████████████████████████████████████████████████████          | 317/362 [19:28<02:38,  3.53
s/it]

Feature = CharacterEducation Health_LifeScience
Prob neg = 0.5     Prob pos = 0.5
count neg = 2     count pos = 2     sum = 4

 88%|██████████████████████████████████████████████████████████████████████▎         | 318/362 [19:31<02:38,  3.60
s/it]

Feature = PerformingArts TeamSports
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4
unique feature = FinancialLiteracy Health_Wellness

 88%|██████████████████████████████████████████████████████████████████████▍         | 319/362 [19:35<02:42,  3.77
s/it]

Feature = FinancialLiteracy Health_Wellness
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Other TeamSports

 88%|██████████████████████████████████████████████████████████████████████▋         | 320/362 [19:39<02:35,  3.71
s/it]

Feature = Other TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 89%|██████████████████████████████████████████████████████████████████████▉         | 321/362 [19:43<02:29,  3.65
s/it]

Feature = Health_Wellness ParentInvolvement
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 89%|███████████████████████████████████████████████████████████████████████▏        | 322/362 [19:46<02:25,  3.63
s/it]

Feature = Health_LifeScience Other
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 89%|███████████████████████████████████████████████████████████████████████▍        | 323/362 [19:50<02:20,  3.61
s/it]
```

```
Feature = CharacterEducation EnvironmentalScience
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 90%|███████████████████████████████████████████████████████████████████████         | 324/362 [19:53<02:15,  3.56
s/it]

Feature = ESL ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3
unique feature = CommunityService FinancialLiteracy

 90%|███████████████████████████████████████████████████████████████████████▊        | 325/362 [19:57<02:11,  3.54
s/it]

Feature = CommunityService FinancialLiteracy
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Extracurricular ForeignLanguages

 90%|████████████████████████████████████████████████████████████████████████        | 326/362 [20:00<02:07,  3.54
s/it]

Feature = Extracurricular ForeignLanguages
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Literacy Warmth Care_Hunger

 90%|████████████████████████████████████████████████████████████████████████▎       | 327/362 [20:04<02:03,  3.53
s/it]

Feature = Literacy Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 91%|████████████████████████████████████████████████████████████████████████▍       | 328/362 [20:07<01:59,  3.51
s/it]

Feature = FinancialLiteracy Literacy
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 5     sum = 5
unique feature = CommunityService ESL

 91%|████████████████████████████████████████████████████████████████████████▋       | 329/362 [20:11<01:55,  3.50
s/it]

Feature = CommunityService ESL
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 91%|████████████████████████████████████████████████████████████████████████▉       | 330/362 [20:14<01:50,  3.47
s/it]

Feature = College_CareerPrep History_Geography
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3
unique feature = Economics Music

 91%|█████████████████████████████████████████████████████████████████████████▏      | 331/362 [20:18<01:49,  3.53
s/it]

Feature = Economics Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 92%|█████████████████████████████████████████████████████████████████████████▎      | 332/362 [20:21<01:45,  3.51
s/it]

Feature = EnvironmentalScience PerformingArts
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2
unique feature = Other PerformingArts

 92%|█████████████████████████████████████████████████████████████████████████       | 333/362 [20:25<01:42,  3.52
s/it]

Feature = Other PerformingArts
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1
unique feature = FinancialLiteracy History_Geography

 92%|█████████████████████████████████████████████████████████████████████████▊      | 334/362 [20:28<01:38,  3.53
s/it]

Feature = FinancialLiteracy History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 93%|██████████████████████████████████████████████████████████████████████████      | 335/362 [20:32<01:34,  3.49
s/it]
```

```
Feature = EarlyDevelopment TeamSports
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 93%|██████████████████████████████████████████████████████████████████████████▍      | 336/362 [20:35<01:30,  3.46
s/it]

Feature = PerformingArts SpecialNeeds
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 93%|██████████████████████████████████████████████████████████████████████████▌      | 337/362 [20:40<01:39,  3.97
s/it]

Feature = Civics_Government VisualArts
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4
unique feature = CharacterEducation Economics

 93%|██████████████████████████████████████████████████████████████████████████▊      | 338/362 [20:45<01:43,  4.31
s/it]

Feature = CharacterEducation Economics
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 94%|██████████████████████████████████████████████████████████████████████████▉      | 339/362 [20:49<01:33,  4.08
s/it]

Feature = Economics VisualArts
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2
unique feature = CharacterEducation FinancialLiteracy

 94%|███████████████████████████████████████████████████████████████████████████▏     | 340/362 [20:52<01:25,  3.90
s/it]

Feature = CharacterEducation FinancialLiteracy
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = College_CareerPrep TeamSports

 94%|███████████████████████████████████████████████████████████████████████████▍     | 341/362 [20:56<01:19,  3.78
s/it]

Feature = College_CareerPrep TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Gym_Fitness Health_LifeScience

 94%|███████████████████████████████████████████████████████████████████████████     | 342/362 [20:59<01:13,  3.69
s/it]

Feature = Gym_Fitness Health_LifeScience
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 95%|███████████████████████████████████████████████████████████████████████████▉     | 343/362 [21:03<01:08,  3.62
s/it]

Feature = College_CareerPrep CommunityService
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = Economics NutritionEducation

 95%|████████████████████████████████████████████████████████████████████████████    | 344/362 [21:06<01:04,  3.60
s/it]

Feature = Economics NutritionEducation
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = CommunityService EarlyDevelopment

 95%|████████████████████████████████████████████████████████████████████████████▏    | 345/362 [21:10<01:00,  3.59
s/it]

Feature = CommunityService EarlyDevelopment
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = ParentInvolvement Warmth Care_Hunger

 96%|████████████████████████████████████████████████████████████████████████████▍    | 346/362 [21:13<00:56,  3.56
s/it]

Feature = ParentInvolvement Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 96%|████████████████████████████████████████████████████████████████████████████▊    | 347/362 [21:17<00:52,  3.52
s/it]
```

```
Feature = CharacterEducation Warmth Care_Hunger
Prob neg = 1.0     Prob pos = 0.0
count neg = 2     count pos = 0     sum = 2
unique feature = AppliedSciences NutritionEducation

 96%|█████████████████████████████████████████████████████████████████████████████▋   | 348/362 [21:20<00:49,  3.52
s/it]

Feature = AppliedSciences NutritionEducation
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 96%|█████████████████████████████████████████████████████████████████████████████▊   | 349/362 [21:24<00:45,  3.51
s/it]

Feature = CharacterEducation Civics_Government
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = EnvironmentalScience Warmth Care_Hunger

 97%|██████████████████████████████████████████████████████████████████████████████   | 350/362 [21:27<00:42,  3.52
s/it]

Feature = EnvironmentalScience Warmth Care_Hunger
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1
unique feature = CommunityService Gym_Fitness

 97%|██████████████████████████████████████████████████████████████████████████████  | 351/362 [21:31<00:39,  3.59
s/it]

Feature = CommunityService Gym_Fitness
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = CommunityService Other

 97%|██████████████████████████████████████████████████████████████████████████████▏  | 352/362 [21:35<00:37,  3.77
s/it]

Feature = CommunityService Other
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = ForeignLanguages PerformingArts

 98%|██████████████████████████████████████████████████████████████████████████████▍ | 353/362 [21:39<00:33,  3.69
s/it]

Feature = ForeignLanguages PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = NutritionEducation VisualArts

 98%|██████████████████████████████████████████████████████████████████████████████▌  | 354/362 [21:42<00:29,  3.67
s/it]

Feature = NutritionEducation VisualArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 98%|██████████████████████████████████████████████████████████████████████████████▊  | 355/362 [21:46<00:25,  3.67
s/it]

Feature = CharacterEducation History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 98%|███████████████████████████████████████████████████████████████████████████████  | 356/362 [21:50<00:21,  3.62
s/it]

unique feature = CommunityService History_Geography
Feature = CommunityService History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 99%|███████████████████████████████████████████████████████████████████████████████▏ | 357/362 [21:53<00:17,  3.59
s/it]

unique feature = ParentInvolvement TeamSports
Feature = ParentInvolvement TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 99%|███████████████████████████████████████████████████████████████████████████████▍ | 358/362 [21:57<00:14,  3.55
s/it]

unique feature = Mathematics Warmth Care_Hunger
Feature = Mathematics Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 99%|███████████████████████████████████████████████████████████████████████████████▋ | 359/362 [22:00<00:10,  3.54
s/it]
```

```
unique feature = ESL Music
Feature = ESL Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 99%|████████████████████████████████████████████████████████████████████████████████ | 360/362 [22:04<00:07,  3.51
s/it]

unique feature = CommunityService PerformingArts
Feature = CommunityService PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

100%|████████████████████████████████████████████████████████████████████████████████▉| 361/362 [22:07<00:03,  3.52
s/it]

unique feature = ESL FinancialLiteracy
Feature = ESL FinancialLiteracy
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

100%|█████████████████████████████████████████████████████████████████████████████████| 362/362 [22:11<00:00,  3.68
s/it]

Feature = EnvironmentalScience Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
```

In [71]:
```
response_clean_subcategories.count()
```

Out[71]:
```
x    30150
y    30150
dtype: int64
```

2.2.2.4 School State

In [72]:
```
X_train.school_state.value_counts()
```

Out[72]:
```
CA    4244
TX    2058
NY    2040
FL    1684
NC    1430
IL    1190
SC    1115
GA    1099
MI     901
PA     853
IN     704
OH     689
MO     684
LA     679
MA     643
WA     643
OK     641
NJ     616
AZ     598
VA     549
WI     497
AL     481
CT     480
UT     478
TN     470
MD     390
NV     387
KY     365
MS     355
OR     349
MN     340
CO     306
AR     257
IA     192
KS     181
ID     177
ME     154
NM     148
DC     147
HI     146
WV     122
DE     102
NE      95
AK      90
SD      83
RI      79
NH      72
MT      62
ND      33
WY      29
VT      23
Name: school_state, dtype: int64
```

In [73]:
```
response_school_state = response(X_train['school_state'],y_train)
```

```
  2%|█▌                                                                              | 1/51 [00:04<03:23,  4.07
s/it]

Feature = MO
Prob neg = 0.12719298245614036     Prob pos = 0.8728070175438597
count neg = 87     count pos = 597     sum = 684

  4%|███▏                                                                            | 2/51 [00:08<03:31,  4.33
s/it]

Feature = NY
Prob neg = 0.15     Prob pos = 0.85
count neg = 306     count pos = 1734     sum = 2040

  6%|████▋                                                                           | 3/51 [00:15<03:57,  4.95
s/it]

Feature = CA
Prob neg = 0.14255419415645618     Prob pos = 0.8574458058435438
count neg = 605     count pos = 3639     sum = 4244

  8%|██████▎                                                                        | 4/51 [00:19<03:39,  4.67
s/it]

Feature = IN
Prob neg = 0.171875     Prob pos = 0.828125
count neg = 121     count pos = 583     sum = 704

 10%|███████▊                                                                        | 5/51 [00:24<03:39,  4.78
s/it]

Feature = TX
Prob neg = 0.19387755102040816     Prob pos = 0.8061224489795918
count neg = 399     count pos = 1659     sum = 2058

 12%|█████████▍                                                                      | 6/51 [00:28<03:21,  4.48
s/it]

Feature = AR
Prob neg = 0.16731517509727625     Prob pos = 0.8326848249027238
count neg = 43     count pos = 214     sum = 257

 14%|██████████▉                                                                     | 7/51 [00:32<03:17,  4.49
s/it]

Feature = NC
Prob neg = 0.13356643356643358     Prob pos = 0.8664335664335664
count neg = 191     count pos = 1239     sum = 1430

 16%|████████████▌                                                                  | 8/51 [00:36<03:05,  4.32
s/it]

Feature = CT
Prob neg = 0.11666666666666667     Prob pos = 0.8833333333333333
count neg = 56     count pos = 424     sum = 480

 18%|██████████████▏                                                                 | 9/51 [00:40<02:51,  4.09
s/it]

Feature = MT
Prob neg = 0.24193548387096775     Prob pos = 0.7580645161290323
count neg = 15     count pos = 47     sum = 62

 20%|███████████████▋                                                               | 10/51 [00:43<02:43,  3.99
s/it]

Feature = MS
Prob neg = 0.19718309859154928     Prob pos = 0.8028169014084507
count neg = 70     count pos = 285     sum = 355

 22%|█████████████████▎                                                              | 11/51 [00:47<02:37,  3.94
s/it]

Feature = NV
Prob neg = 0.12919896640826872     Prob pos = 0.8708010335917312
count neg = 50     count pos = 337     sum = 387

 24%|██████████████████▊                                                             | 12/51 [00:52<02:37,  4.03
s/it]

Feature = MI
Prob neg = 0.16870144284128746     Prob pos = 0.8312985571587126
count neg = 152     count pos = 749     sum = 901

 25%|████████████████████▍                                                           | 13/51 [00:56<02:33,  4.04
s/it]

Feature = MA
Prob neg = 0.14307931570762053     Prob pos = 0.8569206842923794
count neg = 92     count pos = 551     sum = 643
```

```
 27%|██████████████████████▌                                                            | 14/51 [00:59<02:27,  3.99
s/it]

Feature = TN
Prob neg = 0.14893617021276595     Prob pos = 0.851063829787234
count neg = 70     count pos = 400     sum = 470

 29%|████████████████████████                                                           | 15/51 [01:03<02:19,  3.86
s/it]

Feature = KS
Prob neg = 0.11602209944751381     Prob pos = 0.8839779005524862
count neg = 21     count pos = 160     sum = 181

 31%|█████████████████████████▋                                                          | 16/51 [01:08<02:21,  4.05
s/it]

Feature = IL
Prob neg = 0.15126050420168066     Prob pos = 0.8487394957983193
count neg = 180     count pos = 1010     sum = 1190

 33%|███████████████████████████▎                                                        | 17/51 [01:12<02:24,  4.24
s/it]

Feature = FL
Prob neg = 0.18171021377672208     Prob pos = 0.8182897862232779
count neg = 306     count pos = 1378     sum = 1684

 35%|████████████████████████████▉                                                       | 18/51 [01:17<02:21,  4.28
s/it]

Feature = GA
Prob neg = 0.1492265696087352     Prob pos = 0.8507734303912647
count neg = 164     count pos = 935     sum = 1099

 37%|██████████████████████████████▌                                                    | 19/51 [01:21<02:18,  4.31
s/it]

Feature = UT
Prob neg = 0.1694560669456067     Prob pos = 0.8305439330543933
count neg = 81     count pos = 397     sum = 478

 39%|████████████████████████████████▏                                                   | 20/51 [01:25<02:09,  4.19
s/it]

Feature = IA
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 32     count pos = 160     sum = 192

 41%|█████████████████████████████████▊                                                  | 21/51 [01:29<02:05,  4.17
s/it]

Feature = OH
Prob neg = 0.11175616835994194     Prob pos = 0.888243831640058
count neg = 77     count pos = 612     sum = 689

 43%|███████████████████████████████████▍                                                | 22/51 [01:33<01:56,  4.01
s/it]

Feature = DC
Prob neg = 0.19727891156462585     Prob pos = 0.8027210884353742
count neg = 29     count pos = 118     sum = 147

 45%|████████████████████████████████████▉                                               | 23/51 [01:37<01:52,  4.02
s/it]

Feature = AZ
Prob neg = 0.1588628762541806     Prob pos = 0.8411371237458194
count neg = 95     count pos = 503     sum = 598

 47%|██████████████████████████████████████▌                                            | 24/51 [01:40<01:46,  3.93
s/it]

Feature = KY
Prob neg = 0.12876712328767123     Prob pos = 0.8712328767123287
count neg = 47     count pos = 318     sum = 365

 49%|████████████████████████████████████████▏                                           | 25/51 [01:44<01:42,  3.96
s/it]

Feature = LA
Prob neg = 0.17673048600883653     Prob pos = 0.8232695139911634
count neg = 120     count pos = 559     sum = 679

 51%|█████████████████████████████████████████▊                                          | 26/51 [01:48<01:38,  3.96
s/it]

Feature = VA
Prob neg = 0.14754098360655737     Prob pos = 0.8524590163934426
count neg = 81     count pos = 468     sum = 549
```

```
 53%|███████████████████████████████████████████▉                                      | 27/51 [01:52<01:35,  3.98
s/it]

Feature = WA
Prob neg = 0.104199066874028     Prob pos = 0.895800933125972
count neg = 67     count pos = 576     sum = 643

 55%|█████████████████████████████████████████████                                     | 28/51 [01:56<01:31,  3.96
s/it]

Feature = AL
Prob neg = 0.14553014553014554     Prob pos = 0.8544698544698545
count neg = 70     count pos = 411     sum = 481

 57%|██████████████████████████████████████████████▋                                   | 29/51 [02:01<01:29,  4.07
s/it]

Feature = SC
Prob neg = 0.14708520179372198     Prob pos = 0.852914798206278
count neg = 164     count pos = 951     sum = 1115

 59%|████████████████████████████████████████████████▏                                 | 30/51 [02:04<01:23,  3.98
s/it]

Feature = MN
Prob neg = 0.15588235294117647     Prob pos = 0.8441176470588235
count neg = 53     count pos = 287     sum = 340

 61%|█████████████████████████████████████████████████▊                                | 31/51 [02:08<01:19,  3.98
s/it]

Feature = NJ
Prob neg = 0.19318181818181818     Prob pos = 0.8068181818181818
count neg = 119     count pos = 497     sum = 616

 63%|███████████████████████████████████████████████████▍                              | 32/51 [02:12<01:15,  4.00
s/it]

Feature = PA
Prob neg = 0.14419695193434937     Prob pos = 0.8558030480656507
count neg = 123     count pos = 730     sum = 853

 65%|█████████████████████████████████████████████████████                             | 33/51 [02:16<01:10,  3.92
s/it]

Feature = OR
Prob neg = 0.17191977077363896     Prob pos = 0.828080229226361
count neg = 60     count pos = 289     sum = 349

 67%|██████████████████████████████████████████████████████▋                           | 34/51 [02:20<01:05,  3.83
s/it]

Feature = AK
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 15     count pos = 75     sum = 90

 69%|████████████████████████████████████████████████████████▎                         | 35/51 [02:24<01:01,  3.84
s/it]

Feature = MD
Prob neg = 0.1641025641025641     Prob pos = 0.8358974358974359
count neg = 64     count pos = 326     sum = 390

 71%|█████████████████████████████████████████████████████████▉                        | 36/51 [02:27<00:57,  3.80
s/it]

Feature = CO
Prob neg = 0.16013071895424835     Prob pos = 0.8398692810457516
count neg = 49     count pos = 257     sum = 306

 73%|███████████████████████████████████████████████████████████▍                      | 37/51 [02:31<00:51,  3.71
s/it]

Feature = WY
Prob neg = 0.20689655172413793     Prob pos = 0.7931034482758621
count neg = 6     count pos = 23     sum = 29

 75%|█████████████████████████████████████████████████████████████                     | 38/51 [02:35<00:48,  3.69
s/it]

Feature = WV
Prob neg = 0.14754098360655737     Prob pos = 0.8524590163934426
count neg = 18     count pos = 104     sum = 122

 76%|██████████████████████████████████████████████████████████████▋                   | 39/51 [02:38<00:44,  3.67
s/it]

Feature = ID
Prob neg = 0.2033898305084746     Prob pos = 0.7966101694915254
count neg = 36     count pos = 141     sum = 177
```

```
 78%|████████████████████████████████████████████████████████████████▌                  | 40/51 [02:44<00:48,  4.40
s/it]

Feature = NH
Prob neg = 0.1388888888888889     Prob pos = 0.8611111111111112
count neg = 10     count pos = 62     sum = 72

 80%|█████████████████████████████████████████████████████████████████▋                 | 41/51 [02:49<00:43,  4.38
s/it]

Feature = OK
Prob neg = 0.14664586583463338     Prob pos = 0.8533541341653667
count neg = 94     count pos = 547     sum = 641

 82%|███████████████████████████████████████████████████████████████████               | 42/51 [02:53<00:38,  4.25
s/it]

Feature = DE
Prob neg = 0.11764705882352941     Prob pos = 0.8823529411764706
count neg = 12     count pos = 90     sum = 102

 84%|█████████████████████████████████████████████████████████████████████▏             | 43/51 [02:57<00:34,  4.30
s/it]

Feature = ME
Prob neg = 0.16233766233766234     Prob pos = 0.8376623376623377
count neg = 25     count pos = 129     sum = 154

 86%|██████████████████████████████████████████████████████████████████████▊            | 44/51 [03:01<00:30,  4.31
s/it]

Feature = SD
Prob neg = 0.13253012048192772     Prob pos = 0.8674698795180723
count neg = 11     count pos = 72     sum = 83

 88%|████████████████████████████████████████████████████████████████████████▎          | 45/51 [03:05<00:25,  4.21
s/it]

Feature = NM
Prob neg = 0.12162162162162163     Prob pos = 0.8783783783783784
count neg = 18     count pos = 130     sum = 148

 90%|█████████████████████████████████████████████████████████████████████████▉         | 46/51 [03:09<00:20,  4.17
s/it]

Feature = WI
Prob neg = 0.16096579476861167     Prob pos = 0.8390342052313883
count neg = 80     count pos = 417     sum = 497

 92%|███████████████████████████████████████████████████████████████████████████▌      | 47/51 [03:15<00:18,  4.56
s/it]

Feature = HI
Prob neg = 0.1506849315068493     Prob pos = 0.8493150684931506
count neg = 22     count pos = 124     sum = 146

 94%|█████████████████████████████████████████████████████████████████████████████▏     | 48/51 [03:19<00:13,  4.35
s/it]

Feature = NE
Prob neg = 0.16842105263157894     Prob pos = 0.8315789473684211
count neg = 16     count pos = 79     sum = 95

 96%|██████████████████████████████████████████████████████████████████████████████▊    | 49/51 [03:23<00:08,  4.37
s/it]

Feature = ND
Prob neg = 0.09090909090909091     Prob pos = 0.9090909090909091
count neg = 3     count pos = 30     sum = 33

 98%|████████████████████████████████████████████████████████████████████████████████▍  | 50/51 [03:27<00:04,  4.13
s/it]

Feature = VT
Prob neg = 0.21739130434782608     Prob pos = 0.782608695652174
count neg = 5     count pos = 18     sum = 23

100%|██████████████████████████████████████████████████████████████████████████████████| 51/51 [03:30<00:00,  4.13
s/it]

Feature = RI
Prob neg = 0.17721518987341772     Prob pos = 0.8227848101265823
count neg = 14     count pos = 65     sum = 79
```

In [74]:
```
response_school_state.count()
```

Out[74]:
```
x    30150
y    30150
dtype: int64
```

2.2.2.5 Project Grade category

```
In [75]: X_train.clean_project_grade_category.value_counts()
```

```
Out[75]: PreK-2    12293
         3-5       10205
         6-8        4609
         9-12       3043
         Name: clean_project_grade_category, dtype: int64
```

```
In [76]: response_clean_project_grade_category = response(X_train['clean_project_grade_category'],y_train)
```

```
 25%|█████████████████████                                                               | 1/4 [00:15<00:45, 15.05
s/it]

Feature = 3-5
Prob neg = 0.1464968152866242     Prob pos = 0.8535031847133758
count neg = 1495     count pos = 8710     sum = 10205

 50%|██████████████████████████████████████████                                          | 2/4 [00:27<00:28, 14.37
s/it]

Feature = PreK-2
Prob neg = 0.15569836492312697     Prob pos = 0.8443016350768731
count neg = 1914     count pos = 10379     sum = 12293

 75%|███████████████████████████████████████████████████████████████                     | 3/4 [00:37<00:13, 13.04
s/it]

Feature = 6-8
Prob neg = 0.1612063354306791     Prob pos = 0.8387936645693209
count neg = 743     count pos = 3866     sum = 4609

100%|████████████████████████████████████████████████████████████████████████████████████| 4/4 [00:44<00:00, 11.17
s/it]

Feature = 9-12
Prob neg = 0.1616825501150181     Prob pos = 0.8383174498849819
count neg = 492     count pos = 2551     sum = 3043
```

```
In [77]: response_clean_project_grade_category.count()
```

```
Out[77]: x    30150
         y    30150
         dtype: int64
```

```
In [ ]: 
```

2.2.3 Categorical data on test data

2.2.3.1 Teacher Prefix

```
In [78]: X_test['clean_teacher_prefix'].value_counts()
```

```
Out[78]: Mrs        7760
         Ms         5293
         Mr         1497
         Teacher     300
         Name: clean_teacher_prefix, dtype: int64
```

In [79]:
```
response_test_clean_teacher_prefix = response(X_test['clean_teacher_prefix'],y_test)
```

```
 25%|█████████████████████                                                               | 1/4 [00:06<00:18,  6.23
s/it]

Feature = Ms
Prob neg = 0.1581333837143397     Prob pos = 0.8418666162856603
count neg = 837     count pos = 4456     sum = 5293

 50%|██████████████████████████████████████████                                          | 2/4 [00:13<00:13,  6.61
s/it]

Feature = Mrs
Prob neg = 0.1497422680412371     Prob pos = 0.8502577319587629
count neg = 1162     count pos = 6598     sum = 7760

 75%|███████████████████████████████████████████████████████████████                     | 3/4 [00:16<00:05,  5.47
s/it]

Feature = Mr
Prob neg = 0.15297261189044756     Prob pos = 0.8470273881095525
count neg = 229     count pos = 1268     sum = 1497

100%|████████████████████████████████████████████████████████████████████████████████████| 4/4 [00:18<00:00,  4.65
s/it]

Feature = Teacher
Prob neg = 0.2     Prob pos = 0.8
count neg = 60     count pos = 240     sum = 300
```

In [80]:
```
X_test['clean_teacher_prefix'][600:620]
```

Out[80]:
```
10814     Ms
34836    Mrs
23369    Mrs
19844    Mrs
30087    Mrs
24425    Mrs
34891    Mrs
11869     Ms
8159     Mrs
3422     Mrs
6422     Mrs
9980      Ms
6503     Mrs
17339    Mrs
31160    Mrs
8132      Ms
31308     Ms
40351    Mrs
20484    Mrs
33180    Mrs
Name: clean_teacher_prefix, dtype: object
```

```
In [81]: response_test_clean_teacher_prefix[600:620]
```

Out[81]:

	x	y
600	0.158133	0.841867
601	0.149742	0.850258
602	0.149742	0.850258
603	0.149742	0.850258
604	0.149742	0.850258
605	0.149742	0.850258
606	0.149742	0.850258
607	0.158133	0.841867
608	0.149742	0.850258
609	0.149742	0.850258
610	0.149742	0.850258
611	0.158133	0.841867
612	0.149742	0.850258
613	0.149742	0.850258
614	0.149742	0.850258
615	0.158133	0.841867
616	0.158133	0.841867
617	0.149742	0.850258
618	0.149742	0.850258
619	0.149742	0.850258

```
In [ ]:
```

2.2.3.2 Clean categories

In [82]:
```
X_test.clean_categories.value_counts()
```

Out[82]:
```
Literacy_Language                      3232
Math_Science                           2305
Literacy_Language Math_Science         2025
Health_Sports                          1437
Music_Arts                              715
AppliedLearning                         540
SpecialNeeds                            538
Literacy_Language SpecialNeeds          517
Math_Science Literacy_Language          296
AppliedLearning Literacy_Language       291
Literacy_Language Music_Arts            247
Math_Science SpecialNeeds               240
History_Civics                          234
Math_Science Music_Arts                 217
AppliedLearning SpecialNeeds            208
Warmth Care_Hunger                      189
History_Civics Literacy_Language        185
Health_Sports SpecialNeeds              179
Math_Science AppliedLearning            159
AppliedLearning Math_Science            142
AppliedLearning Music_Arts              120
Health_Sports Literacy_Language         117
Literacy_Language History_Civics        104
Literacy_Language AppliedLearning        82
Math_Science History_Civics              81
AppliedLearning Health_Sports            78
History_Civics Math_Science              50
Math_Science Health_Sports               47
History_Civics Music_Arts                38
Health_Sports Math_Science               36
SpecialNeeds Music_Arts                  35
History_Civics SpecialNeeds              29
Health_Sports AppliedLearning            27
Music_Arts SpecialNeeds                  22
AppliedLearning History_Civics           21
Health_Sports Music_Arts                 14
Health_Sports History_Civics             10
History_Civics AppliedLearning           10
Literacy_Language Health_Sports           9
Health_Sports Warmth Care_Hunger          4
Music_Arts History_Civics                 4
History_Civics Health_Sports              3
SpecialNeeds Warmth Care_Hunger           3
SpecialNeeds Health_Sports                3
AppliedLearning Warmth Care_Hunger        2
Math_Science Warmth Care_Hunger           2
Music_Arts Health_Sports                  2
Music_Arts AppliedLearning                1
Name: clean_categories, dtype: int64
```

In [83]:
```
response_test_clean_categories = response(X_test['clean_categories'],y_test)
```

```
  2%|█▊                                                                                     | 1/48 [00:04<03:14,  4.14
s/it]

Feature = Literacy_Language
Prob neg = 0.13675742574257427     Prob pos = 0.8632425742574258
count neg = 442     count pos = 2790     sum = 3232

  4%|███▋                                                                                   | 2/48 [00:05<02:39,  3.46
s/it]

Feature = History_Civics Literacy_Language
Prob neg = 0.07567567567567568     Prob pos = 0.9243243243243243
count neg = 14     count pos = 171     sum = 185

  6%|█████▍                                                                                 | 3/48 [00:09<02:41,  3.58
s/it]

Feature = Literacy_Language Math_Science
Prob neg = 0.13530864197530865     Prob pos = 0.8646913580246913
count neg = 274     count pos = 1751     sum = 2025

  8%|███████▎                                                                               | 4/48 [00:15<03:01,  4.11
s/it]

Feature = Health_Sports
Prob neg = 0.16353514265831592     Prob pos = 0.8364648573416841
count neg = 235     count pos = 1202     sum = 1437

 10%|████████▏                                                                              | 5/48 [00:20<03:09,  4.40
s/it]

Feature = AppliedLearning Music_Arts
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 20     count pos = 100     sum = 120

 12%|██████████▉                                                                            | 6/48 [00:25<03:19,  4.75
s/it]

Feature = Warmth Care_Hunger
Prob neg = 0.09523809523809523     Prob pos = 0.9047619047619048
count neg = 18     count pos = 171     sum = 189

 15%|████████████                                                                          | 7/48 [00:28<02:47,  4.08
s/it]

Feature = Music_Arts
Prob neg = 0.14965034965034965     Prob pos = 0.8503496503496504
count neg = 107     count pos = 608     sum = 715

 17%|██████████████▌                                                                        | 8/48 [00:30<02:21,  3.54
s/it]

Feature = Literacy_Language SpecialNeeds
Prob neg = 0.13539651837524178     Prob pos = 0.8646034816247582
count neg = 70     count pos = 447     sum = 517

 19%|████████████████                                                                      | 9/48 [00:34<02:17,  3.54
s/it]

Feature = SpecialNeeds
Prob neg = 0.19516728624535315     Prob pos = 0.8048327137546468
count neg = 105     count pos = 433     sum = 538

 21%|██████████████████                                                                   | 10/48 [00:37<02:15,  3.57
s/it]

Feature = Math_Science
Prob neg = 0.17136659436008678     Prob pos = 0.8286334056399133
count neg = 395     count pos = 1910     sum = 2305

 23%|███████████████████▉                                                                  | 11/48 [00:39<01:52,  3.04
s/it]

Feature = Health_Sports AppliedLearning
Prob neg = 0.2222222222222222     Prob pos = 0.7777777777777778
count neg = 6     count pos = 21     sum = 27

 25%|█████████████████████                                                                | 12/48 [00:41<01:37,  2.72
s/it]

Feature = Math_Science Literacy_Language
Prob neg = 0.11148648648648649     Prob pos = 0.8885135135135135
count neg = 33     count pos = 263     sum = 296

 27%|██████████████████████▌                                                               | 13/48 [00:43<01:26,  2.48
s/it]

Feature = History_Civics
Prob neg = 0.19658119658119658     Prob pos = 0.8034188034188035
count neg = 46     count pos = 188     sum = 234
```

```
 29%|███████████████████████▋                                                                 | 14/48 [00:45<01:23,  2.45
s/it]

Feature = Literacy_Language Music_Arts
Prob neg = 0.16194331983805668     Prob pos = 0.8380566801619433
count neg = 40     count pos = 207     sum = 247

 31%|█████████████████████████▌                                                               | 15/48 [00:48<01:24,  2.55
s/it]

Feature = Health_Sports Math_Science
Prob neg = 0.3055555555555556     Prob pos = 0.6944444444444444
count neg = 11     count pos = 25     sum = 36

 33%|███████████████████████████▎                                                             | 16/48 [00:51<01:21,  2.55
s/it]

Feature = History_Civics Math_Science
Prob neg = 0.16     Prob pos = 0.84
count neg = 8     count pos = 42     sum = 50

 35%|█████████████████████████████                                                           | 17/48 [00:53<01:19,  2.57
s/it]

Feature = AppliedLearning Literacy_Language
Prob neg = 0.16151202749140894     Prob pos = 0.8384879725085911
count neg = 47     count pos = 244     sum = 291

 38%|██████████████████████████████▊                                                          | 18/48 [00:56<01:14,  2.48
s/it]

Feature = AppliedLearning SpecialNeeds
Prob neg = 0.21634615384615385     Prob pos = 0.7836538461538461
count neg = 45     count pos = 163     sum = 208

 40%|████████████████████████████████▍                                                        | 19/48 [00:59<01:17,  2.69
s/it]

Feature = AppliedLearning
Prob neg = 0.1925925925925926     Prob pos = 0.8074074074074075
count neg = 104     count pos = 436     sum = 540

 42%|██████████████████████████████████▏                                                      | 20/48 [01:02<01:16,  2.73
s/it]

Feature = Math_Science AppliedLearning
Prob neg = 0.1509433962264151     Prob pos = 0.8490566037735849
count neg = 24     count pos = 135     sum = 159

 44%|███████████████████████████████████▋                                                     | 21/48 [01:03<01:06,  2.47
s/it]

Feature = Literacy_Language Health_Sports
Prob neg = 0.1111111111111111     Prob pos = 0.8888888888888888
count neg = 1     count pos = 8     sum = 9

 46%|█████████████████████████████████████                                                   | 22/48 [01:05<00:58,  2.27
s/it]

Feature = Math_Science History_Civics
Prob neg = 0.16049382716049382     Prob pos = 0.8395061728395061
count neg = 13     count pos = 68     sum = 81

 48%|███████████████████████████████████████▍                                                 | 23/48 [01:07<00:53,  2.14
s/it]

Feature = Literacy_Language History_Civics
Prob neg = 0.1346153846153846     Prob pos = 0.8653846153846154
count neg = 14     count pos = 90     sum = 104

 50%|█████████████████████████████████████████                                               | 24/48 [01:09<00:49,  2.07
s/it]

Feature = Math_Science SpecialNeeds
Prob neg = 0.1875     Prob pos = 0.8125
count neg = 45     count pos = 195     sum = 240

 52%|██████████████████████████████████████████▋                                              | 25/48 [01:11<00:47,  2.05
s/it]

Feature = Health_Sports SpecialNeeds
Prob neg = 0.13966480446927373     Prob pos = 0.8603351955307262
count neg = 25     count pos = 154     sum = 179

 54%|████████████████████████████████████████████▌                                            | 26/48 [01:13<00:44,  2.01
s/it]

Feature = AppliedLearning Math_Science
Prob neg = 0.19718309859154928     Prob pos = 0.8028169014084507
count neg = 28     count pos = 114     sum = 142
```

```
 56%|██████████████████████████████████████████████▍                                    | 27/48 [01:15<00:41,  1.99
s/it]

Feature = Health_Sports Literacy_Language
Prob neg = 0.13675213675213677     Prob pos = 0.8632478632478633
count neg = 16     count pos = 101     sum = 117

 58%|███████████████████████████████████████████████▊                                   | 28/48 [01:17<00:40,  2.00
s/it]

Feature = Math_Science Music_Arts
Prob neg = 0.15207373271889402     Prob pos = 0.847926267281106
count neg = 33     count pos = 184     sum = 217

 60%|█████████████████████████████████████████████████                                 | 29/48 [01:19<00:36,  1.95
s/it]

Feature = History_Civics Music_Arts
Prob neg = 0.13157894736842105     Prob pos = 0.868421052631579
count neg = 5     count pos = 33     sum = 38

 62%|███████████████████████████████████████████████████▏                               | 30/48 [01:21<00:34,  1.91
s/it]

Feature = History_Civics Health_Sports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 65%|████████████████████████████████████████████████████▉                              | 31/48 [01:22<00:32,  1.89
s/it]

Feature = AppliedLearning Health_Sports
Prob neg = 0.19230769230769232     Prob pos = 0.8076923076923077
count neg = 15     count pos = 63     sum = 78

 67%|██████████████████████████████████████████████████████▌                            | 32/48 [01:24<00:30,  1.91
s/it]

Feature = SpecialNeeds Music_Arts
Prob neg = 0.14285714285714285     Prob pos = 0.8571428571428571
count neg = 5     count pos = 30     sum = 35

 69%|████████████████████████████████████████████████████████▍                          | 33/48 [01:26<00:28,  1.90
s/it]

Feature = Literacy_Language AppliedLearning
Prob neg = 0.12195121951219512     Prob pos = 0.8780487804878049
count neg = 10     count pos = 72     sum = 82

 71%|██████████████████████████████████████████████████████████                        | 34/48 [01:28<00:26,  1.87
s/it]

Feature = History_Civics AppliedLearning
Prob neg = 0.1     Prob pos = 0.9
count neg = 1     count pos = 9     sum = 10

 73%|███████████████████████████████████████████████████████████▉                       | 35/48 [01:30<00:23,  1.84
s/it]

Feature = AppliedLearning History_Civics
Prob neg = 0.23809523809523808     Prob pos = 0.7619047619047619
count neg = 5     count pos = 16     sum = 21

 75%|█████████████████████████████████████████████████████████████                     | 36/48 [01:32<00:21,  1.81
s/it]

Feature = Health_Sports History_Civics
Prob neg = 0.2     Prob pos = 0.8
count neg = 2     count pos = 8     sum = 10

 77%|███████████████████████████████████████████████████████████████▏                   | 37/48 [01:33<00:20,  1.82
s/it]

Feature = Math_Science Health_Sports
Prob neg = 0.23404255319148937     Prob pos = 0.7659574468085106
count neg = 11     count pos = 36     sum = 47

 79%|████████████████████████████████████████████████████████████████▊                  | 38/48 [01:35<00:18,  1.82
s/it]

Feature = Music_Arts SpecialNeeds
Prob neg = 0.13636363636363635     Prob pos = 0.8636363636363636
count neg = 3     count pos = 19     sum = 22

 81%|██████████████████████████████████████████████████████████████████▊                | 39/48 [01:37<00:16,  1.83
s/it]

Feature = History_Civics SpecialNeeds
Prob neg = 0.20689655172413793     Prob pos = 0.7931034482758621
count neg = 6     count pos = 23     sum = 29
```

```
 83%|████████████████████████████████████████████████████████████████████▌             | 40/48 [01:39<00:14,  1.81
s/it]

Feature = Health_Sports Warmth Care_Hunger
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4

 85%|██████████████████████████████████████████████████████████████████████            | 41/48 [01:41<00:12,  1.79
s/it]

Feature = Music_Arts History_Civics
Prob neg = 0.5     Prob pos = 0.5
count neg = 2     count pos = 2     sum = 4

 88%|███████████████████████████████████████████████████████████████████████▊          | 42/48 [01:42<00:10,  1.79
s/it]

Feature = SpecialNeeds Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 90%|█████████████████████████████████████████████████████████████████████████▍        | 43/48 [01:44<00:08,  1.78
s/it]

Feature = Health_Sports Music_Arts
Prob neg = 0.14285714285714285     Prob pos = 0.8571428571428571
count neg = 2     count pos = 12     sum = 14

 92%|███████████████████████████████████████████████████████████████████████████▏      | 44/48 [01:46<00:07,  1.78
s/it]

Feature = AppliedLearning Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 94%|████████████████████████████████████████████████████████████████████████████▋     | 45/48 [01:48<00:05,  1.78
s/it]

Feature = SpecialNeeds Health_Sports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 96%|██████████████████████████████████████████████████████████████████████████████    | 46/48 [01:49<00:03,  1.78
s/it]

Feature = Math_Science Warmth Care_Hunger
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 98%|████████████████████████████████████████████████████████████████████████████████▍ | 47/48 [01:55<00:02,  2.97
s/it]

Feature = Music_Arts Health_Sports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

100%|██████████████████████████████████████████████████████████████████████████████████| 48/48 [01:58<00:00,  2.99
s/it]

unique feature = Music_Arts AppliedLearning
Feature = Music_Arts AppliedLearning
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

100%|██████████████████████████████████████████████████████████████████████████████████| 48/48 [01:58<00:00,  2.47
s/it]
```

In [84]:
```
response_test_clean_categories.count()
```

Out[84]:
```
x    14850
y    14850
dtype: int64
```

2.2.3.3 Clean Sub categories

```
In [85]: X_test.clean_subcategories.value_counts()
```

```
Out[85]: Literacy                                  1296
         Literacy Mathematics                      1131
         Literature_Writing Mathematics             843
         Literacy Literature_Writing                771
         Mathematics                                749
         Literature_Writing                         625
         SpecialNeeds                               538
         Health_Wellness                            538
         AppliedSciences Mathematics                453
         AppliedSciences                            335
         Literacy SpecialNeeds                      325
         VisualArts                                 307
         ESL Literacy                               302
         Gym_Fitness Health_Wellness                291
         Music                                      213
         Warmth Care_Hunger                         189
         Literature_Writing SpecialNeeds            168
         Mathematics SpecialNeeds                   161
         Health_Wellness SpecialNeeds               156
         Gym_Fitness                                150
         EnvironmentalScience Health_LifeScience    144
         TeamSports                                 142
         EnvironmentalScience                       135
         AppliedSciences EnvironmentalScience       134
         EarlyDevelopment                           123
         Music PerformingArts                       120
         Other                                      118
         Health_LifeScience                         117
         EarlyDevelopment SpecialNeeds              111
         Literature_Writing VisualArts              105
                                                   ...
         College_CareerPrep Warmth Care_Hunger        1
         Civics_Government Economics                  1
         College_CareerPrep ESL                       1
         Civics_Government FinancialLiteracy          1
         EarlyDevelopment TeamSports                  1
         Extracurricular SpecialNeeds                 1
         EnvironmentalScience PerformingArts          1
         Gym_Fitness SocialSciences                   1
         FinancialLiteracy ForeignLanguages           1
         Literacy TeamSports                          1
         AppliedSciences Economics                    1
         TeamSports VisualArts                        1
         CommunityService Economics                   1
         ESL Music                                    1
         ESL Extracurricular                          1
         Civics_Government TeamSports                 1
         CommunityService ParentInvolvement           1
         ForeignLanguages Other                       1
         ESL PerformingArts                           1
         Extracurricular ParentInvolvement            1
         History_Geography ParentInvolvement          1
         CommunityService PerformingArts              1
         Health_LifeScience Music                     1
         AppliedSciences FinancialLiteracy            1
         Music Other                                  1
         CharacterEducation NutritionEducation        1
         EnvironmentalScience TeamSports              1
         CharacterEducation PerformingArts            1
         Gym_Fitness Literature_Writing               1
         FinancialLiteracy Other                      1
         Name: clean_subcategories, Length: 314, dtype: int64
```

In [86]:
```
response_test_clean_subcategories = response(X_test['clean_subcategories'],y_test)
```

```
  0%|▌                                                                               | 1/314 [00:03<17:47,  3.41
s/it]

Feature = Literacy
Prob neg = 0.11882716049382716     Prob pos = 0.8811728395061729
count neg = 154     count pos = 1142     sum = 1296

  1%|▌                                                                              | 2/314 [00:05<15:26,  2.97
s/it]

Feature = History_Geography Literature_Writing
Prob neg = 0.12790697674418605     Prob pos = 0.872093023255814
count neg = 11     count pos = 75     sum = 86

  1%|█                                                                               | 3/314 [00:09<17:41,  3.41
s/it]

Feature = Literacy Mathematics
Prob neg = 0.13704686118479223     Prob pos = 0.8629531388152077
count neg = 155     count pos = 976     sum = 1131

  1%|█                                                                              | 4/314 [00:13<18:49,  3.64
s/it]

Feature = Gym_Fitness Health_Wellness
Prob neg = 0.13745704467353953     Prob pos = 0.8625429553264605
count neg = 40     count pos = 251     sum = 291

  2%|█▌                                                                              | 5/314 [00:16<16:21,  3.18
s/it]

Feature = Extracurricular VisualArts
Prob neg = 0.2     Prob pos = 0.8
count neg = 3     count pos = 12     sum = 15

  2%|█▌                                                                             | 6/314 [00:19<15:58,  3.11
s/it]

Feature = Warmth Care_Hunger
Prob neg = 0.09523809523809523     Prob pos = 0.9047619047619048
count neg = 18     count pos = 171     sum = 189

  2%|██                                                                              | 7/314 [00:25<20:23,  3.98
s/it]

Feature = Literature_Writing
Prob neg = 0.176     Prob pos = 0.824
count neg = 110     count pos = 515     sum = 625

  3%|██                                                                             | 8/314 [00:27<18:41,  3.66
s/it]

Feature = ESL Literacy
Prob neg = 0.13245033112582782     Prob pos = 0.8675496688741722
count neg = 40     count pos = 262     sum = 302

  3%|██▌                                                                             | 9/314 [00:30<16:39,  3.28
s/it]

Feature = Health_Wellness NutritionEducation
Prob neg = 0.17475728155339806     Prob pos = 0.8252427184466019
count neg = 18     count pos = 85     sum = 103

  3%|██▌                                                                           | 10/314 [00:32<15:31,  3.06
s/it]

Feature = Gym_Fitness TeamSports
Prob neg = 0.23232323232323232     Prob pos = 0.7676767676767676
count neg = 23     count pos = 76     sum = 99

  4%|███                                                                            | 11/314 [00:36<15:33,  3.08
s/it]

Feature = Health_Wellness
Prob neg = 0.14312267657992564     Prob pos = 0.8568773234200744
count neg = 77     count pos = 461     sum = 538

  4%|███                                                                           | 12/314 [00:39<15:36,  3.10
s/it]

Feature = VisualArts
Prob neg = 0.1791530944625407     Prob pos = 0.8208469055374593
count neg = 55     count pos = 252     sum = 307

  4%|███▌                                                                           | 13/314 [00:42<16:02,  3.20
s/it]

Feature = Literacy Literature_Writing
Prob neg = 0.14007782101167315     Prob pos = 0.8599221789883269
count neg = 108     count pos = 663     sum = 771
```

```
  4%|███                                                                              | 14/314 [00:46<16:38,  3.33
s/it]

Feature = Literacy SpecialNeeds
Prob neg = 0.10461538461538461     Prob pos = 0.8953846153846153
count neg = 34     count pos = 291     sum = 325

  5%|███▌                                                                              | 15/314 [00:49<16:33,  3.32
s/it]

Feature = NutritionEducation
Prob neg = 0.16279069767441862     Prob pos = 0.8372093023255814
count neg = 7     count pos = 36     sum = 43

  5%|███▊                                                                              | 16/314 [00:52<16:27,  3.31
s/it]

Feature = SpecialNeeds
Prob neg = 0.19516728624535315     Prob pos = 0.8048327137546468
count neg = 105     count pos = 433     sum = 538

  5%|████▏                                                                             | 17/314 [00:56<16:24,  3.31
s/it]

Feature = AppliedSciences Mathematics
Prob neg = 0.1390728476821192     Prob pos = 0.8609271523178808
count neg = 63     count pos = 390     sum = 453

  6%|████▍                                                                             | 18/314 [00:59<15:50,  3.21
s/it]

Feature = Health_Wellness Other
Prob neg = 0.17391304347826086     Prob pos = 0.8260869565217391
count neg = 4     count pos = 19     sum = 23

  6%|████▋                                                                             | 19/314 [01:01<15:10,  3.09
s/it]

Feature = AppliedSciences ESL
Prob neg = 0.11764705882352941     Prob pos = 0.8823529411764706
count neg = 2     count pos = 15     sum = 17

  6%|████▉                                                                             | 20/314 [01:05<15:13,  3.11
s/it]

Feature = Literature_Writing Mathematics
Prob neg = 0.1257413997627521     Prob pos = 0.8742586002372479
count neg = 106     count pos = 737     sum = 843

  7%|█████▏                                                                            | 21/314 [01:07<14:45,  3.02
s/it]

Feature = History_Geography
Prob neg = 0.2571428571428571     Prob pos = 0.7428571428571429
count neg = 18     count pos = 52     sum = 70

  7%|█████▌                                                                            | 22/314 [01:11<14:53,  3.06
s/it]

Feature = Mathematics
Prob neg = 0.17489986648865152     Prob pos = 0.8251001335113485
count neg = 131     count pos = 618     sum = 749

  7%|█████▉                                                                            | 23/314 [01:13<13:31,  2.79
s/it]

Feature = Literacy VisualArts
Prob neg = 0.1506849315068493     Prob pos = 0.8493150684931506
count neg = 11     count pos = 62     sum = 73

  8%|██████                                                                            | 24/314 [01:15<12:27,  2.58
s/it]

Feature = Health_Wellness Mathematics
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 8     count pos = 16     sum = 24

  8%|██████▎                                                                           | 25/314 [01:17<11:33,  2.40
s/it]

Feature = Health_LifeScience Literature_Writing
Prob neg = 0.041666666666666664     Prob pos = 0.9583333333333334
count neg = 1     count pos = 23     sum = 24

  8%|██████▌                                                                           | 26/314 [01:19<11:31,  2.40
s/it]

Feature = Gym_Fitness
Prob neg = 0.21333333333333335     Prob pos = 0.7866666666666666
count neg = 32     count pos = 118     sum = 150
```

```
  9%|██████▉                                                                    | 27/314 [01:21<11:00,  2.30
s/it]

Feature = Music
Prob neg = 0.13615023474178403     Prob pos = 0.863849765258216
count neg = 29     count pos = 184     sum = 213

  9%|███████▏                                                                   | 28/314 [01:24<10:58,  2.30
s/it]

Feature = History_Geography Mathematics
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 3     count pos = 15     sum = 18

  9%|███████▍                                                                   | 29/314 [01:26<10:38,  2.24
s/it]

Feature = AppliedSciences
Prob neg = 0.1761194029850746     Prob pos = 0.8238805970149253
count neg = 59     count pos = 276     sum = 335

 10%|███████▋                                                                   | 30/314 [01:27<09:58,  2.11
s/it]

Feature = ESL Literature_Writing
Prob neg = 0.09278350515463918     Prob pos = 0.9072164948453608
count neg = 9     count pos = 88     sum = 97

 10%|███████▉                                                                   | 31/314 [01:29<09:30,  2.02
s/it]

Feature = College_CareerPrep Literacy
Prob neg = 0.1     Prob pos = 0.9
count neg = 3     count pos = 27     sum = 30

 10%|████████▏                                                                  | 32/314 [01:31<09:14,  1.97
s/it]

Feature = Literature_Writing VisualArts
Prob neg = 0.18095238095238095     Prob pos = 0.819047619047619
count neg = 19     count pos = 86     sum = 105

 11%|████████▍                                                                 | 33/314 [01:33<09:21,  2.00
s/it]

Feature = College_CareerPrep SpecialNeeds
Prob neg = 0.30434782608695654     Prob pos = 0.6956521739130435
count neg = 7     count pos = 16     sum = 23

 11%|████████▋                                                                  | 34/314 [01:35<09:10,  1.97
s/it]

Feature = EnvironmentalScience
Prob neg = 0.15555555555555556     Prob pos = 0.8444444444444444
count neg = 21     count pos = 114     sum = 135

 11%|████████▉                                                                 | 35/314 [01:37<08:50,  1.90
s/it]

Feature = College_CareerPrep Literature_Writing
Prob neg = 0.175     Prob pos = 0.825
count neg = 7     count pos = 33     sum = 40

 11%|█████████▏                                                                 | 36/314 [01:39<08:46,  1.89
s/it]

Feature = EarlyDevelopment
Prob neg = 0.16260162601626016     Prob pos = 0.8373983739837398
count neg = 20     count pos = 103     sum = 123

 12%|█████████▍                                                                | 37/314 [01:41<08:38,  1.87
s/it]

Feature = Other
Prob neg = 0.211864406779661     Prob pos = 0.788135593220339
count neg = 25     count pos = 93     sum = 118

 12%|█████████▋                                                                 | 38/314 [01:42<08:32,  1.86
s/it]

Feature = PerformingArts
Prob neg = 0.11320754716981132     Prob pos = 0.8867924528301887
count neg = 6     count pos = 47     sum = 53

 12%|█████████▉                                                                | 39/314 [01:44<08:26,  1.84
s/it]

Feature = AppliedSciences CharacterEducation
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 2     count pos = 4     sum = 6
```

```
 13%|██████████▏                                                                    | 40/314 [01:46<08:16,  1.81
s/it]

Feature = Mathematics Other
Prob neg = 0.09090909090909091     Prob pos = 0.9090909090909091
count neg = 1     count pos = 10     sum = 11

 13%|██████████▍                                                                   | 41/314 [01:48<08:13,  1.81
s/it]

Feature = EarlyDevelopment Literacy
Prob neg = 0.17307692307692307     Prob pos = 0.8269230769230769
count neg = 18     count pos = 86     sum = 104

 13%|██████████▋                                                                    | 42/314 [01:50<08:15,  1.82
s/it]

Feature = Health_LifeScience
Prob neg = 0.19658119658119658     Prob pos = 0.8034188034188035
count neg = 23     count pos = 94     sum = 117

 14%|██████████▉                                                                   | 43/314 [01:51<08:14,  1.82
s/it]

Feature = Literature_Writing SpecialNeeds
Prob neg = 0.18452380952380953     Prob pos = 0.8154761904761905
count neg = 31     count pos = 137     sum = 168

 14%|███████████▏                                                                   | 44/314 [01:53<08:08,  1.81
s/it]

Feature = History_Geography SocialSciences
Prob neg = 0.1590909090909091     Prob pos = 0.8409090909090909
count neg = 7     count pos = 37     sum = 44

 14%|███████████▍                                                                  | 45/314 [01:55<08:07,  1.81
s/it]

Feature = Music PerformingArts
Prob neg = 0.09166666666666666     Prob pos = 0.9083333333333333
count neg = 11     count pos = 109     sum = 120

 15%|███████████▋                                                                   | 46/314 [01:57<07:57,  1.78
s/it]

Feature = ForeignLanguages Health_Wellness
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 15%|███████████▉                                                                  | 47/314 [01:58<07:53,  1.77
s/it]

Feature = AppliedSciences History_Geography
Prob neg = 0.2     Prob pos = 0.8
count neg = 2     count pos = 8     sum = 10

 15%|████████████▏                                                                  | 48/314 [02:00<07:58,  1.80
s/it]

Feature = Literacy SocialSciences
Prob neg = 0.15517241379310345     Prob pos = 0.8448275862068966
count neg = 9     count pos = 49     sum = 58

 16%|████████████▍                                                                  | 49/314 [02:02<07:54,  1.79
s/it]

Feature = EnvironmentalScience SpecialNeeds
Prob neg = 0.22727272727272727     Prob pos = 0.7727272727272727
count neg = 5     count pos = 17     sum = 22

 16%|████████████▋                                                                  | 50/314 [02:04<08:03,  1.83
s/it]

Feature = Mathematics SpecialNeeds
Prob neg = 0.2111801242236025     Prob pos = 0.7888198757763976
count neg = 34     count pos = 127     sum = 161

 16%|████████████▉                                                                  | 51/314 [02:06<07:51,  1.79
s/it]

Feature = Civics_Government SocialSciences
Prob neg = 0.15384615384615385     Prob pos = 0.8461538461538461
count neg = 2     count pos = 11     sum = 13

 17%|█████████████▏                                                                 | 52/314 [02:08<07:54,  1.81
s/it]

Feature = Health_Wellness SpecialNeeds
Prob neg = 0.1346153846153846     Prob pos = 0.8653846153846154
count neg = 21     count pos = 135     sum = 156
```

```
 17%|█████████████▎                                                                 | 53/314 [02:09<07:46,  1.79
s/it]

Feature = Other SpecialNeeds
Prob neg = 0.20408163265306123     Prob pos = 0.7959183673469388
count neg = 10     count pos = 39     sum = 49

 17%|█████████████▋                                                                 | 54/314 [02:11<07:41,  1.77
s/it]

Feature = College_CareerPrep Health_LifeScience
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 18%|██████████████▏                                                                | 55/314 [02:14<08:37,  2.00
s/it]

Feature = ESL SpecialNeeds
Prob neg = 0.20833333333333334     Prob pos = 0.7916666666666666
count neg = 5     count pos = 19     sum = 24

 18%|██████████████▍                                                                | 56/314 [02:16<09:21,  2.18
s/it]

Feature = CharacterEducation College_CareerPrep
Prob neg = 0.18181818181818182     Prob pos = 0.8181818181818182
count neg = 2     count pos = 9     sum = 11

 18%|██████████████▊                                                                | 57/314 [02:19<09:41,  2.26
s/it]

Feature = College_CareerPrep Extracurricular
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 8     sum = 8

 18%|███████████████▏                                                               | 58/314 [02:21<10:04,  2.36
s/it]

Feature = Health_LifeScience Mathematics
Prob neg = 0.189873417721519     Prob pos = 0.810126582278481
count neg = 15     count pos = 64     sum = 79

 19%|███████████████▏                                                               | 59/314 [02:23<09:40,  2.28
s/it]

Feature = Health_Wellness Literature_Writing
Prob neg = 0.15789473684210525     Prob pos = 0.8421052631578947
count neg = 6     count pos = 32     sum = 38

 19%|███████████████▍                                                               | 60/314 [02:25<09:02,  2.13
s/it]

Feature = ForeignLanguages Literacy
Prob neg = 0.13333333333333333     Prob pos = 0.8666666666666667
count neg = 4     count pos = 26     sum = 30

 19%|███████████████▊                                                               | 61/314 [02:27<08:27,  2.01
s/it]

Feature = EarlyDevelopment Mathematics
Prob neg = 0.18181818181818182     Prob pos = 0.8181818181818182
count neg = 8     count pos = 36     sum = 44

 20%|████████████████▏                                                              | 62/314 [02:29<08:08,  1.94
s/it]

Feature = Extracurricular PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 20%|████████████████▍                                                              | 63/314 [02:30<07:56,  1.90
s/it]

Feature = Health_LifeScience Literacy
Prob neg = 0.10526315789473684     Prob pos = 0.8947368421052632
count neg = 4     count pos = 34     sum = 38

 20%|████████████████▌                                                            | 64/314 [02:32<07:43,  1.85
s/it]

Feature = AppliedSciences VisualArts
Prob neg = 0.1414141414141414     Prob pos = 0.8585858585858586
count neg = 14     count pos = 85     sum = 99

 21%|█████████████████▏                                                             | 65/314 [02:34<07:38,  1.84
s/it]

Feature = History_Geography Literacy
Prob neg = 0.014925373134328358     Prob pos = 0.9850746268656716
count neg = 1     count pos = 66     sum = 67
```

```
 21%|█████████████████                                                                | 66/314 [02:36<07:29,  1.81
s/it]

Feature = Mathematics Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 6     sum = 6

 21%|█████████████████▏                                                                | 67/314 [02:37<07:23,  1.80
s/it]

Feature = AppliedSciences Health_LifeScience
Prob neg = 0.13432835820895522     Prob pos = 0.8656716417910447
count neg = 9     count pos = 58     sum = 67

 22%|█████████████████▌                                                               | 68/314 [02:39<07:13,  1.76
s/it]

Feature = Economics FinancialLiteracy
Prob neg = 0.3076923076923077     Prob pos = 0.6923076923076923
count neg = 4     count pos = 9     sum = 13

 22%|█████████████████▋                                                                | 69/314 [02:41<07:17,  1.79
s/it]

Feature = EnvironmentalScience Health_LifeScience
Prob neg = 0.22916666666666666     Prob pos = 0.7708333333333334
count neg = 33     count pos = 111     sum = 144

 22%|██████████████████                                                               | 70/314 [02:43<07:11,  1.77
s/it]

Feature = EarlyDevelopment EnvironmentalScience
Prob neg = 0.4     Prob pos = 0.6
count neg = 2     count pos = 3     sum = 5

 23%|██████████████████▏                                                               | 71/314 [02:44<07:09,  1.77
s/it]

Feature = CharacterEducation
Prob neg = 0.23529411764705882     Prob pos = 0.7647058823529411
count neg = 12     count pos = 39     sum = 51

 23%|██████████████████▌                                                              | 72/314 [02:46<07:27,  1.85
s/it]

Feature = History_Geography VisualArts
Prob neg = 0.1     Prob pos = 0.9
count neg = 2     count pos = 18     sum = 20

 23%|██████████████████▋                                                               | 73/314 [02:49<08:02,  2.00
s/it]

Feature = Other VisualArts
Prob neg = 0.25     Prob pos = 0.75
count neg = 2     count pos = 6     sum = 8

 24%|███████████████████                                                              | 74/314 [02:51<07:45,  1.94
s/it]

Feature = AppliedSciences Literacy
Prob neg = 0.15789473684210525     Prob pos = 0.8421052631578947
count neg = 12     count pos = 64     sum = 76
unique feature = Civics_Government Health_Wellness

 24%|███████████████████▏                                                              | 75/314 [02:52<07:30,  1.88
s/it]

Feature = Civics_Government Health_Wellness
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 24%|███████████████████▌                                                             | 76/314 [02:54<07:28,  1.88
s/it]

Feature = Health_Wellness TeamSports
Prob neg = 0.15254237288135594     Prob pos = 0.847457627118644
count neg = 9     count pos = 50     sum = 59

 25%|███████████████████▋                                                              | 77/314 [02:56<07:20,  1.86
s/it]

Feature = EarlyDevelopment Health_Wellness
Prob neg = 0.10810810810810811     Prob pos = 0.8918918918918919
count neg = 4     count pos = 33     sum = 37

 25%|████████████████████                                                             | 78/314 [02:58<07:16,  1.85
s/it]

Feature = TeamSports
Prob neg = 0.18309859154929578     Prob pos = 0.8169014084507042
count neg = 26     count pos = 116     sum = 142
```

```
 25%|████████████████████▍                                                            | 79/314 [03:00<07:11,  1.84
s/it]

Feature = EnvironmentalScience History_Geography
Prob neg = 0.16     Prob pos = 0.84
count neg = 4     count pos = 21     sum = 25

 25%|████████████████████▋                                                            | 80/314 [03:01<07:04,  1.81
s/it]

Feature = EnvironmentalScience Literature_Writing
Prob neg = 0.11428571428571428     Prob pos = 0.8857142857142857
count neg = 4     count pos = 31     sum = 35

 26%|████████████████████▉                                                            | 81/314 [03:03<07:01,  1.81
s/it]

Feature = Health_Wellness Literacy
Prob neg = 0.12162162162162163     Prob pos = 0.8783783783783784
count neg = 9     count pos = 65     sum = 74

 26%|█████████████████████▏                                                           | 82/314 [03:05<06:52,  1.78
s/it]

Feature = EarlyDevelopment VisualArts
Prob neg = 0.2916666666666667     Prob pos = 0.7083333333333334
count neg = 7     count pos = 17     sum = 24

 26%|█████████████████████▍                                                           | 83/314 [03:07<06:49,  1.77
s/it]

Feature = FinancialLiteracy
Prob neg = 0.11764705882352941     Prob pos = 0.8823529411764706
count neg = 2     count pos = 15     sum = 17

 27%|█████████████████████▊                                                           | 84/314 [03:08<06:45,  1.76
s/it]

Feature = College_CareerPrep Mathematics
Prob neg = 0.23255813953488372     Prob pos = 0.7674418604651163
count neg = 10     count pos = 33     sum = 43

 27%|██████████████████████                                                           | 85/314 [03:10<06:42,  1.76
s/it]

Feature = SpecialNeeds VisualArts
Prob neg = 0.14285714285714285     Prob pos = 0.8571428571428571
count neg = 5     count pos = 30     sum = 35

 27%|██████████████████████▎                                                          | 86/314 [03:12<06:34,  1.73
s/it]

Feature = ESL EnvironmentalScience
Prob neg = 0.4     Prob pos = 0.6
count neg = 2     count pos = 3     sum = 5

 28%|██████████████████████▌                                                          | 87/314 [03:14<06:32,  1.73
s/it]

Feature = CharacterEducation Gym_Fitness
Prob neg = 0.6666666666666666     Prob pos = 0.3333333333333333
count neg = 2     count pos = 1     sum = 3

 28%|██████████████████████▊                                                          | 88/314 [03:15<06:29,  1.72
s/it]

Feature = CharacterEducation Literacy
Prob neg = 0.2     Prob pos = 0.8
count neg = 9     count pos = 36     sum = 45

 28%|███████████████████████                                                          | 89/314 [03:17<06:28,  1.72
s/it]

Feature = FinancialLiteracy Mathematics
Prob neg = 0.15789473684210525     Prob pos = 0.8421052631578947
count neg = 3     count pos = 16     sum = 19

 29%|███████████████████████▏                                                         | 90/314 [03:19<06:35,  1.76
s/it]

Feature = Gym_Fitness SpecialNeeds
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 3     count pos = 15     sum = 18

 29%|███████████████████████▌                                                         | 91/314 [03:21<06:28,  1.74
s/it]

Feature = CharacterEducation Literature_Writing
Prob neg = 0.08695652173913043     Prob pos = 0.9130434782608695
count neg = 2     count pos = 21     sum = 23
```

```
 29%|███████████████████████▍                                                        | 92/314 [03:22<06:28,  1.75
s/it]

Feature = Literature_Writing SocialSciences
Prob neg = 0.10526315789473684     Prob pos = 0.8947368421052632
count neg = 4     count pos = 34     sum = 38

 30%|███████████████████████▋                                                        | 93/314 [03:24<06:28,  1.76
s/it]

Feature = Extracurricular Health_LifeScience
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 30%|████████████████████████▏                                                       | 94/314 [03:26<06:29,  1.77
s/it]

Feature = CharacterEducation EarlyDevelopment
Prob neg = 0.23529411764705882     Prob pos = 0.7647058823529411
count neg = 8     count pos = 26     sum = 34

 30%|████████████████████████                                                       | 95/314 [03:28<06:26,  1.77
s/it]

Feature = Literacy ParentInvolvement
Prob neg = 0.125     Prob pos = 0.875
count neg = 3     count pos = 21     sum = 24

 31%|████████████████████████▋                                                       | 96/314 [03:30<06:30,  1.79
s/it]

Feature = EnvironmentalScience Mathematics
Prob neg = 0.14130434782608695     Prob pos = 0.8586956521739131
count neg = 13     count pos = 79     sum = 92

 31%|█████████████████████████                                                      | 97/314 [03:31<06:22,  1.76
s/it]

Feature = AppliedSciences SocialSciences
Prob neg = 0.1111111111111111     Prob pos = 0.8888888888888888
count neg = 1     count pos = 8     sum = 9

 31%|█████████████████████████▎                                                      | 98/314 [03:33<06:23,  1.78
s/it]

Feature = ForeignLanguages
Prob neg = 0.13333333333333333     Prob pos = 0.8666666666666667
count neg = 4     count pos = 26     sum = 30

 32%|█████████████████████████                                                      | 99/314 [03:35<06:16,  1.75
s/it]

Feature = EnvironmentalScience ForeignLanguages
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 32%|█████████████████████████▌                                                     | 100/314 [03:37<06:16,  1.76
s/it]

Feature = Health_LifeScience SpecialNeeds
Prob neg = 0.06666666666666667     Prob pos = 0.9333333333333333
count neg = 1     count pos = 14     sum = 15

 32%|█████████████████████████▊                                                     | 101/314 [03:39<06:54,  1.94
s/it]

Feature = CharacterEducation Health_Wellness
Prob neg = 0.2727272727272727     Prob pos = 0.7272727272727273
count neg = 3     count pos = 8     sum = 11

 32%|█████████████████████████▉                                                     | 102/314 [03:41<06:35,  1.86
s/it]

Feature = College_CareerPrep VisualArts
Prob neg = 0.03225806451612903     Prob pos = 0.967741935483871
count neg = 1     count pos = 30     sum = 31

 33%|██████████████████████████▏                                                    | 103/314 [03:42<06:23,  1.82
s/it]

Feature = AppliedSciences EarlyDevelopment
Prob neg = 0.18181818181818182     Prob pos = 0.8181818181818182
count neg = 4     count pos = 18     sum = 22

 33%|██████████████████████████▌                                                    | 104/314 [03:44<06:25,  1.83
s/it]

Feature = AppliedSciences EnvironmentalScience
Prob neg = 0.208955223880597     Prob pos = 0.7910447761194029
count neg = 28     count pos = 106     sum = 134
```

```
 33%|██████████████████████████▊                                                     | 105/314 [03:46<06:20,  1.82
s/it]

Feature = AppliedSciences SpecialNeeds
Prob neg = 0.11904761904761904     Prob pos = 0.8809523809523809
count neg = 5     count pos = 37     sum = 42

 34%|███████████████████████████                                                     | 106/314 [03:48<06:17,  1.81
s/it]

Feature = College_CareerPrep
Prob neg = 0.19642857142857142     Prob pos = 0.8035714285714286
count neg = 11     count pos = 45     sum = 56

 34%|███████████████████████████▎                                                    | 107/314 [03:50<06:13,  1.80
s/it]

Feature = EnvironmentalScience Literacy
Prob neg = 0.11764705882352941     Prob pos = 0.8823529411764706
count neg = 6     count pos = 45     sum = 51

 34%|███████████████████████████▌                                                    | 108/314 [03:52<06:27,  1.88
s/it]

Feature = Extracurricular Literacy
Prob neg = 0.375     Prob pos = 0.625
count neg = 3     count pos = 5     sum = 8

 35%|███████████████████████████▊                                                    | 109/314 [03:54<06:47,  1.99
s/it]

Feature = Gym_Fitness Literacy
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4

 35%|████████████████████████████                                                    | 110/314 [03:56<07:07,  2.09
s/it]

Feature = Literacy Other
Prob neg = 0.17391304347826086     Prob pos = 0.8260869565217391
count neg = 4     count pos = 19     sum = 23

 35%|████████████████████████████▎                                                   | 111/314 [03:59<07:24,  2.19
s/it]

Feature = Health_LifeScience ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 36%|████████████████████████████▌                                                   | 112/314 [04:01<07:40,  2.28
s/it]

Feature = ESL
Prob neg = 0.125     Prob pos = 0.875
count neg = 8     count pos = 56     sum = 64

 36%|████████████████████████████▊                                                   | 113/314 [04:04<07:48,  2.33
s/it]

Feature = Mathematics VisualArts
Prob neg = 0.1896551724137931     Prob pos = 0.8103448275862069
count neg = 11     count pos = 47     sum = 58

 36%|█████████████████████████████                                                   | 114/314 [04:06<07:52,  2.36
s/it]

Feature = ForeignLanguages History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 37%|█████████████████████████████▎                                                  | 115/314 [04:08<07:48,  2.35
s/it]

Feature = Literature_Writing Music
Prob neg = 0.1     Prob pos = 0.9
count neg = 1     count pos = 9     sum = 10

 37%|█████████████████████████████▌                                                  | 116/314 [04:10<07:17,  2.21
s/it]

Feature = CommunityService VisualArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 37%|█████████████████████████████▊                                                  | 117/314 [04:12<06:46,  2.06
s/it]

Feature = College_CareerPrep EnvironmentalScience
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4
```

```
 38%|██████████████████████████████                                                  | 118/314 [04:14<06:29,  1.99
s/it]

Feature = CharacterEducation CommunityService
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 8     sum = 8

 38%|██████████████████████████████▍                                                 | 119/314 [04:15<06:12,  1.91
s/it]

Feature = EarlyDevelopment Literature_Writing
Prob neg = 0.10714285714285714     Prob pos = 0.8928571428571429
count neg = 3     count pos = 25     sum = 28

 38%|██████████████████████████████▌                                                 | 120/314 [04:17<05:59,  1.85
s/it]

Feature = CharacterEducation Mathematics
Prob neg = 0.3     Prob pos = 0.7
count neg = 3     count pos = 7     sum = 10

 39%|██████████████████████████████▉                                                 | 121/314 [04:19<05:50,  1.82
s/it]

Feature = AppliedSciences Other
Prob neg = 0.05555555555555555     Prob pos = 0.9444444444444444
count neg = 1     count pos = 17     sum = 18

 39%|███████████████████████████████                                                 | 122/314 [04:21<05:44,  1.79
s/it]

Feature = AppliedSciences College_CareerPrep
Prob neg = 0.18333333333333332     Prob pos = 0.8166666666666667
count neg = 11     count pos = 49     sum = 60

 39%|███████████████████████████████▍                                                | 123/314 [04:22<05:39,  1.78
s/it]

Feature = Extracurricular Other
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 6     sum = 6

 39%|███████████████████████████████▌                                                | 124/314 [04:24<05:34,  1.76
s/it]

Feature = Literature_Writing PerformingArts
Prob neg = 0.2     Prob pos = 0.8
count neg = 3     count pos = 12     sum = 15

 40%|███████████████████████████████▉                                                | 125/314 [04:26<05:30,  1.75
s/it]

Feature = Civics_Government College_CareerPrep
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5
unique feature = Extracurricular History_Geography

 40%|████████████████████████████████                                                | 126/314 [04:28<05:27,  1.74
s/it]

Feature = Extracurricular History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 40%|████████████████████████████████▍                                               | 127/314 [04:29<05:23,  1.73
s/it]

Feature = Gym_Fitness Mathematics
Prob neg = 0.3     Prob pos = 0.7
count neg = 3     count pos = 7     sum = 10

 41%|████████████████████████████████▌                                               | 128/314 [04:31<05:20,  1.72
s/it]

Feature = Civics_Government Literacy
Prob neg = 0.06666666666666667     Prob pos = 0.9333333333333333
count neg = 1     count pos = 14     sum = 15

 41%|████████████████████████████████▉                                               | 129/314 [04:33<05:21,  1.74
s/it]

Feature = AppliedSciences Literature_Writing
Prob neg = 0.07692307692307693     Prob pos = 0.9230769230769231
count neg = 4     count pos = 48     sum = 52

 41%|█████████████████████████████████                                               | 130/314 [04:34<05:16,  1.72
s/it]

Feature = SocialSciences
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 4     count pos = 20     sum = 24
```

```
 42%|█████████████████████████████████▍                                              | 131/314 [04:36<05:12,  1.71
s/it]

Feature = Health_Wellness History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 42%|█████████████████████████████████▋                                              | 132/314 [04:38<05:14,  1.73
s/it]

Feature = Mathematics TeamSports
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 42%|█████████████████████████████████▊                                              | 133/314 [04:40<05:20,  1.77
s/it]

Feature = Extracurricular
Prob neg = 0.1875     Prob pos = 0.8125
count neg = 3     count pos = 13     sum = 16

 43%|██████████████████████████████████▏                                             | 134/314 [04:42<05:21,  1.79
s/it]

Feature = EarlyDevelopment SpecialNeeds
Prob neg = 0.18018018018018017     Prob pos = 0.8198198198198198
count neg = 20     count pos = 91     sum = 111

 43%|██████████████████████████████████▍                                             | 135/314 [04:43<05:16,  1.77
s/it]

Feature = CommunityService
Prob neg = 0.18181818181818182     Prob pos = 0.8181818181818182
count neg = 2     count pos = 9     sum = 11

 43%|██████████████████████████████████▋                                             | 136/314 [04:45<05:14,  1.77
s/it]

Feature = CommunityService Health_Wellness
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 44%|██████████████████████████████████▊                                             | 137/314 [04:47<05:39,  1.92
s/it]

Feature = Extracurricular Music
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 44%|███████████████████████████████████▏                                            | 138/314 [04:49<05:49,  1.99
s/it]

Feature = Literature_Writing Other
Prob neg = 0.07142857142857142     Prob pos = 0.9285714285714286
count neg = 1     count pos = 13     sum = 14

 44%|███████████████████████████████████▍                                            | 139/314 [04:51<05:35,  1.92
s/it]

Feature = Extracurricular Health_Wellness
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 45%|███████████████████████████████████▋                                            | 140/314 [04:53<05:22,  1.85
s/it]

Feature = Literature_Writing ParentInvolvement
Prob neg = 0.1     Prob pos = 0.9
count neg = 1     count pos = 9     sum = 10

 45%|███████████████████████████████████▊                                            | 141/314 [04:55<05:16,  1.83
s/it]

Feature = CharacterEducation Music
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5

 45%|████████████████████████████████████▏                                           | 142/314 [04:56<05:10,  1.81
s/it]

Feature = Economics Mathematics
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 7     sum = 7

 46%|████████████████████████████████████▍                                           | 143/314 [04:58<05:05,  1.78
s/it]

Feature = Music SpecialNeeds
Prob neg = 0.06666666666666667     Prob pos = 0.9333333333333333
count neg = 1     count pos = 14     sum = 15
```

```
 46%|████████████████████████████████████▌                                           | 144/314 [05:00<05:06,  1.81
s/it]

Feature = Health_LifeScience Health_Wellness
Prob neg = 0.21052631578947367     Prob pos = 0.7894736842105263
count neg = 4     count pos = 15     sum = 19

 46%|████████████████████████████████████▋                                           | 145/314 [05:02<05:03,  1.79
s/it]

Feature = ESL VisualArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 7     sum = 7

 46%|█████████████████████████████████████▏                                          | 146/314 [05:04<04:58,  1.78
s/it]

Feature = CharacterEducation Other
Prob neg = 0.1875     Prob pos = 0.8125
count neg = 3     count pos = 13     sum = 16

 47%|█████████████████████████████████████▌                                          | 147/314 [05:05<04:53,  1.76
s/it]

Feature = AppliedSciences Extracurricular
Prob neg = 0.1     Prob pos = 0.9
count neg = 1     count pos = 9     sum = 10

 47%|█████████████████████████████████████▊                                          | 148/314 [05:07<04:51,  1.75
s/it]

Feature = CharacterEducation SpecialNeeds
Prob neg = 0.30434782608695654     Prob pos = 0.6956521739130435
count neg = 7     count pos = 16     sum = 23

 47%|█████████████████████████████████████▉                                          | 149/314 [05:09<04:42,  1.71
s/it]

Feature = ESL EarlyDevelopment
Prob neg = 0.125     Prob pos = 0.875
count neg = 1     count pos = 7     sum = 8

 48%|██████████████████████████████████████▏                                         | 150/314 [05:10<04:41,  1.72
s/it]

Feature = CharacterEducation EnvironmentalScience
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4

 48%|██████████████████████████████████████▌                                         | 151/314 [05:12<04:41,  1.73
s/it]

Feature = ESL Mathematics
Prob neg = 0.16216216216216217     Prob pos = 0.8378378378378378
count neg = 6     count pos = 31     sum = 37

 48%|██████████████████████████████████████▊                                         | 152/314 [05:14<04:41,  1.74
s/it]

Feature = Civics_Government Health_LifeScience
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4

 49%|██████████████████████████████████████▉                                         | 153/314 [05:16<04:44,  1.76
s/it]

Feature = ESL ForeignLanguages
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5

 49%|███████████████████████████████████████▏                                        | 154/314 [05:18<04:53,  1.84
s/it]

Feature = Gym_Fitness NutritionEducation
Prob neg = 0.25     Prob pos = 0.75
count neg = 3     count pos = 9     sum = 12

 49%|███████████████████████████████████████▌                                        | 155/314 [05:20<05:15,  1.98
s/it]

Feature = Civics_Government SpecialNeeds
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5

 50%|███████████████████████████████████████▊                                        | 156/314 [05:22<05:00,  1.90
s/it]

Feature = Mathematics SocialSciences
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 3     count pos = 6     sum = 9
unique feature = History_Geography ParentInvolvement
```

```
 50%|████████████████████████████████████████                                        | 157/314 [05:23<04:49,  1.84
s/it]

Feature = History_Geography ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 50%|████████████████████████████████████████▎                                       | 158/314 [05:25<04:41,  1.81
s/it]

Feature = Health_LifeScience VisualArts
Prob neg = 0.26666666666666666     Prob pos = 0.7333333333333333
count neg = 4     count pos = 11     sum = 15

 51%|████████████████████████████████████████▌                                       | 159/314 [05:27<04:38,  1.80
s/it]

Feature = Literacy Music
Prob neg = 0.06666666666666667     Prob pos = 0.9333333333333333
count neg = 1     count pos = 14     sum = 15

 51%|████████████████████████████████████████▊                                       | 160/314 [05:29<04:35,  1.79
s/it]

Feature = NutritionEducation SpecialNeeds
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5

 51%|█████████████████████████████████████████                                       | 161/314 [05:30<04:29,  1.76
s/it]

Feature = CharacterEducation TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 5     sum = 5

 52%|█████████████████████████████████████████▎                                      | 162/314 [05:32<04:23,  1.74
s/it]

Feature = Civics_Government Literature_Writing
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 11     sum = 11

 52%|█████████████████████████████████████████▌                                      | 163/314 [05:34<04:21,  1.73
s/it]

Feature = FinancialLiteracy SpecialNeeds
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 52%|█████████████████████████████████████████▊                                      | 164/314 [05:36<04:20,  1.74
s/it]

Feature = ForeignLanguages VisualArts
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 53%|██████████████████████████████████████████                                      | 165/314 [05:37<04:18,  1.74
s/it]

Feature = Civics_Government History_Geography
Prob neg = 0.20833333333333334     Prob pos = 0.7916666666666666
count neg = 5     count pos = 19     sum = 24

 53%|██████████████████████████████████████████▎                                     | 166/314 [05:39<04:17,  1.74
s/it]

Feature = EarlyDevelopment PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 53%|██████████████████████████████████████████▌                                     | 167/314 [05:41<04:14,  1.73
s/it]

Feature = AppliedSciences Gym_Fitness
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 54%|██████████████████████████████████████████▊                                     | 168/314 [05:42<04:12,  1.73
s/it]

Feature = College_CareerPrep EarlyDevelopment
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 54%|███████████████████████████████████████████                                     | 169/314 [05:44<04:11,  1.73
s/it]

Feature = ESL Health_LifeScience
Prob neg = 0.75     Prob pos = 0.25
count neg = 3     count pos = 1     sum = 4
```

```
 54%|███████████████████████████████████████████▎                                    | 170/314 [05:46<04:09,  1.73
s/it]

Feature = Extracurricular Literature_Writing
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 54%|███████████████████████████████████████████▌                                    | 171/314 [05:48<04:08,  1.74
s/it]

Feature = EnvironmentalScience SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 9     sum = 9

 55%|███████████████████████████████████████████▊                                    | 172/314 [05:49<04:07,  1.74
s/it]

Feature = FinancialLiteracy History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = FinancialLiteracy ForeignLanguages

 55%|████████████████████████████████████████████                                    | 173/314 [05:51<04:03,  1.72
s/it]

Feature = FinancialLiteracy ForeignLanguages
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 55%|████████████████████████████████████████████▎                                   | 174/314 [05:53<04:00,  1.71
s/it]

Feature = Mathematics ParentInvolvement
Prob neg = 0.08333333333333333     Prob pos = 0.9166666666666666
count neg = 1     count pos = 11     sum = 12

 56%|████████████████████████████████████████████▌                                   | 175/314 [05:54<03:58,  1.72
s/it]

Feature = AppliedSciences ParentInvolvement
Prob neg = 0.125     Prob pos = 0.875
count neg = 1     count pos = 7     sum = 8

 56%|████████████████████████████████████████████▊                                   | 176/314 [05:56<03:59,  1.73
s/it]

Feature = CharacterEducation Health_LifeScience
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 56%|█████████████████████████████████████████████                                   | 177/314 [05:58<03:55,  1.72
s/it]

Feature = EarlyDevelopment NutritionEducation
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 57%|█████████████████████████████████████████████▎                                  | 178/314 [06:00<03:57,  1.75
s/it]

Feature = ForeignLanguages Literature_Writing
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 4     count pos = 8     sum = 12

 57%|█████████████████████████████████████████████▌                                  | 179/314 [06:02<03:57,  1.76
s/it]

Feature = EnvironmentalScience Gym_Fitness
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 57%|█████████████████████████████████████████████▊                                  | 180/314 [06:03<03:57,  1.77
s/it]

Feature = EarlyDevelopment Extracurricular
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3
unique feature = ParentInvolvement PerformingArts

 58%|██████████████████████████████████████████████                                  | 181/314 [06:05<03:55,  1.77
s/it]

Feature = ParentInvolvement PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 58%|██████████████████████████████████████████████▎                                 | 182/314 [06:07<03:52,  1.76
s/it]
```

```
Feature = History_Geography SpecialNeeds
Prob neg = 0.3125     Prob pos = 0.6875
count neg = 5     count pos = 11     sum = 16
unique feature = ESL Music

 58%|██████████████████████████████████████████████▋                                 | 183/314 [06:09<03:49,  1.75
s/it]

Feature = ESL Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 59%|██████████████████████████████████████████████▉                                 | 184/314 [06:10<03:52,  1.79
s/it]

Feature = Health_Wellness Warmth Care_Hunger
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4

 59%|███████████████████████████████████████████████▏                                | 185/314 [06:12<03:48,  1.77
s/it]

Feature = Music SocialSciences
Prob neg = 0.5     Prob pos = 0.5
count neg = 2     count pos = 2     sum = 4

 59%|███████████████████████████████████████████████▍                                | 186/314 [06:14<03:46,  1.77
s/it]

Feature = Extracurricular Mathematics
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 8     sum = 8

 60%|███████████████████████████████████████████████▋                                | 187/314 [06:16<03:41,  1.74
s/it]

Feature = CommunityService EnvironmentalScience
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5

 60%|███████████████████████████████████████████████▉                                | 188/314 [06:17<03:38,  1.73
s/it]

Feature = AppliedSciences Health_Wellness
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6

 60%|████████████████████████████████████████████████▏                               | 189/314 [06:19<03:35,  1.72
s/it]

Feature = College_CareerPrep ParentInvolvement
Prob neg = 0.5     Prob pos = 0.5
count neg = 2     count pos = 2     sum = 4

 61%|████████████████████████████████████████████████▍                               | 190/314 [06:21<03:33,  1.72
s/it]

Feature = CharacterEducation ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 61%|████████████████████████████████████████████████▋                               | 191/314 [06:22<03:28,  1.70
s/it]

Feature = SpecialNeeds Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 61%|████████████████████████████████████████████████▉                               | 192/314 [06:24<03:28,  1.71
s/it]

Feature = ParentInvolvement
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5

 61%|█████████████████████████████████████████████████▏                              | 193/314 [06:26<03:27,  1.72
s/it]

Feature = EnvironmentalScience VisualArts
Prob neg = 0.13043478260869565     Prob pos = 0.8695652173913043
count neg = 3     count pos = 20     sum = 23

 62%|█████████████████████████████████████████████████▍                              | 194/314 [06:28<03:26,  1.72
s/it]

Feature = Civics_Government
Prob neg = 0.2727272727272727     Prob pos = 0.7272727272727273
count neg = 3     count pos = 8     sum = 11

 62%|█████████████████████████████████████████████████▋                              | 195/314 [06:29<03:25,  1.72
s/it]
```

```
Feature = Health_LifeScience SocialSciences
Prob neg = 0.1111111111111111     Prob pos = 0.8888888888888888
count neg = 1     count pos = 8     sum = 9

 62%|█████████████████████████████████████████████████▊                               | 196/314 [06:31<03:22,  1.72
s/it]

Feature = PerformingArts VisualArts
Prob neg = 0.38461538461538464     Prob pos = 0.6153846153846154
count neg = 5     count pos = 8     sum = 13

 63%|██████████████████████████████████████████████████▏                              | 197/314 [06:33<03:20,  1.71
s/it]

Feature = CommunityService Extracurricular
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 63%|██████████████████████████████████████████████████▌                              | 198/314 [06:34<03:17,  1.70
s/it]

Feature = ESL Health_Wellness
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 63%|██████████████████████████████████████████████████▋                              | 199/314 [06:36<03:15,  1.70
s/it]

Feature = AppliedSciences Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 6     sum = 6

 64%|██████████████████████████████████████████████████▊                              | 200/314 [06:38<03:13,  1.70
s/it]

Feature = EarlyDevelopment History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 64%|███████████████████████████████████████████████████▏                             | 201/314 [06:39<03:10,  1.69
s/it]

Feature = College_CareerPrep FinancialLiteracy
Prob neg = 0.6666666666666666     Prob pos = 0.3333333333333333
count neg = 2     count pos = 1     sum = 3

 64%|███████████████████████████████████████████████████▌                             | 202/314 [06:41<03:10,  1.70
s/it]

Feature = Literacy PerformingArts
Prob neg = 0.2222222222222222     Prob pos = 0.7777777777777778
count neg = 4     count pos = 14     sum = 18
unique feature = Other PerformingArts

 65%|███████████████████████████████████████████████████▋                             | 203/314 [06:43<03:10,  1.71
s/it]

Feature = Other PerformingArts
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 65%|███████████████████████████████████████████████████▉                             | 204/314 [06:45<03:06,  1.70
s/it]

Feature = EnvironmentalScience ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 65%|████████████████████████████████████████████████████▏                            | 205/314 [06:47<03:14,  1.79
s/it]

Feature = EnvironmentalScience Health_Wellness
Prob neg = 0.42857142857142855     Prob pos = 0.5714285714285714
count neg = 3     count pos = 4     sum = 7

 66%|████████████████████████████████████████████████████▌                            | 206/314 [06:49<03:31,  1.96
s/it]

Feature = Mathematics PerformingArts
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6

 66%|████████████████████████████████████████████████████▊                            | 207/314 [06:51<03:24,  1.91
s/it]

Feature = Music VisualArts
Prob neg = 0.1111111111111111     Prob pos = 0.8888888888888888
count neg = 1     count pos = 8     sum = 9

 66%|█████████████████████████████████████████████████████▏                           | 208/314 [06:52<03:17,  1.87
s/it]
```

```
Feature = EarlyDevelopment Other
Prob neg = 0.2608695652173913     Prob pos = 0.7391304347826086
count neg = 6     count pos = 17     sum = 23

 67%|█████████████████████████████████████████████████████▏                          | 209/314 [06:54<03:13,  1.84
s/it]

Feature = CharacterEducation VisualArts
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 3     count pos = 6     sum = 9

 67%|██████████████████████████████████████████████████████                          | 210/314 [06:56<03:07,  1.80
s/it]

Feature = Other TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 67%|██████████████████████████████████████████████████████▍                         | 211/314 [06:58<03:02,  1.77
s/it]

Feature = Health_Wellness SocialSciences
Prob neg = 0.4     Prob pos = 0.6
count neg = 2     count pos = 3     sum = 5
unique feature = ESL Gym_Fitness

 68%|██████████████████████████████████████████████████████                          | 212/314 [07:00<03:02,  1.79
s/it]

Feature = ESL Gym_Fitness
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 68%|██████████████████████████████████████████████████████▏                         | 213/314 [07:01<02:57,  1.76
s/it]

Feature = College_CareerPrep PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 68%|███████████████████████████████████████████████████████                         | 214/314 [07:03<02:54,  1.75
s/it]

Feature = Civics_Government VisualArts
Prob neg = 0.25     Prob pos = 0.75
count neg = 1     count pos = 3     sum = 4

 68%|██████████████████████████████████████████████████████▌                         | 215/314 [07:05<02:50,  1.72
s/it]

Feature = Economics SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 69%|███████████████████████████████████████████████████████                         | 216/314 [07:06<02:48,  1.72
s/it]

Feature = Other ParentInvolvement
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2
unique feature = Extracurricular TeamSports

 69%|███████████████████████████████████████████████████████▏                        | 217/314 [07:08<02:45,  1.71
s/it]

Feature = Extracurricular TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 69%|████████████████████████████████████████████████████████                        | 218/314 [07:10<02:42,  1.70
s/it]

Feature = SocialSciences VisualArts
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6

 70%|███████████████████████████████████████████████████████▌                        | 219/314 [07:11<02:41,  1.70
s/it]

Feature = PerformingArts SpecialNeeds
Prob neg = 0.2857142857142857     Prob pos = 0.7142857142857143
count neg = 2     count pos = 5     sum = 7
unique feature = EnvironmentalScience TeamSports

 70%|████████████████████████████████████████████████████████                        | 220/314 [07:13<02:40,  1.71
s/it]

Feature = EnvironmentalScience TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
```

```
 70%|████████████████████████████████████████████████████████▎                       | 221/314 [07:15<02:39,  1.71
s/it]

Feature = College_CareerPrep Health_Wellness
Prob neg = 0.5     Prob pos = 0.5
count neg = 2     count pos = 2     sum = 4

 71%|████████████████████████████████████████████████████████▌                       | 222/314 [07:16<02:36,  1.70
s/it]

Feature = Health_LifeScience NutritionEducation
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 71%|████████████████████████████████████████████████████████▋                       | 223/314 [07:18<02:34,  1.70
s/it]

Feature = EarlyDevelopment Health_LifeScience
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 2     count pos = 4     sum = 6

 71%|█████████████████████████████████████████████████████████                       | 224/314 [07:20<02:33,  1.71
s/it]

Feature = AppliedSciences PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 72%|█████████████████████████████████████████████████████████▎                      | 225/314 [07:22<02:32,  1.71
s/it]

Feature = Gym_Fitness Health_LifeScience
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = CharacterEducation PerformingArts

 72%|█████████████████████████████████████████████████████████▌                      | 226/314 [07:24<02:44,  1.87
s/it]

Feature = CharacterEducation PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 72%|█████████████████████████████████████████████████████████▋                      | 227/314 [07:26<03:00,  2.07
s/it]

Feature = Gym_Fitness PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4

 73%|██████████████████████████████████████████████████████████                      | 228/314 [07:29<03:14,  2.27
s/it]

Feature = Economics
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6

 73%|██████████████████████████████████████████████████████████▎                     | 229/314 [07:32<03:27,  2.44
s/it]

Feature = NutritionEducation Other
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 73%|██████████████████████████████████████████████████████████▌                     | 230/314 [07:34<03:08,  2.25
s/it]

Feature = ESL History_Geography
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6
unique feature = EarlyDevelopment Warmth Care_Hunger

 74%|██████████████████████████████████████████████████████████▊                     | 231/314 [07:36<02:54,  2.10
s/it]

Feature = EarlyDevelopment Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 74%|███████████████████████████████████████████████████████████                     | 232/314 [07:37<02:43,  2.00
s/it]

Feature = College_CareerPrep CommunityService
Prob neg = 0.14285714285714285     Prob pos = 0.8571428571428571
count neg = 1     count pos = 6     sum = 7

 74%|███████████████████████████████████████████████████████████▎                    | 233/314 [07:39<02:34,  1.91
s/it]

Feature = ForeignLanguages Mathematics
Prob neg = 0.4     Prob pos = 0.6
count neg = 2     count pos = 3     sum = 5
```

```
 75%|███████████████████████████████████████████████████████████                     | 234/314 [07:41<02:28,  1.85
s/it]

Feature = EarlyDevelopment Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 75%|███████████████████████████████████████████████████████████▉                    | 235/314 [07:42<02:21,  1.79
s/it]

Feature = Economics History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 5     sum = 5

 75%|████████████████████████████████████████████████████████████▏                   | 236/314 [07:44<02:19,  1.79
s/it]

Feature = College_CareerPrep Other
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 5     count pos = 10     sum = 15
unique feature = FinancialLiteracy VisualArts

 75%|████████████████████████████████████████████████████████████▍                   | 237/314 [07:46<02:17,  1.78
s/it]

Feature = FinancialLiteracy VisualArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 76%|████████████████████████████████████████████████████████████▋                   | 238/314 [07:48<02:13,  1.76
s/it]

Feature = History_Geography Music
Prob neg = 0.2     Prob pos = 0.8
count neg = 1     count pos = 4     sum = 5

 76%|████████████████████████████████████████████████████████████▉                   | 239/314 [07:49<02:11,  1.75
s/it]

Feature = CharacterEducation Extracurricular
Prob neg = 0.1111111111111111     Prob pos = 0.8888888888888888
count neg = 1     count pos = 8     sum = 9

 76%|█████████████████████████████████████████████████████████████▏                  | 240/314 [07:51<02:09,  1.75
s/it]

Feature = Health_Wellness Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 77%|█████████████████████████████████████████████████████████████▍                  | 241/314 [07:53<02:06,  1.73
s/it]

Feature = SpecialNeeds TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3
unique feature = ESL PerformingArts

 77%|█████████████████████████████████████████████████████████████▋                  | 242/314 [07:54<02:03,  1.71
s/it]

Feature = ESL PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Extracurricular ParentInvolvement

 77%|█████████████████████████████████████████████████████████████▉                  | 243/314 [07:56<02:01,  1.71
s/it]

Feature = Extracurricular ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Extracurricular SocialSciences

 78%|██████████████████████████████████████████████████████████████▏                 | 244/314 [07:58<01:59,  1.71
s/it]

Feature = Extracurricular SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 78%|██████████████████████████████████████████████████████████████▍                 | 245/314 [08:00<02:00,  1.75
s/it]

Feature = Health_Wellness VisualArts
Prob neg = 0.5     Prob pos = 0.5
count neg = 2     count pos = 2     sum = 4

 78%|██████████████████████████████████████████████████████████████▋                 | 246/314 [08:01<01:58,  1.74
s/it]
```

```
Feature = SocialSciences SpecialNeeds
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 5     sum = 5

 79%|██████████████████████████████████████████████████████████████▋                | 247/314 [08:03<01:55,  1.73
s/it]

Feature = Civics_Government CommunityService
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3
unique feature = EnvironmentalScience Warmth Care_Hunger

 79%|███████████████████████████████████████████████████████████████▏               | 248/314 [08:05<01:54,  1.73
s/it]

Feature = EnvironmentalScience Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = College_CareerPrep ESL

 79%|███████████████████████████████████████████████████████████████▍               | 249/314 [08:07<01:51,  1.72
s/it]

Feature = College_CareerPrep ESL
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Literacy TeamSports

 80%|███████████████████████████████████████████████████████████████▋               | 250/314 [08:08<01:50,  1.72
s/it]

Feature = Literacy TeamSports
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 80%|███████████████████████████████████████████████████████████████▉               | 251/314 [08:10<01:47,  1.71
s/it]

Feature = ParentInvolvement VisualArts
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6

 80%|████████████████████████████████████████████████████████████████▏              | 252/314 [08:12<01:46,  1.71
s/it]

Feature = ParentInvolvement SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = Gym_Fitness ParentInvolvement

 81%|████████████████████████████████████████████████████████████████▍              | 253/314 [08:13<01:44,  1.71
s/it]

Feature = Gym_Fitness ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Health_LifeScience Music

 81%|████████████████████████████████████████████████████████████████▋              | 254/314 [08:15<01:43,  1.72
s/it]

Feature = Health_LifeScience Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 81%|████████████████████████████████████████████████████████████████▉              | 255/314 [08:17<01:41,  1.71
s/it]

Feature = EnvironmentalScience NutritionEducation
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3

 82%|█████████████████████████████████████████████████████████████████▏             | 256/314 [08:18<01:38,  1.70
s/it]

Feature = EarlyDevelopment Gym_Fitness
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2

 82%|█████████████████████████████████████████████████████████████████▍             | 257/314 [08:20<01:37,  1.71
s/it]

Feature = CommunityService Mathematics
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3
unique feature = Civics_Government EnvironmentalScience

 82%|█████████████████████████████████████████████████████████████████▋             | 258/314 [08:22<01:35,  1.71
s/it]
```

```
Feature = Civics_Government EnvironmentalScience
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Economics SpecialNeeds

 82%|█████████████████████████████████████████████████████████████████▉              | 259/314 [08:24<01:34,  1.72
s/it]

Feature = Economics SpecialNeeds
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Civics_Government Mathematics

 83%|██████████████████████████████████████████████████████████████████▏             | 260/314 [08:25<01:33,  1.73
s/it]

Feature = Civics_Government Mathematics
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 83%|██████████████████████████████████████████████████████████████████▌             | 261/314 [08:27<01:31,  1.73
s/it]

Feature = Economics Literacy
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 83%|██████████████████████████████████████████████████████████████████▊             | 262/314 [08:29<01:29,  1.73
s/it]

Feature = AppliedSciences CommunityService
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3

 84%|███████████████████████████████████████████████████████████████████             | 263/314 [08:31<01:27,  1.72
s/it]

Feature = College_CareerPrep History_Geography
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2
unique feature = ForeignLanguages Other

 84%|███████████████████████████████████████████████████████████████████▎            | 264/314 [08:32<01:26,  1.72
s/it]

Feature = ForeignLanguages Other
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 84%|███████████████████████████████████████████████████████████████████             | 265/314 [08:34<01:24,  1.72
s/it]

Feature = Health_LifeScience History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4
unique feature = CommunityService Literacy

 85%|███████████████████████████████████████████████████████████████████▊            | 266/314 [08:36<01:22,  1.73
s/it]

Feature = CommunityService Literacy
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 85%|████████████████████████████████████████████████████████████████████            | 267/314 [08:38<01:21,  1.73
s/it]

Feature = CommunityService History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 3     sum = 3
unique feature = CharacterEducation History_Geography

 85%|████████████████████████████████████████████████████████████████████▎           | 268/314 [08:39<01:19,  1.72
s/it]

Feature = CharacterEducation History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 86%|████████████████████████████████████████████████████████████████████            | 269/314 [08:41<01:18,  1.74
s/it]

Feature = EarlyDevelopment ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 86%|████████████████████████████████████████████████████████████████████▊           | 270/314 [08:43<01:16,  1.73
s/it]

Feature = AppliedSciences Civics_Government
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 4     sum = 4
```

```
 86%|█████████████████████████████████████████████████████████████████████           | 271/314 [08:45<01:16,  1.78
s/it]

Feature = Other SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 87%|█████████████████████████████████████████████████████████████████████▎          | 272/314 [08:48<01:32,  2.21
s/it]

Feature = History_Geography PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = CommunityService ParentInvolvement

 87%|█████████████████████████████████████████████████████████████████████▌          | 273/314 [08:50<01:36,  2.34
s/it]

Feature = CommunityService ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Civics_Government Economics

 87%|█████████████████████████████████████████████████████████████████████▊          | 274/314 [08:53<01:31,  2.29
s/it]

Feature = Civics_Government Economics
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = EnvironmentalScience PerformingArts

 88%|██████████████████████████████████████████████████████████████████████          | 275/314 [08:55<01:25,  2.19
s/it]

Feature = EnvironmentalScience PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 88%|██████████████████████████████████████████████████████████████████████▎         | 276/314 [08:57<01:20,  2.11
s/it]

Feature = Health_Wellness PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2
unique feature = CharacterEducation ForeignLanguages

 88%|██████████████████████████████████████████████████████████████████████▌         | 277/314 [08:58<01:14,  2.03
s/it]

Feature = CharacterEducation ForeignLanguages
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1
unique feature = Literature_Writing TeamSports

 89%|██████████████████████████████████████████████████████████████████████▊         | 278/314 [09:00<01:12,  2.01
s/it]

Feature = Literature_Writing TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Mathematics Warmth Care_Hunger

 89%|███████████████████████████████████████████████████████████████████████         | 279/314 [09:02<01:09,  1.99
s/it]

Feature = Mathematics Warmth Care_Hunger
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1
unique feature = EarlyDevelopment TeamSports

 89%|███████████████████████████████████████████████████████████████████████▎        | 280/314 [09:04<01:07,  1.98
s/it]

Feature = EarlyDevelopment TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = FinancialLiteracy Health_Wellness

 89%|███████████████████████████████████████████████████████████████████████▌        | 281/314 [09:06<01:04,  1.96
s/it]

Feature = FinancialLiteracy Health_Wellness
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = AppliedSciences Economics

 90%|███████████████████████████████████████████████████████████████████████▊        | 282/314 [09:08<01:01,  1.93
s/it]

Feature = AppliedSciences Economics
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1
```

```
 90%|████████████████████████████████████████████████████████████████████████        | 283/314 [09:10<00:58,  1.88
s/it]

Feature = Music TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 90%|████████████████████████████████████████████████████████████████████████▏       | 284/314 [09:11<00:54,  1.82
s/it]

Feature = EnvironmentalScience Other
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2
unique feature = FinancialLiteracy Other

 91%|████████████████████████████████████████████████████████████████████████▋       | 285/314 [09:13<00:52,  1.80
s/it]

Feature = FinancialLiteracy Other
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 91%|█████████████████████████████████████████████████████████████████████████       | 286/314 [09:15<00:50,  1.80
s/it]

Feature = EarlyDevelopment ForeignLanguages
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 91%|█████████████████████████████████████████████████████████████████████████       | 287/314 [09:17<00:47,  1.77
s/it]

Feature = CommunityService Literature_Writing
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 92%|█████████████████████████████████████████████████████████████████████████▌      | 288/314 [09:18<00:45,  1.76
s/it]

Feature = CharacterEducation ESL
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2
unique feature = ParentInvolvement SpecialNeeds

 92%|█████████████████████████████████████████████████████████████████████████▊      | 289/314 [09:20<00:43,  1.75
s/it]

Feature = ParentInvolvement SpecialNeeds
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 92%|██████████████████████████████████████████████████████████████████████████      | 290/314 [09:22<00:41,  1.72
s/it]

Feature = FinancialLiteracy Literacy
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2
unique feature = AppliedSciences TeamSports

 93%|██████████████████████████████████████████████████████████████████████████▏     | 291/314 [09:24<00:39,  1.73
s/it]

Feature = AppliedSciences TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = CharacterEducation NutritionEducation

 93%|██████████████████████████████████████████████████████████████████████████▌     | 292/314 [09:25<00:37,  1.72
s/it]

Feature = CharacterEducation NutritionEducation
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 93%|██████████████████████████████████████████████████████████████████████████▊     | 293/314 [09:27<00:36,  1.73
s/it]

Feature = College_CareerPrep SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 2     sum = 2

 94%|███████████████████████████████████████████████████████████████████████████▋    | 294/314 [09:29<00:34,  1.74
s/it]

Feature = Health_LifeScience Other
Prob neg = 0.3333333333333333     Prob pos = 0.6666666666666666
count neg = 1     count pos = 2     sum = 3
unique feature = Economics Literature_Writing

 94%|███████████████████████████████████████████████████████████████████████████▏    | 295/314 [09:31<00:33,  1.76
s/it]
```

```
Feature = Economics Literature_Writing
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = ESL Extracurricular

 94%|█████████████████████████████████████████████████████████████████████████▉     | 296/314 [09:32<00:32,  1.80
s/it]

Feature = ESL Extracurricular
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 95%|██████████████████████████████████████████████████████████████████████████▎    | 297/314 [09:34<00:31,  1.84
s/it]

Feature = College_CareerPrep NutritionEducation
Prob neg = 0.5     Prob pos = 0.5
count neg = 1     count pos = 1     sum = 2
unique feature = CommunityService Economics

 95%|██████████████████████████████████████████████████████████████████████████▌    | 298/314 [09:36<00:29,  1.82
s/it]

Feature = CommunityService Economics
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1
unique feature = Gym_Fitness Other

 95%|██████████████████████████████████████████████████████████████████████████▊    | 299/314 [09:38<00:26,  1.79
s/it]

Feature = Gym_Fitness Other
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1
unique feature = Gym_Fitness History_Geography

 96%|███████████████████████████████████████████████████████████████████████████    | 300/314 [09:40<00:25,  1.79
s/it]

Feature = Gym_Fitness History_Geography
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
unique feature = Gym_Fitness Music

 96%|███████████████████████████████████████████████████████████████████████████▎   | 301/314 [09:42<00:23,  1.81
s/it]

Feature = Gym_Fitness Music
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 96%|███████████████████████████████████████████████████████████████████████████▌   | 302/314 [09:43<00:21,  1.82
s/it]

unique feature = AppliedSciences FinancialLiteracy
Feature = AppliedSciences FinancialLiteracy
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 96%|███████████████████████████████████████████████████████████████████████████▊   | 303/314 [09:45<00:20,  1.83
s/it]

unique feature = Civics_Government FinancialLiteracy
Feature = Civics_Government FinancialLiteracy
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 97%|████████████████████████████████████████████████████████████████████████████   | 304/314 [09:47<00:18,  1.83
s/it]

unique feature = ESL ParentInvolvement
Feature = ESL ParentInvolvement
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 97%|████████████████████████████████████████████████████████████████████████████▎  | 305/314 [09:49<00:16,  1.87
s/it]

unique feature = Extracurricular SpecialNeeds
Feature = Extracurricular SpecialNeeds
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 97%|████████████████████████████████████████████████████████████████████████████▌  | 306/314 [09:51<00:14,  1.87
s/it]

unique feature = Music Other
Feature = Music Other
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
```

```
 98%|██████████████████████████████████████████████████████████████████████████████▏ | 307/314 [09:53<00:12,  1.83
s/it]

unique feature = College_CareerPrep Warmth Care_Hunger
Feature = College_CareerPrep Warmth Care_Hunger
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 98%|██████████████████████████████████████████████████████████████████████████████▍ | 308/314 [09:54<00:11,  1.84
s/it]

unique feature = Civics_Government TeamSports
Feature = Civics_Government TeamSports
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 98%|██████████████████████████████████████████████████████████████████████████████▋ | 309/314 [09:56<00:09,  1.81
s/it]

unique feature = Gym_Fitness SocialSciences
Feature = Gym_Fitness SocialSciences
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 99%|██████████████████████████████████████████████████████████████████████████████▉ | 310/314 [09:58<00:07,  1.79
s/it]

unique feature = CommunityService PerformingArts
Feature = CommunityService PerformingArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

 99%|███████████████████████████████████████████████████████████████████████████████▏| 311/314 [10:00<00:05,  1.82
s/it]

unique feature = EarlyDevelopment SocialSciences
Feature = EarlyDevelopment SocialSciences
Prob neg = 1.0     Prob pos = 0.0
count neg = 1     count pos = 0     sum = 1

 99%|███████████████████████████████████████████████████████████████████████████████▍| 312/314 [10:02<00:03,  1.80
s/it]

unique feature = TeamSports VisualArts
Feature = TeamSports VisualArts
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

100%|███████████████████████████████████████████████████████████████████████████████▋| 313/314 [10:03<00:01,  1.81
s/it]

unique feature = AppliedSciences ForeignLanguages
Feature = AppliedSciences ForeignLanguages
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1

100%|████████████████████████████████████████████████████████████████████████████████| 314/314 [10:05<00:00,  1.93
s/it]

unique feature = Gym_Fitness Literature_Writing
Feature = Gym_Fitness Literature_Writing
Prob neg = 0.0     Prob pos = 1.0
count neg = 0     count pos = 1     sum = 1
```

In [87]:
```
response_test_clean_subcategories.count()
```

Out[87]:
```
x    14850
y    14850
dtype: int64
```

2.2.3.4 School State

```
In [88]: X_test.school_state.value_counts()
```

```
Out[88]: CA    2078
         NY    1020
         TX     946
         FL     853
         NC     677
         IL     593
         GA     539
         SC     523
         MI     430
         PA     415
         OH     377
         WA     350
         IN     350
         MO     346
         MA     326
         LA     315
         OK     312
         NJ     304
         AZ     300
         VA     286
         WI     255
         TN     235
         UT     230
         CT     226
         AL     223
         NV     213
         MD     199
         KY     190
         CO     171
         MS     169
         OR     169
         MN     159
         AR     131
         ID      99
         IA      84
         DC      79
         KS      77
         WV      75
         NM      68
         HI      66
         NH      53
         ME      50
         AK      47
         SD      45
         DE      42
         RI      40
         NE      37
         MT      36
         ND      24
         WY      12
         VT       6
         Name: school_state, dtype: int64
```

In [89]:
```
response_test_school_state = response(X_test['school_state'],y_test)
```

```
  2%|█▋                                                                                   | 1/51 [00:02<01:46,  2.14
s/it]

Feature = OH
Prob neg = 0.14058355437665782     Prob pos = 0.8594164456233422
count neg = 53     count pos = 324     sum = 377

  4%|███▍                                                                                 | 2/51 [00:04<01:49,  2.24
s/it]

Feature = TX
Prob neg = 0.20190274841437633     Prob pos = 0.7980972515856237
count neg = 191     count pos = 755     sum = 946

  6%|█████▏                                                                               | 3/51 [00:06<01:43,  2.16
s/it]

Feature = OK
Prob neg = 0.17307692307692307     Prob pos = 0.8269230769230769
count neg = 54     count pos = 258     sum = 312

  8%|██████▉                                                                             | 4/51 [00:08<01:42,  2.19
s/it]

Feature = GA
Prob neg = 0.20222634508348794     Prob pos = 0.7977736549165121
count neg = 109     count pos = 430     sum = 539

 10%|████████▌                                                                            | 5/51 [00:10<01:38,  2.14
s/it]

Feature = OR
Prob neg = 0.20118343195266272     Prob pos = 0.7988165680473372
count neg = 34     count pos = 135     sum = 169

 12%|██████████▎                                                                          | 6/51 [00:12<01:35,  2.11
s/it]

Feature = MA
Prob neg = 0.147239263803681     Prob pos = 0.852760736196319
count neg = 48     count pos = 278     sum = 326

 14%|████████████                                                                         | 7/51 [00:15<01:37,  2.21
s/it]

Feature = FL
Prob neg = 0.17233294255568582     Prob pos = 0.8276670574443142
count neg = 147     count pos = 706     sum = 853

 16%|█████████████▊                                                                      | 8/51 [00:17<01:34,  2.20
s/it]

Feature = SC
Prob neg = 0.1491395793499044     Prob pos = 0.8508604206500956
count neg = 78     count pos = 445     sum = 523

 18%|███████████████▌                                                                     | 9/51 [00:19<01:30,  2.15
s/it]

Feature = NJ
Prob neg = 0.1611842105263158     Prob pos = 0.8388157894736842
count neg = 49     count pos = 255     sum = 304

 20%|████████████████▉                                                                  | 10/51 [00:22<01:43,  2.52
s/it]

Feature = CA
Prob neg = 0.14485081809432146     Prob pos = 0.8551491819056786
count neg = 301     count pos = 1777     sum = 2078

 22%|██████████████████▋                                                                 | 11/51 [00:24<01:32,  2.32
s/it]

Feature = KY
Prob neg = 0.12631578947368421     Prob pos = 0.8736842105263158
count neg = 24     count pos = 166     sum = 190

 24%|████████████████████▎                                                               | 12/51 [00:26<01:26,  2.22
s/it]

Feature = MS
Prob neg = 0.09467455621301775     Prob pos = 0.9053254437869822
count neg = 16     count pos = 153     sum = 169

 25%|██████████████████████                                                              | 13/51 [00:28<01:21,  2.15
s/it]

Feature = CT
Prob neg = 0.11504424778761062     Prob pos = 0.8849557522123894
count neg = 26     count pos = 200     sum = 226
```

```
 27%|██████████████████████                                                                   | 14/51 [00:30<01:16,  2.06
s/it]

Feature = AR
Prob neg = 0.17557251908396945     Prob pos = 0.8244274809160306
count neg = 23     count pos = 108     sum = 131

 29%|████████████████████████                                                                 | 15/51 [00:32<01:15,  2.09
s/it]

Feature = MO
Prob neg = 0.14450867052023122     Prob pos = 0.8554913294797688
count neg = 50     count pos = 296     sum = 346

 31%|█████████████████████████▌                                                                | 16/51 [00:34<01:14,  2.12
s/it]

Feature = PA
Prob neg = 0.14457831325301204     Prob pos = 0.8554216867469879
count neg = 60     count pos = 355     sum = 415

 33%|███████████████████████████▎                                                              | 17/51 [00:36<01:10,  2.08
s/it]

Feature = CO
Prob neg = 0.18128654970760233     Prob pos = 0.8187134502923976
count neg = 31     count pos = 140     sum = 171

 35%|████████████████████████████▊                                                             | 18/51 [00:38<01:07,  2.04
s/it]

Feature = MN
Prob neg = 0.11320754716981132     Prob pos = 0.8867924528301887
count neg = 18     count pos = 141     sum = 159

 37%|██████████████████████████████                                                           | 19/51 [00:41<01:12,  2.26
s/it]

Feature = NC
Prob neg = 0.1654357459379616     Prob pos = 0.8345642540620384
count neg = 112     count pos = 565     sum = 677

 39%|████████████████████████████████▏                                                         | 20/51 [00:44<01:12,  2.35
s/it]

Feature = NV
Prob neg = 0.14084507042253522     Prob pos = 0.8591549295774648
count neg = 30     count pos = 183     sum = 213

 41%|█████████████████████████████████▊                                                        | 21/51 [00:46<01:08,  2.29
s/it]

Feature = WA
Prob neg = 0.11142857142857143     Prob pos = 0.8885714285714286
count neg = 39     count pos = 311     sum = 350

 43%|███████████████████████████████████▍                                                      | 22/51 [00:48<01:08,  2.35
s/it]

Feature = IL
Prob neg = 0.17706576728499157     Prob pos = 0.8229342327150084
count neg = 105     count pos = 488     sum = 593

 45%|████████████████████████████████████▊                                                     | 23/51 [00:51<01:06,  2.38
s/it]

Feature = TN
Prob neg = 0.14042553191489363     Prob pos = 0.8595744680851064
count neg = 33     count pos = 202     sum = 235

 47%|██████████████████████████████████████                                                   | 24/51 [00:54<01:08,  2.55
s/it]

Feature = NY
Prob neg = 0.1264705882352941     Prob pos = 0.8735294117647059
count neg = 129     count pos = 891     sum = 1020

 49%|████████████████████████████████████████▏                                                 | 25/51 [00:56<01:03,  2.45
s/it]

Feature = LA
Prob neg = 0.1619047619047619     Prob pos = 0.8380952380952381
count neg = 51     count pos = 264     sum = 315

 51%|█████████████████████████████████████████▊                                                | 26/51 [00:58<00:59,  2.37
s/it]

Feature = IN
Prob neg = 0.12857142857142856     Prob pos = 0.8714285714285714
count neg = 45     count pos = 305     sum = 350
```

```
 53%|███████████████████████████████████████████▌                                       | 27/51 [01:01<00:57,  2.41
s/it]

Feature = AZ
Prob neg = 0.15333333333333332     Prob pos = 0.8466666666666667
count neg = 46     count pos = 254     sum = 300

 55%|█████████████████████████████████████████████                                      | 28/51 [01:03<00:53,  2.31
s/it]

Feature = UT
Prob neg = 0.15217391304347827     Prob pos = 0.8478260869565217
count neg = 35     count pos = 195     sum = 230

 57%|███████████████████████████████████████████████▌                                   | 29/51 [01:05<00:48,  2.22
s/it]

Feature = NE
Prob neg = 0.13513513513513514     Prob pos = 0.8648648648648649
count neg = 5     count pos = 32     sum = 37

 59%|████████████████████████████████████████████████▏                                  | 30/51 [01:08<00:50,  2.39
s/it]

Feature = VA
Prob neg = 0.12937062937062938     Prob pos = 0.8706293706293706
count neg = 37     count pos = 249     sum = 286

 61%|██████████████████████████████████████████████████▌                                | 31/51 [01:11<00:53,  2.67
s/it]

Feature = AL
Prob neg = 0.12556053811659193     Prob pos = 0.874439461883408
count neg = 28     count pos = 195     sum = 223

 63%|███████████████████████████████████████████████████▌                               | 32/51 [01:14<00:53,  2.82
s/it]

Feature = KS
Prob neg = 0.12987012987012986     Prob pos = 0.8701298701298701
count neg = 10     count pos = 67     sum = 77

 65%|█████████████████████████████████████████████████████                              | 33/51 [01:17<00:51,  2.88
s/it]

Feature = WI
Prob neg = 0.16862745098039217     Prob pos = 0.8313725490196079
count neg = 43     count pos = 212     sum = 255

 67%|███████████████████████████████████████████████████████▌                           | 34/51 [01:20<00:49,  2.90
s/it]

Feature = MI
Prob neg = 0.15813953488372093     Prob pos = 0.8418604651162791
count neg = 68     count pos = 362     sum = 430

 69%|████████████████████████████████████████████████████████▏                          | 35/51 [01:24<00:49,  3.12
s/it]

Feature = MD
Prob neg = 0.1407035175879397     Prob pos = 0.8592964824120602
count neg = 28     count pos = 171     sum = 199

 71%|██████████████████████████████████████████████████████████▊                        | 36/51 [01:26<00:43,  2.87
s/it]

Feature = MT
Prob neg = 0.2222222222222222     Prob pos = 0.7777777777777778
count neg = 8     count pos = 28     sum = 36

 73%|███████████████████████████████████████████████████████████▌                       | 37/51 [01:29<00:41,  2.93
s/it]

Feature = ID
Prob neg = 0.16161616161616163     Prob pos = 0.8383838383838383
count neg = 16     count pos = 83     sum = 99

 75%|█████████████████████████████████████████████████████████████                      | 38/51 [01:32<00:37,  2.86
s/it]

Feature = ME
Prob neg = 0.2     Prob pos = 0.8
count neg = 10     count pos = 40     sum = 50

 76%|███████████████████████████████████████████████████████████████▌                   | 39/51 [01:34<00:31,  2.64
s/it]

Feature = ND
Prob neg = 0.125     Prob pos = 0.875
count neg = 3     count pos = 21     sum = 24
```

```
 78%|████████████████████████████████████████████████████████████████▎                 | 40/51 [01:36<00:28,  2.58
s/it]

Feature = DC
Prob neg = 0.24050632911392406     Prob pos = 0.759493670886076
count neg = 19     count pos = 60     sum = 79

 80%|█████████████████████████████████████████████████████████████████▉                | 41/51 [01:38<00:23,  2.39
s/it]

Feature = HI
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 11     count pos = 55     sum = 66

 82%|███████████████████████████████████████████████████████████████████               | 42/51 [01:40<00:20,  2.26
s/it]

Feature = NM
Prob neg = 0.14705882352941177     Prob pos = 0.8529411764705882
count neg = 10     count pos = 58     sum = 68

 84%|█████████████████████████████████████████████████████████████████████▏            | 43/51 [01:42<00:17,  2.13
s/it]

Feature = RI
Prob neg = 0.1     Prob pos = 0.9
count neg = 4     count pos = 36     sum = 40

 86%|██████████████████████████████████████████████████████████████████████▊           | 44/51 [01:44<00:14,  2.05
s/it]

Feature = SD
Prob neg = 0.15555555555555556     Prob pos = 0.8444444444444444
count neg = 7     count pos = 38     sum = 45

 88%|████████████████████████████████████████████████████████████████████████▍         | 45/51 [01:46<00:12,  2.10
s/it]

Feature = WV
Prob neg = 0.16     Prob pos = 0.84
count neg = 12     count pos = 63     sum = 75

 90%|█████████████████████████████████████████████████████████████████████████▉        | 46/51 [01:48<00:10,  2.11
s/it]

Feature = AK
Prob neg = 0.1702127659574468     Prob pos = 0.8297872340425532
count neg = 8     count pos = 39     sum = 47

 92%|███████████████████████████████████████████████████████████████████████████       | 47/51 [01:50<00:08,  2.14
s/it]

Feature = DE
Prob neg = 0.11904761904761904     Prob pos = 0.8809523809523809
count neg = 5     count pos = 37     sum = 42

 94%|█████████████████████████████████████████████████████████████████████████████▏    | 48/51 [01:52<00:06,  2.08
s/it]

Feature = NH
Prob neg = 0.09433962264150944     Prob pos = 0.9056603773584906
count neg = 5     count pos = 48     sum = 53

 96%|██████████████████████████████████████████████████████████████████████████████▊   | 49/51 [01:55<00:04,  2.34
s/it]

Feature = WY
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 2     count pos = 10     sum = 12

 98%|████████████████████████████████████████████████████████████████████████████████▍ | 50/51 [02:00<00:03,  3.04
s/it]

Feature = IA
Prob neg = 0.13095238095238096     Prob pos = 0.8690476190476191
count neg = 11     count pos = 73     sum = 84

100%|██████████████████████████████████████████████████████████████████████████████████| 51/51 [02:02<00:00,  2.40
s/it]

Feature = VT
Prob neg = 0.16666666666666666     Prob pos = 0.8333333333333334
count neg = 1     count pos = 5     sum = 6
```

```
In [90]: response_test_school_state.count()
```

```
Out[90]: x     14850
         y     14850
         dtype: int64
```

2.2.2.5 Project Grade category

```
In [91]: X_test.clean_project_grade_category.value_counts()
```

```
Out[91]: PreK-2    5958
         3-5       5117
         6-8       2338
         9-12      1437
         Name: clean_project_grade_category, dtype: int64
```

```
In [92]: response_test_clean_project_grade_category = response(X_test['clean_project_grade_category'],y_test)
```

```
 25%|█████████████████████                                                                    | 1/4 [00:06<00:20,  6.71
s/it]

Feature = PreK-2
Prob neg = 0.15256797583081572     Prob pos = 0.8474320241691843
count neg = 909     count pos = 5049     sum = 5958

 50%|██████████████████████████████████████████                                               | 2/4 [00:12<00:12,  6.49
s/it]

Feature = 3-5
Prob neg = 0.146765683017393     Prob pos = 0.853234316982607
count neg = 751     count pos = 4366     sum = 5117

 75%|███████████████████████████████████████████████████████████████                          | 3/4 [00:16<00:05,  5.77
s/it]

Feature = 6-8
Prob neg = 0.16167664670658682     Prob pos = 0.8383233532934131
count neg = 378     count pos = 1960     sum = 2338

100%|████████████████████████████████████████████████████████████████████████████████████████| 4/4 [00:20<00:00,  5.08
s/it]

Feature = 9-12
Prob neg = 0.17397355601948503     Prob pos = 0.826026443980515
count neg = 250     count pos = 1187     sum = 1437
```

```
In [93]: response_test_clean_project_grade_category.count()
```

```
Out[93]: x    14850
         y    14850
         dtype: int64
```

```
In [ ]:
```

```
In [ ]:
```

2.3 Make Data Model Ready: encoding eassay, and project_title

```
In [94]: # 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
```

Ecoding Essay and Project title

```
2.3.1 BOW
2.3.2 TFIDF
2.3.3 AVG W2V
2.3.4 TFIDF W2V
```

2.3.1 BOW Essays and Title

2.3.1.1 BOW Essay

In [95]:
```python
print(X_train.shape, y_train.shape)
print(X_test.shape, y_test.shape)

print("="*100)


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

# we use the fitted CountVectorizer to convert the text to vector
X_train_essay_bow = vectorizer.transform(X_train['preprocessed_essays'].values)
X_test_essay_bow = vectorizer.transform(X_test['preprocessed_essays'].values)

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

```
(30150, 10) (30150,)
(14850, 10) (14850,)
====================================================================================================
After vectorizations
(30150, 5000) (30150,)
(14850, 5000) (14850,)
====================================================================================================
```

2.3.1.2 BOW Title

In [96]:
```python
print(X_train.shape, y_train.shape)
print(X_test.shape, y_test.shape)

print("="*100)


vectorizer = CountVectorizer(min_df=10, max_features=5000)
vectorizer.fit(X_train['preprocessed_titles'].values) # fit has to happen only on train data

# we use the fitted CountVectorizer to convert the text to vector
X_train_title_bow = vectorizer.transform(X_train['preprocessed_titles'].values)
X_test_title_bow = vectorizer.transform(X_test['preprocessed_titles'].values)

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

```
(30150, 10) (30150,)
(14850, 10) (14850,)
====================================================================================================
After vectorizations
(30150, 1512) (30150,)
(14850, 1512) (14850,)
====================================================================================================
```

2.3.2 TF IDF Essay and Title

2.3.2.1 TF IDF Essay

In [97]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

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

print("="*100)


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

# we use the fitted CountVectorizer to convert the text to vector
X_train_essay_tfidf = vectorizer.transform(X_train['preprocessed_essays'].values)
X_test_essay_tfidf = vectorizer.transform(X_test['preprocessed_essays'].values)

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

```
(30150, 10) (30150,)
(14850, 10) (14850,)
====================================================================================================
After vectorizations
(30150, 5000) (30150,)
(14850, 5000) (14850,)
====================================================================================================
```

2.3.2.2 TF IDF Title

In [98]:
```
print(X_train.shape, y_train.shape)
print(X_test.shape, y_test.shape)

print("="*100)


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

# we use the fitted CountVectorizer to convert the text to vector
X_train_title_tfidf = vectorizer.transform(X_train['preprocessed_titles'].values)
X_test_title_tfidf = vectorizer.transform(X_test['preprocessed_titles'].values)

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

```
(30150, 10) (30150,)
(14850, 10) (14850,)
====================================================================================================
After vectorizations
(30150, 1512) (30150,)
(14850, 1512) (14850,)
====================================================================================================
```

2.3.3 AVG W2V Essay and Title

2.3.3.1 AVG W2V Essay

In [99]:
```
# stronging variables into pickle files python: http://www.jessicayung.com/how-to-use-pickle-to-save-and-load-varia
bles-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 [100]: # average Word2Vec
          # compute average word2vec for each review.
          avg_w2v_vectors_train = []; # the avg-w2v for each sentence/review is stored in this list
          for sentence in tqdm(X_train['preprocessed_essays'].values): # for each review/sentence
              vector = np.zeros(300) # as word vectors are of zero length
              cnt_words =0; # num of words with a valid vector in the sentence/review
              for word in sentence.split(): # for each word in a review/sentence
                  if word in glove_words:
                      vector += model[word]
                      cnt_words += 1
              if cnt_words != 0:
                  vector /= cnt_words
              avg_w2v_vectors_train.append(vector)

          print(len(avg_w2v_vectors_train))
          print(len(avg_w2v_vectors_train[0]))
          print(avg_w2v_vectors_train[0])
```

```
100%|███████████████████████████████████████████████████████████████████████████| 30150/30150 [00:30<00:00, 983.59
it/s]

30150
300
[ 6.03663703e-02  1.80357668e-02  3.15515758e-03 -1.04732678e-01
 -1.79821898e-02  9.56745273e-03 -3.02265479e+00  1.32809467e-01
 -8.69166121e-03 -2.74705697e-02  5.51042855e-02  6.59065341e-02
  1.57215642e-01 -2.28268739e-02 -3.40212103e-02 -2.29368303e-03
  4.55836364e-02 -5.95546461e-02  5.92127194e-02  7.38102200e-03
  7.48897327e-02  4.64619152e-02 -7.62177612e-02  3.14140000e-04
 -4.22899273e-02 -3.39375394e-03  9.68373842e-02 -8.95884752e-02
 -6.50876739e-02 -5.10648576e-02 -1.77028622e-01 -4.59837964e-02
  2.07798091e-02  1.60771218e-02 -1.64737788e-02 -1.83907582e-02
 -7.14229091e-04 -7.54712121e-03 -3.63153080e-02 -6.69749376e-02
 -9.37662294e-02  8.21647636e-03 -3.69618503e-03 -1.14859694e-01
 -6.20851236e-03  5.34671030e-02  2.33277945e-02 -5.25274559e-02
 -1.87031879e-03 -7.81031067e-02  7.88129576e-03 -3.04058097e-02
  5.54023709e-02  7.59857212e-03 -8.58008606e-03 -6.15183030e-04
  1.15562213e-01  1.28996752e-02 -7.54857739e-02  4.64970788e-02
 -3.97331855e-02  1.66540970e-02  4.83183661e-02 -4.83502424e-04
 -1.13768747e-01  5.96814627e-02  7.99450036e-02 -3.85320782e-02
  1.51412053e-01 -1.25391520e-01 -4.03803784e-02 -7.22552000e-03
  3.20000885e-02 -5.68933248e-02  7.12134788e-03 -1.31873590e-01
  1.90602921e-02  1.01376697e-02  6.63501188e-02 -2.30053576e-02
  1.11786321e-02 -3.87882630e-01  4.98287130e-02 -5.71177436e-02
 -7.35391564e-02 -9.79018061e-04  1.44921036e-01 -7.30829745e-02
  8.78723109e-02 -1.82733219e-02  6.99676835e-02 -7.86330848e-03
  3.60461848e-02  9.41429539e-02 -2.29369636e-03 -3.01215024e-01
 -2.19216030e+00 -9.21735030e-03  4.79848315e-02  6.35802406e-02
 -8.35392215e-02  1.68145527e-02  1.13495926e-01 -1.71403636e-04
  5.09827406e-02 -2.59182164e-02  5.40112036e-02 -1.69917625e-01
 -1.02474612e-03  3.02922933e-02 -9.46909091e-04  6.73906424e-03
  3.75585573e-02  1.73469830e-01 -7.99429697e-03  2.66149727e-02
 -8.14717035e-02  7.50661230e-02  7.80216863e-02  3.17567842e-02
  2.17011776e-02  6.37558223e-02 -1.53692533e-03 -1.25898069e-01
  8.69042339e-02  3.81207067e-02  8.82430573e-02 -3.62061012e-02
  4.53016738e-02  1.40823108e-01  6.65890581e-02  1.81422055e-02
  4.03704479e-02 -5.57535855e-02  3.47837339e-02 -2.64621532e-02
  2.00530463e-01 -3.03405067e-02  1.05881362e-01  3.25648363e-01
  7.74697891e-02 -1.69235706e-02  1.88543527e-02 -4.64030273e-02
  1.15922584e-02 -9.28436364e-03  6.41507142e-02 -3.38219758e-02
  1.03772382e-01  2.87533873e-02 -4.38730909e-03 -2.75668976e-02
  3.97403509e-02 -6.67318909e-03  3.89856781e-02 -8.68695758e-03
 -1.23950960e-02 -9.02351935e-02 -8.93452944e-02  2.86166570e-02
  4.40412733e-02 -1.57587697e-04  7.58013000e-02  3.35279739e-02
 -1.89131630e-02  9.46441642e-02 -6.63584309e-02  3.69610933e-02
  1.16869763e-01 -5.36030319e-02 -1.48890061e-02 -2.84750715e-02
 -6.63892412e-02 -1.04390782e-01 -8.49304115e-03  6.69286053e-02
  2.30175150e-02  7.69093939e-04 -7.53266093e-02 -8.66918424e-02
  1.65817059e-02  2.53495826e-01 -1.82576606e-02 -4.42825600e-02
 -6.30473430e-02 -7.12471145e-02 -3.06051758e-03 -1.08713487e-01
  1.24510255e-01 -4.30205352e-02 -1.27755333e-02 -9.36726358e-02
 -6.72364381e-02 -6.81140127e-02 -2.47674303e-03 -1.48399400e-01
 -4.13821152e-02  5.72004139e-02 -2.05502103e-02 -2.34503461e-02
  8.69174085e-02 -3.21559321e-02  4.11321976e-03  7.45151833e-02
 -4.43542394e-02  1.11177139e-02  6.30421303e-02 -9.25360842e-02
  1.23839197e-01  3.17541073e-02 -1.11802988e-02  3.88824370e-02
  2.62183648e-02 -1.22908156e-01 -9.22959104e-03  3.35727073e-02
 -3.95522588e-02 -7.85739182e-02 -3.89616782e-02  2.35474675e-02
 -1.21659675e-01 -4.06486285e-02 -5.34570796e-02 -9.06211624e-02
 -2.32337142e+00  1.78750618e-01  3.87736236e-02  3.66324455e-02
 -3.29713885e-02 -1.08320028e-01 -4.07908242e-03 -2.69775297e-02
 -2.29992976e-02 -3.13711018e-02 -4.92665758e-02  8.06624671e-02
  9.00577176e-02 -5.77090339e-02  1.42843103e-02  6.46711152e-02
 -6.17939697e-03  5.13598001e-02 -1.97439287e-01 -4.33795091e-03
 -6.40560630e-02  8.94114848e-03 -8.70763879e-03 -1.18347839e-01
 -5.80593199e-02 -1.01852168e-02 -1.05146364e-02  1.03923228e-01
  8.57433624e-02 -1.74722194e-02  9.89866827e-02 -3.94328970e-02
  7.17072764e-02 -9.90313297e-02  2.55794727e-02 -4.63229230e-02
 -1.45581552e-02  1.04945352e-02  4.91588618e-02  3.49634933e-03
  6.89065109e-02 -1.13079510e-01 -9.55692515e-02  1.99901133e-02
  6.25565030e-03  2.11005641e-02 -2.03946733e-02 -2.71862265e-02
 -6.12519079e-02  3.08580655e-02 -5.98302994e-02  8.31249915e-02
  9.66705503e-02  5.64963924e-02 -3.19526897e-02  2.80629105e-02
  1.02855123e-01 -4.66336545e-03 -5.33676732e-02  9.66969891e-02
  8.35056406e-03  9.09357376e-02  6.48731576e-03 -3.05745394e-03
  4.34447406e-02 -2.69240073e-02  2.19264891e-02 -2.12284418e-02
 -4.80163242e-02 -7.89752091e-02  3.29369612e-02  1.58385248e-02
  4.95401897e-02  8.76077006e-02  7.06253455e-02 -1.82354550e-02]
```

```
In [101]: avg_w2v_vectors_test = []; # the avg-w2v for each sentence/review is stored in this list
          for sentence in tqdm(X_test['preprocessed_essays'].values): # for each review/sentence
              vector = np.zeros(300) # as word vectors are of zero length
              cnt_words =0; # num of words with a valid vector in the sentence/review
              for word in sentence.split(): # for each word in a review/sentence
                  if word in glove_words:
                      vector += model[word]
                      cnt_words += 1
              if cnt_words != 0:
                  vector /= cnt_words
              avg_w2v_vectors_test.append(vector)
```

```
100%|██████████████████████████████████████████████████████████████████████████| 14850/14850 [00:14<00:00, 1022.43
it/s]
```

2.3.3.2 AVG W2V Title

```
In [102]: # average Word2Vec
          # compute average word2vec for each review.
          avg_w2v_vectors_train_title = []; # the avg-w2v for each sentence/review is stored in this list
          for sentence in tqdm(X_train['preprocessed_titles'].values): # for each review/sentence
              vector = np.zeros(300) # as word vectors are of zero length
              cnt_words =0; # num of words with a valid vector in the sentence/review
              for word in sentence.split(): # for each word in a review/sentence
                  if word in glove_words:
                      vector += model[word]
                      cnt_words += 1
              if cnt_words != 0:
                  vector /= cnt_words
              avg_w2v_vectors_train_title.append(vector)

          print(len(avg_w2v_vectors_train_title))
          print(len(avg_w2v_vectors_train_title[0]))
          print(avg_w2v_vectors_train_title[0])
```

```
100%|██████████████████████████████████████████████████████████████████████████| 30150/30150 [00:01<00:00, 15614.98
it/s]

30150
300
[ 8.09192500e-02  2.25725750e-01  1.23810750e-01 -1.12325000e-01
 -2.79852500e-02 -2.62505250e-01 -2.71047500e+00  8.27350000e-02
  1.73547500e-01  1.20847500e-01 -9.43315000e-02  3.36747500e-01
 -2.45000000e-03 -2.37781250e-01 -7.59747750e-02 -2.60997000e-01
  1.23182500e-01  1.25437500e-02  2.67428000e-01 -2.25247500e-02
  1.11480000e-01 -1.80011250e-01 -2.32323250e-01 -4.68972500e-01
  1.00950000e-02 -1.21665000e-01 -2.46630000e-02 -2.09951500e-01
 -2.26317500e-01 -3.43798500e-01 -3.52667750e-02 -3.80957500e-01
 -1.86084000e-01  2.70081750e-01 -1.44945250e-01 -1.17561800e-01
  3.75482500e-01  1.62092500e-02 -2.68575000e-01 -2.89171658e-01
  8.56400000e-03  3.82320000e-01 -1.69261250e-01 -4.97175000e-02
  5.87237500e-02 -3.56960000e-02  1.10898500e-01 -1.06840000e-01
 -1.47327500e-02 -1.83545050e-01 -1.01819500e-01 -1.22627500e-01
 -9.16967250e-02 -2.82615000e-01  1.54660500e-01 -3.60950000e-01
 -2.55475250e-01 -3.31797500e-02 -2.62308000e-01 -3.35335000e-01
 -8.21320000e-02 -2.13988000e-01  2.68170850e-01 -7.27467750e-02
 -4.18037500e-01  1.36050000e-01 -1.24332100e-01  3.77004750e-02
 -1.00504500e-01  2.90213500e-01 -2.52172250e-01 -1.51427500e-02
 -2.16526500e-02 -7.96100000e-02 -9.48175000e-02 -5.60175000e-02
 -1.42457250e-01 -4.26905000e-01  2.26562500e-01  1.46657500e-02
 -3.54375000e-02 -3.24466500e-01  1.30655000e-01  1.28730000e-02
 -3.38460000e-02 -1.60330000e-02 -3.29347250e-02 -3.23240000e-02
  1.23099950e-01 -1.07882250e-01 -8.86112500e-02  2.09221500e-01
 -1.68955000e-01  3.11000000e-03 -1.67937000e-01 -9.84500000e-02
 -2.61160000e+00  4.79739250e-02  4.30740000e-02  1.70487000e-01
  2.80900000e-02 -8.55630000e-02 -1.64765500e-01 -1.49010750e-01
  2.59935000e-02  4.65755000e-02  3.08275000e-02 -1.30336000e-01
  8.29325000e-02  9.13720000e-02 -2.50340000e-02  1.58765000e-01
  2.65534000e-01  2.62134000e-01  1.14562525e-01 -2.44292500e-02
  8.62655000e-02 -7.43650000e-03  3.50975500e-02  4.30850000e-02
 -2.74980745e-01  1.75480750e-02  3.59825000e-02 -4.63547500e-02
 -6.26280250e-02 -6.80650000e-02 -1.02775000e-01 -4.57535000e-02
 -1.18027250e-01  7.14670000e-02  6.27175000e-02 -8.73900000e-02
 -1.36719000e-01  1.41647250e-01  1.14339500e-01 -1.00730750e-01
  4.00485000e-01 -2.80815000e-03  6.39505000e-02  5.25776250e-01
  1.40375250e-01  1.22438000e-01  2.05854500e-01  1.07297500e-01
 -8.76750000e-03  6.07640000e-02  9.10130000e-02 -2.07137700e-01
  1.90195000e-01 -2.02525000e-03 -5.82625000e-02  1.08279475e-01
 -1.96215000e-02 -2.09169250e-01  3.74632500e-02 -1.07132500e-01
  1.11195250e-01 -2.06273000e-01 -3.99687500e-02 -6.25425000e-03
 -1.73385000e-02  1.30416000e-01  2.76878000e-01  1.44389750e-01
  7.90350000e-02  1.69810500e-01  3.24967250e-02 -2.49205000e-02
 -3.27700000e-02 -2.94869250e-01  2.89200000e-02 -3.32725000e-02
 -6.61415000e-02 -2.47425500e-01  2.44144000e-01 -3.51977500e-02
 -2.61317500e-01 -3.39562500e-02 -1.69233000e-01 -1.78038750e-01
 -9.28350000e-02  1.75994750e-01  3.20100000e-02 -2.41027750e-01
  4.70100000e-02 -1.00090000e-01  6.25418575e-02 -3.29405000e-01
 -2.11350000e-01 -1.35214250e-01  2.36130000e-01  1.26630250e-01
  1.32543750e-01 -8.97402500e-02 -1.63727500e-02  8.40520000e-02
 -8.19550000e-02 -9.61660000e-02  2.92652250e-01 -1.63467500e-01
 -1.40985000e-01 -8.94312500e-02 -3.27874000e-01 -1.55785500e-01
 -1.18813250e-01 -1.72396000e-01  8.66525000e-02 -3.25992500e-02
  3.99301750e-01  1.31446275e-01 -5.08000000e-03  1.23748750e-01
  2.49240250e-01 -1.38302500e-01 -8.28400000e-02 -1.44137750e-01
  2.32846750e-01 -2.06520000e-01  5.68900000e-02 -1.11782500e-01
 -1.38692500e-01 -5.79640000e-02 -2.68050000e-01 -6.66887500e-02
 -2.50507500e+00  2.92281750e-01 -1.96867500e-01  1.66687500e-01
  8.17602500e-02 -1.10094750e-01  6.94025000e-02  1.72492000e-01
  1.27263750e-01 -1.33710500e-01 -1.67107500e-01  1.19310000e-01
 -9.17055750e-02 -2.05655000e-01 -4.00232500e-02  2.87644750e-01
 -2.72990000e-02  1.25467500e-01 -1.04175000e-01  8.11000000e-03
 -9.62575000e-02  1.70330750e-01  1.12170500e-01 -5.13150000e-02
  7.59440000e-02 -4.01062500e-02 -4.84118000e-01  1.18736750e-01
  2.61448750e-01  2.39340000e-01  3.31425250e-02  4.32915000e-02
 -1.72322500e-01  1.62700000e-02  1.79570000e-01  2.46457500e-01
  9.01272500e-02  5.84817500e-02  2.38062500e-01  2.48922500e-01
 -1.02975000e-02 -3.95885000e-02 -1.11612500e-01  1.30254500e-01
 -6.04825000e-02 -1.33092750e-01 -8.06890000e-02  1.27213000e-01
 -3.94580750e-01  1.79165000e-01 -1.56088455e-01  1.64944500e-01
  8.13125000e-03 -1.00271000e-01 -1.88676000e-01 -1.35252500e-02
  7.49275000e-02 -1.43927500e-01  1.65800000e-02  7.74105000e-02
 -1.42055250e-01 -1.55401750e-01  5.53625000e-03 -6.13671000e-02
 -2.28220000e-02  2.00658750e-01 -1.78083500e-01 -1.62824000e-01
  9.60645000e-02  9.08810000e-02  8.41655000e-02  1.81762500e-02
  1.63742500e-01  2.37938500e-01 -2.84600000e-03  6.00440000e-02]
```

In [103]:
```
avg_w2v_vectors_test_title = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_test['preprocessed_titles'].values): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    cnt_words =0; # num of words with a valid vector in the sentence/review
    for word in sentence.split(): # for each word in a review/sentence
        if word in glove_words:
            vector += model[word]
            cnt_words += 1
    if cnt_words != 0:
        vector /= cnt_words
    avg_w2v_vectors_test_title.append(vector)
```

```
100%|████████████████████████████████████████████████████████████████████████| 14850/14850 [00:00<00:00, 16617.24
it/s]
```

2.3.4 TF IDF W2V Essay and Title

2.3.4.1 TF IDF W2V Essay

In [104]:
```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train['preprocessed_essays'].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 [105]:
```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_train = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_train['preprocessed_essays'].values): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    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(senten
ce.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tfidf value for ea
ch 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_train.append(vector)

print(len(tfidf_w2v_vectors_train))
print(len(tfidf_w2v_vectors_train[0]))
```

```
100%|████████████████████████████████████████████████████████████████████████████| 30150/30150 [02:45<00:00, 182.38
it/s]

30150
300
```

In [106]:
```
tfidf_w2v_vectors_test = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_test['preprocessed_essays'].values): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    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(senten
ce.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tfidf value for ea
ch 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_test.append(vector)
```

```
100%|████████████████████████████████████████████████████████████████████████████| 14850/14850 [01:21<00:00, 181.19
it/s]
```

2.3.4.2 TF IDF W2V Title

In [107]:
```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
tfidf_model = TfidfVectorizer()
tfidf_model.fit(X_train['preprocessed_titles'].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 [108]:
```
# average Word2Vec
# compute average word2vec for each review.
tfidf_w2v_vectors_train_title = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_train['preprocessed_titles'].values): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    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(senten
ce.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tfidf value for ea
ch 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_train_title.append(vector)

print(len(tfidf_w2v_vectors_train_title))
print(len(tfidf_w2v_vectors_train_title[0]))
```

```
100%|████████████████████████████████████████████████████████████████████████| 30150/30150 [00:02<00:00, 11582.62
it/s]

30150
300
```

In [109]:
```
tfidf_w2v_vectors_test_title = []; # the avg-w2v for each sentence/review is stored in this list
for sentence in tqdm(X_test['preprocessed_titles'].values): # for each review/sentence
    vector = np.zeros(300) # as word vectors are of zero length
    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(senten
ce.split())))
            tf_idf = dictionary[word]*(sentence.count(word)/len(sentence.split())) # getting the tfidf value for ea
ch 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_test_title.append(vector)
```

```
100%|████████████████████████████████████████████████████████████████████████| 14850/14850 [00:01<00:00, 10696.65
it/s]
```

In [ ]:

Concatinating all the features

1. SET 1 BOW

In [110]:
```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
X_tr_BOW = hstack((X_train_essay_bow, X_train_title_bow, response_clean_teacher_prefix, response_clean_categories,
response_clean_subcategories, response_school_state, response_clean_project_grade_category, X_train_price_norm, X_t
rain_quantity_norm, X_train_TPPP_norm)).tocsr()
X_te_BOW = hstack((X_test_essay_bow, X_test_title_bow, response_test_clean_teacher_prefix, response_test_clean_cate
gories, response_test_clean_subcategories, response_test_school_state, response_test_clean_project_grade_category,
X_test_price_norm, X_test_quantity_norm, X_test_TPPP_norm)).tocsr()

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

```
Final Data matrix
(30150, 6525) (30150,)
(14850, 6525) (14850,)
====================================================================================================
```

2. SET 2 TF IDF

In [111]:
```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
X_tr_TFIDF = hstack((X_train_essay_tfidf, X_train_title_tfidf, response_clean_teacher_prefix, response_clean_catego
ries, response_clean_subcategories, response_school_state, response_clean_project_grade_category, X_train_price_nor
m, X_train_quantity_norm, X_train_TPPP_norm)).tocsr()
X_te_TFIDF = hstack((X_test_essay_tfidf, X_test_title_tfidf, response_test_clean_teacher_prefix, response_test_clea
n_categories, response_test_clean_subcategories, response_test_school_state, response_test_clean_project_grade_cate
gory, X_test_price_norm, X_test_quantity_norm, X_test_TPPP_norm)).tocsr()

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

```
Final Data matrix
(30150, 6525) (30150,)
(14850, 6525) (14850,)
====================================================================================================
```

3. SET 3 AVG W2V

In [112]:
```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
X_tr_AVG_W2V = hstack((avg_w2v_vectors_train, avg_w2v_vectors_train_title, response_clean_teacher_prefix, response_
clean_categories, response_clean_subcategories, response_school_state, response_clean_project_grade_category, X_tra
in_price_norm, X_train_quantity_norm, X_train_TPPP_norm)).tocsr()
X_te_AVG_W2V = hstack((avg_w2v_vectors_test, avg_w2v_vectors_test_title, response_test_clean_teacher_prefix, respon
se_test_clean_categories, response_test_clean_subcategories, response_test_school_state, response_test_clean_projec
t_grade_category, X_test_price_norm, X_test_quantity_norm, X_test_TPPP_norm)).tocsr()

print("Final Data matrix")
print(X_tr_AVG_W2V.shape, y_train.shape)
print(X_te_AVG_W2V.shape, y_test.shape)
print("="*100)
```

```
Final Data matrix
(30150, 613) (30150,)
(14850, 613) (14850,)
====================================================================================================
```

4. SET 4 TF IDF W2V

In [113]:
```
# merge two sparse matrices: https://stackoverflow.com/a/19710648/4084039
from scipy.sparse import hstack
X_tr_TFIDF_W2V = hstack((tfidf_w2v_vectors_train, tfidf_w2v_vectors_train_title,response_clean_teacher_prefix, resp
onse_clean_categories, response_clean_subcategories, response_school_state, response_clean_project_grade_category,
X_train_price_norm, X_train_quantity_norm, X_train_TPPP_norm)).tocsr()
X_te_TFIDF_W2V = hstack((tfidf_w2v_vectors_test, tfidf_w2v_vectors_test_title, response_test_clean_teacher_prefix,
response_test_clean_categories, response_test_clean_subcategories, response_test_school_state, response_test_clean_
project_grade_category, X_test_price_norm, X_test_quantity_norm, X_test_TPPP_norm)).tocsr()

print("Final Data matrix")
print(X_tr_TFIDF_W2V.shape, y_train.shape)
print(X_te_TFIDF_W2V.shape, y_test.shape)
print("="*100)
```

```
Final Data matrix
(30150, 613) (30150,)
(14850, 613) (14850,)
====================================================================================================
```

In [ ]:

2.4 Applying Random Forest

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

2.4.1 Applying Random Forests on BOW, SET 1

In [114]:
```
# Please write all the code with proper documentation
```

In [115]:
```
import warnings
warnings.filterwarnings('ignore')
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt
#from sklearn.grid_search import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import learning_curve, GridSearchCV

#n_estimators = [10, 50, 100, 150, 200, 300, 500, 1000], max_depth = [2, 3, 4, 5, 6, 7, 8, 9, 10]

clf = RandomForestClassifier(class_weight='balanced')
parameters ={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]}
set1 =GridSearchCV(clf, parameters, cv=5, scoring='roc_auc',return_train_score=True)
set1.fit(X_tr_BOW, y_train)
```

Out[115]:
```
GridSearchCV(cv=5, error_score='raise',
       estimator=RandomForestClassifier(bootstrap=True, class_weight='balanced',
            criterion='gini', max_depth=None, max_features='auto',
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=10, n_jobs=1, oob_score=False, random_state=None,
            verbose=0, warm_start=False),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
       scoring='roc_auc', verbose=0)
```

In [116]:
```
print(set1.cv_results_)
```

```
{'mean_fit_time': array([ 0.30039654,  0.76196179,  1.16388173,  1.65357752,  2.11833329,
        3.03269095,  5.13307285,  0.32014251,  0.83456836,  1.37073483,
        1.93741269,  2.67723908,  3.66857982,  6.78026919,  0.34447908,
        0.93489819,  1.65238781,  2.35250893,  3.11966391,  4.54365506,
        7.66848507,  0.39594169,  1.08011932,  2.00563493,  2.98980455,
        3.75256476,  6.13100371, 10.2737278 ,  0.42626829,  1.2666254 ,
        2.4989182 ,  3.61453395,  5.47934685,  7.45108247, 11.59040475,
        0.49188375,  1.77365808,  2.94013567,  4.23926253,  5.65127959,
        8.34588046, 13.91457443,  0.54893265,  2.05709791,  3.52457409,
        5.15321312,  6.79782233, 10.33734798, 16.7803257 ]), 'std_fit_time': array([0.01544335, 0.09463311, 0.0121
4413, 0.01040326, 0.04945349,
       0.02188057, 0.44005595, 0.01291144, 0.06704364, 0.02319137,
       0.02294358, 0.28697989, 0.05286355, 1.03527446, 0.00794287,
       0.02064697, 0.02722317, 0.01676522, 0.0458745 , 0.04600046,
       0.42041819, 0.02389264, 0.02297892, 0.03882294, 0.18203374,
       0.04311213, 0.6611406 , 0.75031686, 0.00622063, 0.01130101,
       0.19263915, 0.14951075, 0.39591793, 0.79206606, 0.56643555,
       0.01573947, 0.21517587, 0.02626613, 0.06511739, 0.06147806,
       0.10251301, 0.39722547, 0.02940678, 0.21438413, 0.16910524,
       0.07485575, 0.09816269, 0.28602237, 0.07702868]), 'mean_score_time': array([0.07839932, 0.23517275, 0.38377
967, 0.56010218, 0.7426373 ,
       1.10145521, 1.87518606, 0.07360382, 0.2451509 , 0.41470599,
       0.58205199, 0.84634461, 1.10823836, 1.9647665 , 0.07480011,
       0.21702118, 0.40831704, 0.58665013, 0.76336074, 1.12540669,
       1.96096306, 0.07680798, 0.24376206, 0.40652876, 0.61118116,
       0.77294283, 1.15851083, 2.13589387, 0.08078456, 0.21962142,
       0.39574938, 0.59741845, 0.84773355, 1.1966095 , 1.93602996,
       0.07958989, 0.2760632 , 0.41410851, 0.59961176, 0.78271523,
       1.1936101 , 1.90971589, 0.07561226, 0.26928129, 0.42446604,
       0.60538855, 0.79068618, 1.16848536, 2.0028511 ]), 'std_score_time': array([0.01078722, 0.01946526, 0.008151
53, 0.01581277, 0.01555876,
       0.01049177, 0.14368557, 0.00222379, 0.03250394, 0.00559259,
       0.00796629, 0.08992769, 0.00424927, 0.20665757, 0.00260214,
       0.00376453, 0.01175971, 0.01117457, 0.00925067, 0.00866283,
       0.17616381, 0.00554566, 0.00361278, 0.0143411 , 0.04366763,
       0.00790213, 0.04642407, 0.31901821, 0.00567733, 0.01103875,
       0.01072507, 0.01199905, 0.06230652, 0.07650343, 0.11972388,
       0.00498334, 0.040904  , 0.01236756, 0.01326521, 0.012801  ,
       0.05993139, 0.009699  , 0.00230998, 0.03115082, 0.01744232,
       0.01553268, 0.01493261, 0.00739234, 0.19816856]), 'param_max_depth': masked_array(data=[2, 2, 2, 2, 2, 2,
2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4,
                   4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7,
                   7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'param_n_estimators': masked_array(data=[10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
150, 200,
                   300, 500, 10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
                   150, 200, 300, 500, 10, 50, 100, 150, 200, 300, 500,
                   10, 50, 100, 150, 200, 300, 500, 10, 50, 100, 150, 200,
                   300, 500],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'params': [{'max_depth': 2, 'n_estimators': 10}, {'max_depth': 2, 'n_estimators': 50},
{'max_depth': 2, 'n_estimators': 100}, {'max_depth': 2, 'n_estimators': 150}, {'max_depth': 2, 'n_estimators': 20
0}, {'max_depth': 2, 'n_estimators': 300}, {'max_depth': 2, 'n_estimators': 500}, {'max_depth': 3, 'n_estimators':
10}, {'max_depth': 3, 'n_estimators': 50}, {'max_depth': 3, 'n_estimators': 100}, {'max_depth': 3, 'n_estimators':
150}, {'max_depth': 3, 'n_estimators': 200}, {'max_depth': 3, 'n_estimators': 300}, {'max_depth': 3, 'n_estimator
s': 500}, {'max_depth': 4, 'n_estimators': 10}, {'max_depth': 4, 'n_estimators': 50}, {'max_depth': 4, 'n_estimato
rs': 100}, {'max_depth': 4, 'n_estimators': 150}, {'max_depth': 4, 'n_estimators': 200}, {'max_depth': 4, 'n_estim
ators': 300}, {'max_depth': 4, 'n_estimators': 500}, {'max_depth': 5, 'n_estimators': 10}, {'max_depth': 5, 'n_est
imators': 50}, {'max_depth': 5, 'n_estimators': 100}, {'max_depth': 5, 'n_estimators': 150}, {'max_depth': 5, 'n_e
stimators': 200}, {'max_depth': 5, 'n_estimators': 300}, {'max_depth': 5, 'n_estimators': 500}, {'max_depth': 6,
'n_estimators': 10}, {'max_depth': 6, 'n_estimators': 50}, {'max_depth': 6, 'n_estimators': 100}, {'max_depth': 6,
'n_estimators': 150}, {'max_depth': 6, 'n_estimators': 200}, {'max_depth': 6, 'n_estimators': 300}, {'max_depth':
6, 'n_estimators': 500}, {'max_depth': 7, 'n_estimators': 10}, {'max_depth': 7, 'n_estimators': 50}, {'max_depth':
7, 'n_estimators': 100}, {'max_depth': 7, 'n_estimators': 150}, {'max_depth': 7, 'n_estimators': 200}, {'max_dept
h': 7, 'n_estimators': 300}, {'max_depth': 7, 'n_estimators': 500}, {'max_depth': 8, 'n_estimators': 10}, {'max_de
pth': 8, 'n_estimators': 50}, {'max_depth': 8, 'n_estimators': 100}, {'max_depth': 8, 'n_estimators': 150}, {'max_
depth': 8, 'n_estimators': 200}, {'max_depth': 8, 'n_estimators': 300}, {'max_depth': 8, 'n_estimators': 500}], 's
plit0_test_score': array([0.58237942, 0.65504505, 0.67011174, 0.66637516, 0.68577045,
       0.68434675, 0.6829028 , 0.62614051, 0.64818963, 0.68131959,
       0.69144796, 0.68824991, 0.69318792, 0.69597351, 0.63297936,
       0.67340611, 0.68587974, 0.6917311 , 0.67949672, 0.69093696,
       0.69870635, 0.62393048, 0.68201478, 0.68732391, 0.69299487,
       0.69080405, 0.69916354, 0.69791137, 0.64480687, 0.67899099,
       0.6965383 , 0.69535786, 0.6950973 , 0.70094634, 0.70187697,
       0.64313526, 0.68601224, 0.69970956, 0.70374015, 0.70026887,
       0.69734678, 0.70076574, 0.62825729, 0.69358689, 0.69608512,
       0.70168034, 0.70029483, 0.70763317, 0.70547441]), 'split1_test_score': array([0.57461474, 0.65424296, 0.672
93671, 0.66336578, 0.67474201,
       0.67614763, 0.67186894, 0.63112997, 0.68552157, 0.66904039,
       0.67593935, 0.66059104, 0.68872352, 0.69112707, 0.62446672,
       0.68176843, 0.66881945, 0.69729231, 0.68837723, 0.69221763,
       0.68982569, 0.61533957, 0.67455568, 0.70736737, 0.67627825,
       0.68548876, 0.69906933, 0.6900939 , 0.64198075, 0.66426854,
       0.66943099, 0.69460451, 0.68176991, 0.68904765, 0.69095994,
       0.67085212, 0.67535925, 0.69329512, 0.68954567, 0.69614751,
       0.69896614, 0.70159949, 0.64204859, 0.68308584, 0.69543678,
       0.69496557, 0.71237346, 0.70191138, 0.70458926]), 'split2_test_score': array([0.6151851 , 0.66553098, 0.691
99184, 0.68199   , 0.70051905,
       0.69328267, 0.70098921, 0.6473632 , 0.68847515, 0.68683909,
       0.69227567, 0.69830521, 0.7016955 , 0.70162186, 0.61735885,
       0.68439925, 0.69266099, 0.69715999, 0.69908304, 0.7107374 ,
       0.70742223, 0.64473175, 0.69350804, 0.70545614, 0.70666804,
       0.71699738, 0.71589796, 0.70843472, 0.64523525, 0.69009369,
       0.69297204, 0.70584822, 0.71035608, 0.71701469, 0.71411207,
       0.66999769, 0.69697282, 0.70824037, 0.71081295, 0.71102249,
       0.72053877, 0.72263464, 0.65105842, 0.7089378 , 0.70947359,
       0.70501721, 0.71266129, 0.71607247, 0.72436967]), 'split3_test_score': array([0.56888347, 0.68961066, 0.678
4406 , 0.69153139, 0.68983635,
       0.70038526, 0.69555548, 0.6085974 , 0.68795361, 0.69758584,
       0.69917948, 0.69106007, 0.69084683, 0.69935969, 0.65217135,
       0.67862725, 0.69838065, 0.70478973, 0.69445574, 0.69855696,
       0.70749145, 0.61334593, 0.67415948, 0.69916809, 0.71153021,
       0.70444143, 0.70588831, 0.70902854, 0.66013724, 0.68970636,
       0.70112469, 0.7013336 , 0.71348597, 0.70847545, 0.71146395,
       0.63410992, 0.69842877, 0.6971104 , 0.71581545, 0.71337202,
       0.7138293 , 0.7118516 , 0.65130858, 0.70033567, 0.70921445,
       0.71723394, 0.71999116, 0.71563376, 0.71101827]), 'split4_test_score': array([0.59707   , 0.68917595, 0.674
23498, 0.68943526, 0.68014871,
       0.6957603 , 0.68005027, 0.64957186, 0.6965265 , 0.68643382,
       0.68517435, 0.6884502 , 0.69991284, 0.69463412, 0.63806655,
       0.67498449, 0.70039787, 0.70024366, 0.69041187, 0.70580756,
       0.69974384, 0.6344189 , 0.6912461 , 0.69512908, 0.70244488,
       0.70342677, 0.71070264, 0.70505128, 0.65789606, 0.68806108,
       0.71732977, 0.71436086, 0.71151131, 0.70662594, 0.70593135,
       0.66397774, 0.69907502, 0.70720624, 0.70422994, 0.71523142,
       0.71240257, 0.71758453, 0.64944944, 0.71151976, 0.70949852,
       0.70796041, 0.71971119, 0.71771297, 0.72045014]), 'mean_test_score': array([0.58762606, 0.67071999, 0.67754
304, 0.67853875, 0.6862035 ,
       0.68998414, 0.68627343, 0.63255981, 0.68133169, 0.68424358,
       0.68880357, 0.68533128, 0.6948731 , 0.69654329, 0.6330084 ,
       0.67863705, 0.68922726, 0.69824307, 0.69036456, 0.69965081,
       0.70063788, 0.62635298, 0.68309651, 0.69888866, 0.69798294,
       0.70023126, 0.70614397, 0.70210373, 0.6500108 , 0.68222383,
       0.69547847, 0.70230038, 0.70244357, 0.70442182, 0.70486872,
       0.65641386, 0.69116919, 0.70111209, 0.70482881, 0.70720797,
       0.70861621, 0.71088664, 0.64442376, 0.6994926 , 0.70394125,
       0.70537129, 0.71300574, 0.71179242, 0.71317985]), 'std_test_score': array([0.01671398, 0.01575809, 0.007706
7 , 0.01164157, 0.00879023,
       0.00866783, 0.01058569, 0.01500639, 0.0169815 , 0.00926539,
       0.00781494, 0.01289548, 0.0050756 , 0.00366138, 0.01192199,
       0.00409182, 0.01138441, 0.00427435, 0.00655486, 0.00765773,
       0.00654878, 0.01182519, 0.00810915, 0.00725253, 0.01244864,
       0.01109294, 0.00655391, 0.00719077, 0.00747108, 0.0098472 ,
       0.01546064, 0.00729912, 0.01223217, 0.00925544, 0.00815441,
       0.01499335, 0.00922419, 0.00577981, 0.00885025, 0.00758161,
       0.00898669, 0.00863136, 0.00876028, 0.01037898, 0.00668351,
       0.00734104, 0.00715352, 0.00605077, 0.00794877]), 'rank_test_score': array([49, 42, 41, 40, 33, 29, 32, 47,
38, 35, 31, 34, 26, 24, 46, 39, 30,
       22, 28, 19, 17, 48, 36, 21, 23, 18,  7, 15, 44, 37, 25, 14, 13, 11,
        9, 43, 27, 16, 10,  6,  5,  4, 45, 20, 12,  8,  2,  3,  1]), 'split0_train_score': array([0.61063298, 0.68
568082, 0.70710099, 0.7134753 , 0.72957268,
       0.73082774, 0.71982317, 0.64506074, 0.70499323, 0.73504981,
       0.73642614, 0.73535155, 0.74578915, 0.74866442, 0.66898936,
       0.73947778, 0.74031541, 0.75851967, 0.74181156, 0.75215325,
       0.76363423, 0.68896402, 0.75702892, 0.76427522, 0.76828077,
       0.7712031 , 0.78016388, 0.77612007, 0.71741321, 0.77517386,
       0.78914935, 0.78962952, 0.78832924, 0.79917645, 0.79972507,
       0.73025324, 0.79213589, 0.8068392 , 0.81351553, 0.81007572,
       0.80890845, 0.81233177, 0.72904998, 0.80797205, 0.83742839,
       0.83277557, 0.83667881, 0.84307698, 0.84024205]), 'split1_train_score': array([0.60594531, 0.71721729, 0.71
737011, 0.70973782, 0.71597721,
       0.72580396, 0.71985085, 0.65418753, 0.73191496, 0.72206015,
       0.73349837, 0.7223375 , 0.74483584, 0.74560841, 0.67057901,
       0.74765952, 0.73753901, 0.76017911, 0.75374631, 0.76042648,
       0.7587017 , 0.68601167, 0.7526542 , 0.78366576, 0.76646151,
       0.7686037 , 0.78104612, 0.78196193, 0.70910204, 0.76374945,
       0.76472328, 0.79641863, 0.78364421, 0.79219439, 0.79575506,
       0.73938554, 0.78780702, 0.81264327, 0.80982009, 0.81818144,
       0.81674533, 0.82327162, 0.74754803, 0.81403003, 0.83229004,
       0.83536258, 0.84826925, 0.83912574, 0.84372332]), 'split2_train_score': array([0.62151625, 0.6909055 , 0.71
276295, 0.69939357, 0.72608048,
       0.71931899, 0.72573164, 0.6558565 , 0.71704863, 0.71865885,
       0.72235679, 0.73208655, 0.72716787, 0.73663301, 0.65425533,
       0.74135952, 0.74495147, 0.74885555, 0.75271005, 0.76102938,
       0.75384447, 0.6938198 , 0.75292968, 0.77094886, 0.77785319,
       0.78279534, 0.77862617, 0.77086339, 0.71200799, 0.76550354,
       0.77766344, 0.78984261, 0.79257847, 0.79995814, 0.79526997,
       0.744249  , 0.79548929, 0.80078454, 0.80709369, 0.81164472,
       0.82138659, 0.82507015, 0.74752612, 0.82355801, 0.83526133,
       0.82368585, 0.82954003, 0.83593019, 0.84266518]), 'split3_train_score': array([0.58805423, 0.7151461 , 0.70
91642 , 0.71942282, 0.71596671,
       0.72823369, 0.72066075, 0.64099231, 0.72315823, 0.73829481,
       0.73675321, 0.72878874, 0.73165192, 0.74273115, 0.68266332,
       0.72883714, 0.75597696, 0.75950451, 0.75200723, 0.75386468,
       0.76239286, 0.67781549, 0.74489106, 0.77102428, 0.77562941,
       0.77358074, 0.77672026, 0.78226008, 0.71781751, 0.77653416,
       0.78220196, 0.79030041, 0.79700108, 0.7979946 , 0.80238995,
       0.71720291, 0.79528829, 0.78973512, 0.81730772, 0.81341961,
       0.82053797, 0.81760038, 0.73774689, 0.81169747, 0.8311764 ,
       0.84537245, 0.83858753, 0.83578499, 0.83643461]), 'split4_train_score': array([0.61042075, 0.70462791, 0.70
366281, 0.70952046, 0.70517572,
       0.72128471, 0.70609934, 0.66703567, 0.72927049, 0.71992543,
       0.72311515, 0.71958851, 0.7364624 , 0.73079756, 0.68000087,
       0.72105592, 0.75481306, 0.75152687, 0.73447793, 0.75831945,
       0.75262189, 0.68214847, 0.75100883, 0.7559404 , 0.76705072,
       0.7664622 , 0.77319261, 0.77179282, 0.7094158 , 0.7716319 ,
       0.79229329, 0.79361691, 0.79245902, 0.791356  , 0.78825569,
       0.72660307, 0.79192138, 0.80097113, 0.79976389, 0.80979669,
       0.81379553, 0.81853584, 0.7463991 , 0.81655893, 0.82574085,
       0.83369267, 0.83200726, 0.84105192, 0.84352518]), 'mean_train_score': array([0.6073139 , 0.70271552, 0.7100
1221, 0.71030999, 0.71855456,
       0.72509382, 0.71843315, 0.65262655, 0.72127711, 0.72679781,
       0.73042993, 0.72763057, 0.73718144, 0.74088691, 0.67129758,
       0.73567797, 0.74671918, 0.75571714, 0.74695062, 0.75715865,
       0.75823903, 0.68575189, 0.75170254, 0.7691709 , 0.77105512,
       0.77252902, 0.77794981, 0.77659966, 0.71315131, 0.77051858,
       0.78120627, 0.79196161, 0.7908024 , 0.79613592, 0.79627915,
       0.73153875, 0.79252837, 0.80219465, 0.80950018, 0.81262364,
       0.81627478, 0.81936195, 0.74165402, 0.8147633 , 0.8323794 ,
       0.83417782, 0.83701657, 0.83899396, 0.84131807]), 'std_train_score': array([0.01091039, 0.01263436, 0.00471
624, 0.00652965, 0.00860664,
       0.00426847, 0.00654668, 0.00908943, 0.00962955, 0.00819993,
       0.00638804, 0.00589071, 0.007267  , 0.00641988, 0.01001209,
       0.00949608, 0.00747832, 0.00462055, 0.00757495, 0.00354719,
       0.00441454, 0.00550359, 0.00394105, 0.00911823, 0.00473226,
       0.00566526, 0.00279605, 0.00483807, 0.00378405, 0.00510039,
       0.00970821, 0.00265682, 0.00450952, 0.0036247 , 0.00479344,
       0.00953575, 0.00279992, 0.00760882, 0.00596271, 0.00306478,
       0.00458118, 0.00449618, 0.00729197, 0.00522742, 0.00398305,
       0.00691385, 0.00648377, 0.00284983, 0.00273778])}
```

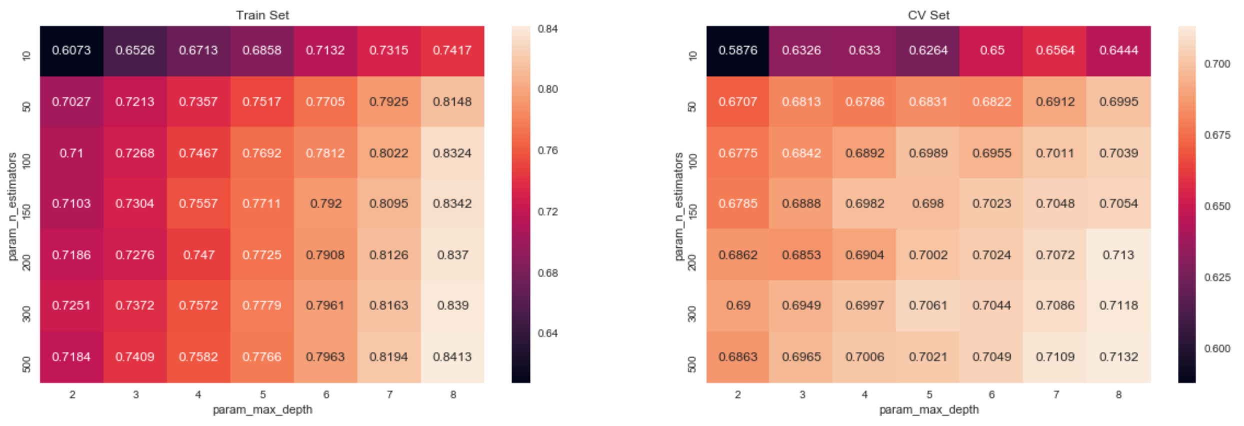
In [117]:
```
import seaborn as sns; sns.set()
max_scores1 = pd.DataFrame(set1.cv_results_).groupby(['param_n_estimators', 'param_max_depth']).max().unstack()
max_scores1
```

Out[117]:

	mean_fit_time							mean_score_time			...	std_test_scor	
param_max_depth	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**2**	**3**	**4**	**...**	**6**	**7**
param_n_estimators													
10	0.300397	0.320143	0.344479	0.395942	0.426268	0.491884	0.548933	0.078399	0.073604	0.074800	...	0.007471	0.0
50	0.761962	0.834568	0.934898	1.080119	1.266625	1.773658	2.057098	0.235173	0.245151	0.217021	...	0.009847	0.0
100	1.163882	1.370735	1.652388	2.005635	2.498918	2.940136	3.524574	0.383780	0.414706	0.408317	...	0.015461	0.0
150	1.653578	1.937413	2.352509	2.989805	3.614534	4.239263	5.153213	0.560102	0.582052	0.586650	...	0.007299	0.0
200	2.118333	2.677239	3.119664	3.752565	5.479347	5.651280	6.797822	0.742637	0.846345	0.763361	...	0.012232	0.0
300	3.032691	3.668580	4.543655	6.131004	7.451082	8.345880	10.337348	1.101455	1.108238	1.125407	...	0.009255	0.0
500	5.133073	6.780269	7.668485	10.273728	11.590405	13.914574	16.780326	1.875186	1.964767	1.960963	...	0.008154	0.0

7 rows × 140 columns

```
In [118]: fig, ax = plt.subplots(1,2, figsize=(20,6))
          sns.heatmap(max_scores1.mean_train_score, annot = True, fmt='.4g', ax=ax[0])
          sns.heatmap(max_scores1.mean_test_score, annot = True, fmt='.4g', ax=ax[1])
          ax[0].set_title('Train Set')
          ax[1].set_title('CV Set')
          plt.show()
```



```
In [119]: print(set1.best_estimator_)

RandomForestClassifier(bootstrap=True, class_weight='balanced',
            criterion='gini', max_depth=8, max_features='auto',
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=500, n_jobs=1, oob_score=False, random_state=None,
            verbose=0, warm_start=False)
```

Training our model with best Hyperparameters

```
In [123]: def pred_prob(clf, data):
              y_pred = []
              y_pred = clf.predict_proba(data)[:,1]
              return y_pred
```

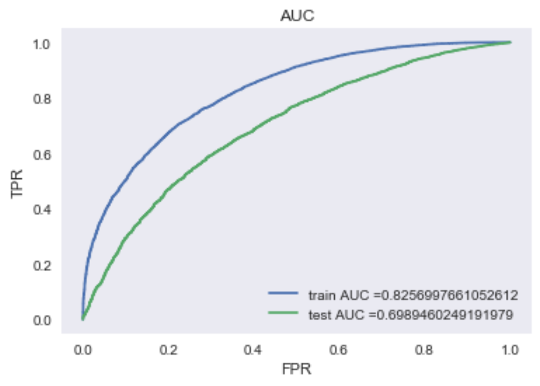
```
In [124]: # https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
          from sklearn.metrics import roc_curve, auc
          model = RandomForestClassifier(max_depth = 8, n_estimators = 500)

          model.fit(X_tr_BOW, y_train)

          y_train_pred = pred_prob(model,X_tr_BOW)
          y_test_pred = pred_prob(model,X_te_BOW)

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

          plt.close
          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("FPR")
          plt.ylabel("TPR")
          plt.title("AUC")
          plt.grid()
          plt.show()
```



Confusion Matrix

In [128]:
```
# we are writing our own function for predict, with defined threshold
# we will pick a threshold that will give the least fpr
def find_best_threshold(threshold, fpr, tpr):
    t = threshold[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, threshold):
    predictions = []
    for i in proba:
        if i>=threshold:
            predictions.append(1)
        else:
            predictions.append(0)
    return predictions
```

In [129]:
```
#our objective here is to make auc the maximum
#so we find  the best threshold that will give the least fpr
best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")
print(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)))
```

```
the maximum value of tpr*(1-fpr) 0.5474011181642218 for threshold 0.842
Train confusion matrix
[[ 3603  1041]
 [ 7510 17996]]
```

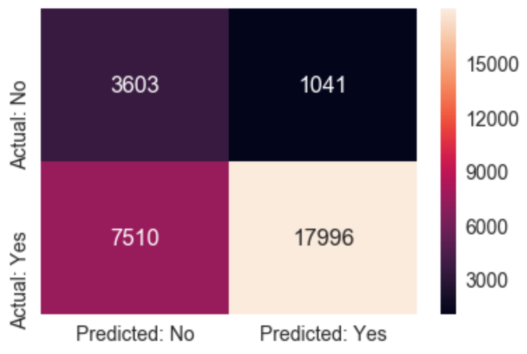
In [130]:
```
#plotting confusion matrix using seaborn's heatmap
# https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix

print("Train data confusion matrix")

confusion_matrix_df_train = pd.DataFrame(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)), ['Ac
tual: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Train data confusion matrix
```

Out[130]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c385fae588>
```



In [131]:
```
print("Test confusion matrix")
print(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)))
```

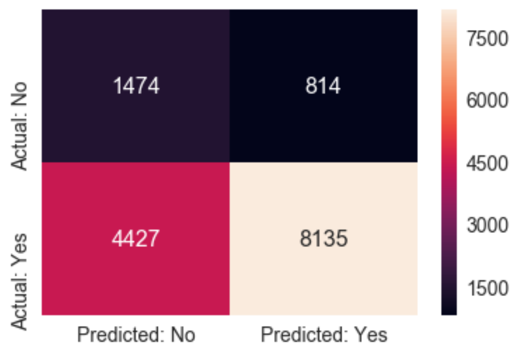
```
Test confusion matrix
[[1474  814]
 [4427 8135]]
```

In [132]:
```
print("Test data confusion matrix")

confusion_matrix_df_test = pd.DataFrame(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)), ['Actua
l: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_test, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Test data confusion matrix
```

Out[132]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c387735b38>
```



In [ ]:

2.4.2 Applying Random Forests on TFIDF, SET 2

In [120]:
```
# Please write all the code with proper documentation
```

In [133]:
```
import warnings
warnings.filterwarnings('ignore')
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt
#from sklearn.grid_search import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import learning_curve, GridSearchCV

#n_estimators = [10, 50, 100, 150, 200, 300, 500, 1000], max_depth = [2, 3, 4, 5, 6, 7, 8, 9, 10]

clf = RandomForestClassifier(class_weight='balanced')
parameters ={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]}
set2 = GridSearchCV(clf, parameters, cv=5, scoring='roc_auc',return_train_score=True)
set2.fit(X_tr_TFIDF, y_train)
```

Out[133]:
```
GridSearchCV(cv=5, error_score='raise',
       estimator=RandomForestClassifier(bootstrap=True, class_weight='balanced',
            criterion='gini', max_depth=None, max_features='auto',
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=10, n_jobs=1, oob_score=False, random_state=None,
            verbose=0, warm_start=False),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
       scoring='roc_auc', verbose=0)
```

In [134]:
```
print(set2.cv_results_)
```

```
{'mean_fit_time': array([ 0.4076952 ,  1.16727767,  2.33196354,  2.99558949,  4.33202219,
        6.12521901,  8.74461374,  0.41348557,  1.26003013,  2.41334519,
        3.40130334,  4.52350416,  6.89954138, 10.94352593,  0.47871876,
        1.59712043,  3.07917299,  4.48262038,  5.70853405,  8.80983176,
       14.35780282,  0.53516898,  1.92305918,  3.59339061,  5.25654263,
        6.9629869 , 10.73708758, 17.41343155,  0.59839835,  2.28629332,
        4.34577675,  6.44795704,  8.79827929, 12.69524107, 21.16878352,
        0.71409078,  2.73548579,  5.46937947,  7.78020101, 10.12671895,
       15.4546618 , 25.24667745,  0.80145645,  3.15077386,  6.1425745 ,
        9.15171776, 12.40321641, 18.29309244, 30.39770827]), 'std_fit_time': array([0.03426796, 0.10991101, 0.4289
0565, 0.32914913, 0.96496716,
       0.39070317, 0.98083958, 0.00865933, 0.03619819, 0.13269056,
       0.02942185, 0.06110721, 0.30598282, 0.06592688, 0.01072289,
       0.04126672, 0.27238769, 0.29023146, 0.09511772, 0.67350047,
       0.73882345, 0.01824359, 0.01835603, 0.0452354 , 0.07094357,
       0.03529302, 0.64508143, 0.20238695, 0.01526869, 0.04073326,
       0.05727095, 0.07431261, 0.56928563, 0.11789221, 0.29433771,
       0.02821585, 0.04611296, 0.63831395, 0.24261047, 0.09772049,
       0.3099549 , 0.28404538, 0.02695052, 0.03359468, 0.07774381,
       0.14437869, 0.80986628, 0.45156134, 0.70846613]), 'mean_score_time': array([0.08577895, 0.23397527, 0.48689
885, 0.59521046, 0.83756695,
       1.31388655, 1.91907697, 0.07360592, 0.21383619, 0.39116917,
       0.56488934, 0.73864627, 1.16848254, 1.78823404, 0.0728066 ,
       0.21043925, 0.39952569, 0.57388611, 0.75897889, 1.10804491,
       1.82232957, 0.07640367, 0.21502481, 0.38358307, 0.5764607 ,
       0.75437703, 1.23828921, 1.84148369, 0.08158398, 0.2162137 ,
       0.40013971, 0.58205853, 0.80305362, 1.12839026, 1.84648285,
       0.07580528, 0.23039155, 0.42546425, 0.59720473, 0.79788871,
       1.1485425 , 1.99766521, 0.07460899, 0.21843281, 0.4186738 ,
       0.60320325, 0.78670444, 1.18283672, 1.93124361]), 'std_score_time': array([0.01157176, 0.01626087, 0.061260
88, 0.03529074, 0.17191734,
       0.10463615, 0.22980543, 0.00353423, 0.01205511, 0.02066354,
       0.01307506, 0.01609905, 0.12901427, 0.02237808, 0.00165247,
       0.00333691, 0.01847885, 0.01654094, 0.00535666, 0.00966887,
       0.03971036, 0.00868121, 0.00586762, 0.00430178, 0.01712395,
       0.01326911, 0.19766863, 0.02038736, 0.01459346, 0.00395011,
       0.01537596, 0.01490048, 0.05482295, 0.00232832, 0.0201999 ,
       0.00269176, 0.01502625, 0.04319615, 0.0134722 , 0.06676378,
       0.00204233, 0.2230724 , 0.00116982, 0.00166821, 0.01713371,
       0.01350076, 0.01469486, 0.01866192, 0.01305054]), 'param_max_depth': masked_array(data=[2, 2, 2, 2, 2, 2,
2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4,
                   4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7,
                   7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'param_n_estimators': masked_array(data=[10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
150, 200,
                   300, 500, 10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
                   150, 200, 300, 500, 10, 50, 100, 150, 200, 300, 500,
                   10, 50, 100, 150, 200, 300, 500, 10, 50, 100, 150, 200,
                   300, 500],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'params': [{'max_depth': 2, 'n_estimators': 10}, {'max_depth': 2, 'n_estimators': 50},
{'max_depth': 2, 'n_estimators': 100}, {'max_depth': 2, 'n_estimators': 150}, {'max_depth': 2, 'n_estimators': 20
0}, {'max_depth': 2, 'n_estimators': 300}, {'max_depth': 2, 'n_estimators': 500}, {'max_depth': 3, 'n_estimators':
10}, {'max_depth': 3, 'n_estimators': 50}, {'max_depth': 3, 'n_estimators': 100}, {'max_depth': 3, 'n_estimators':
150}, {'max_depth': 3, 'n_estimators': 200}, {'max_depth': 3, 'n_estimators': 300}, {'max_depth': 3, 'n_estimator
s': 500}, {'max_depth': 4, 'n_estimators': 10}, {'max_depth': 4, 'n_estimators': 50}, {'max_depth': 4, 'n_estimato
rs': 100}, {'max_depth': 4, 'n_estimators': 150}, {'max_depth': 4, 'n_estimators': 200}, {'max_depth': 4, 'n_estim
ators': 300}, {'max_depth': 4, 'n_estimators': 500}, {'max_depth': 5, 'n_estimators': 10}, {'max_depth': 5, 'n_est
imators': 50}, {'max_depth': 5, 'n_estimators': 100}, {'max_depth': 5, 'n_estimators': 150}, {'max_depth': 5, 'n_e
stimators': 200}, {'max_depth': 5, 'n_estimators': 300}, {'max_depth': 5, 'n_estimators': 500}, {'max_depth': 6,
'n_estimators': 10}, {'max_depth': 6, 'n_estimators': 50}, {'max_depth': 6, 'n_estimators': 100}, {'max_depth': 6,
'n_estimators': 150}, {'max_depth': 6, 'n_estimators': 200}, {'max_depth': 6, 'n_estimators': 300}, {'max_depth':
6, 'n_estimators': 500}, {'max_depth': 7, 'n_estimators': 10}, {'max_depth': 7, 'n_estimators': 50}, {'max_depth':
7, 'n_estimators': 100}, {'max_depth': 7, 'n_estimators': 150}, {'max_depth': 7, 'n_estimators': 200}, {'max_dept
h': 7, 'n_estimators': 300}, {'max_depth': 7, 'n_estimators': 500}, {'max_depth': 8, 'n_estimators': 10}, {'max_de
pth': 8, 'n_estimators': 50}, {'max_depth': 8, 'n_estimators': 100}, {'max_depth': 8, 'n_estimators': 150}, {'max_
depth': 8, 'n_estimators': 200}, {'max_depth': 8, 'n_estimators': 300}, {'max_depth': 8, 'n_estimators': 500}], 's
plit0_test_score': array([0.60709112, 0.65996608, 0.65693607, 0.68059994, 0.68070058,
       0.68307791, 0.69690689, 0.62285036, 0.67148492, 0.69363183,
       0.6822842 , 0.69013228, 0.69554796, 0.69453335, 0.6197751 ,
       0.67387618, 0.70328844, 0.69725606, 0.69188554, 0.70285403,
       0.7049784 , 0.66059037, 0.69065108, 0.69690562, 0.70197297,
          0.70596073, 0.69216023, 0.70327325, 0.64122915, 0.66919197,
          0.69564301, 0.70034061, 0.70562506, 0.70379711, 0.70655422,
          0.62901313, 0.67718563, 0.6896504 , 0.69939436, 0.71296762,
          0.70969172, 0.70673165, 0.64653754, 0.70666435, 0.70582063,
          0.70637552, 0.70458239, 0.70825599, 0.71399742]), 'split1_test_score': array([0.58874365, 0.64744581, 0.673
15459, 0.67797319, 0.67833214,
          0.67892194, 0.67948305, 0.62217523, 0.67969629, 0.66181287,
          0.68390545, 0.68859944, 0.68122272, 0.69203784, 0.62051574,
          0.65313097, 0.68652024, 0.68819217, 0.69321619, 0.68671839,
          0.69473577, 0.60940741, 0.69302058, 0.68665149, 0.69501263,
          0.68698765, 0.69729378, 0.69593058, 0.63872499, 0.67732134,
          0.68508824, 0.69615869, 0.7024153 , 0.6996853 , 0.69575754,
          0.66625057, 0.68789758, 0.69198699, 0.69519938, 0.69486618,
          0.70242438, 0.7020477 , 0.63824565, 0.6989731 , 0.69134928,
          0.70382683, 0.7087464 , 0.70417502, 0.70218719]), 'split2_test_score': array([0.59439716, 0.6910779 , 0.703
17161, 0.7067554 , 0.69571681,
          0.69720199, 0.70971943, 0.65026951, 0.67574078, 0.6931531 ,
          0.70674717, 0.70006303, 0.71220211, 0.71786933, 0.63608647,
          0.68817887, 0.7031486 , 0.70137243, 0.71061775, 0.70304352,
          0.71761188, 0.64479801, 0.69276714, 0.69424556, 0.71760576,
          0.7105671 , 0.71413676, 0.71947774, 0.65195684, 0.69185657,
          0.7001831 , 0.71396098, 0.71521699, 0.71516128, 0.71813459,
          0.63088856, 0.69326705, 0.710121  , 0.72050859, 0.71834054,
          0.72205834, 0.72236981, 0.62816415, 0.7178248 , 0.72803155,
          0.72165086, 0.71754457, 0.72771712, 0.72631846]), 'split3_test_score': array([0.62586137, 0.67015026, 0.693
23191, 0.69269613, 0.68409717,
          0.68567857, 0.69358823, 0.60156148, 0.66102565, 0.68852083,
          0.70071488, 0.69486059, 0.69882496, 0.6985519 , 0.61282914,
          0.67747486, 0.69760356, 0.69230268, 0.70535685, 0.70180439,
          0.70282532, 0.6280656 , 0.69317473, 0.6917077 , 0.70110485,
          0.70594022, 0.70706898, 0.70824649, 0.64167234, 0.67828856,
          0.70221968, 0.69972919, 0.70835116, 0.70373061, 0.70370887,
          0.65365927, 0.68828132, 0.69646763, 0.71471307, 0.70905639,
          0.71261044, 0.71300927, 0.64855695, 0.69696427, 0.70351705,
          0.7034318 , 0.71356658, 0.7128759 , 0.7142657 ]), 'split4_test_score': array([0.62961518, 0.67965555, 0.681
4555 , 0.68989008, 0.69278315,
          0.69203237, 0.69505958, 0.62636288, 0.66868143, 0.7013299 ,
          0.70362049, 0.6971011 , 0.69668114, 0.69994622, 0.62379102,
          0.68627898, 0.69443703, 0.68676527, 0.70802294, 0.71122549,
          0.70520317, 0.62770653, 0.68874785, 0.7024472 , 0.70592839,
          0.71084925, 0.70736489, 0.71099037, 0.63516472, 0.69338965,
          0.69579853, 0.70898645, 0.71803893, 0.71273888, 0.71012509,
          0.63877456, 0.70305497, 0.69512993, 0.71054146, 0.71562435,
          0.71375668, 0.71772671, 0.64983803, 0.6996868 , 0.71250799,
          0.71824363, 0.71743032, 0.72112762, 0.71949106]), 'mean_test_score': array([0.60914095, 0.66965847, 0.68158
912, 0.68958264, 0.68632557,
          0.68738226, 0.6949515 , 0.62464378, 0.67132591, 0.68768945,
          0.69545373, 0.69415106, 0.69689574, 0.70058755, 0.62259936,
          0.67578756, 0.69699987, 0.69317807, 0.70181932, 0.70112888,
          0.7050709 , 0.63411468, 0.69167234, 0.69439133, 0.70432479,
          0.70406083, 0.70360443, 0.70758343, 0.64174981, 0.68200882,
          0.69578651, 0.7038349 , 0.70992907, 0.70702234, 0.70685594,
          0.64371689, 0.68993645, 0.69667101, 0.708071  , 0.71017093,
          0.71210818, 0.71237666, 0.64226836, 0.7040229 , 0.70824508,
          0.71070533, 0.71237362, 0.7148299 , 0.71525178]), 'std_test_score': array([0.01634826, 0.01514146, 0.016002
44, 0.01019974, 0.00678816,
          0.00649703, 0.0096235 , 0.01550313, 0.00636659, 0.01357563,
          0.01028333, 0.00426919, 0.00985407, 0.00908815, 0.00763018,
          0.01251733, 0.00623079, 0.00548677, 0.00776035, 0.00795846,
          0.00733703, 0.01733961, 0.0017244 , 0.00525999, 0.00750429,
          0.00879821, 0.00784713, 0.00784716, 0.00560297, 0.00923679,
          0.00591928, 0.00658885, 0.00585141, 0.00589894, 0.00736292,
          0.01422563, 0.00839841, 0.0071352 , 0.00944339, 0.00824352,
          0.00634967, 0.00731355, 0.00812621, 0.00763668, 0.01202809,
          0.00768882, 0.00505169, 0.00855924, 0.00792048]), 'rank_test_score': array([49, 42, 39, 34, 37, 36, 28, 47,
41, 35, 27, 30, 24, 22, 48, 40, 23,
          31, 20, 21, 14, 46, 32, 29, 15, 16, 19, 11, 45, 38, 26, 18,  8, 12,
          13, 43, 33, 25, 10,  7,  5,  3, 44, 17,  9,  6,  4,  2,  1]), 'split0_train_score': array([0.62199605, 0.69
261902, 0.68940337, 0.72635312, 0.72421576,
          0.73409359, 0.73854176, 0.67774106, 0.73229192, 0.75344117,
          0.73678583, 0.75087438, 0.75460314, 0.74756329, 0.67645014,
          0.74815777, 0.76944217, 0.76228239, 0.77660589, 0.77872616,
          0.77898032, 0.71293973, 0.77061531, 0.78849306, 0.78914537,
          0.79709075, 0.78803312, 0.79761429, 0.72733945, 0.786192  ,
          0.80817274, 0.80753997, 0.80622257, 0.81458778, 0.81930834,
          0.73134805, 0.80578292, 0.82516592, 0.83077874, 0.84052   ,
          0.83883212, 0.84086225, 0.76089089, 0.84535067, 0.85254283,
          0.85204931, 0.85844889, 0.86626734, 0.86688577]), 'split1_train_score': array([0.62453712, 0.69335799, 0.72
090997, 0.73006822, 0.72854897,
          0.72995487, 0.72993784, 0.66615993, 0.73497106, 0.72238187,
          0.75677394, 0.751959  , 0.74877772, 0.75700919, 0.67523135,
          0.73234443, 0.76502255, 0.77049551, 0.77270212, 0.76790238,
          0.77709756, 0.6762487 , 0.77896886, 0.78181271, 0.79517713,
          0.78676542, 0.79333332, 0.79591088, 0.7333879 , 0.79329065,
          0.80209381, 0.80764523, 0.81775922, 0.81606141, 0.81923979,
          0.74598757, 0.8119533 , 0.8343064 , 0.83383883, 0.83042749,
          0.84385102, 0.84032646, 0.76925082, 0.84421782, 0.85708232,
       0.85606177, 0.86761899, 0.86320359, 0.86406942]), 'split2_train_score': array([0.63098364, 0.70712902, 0.74
00121 , 0.73440031, 0.72557712,
       0.72263284, 0.74029989, 0.6609549 , 0.71599755, 0.73314332,
       0.74834906, 0.74511224, 0.75496787, 0.75956017, 0.68892432,
       0.75244354, 0.75939353, 0.77028313, 0.77229345, 0.76353787,
       0.77435741, 0.70009032, 0.77151613, 0.77775752, 0.78831656,
       0.79251616, 0.79324541, 0.79872303, 0.7333589 , 0.78784354,
       0.80174299, 0.81121928, 0.80987413, 0.81325254, 0.81293827,
       0.7396951 , 0.80234983, 0.8342375 , 0.83715094, 0.83718056,
       0.84246733, 0.83887843, 0.76532908, 0.84828381, 0.85197397,
       0.85702519, 0.8550176 , 0.86772284, 0.86306047]), 'split3_train_score': array([0.62968306, 0.70906232, 0.72
155135, 0.73239791, 0.73511138,
       0.72647108, 0.73901672, 0.64243619, 0.70642736, 0.74234572,
       0.75394139, 0.75321507, 0.75669579, 0.75792526, 0.65875534,
       0.74148965, 0.77147353, 0.75882053, 0.77807228, 0.77994764,
       0.77551701, 0.69738135, 0.778255  , 0.78094718, 0.79671174,
       0.79698148, 0.79497481, 0.7963314 , 0.72941207, 0.78449631,
       0.80664346, 0.80627666, 0.81412453, 0.81690928, 0.81362102,
       0.75318134, 0.80216046, 0.82665405, 0.84051381, 0.84065786,
       0.83852529, 0.84281595, 0.77729483, 0.83952171, 0.84569749,
       0.85491012, 0.85894536, 0.86021393, 0.86637146]), 'split4_train_score': array([0.64319485, 0.7080278 , 0.71
516836, 0.72551947, 0.72998474,
       0.72633209, 0.73134827, 0.64544982, 0.71214277, 0.73611941,
       0.74740496, 0.74558508, 0.74553179, 0.7457128 , 0.67221592,
       0.76307981, 0.75114309, 0.75764058, 0.77024032, 0.77437152,
       0.7724447 , 0.70209531, 0.76965853, 0.78153244, 0.79012895,
       0.79374822, 0.78902169, 0.79675477, 0.70569387, 0.78818538,
       0.79265755, 0.81138987, 0.8194473 , 0.81087346, 0.81169534,
       0.72996193, 0.81720015, 0.81686934, 0.83406652, 0.84072599,
       0.83329994, 0.84236391, 0.76586877, 0.83577256, 0.85498912,
       0.85691263, 0.85646319, 0.86181142, 0.85868766]), 'mean_train_score': array([0.63007894, 0.70203923, 0.7174
0903, 0.72974781, 0.72868759,
       0.7278969 , 0.7358289 , 0.65854838, 0.72036613, 0.7374863 ,
       0.74865104, 0.74934915, 0.75211526, 0.75355414, 0.67431541,
       0.74750304, 0.76329498, 0.76390443, 0.77398281, 0.77289711,
       0.7756794 , 0.69775108, 0.77380277, 0.78210858, 0.79189595,
       0.79342041, 0.79172167, 0.79706688, 0.72583844, 0.78800158,
       0.80226211, 0.8088142 , 0.81348555, 0.81433689, 0.81536055,
       0.74003479, 0.80788933, 0.82744664, 0.83526977, 0.83790238,
       0.83939514, 0.8410494 , 0.76772688, 0.84262932, 0.85245714,
       0.8553918 , 0.85929881, 0.86384382, 0.86381496]), 'std_train_score': array([0.00733549, 0.00741885, 0.01630
737, 0.00341099, 0.00381154,
       0.00386849, 0.00429633, 0.01313972, 0.01128296, 0.01026689,
       0.00687599, 0.00335269, 0.00423767, 0.00573574, 0.00964717,
       0.01032877, 0.00735596, 0.00551077, 0.00290188, 0.00629927,
       0.00224252, 0.01197939, 0.00397679, 0.00350604, 0.00338993,
       0.003778  , 0.00269805, 0.00100172, 0.01033808, 0.00295278,
       0.00541622, 0.00209035, 0.00489915, 0.0021366 , 0.00325456,
       0.00877844, 0.00585064, 0.00649432, 0.00330767, 0.0039702 ,
       0.00367473, 0.00142264, 0.00547359, 0.00444015, 0.00384023,
       0.00183489, 0.00439186, 0.00277982, 0.00292878])}
```

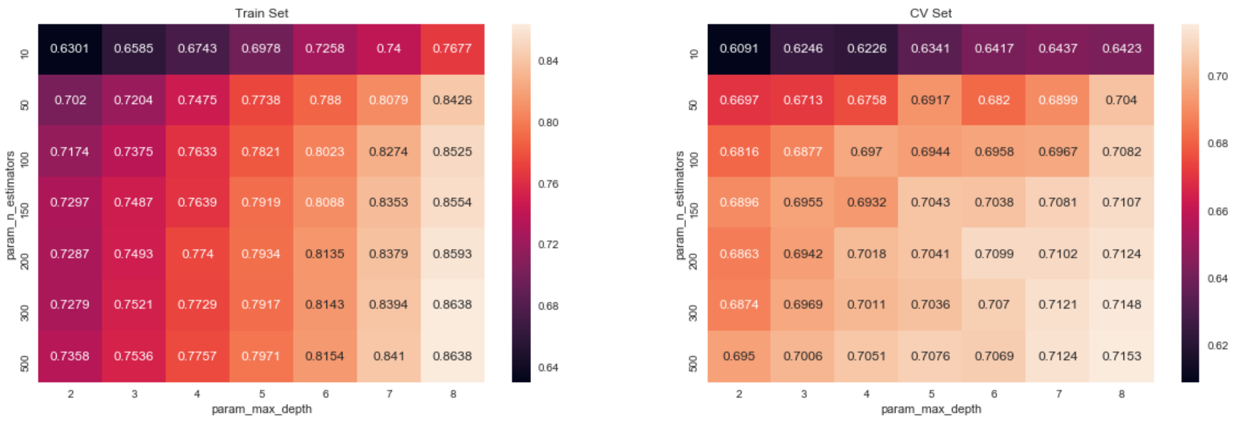
In [135]:
```python
import seaborn as sns; sns.set()
max_scores1 = pd.DataFrame(set2.cv_results_).groupby(['param_n_estimators', 'param_max_depth']).max().unstack()
max_scores1
```

Out[135]:

	mean_fit_time							mean_score_time			...	std_test_sc
param_max_depth	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**2**	**3**	**4**	**...**	**6**
param_n_estimators												
10	0.407695	0.413486	0.478719	0.535169	0.598398	0.714091	0.801456	0.085779	0.073606	0.072807	...	0.005603
50	1.167278	1.260030	1.597120	1.923059	2.286293	2.735486	3.150774	0.233975	0.213836	0.210439	...	0.009237
100	2.331964	2.413345	3.079173	3.593391	4.345777	5.469379	6.142575	0.486899	0.391169	0.399526	...	0.005919
150	2.995589	3.401303	4.482620	5.256543	6.447957	7.780201	9.151718	0.595210	0.564889	0.573886	...	0.006589
200	4.332022	4.523504	5.708534	6.962987	8.798279	10.126719	12.403216	0.837567	0.738646	0.758979	...	0.005851
300	6.125219	6.899541	8.809832	10.737088	12.695241	15.454662	18.293092	1.313887	1.168483	1.108045	...	0.005899
500	8.744614	10.943526	14.357803	17.413432	21.168784	25.246677	30.397708	1.919077	1.788234	1.822330	...	0.007363

7 rows × 140 columns

In [136]:
```
fig, ax = plt.subplots(1,2, figsize=(20,6))
sns.heatmap(max_scores1.mean_train_score, annot = True, fmt='.4g', ax=ax[0])
sns.heatmap(max_scores1.mean_test_score, annot = True, fmt='.4g', ax=ax[1])
ax[0].set_title('Train Set')
ax[1].set_title('CV Set')
plt.show()
```



In [137]:
```
print(set2.best_estimator_)
```

```
RandomForestClassifier(bootstrap=True, class_weight='balanced',
            criterion='gini', max_depth=8, max_features='auto',
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=500, n_jobs=1, oob_score=False, random_state=None,
            verbose=0, warm_start=False)
```

Training our model with best Hyperparameters

In [138]:
```
def pred_prob(clf, data):
    y_pred = []
    y_pred = clf.predict_proba(data)[:,1]
    return y_pred
```

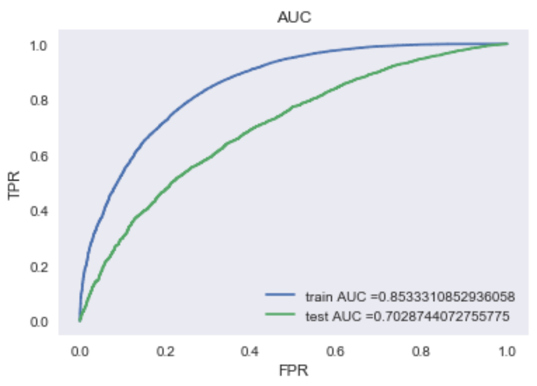
In [140]:
```
# https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
from sklearn.metrics import roc_curve, auc
model = RandomForestClassifier(max_depth = 8, n_estimators = 500)

model.fit(X_tr_TFIDF, y_train)

y_train_pred = pred_prob(model,X_tr_TFIDF)
y_test_pred = pred_prob(model,X_te_TFIDF)

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

plt.close
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("FPR")
plt.ylabel("TPR")
plt.title("AUC")
plt.grid()
plt.show()
```



Confusion Matrix

In [141]:
```
#our objective here is to make auc the maximum
#so we find  the best threshold that will give the least fpr
best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")
print(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)))
```

```
the maximum value of tpr*(1-fpr) 0.5926446990264167 for threshold 0.841
Train confusion matrix
[[ 3469  1175]
 [ 5270 20236]]
```

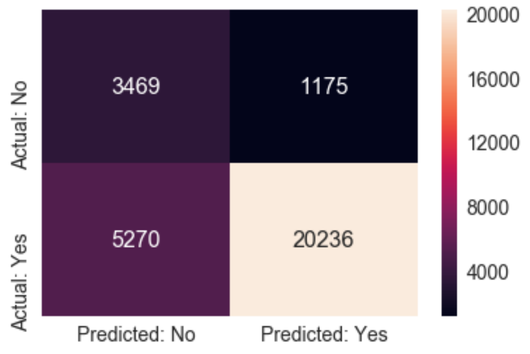
In [142]:
```
# https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix

print("Train data confusion matrix")

confusion_matrix_df_train = pd.DataFrame(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)), ['Ac
tual: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Train data confusion matrix
```

Out[142]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c387733e80>
```



In [143]:
```
print("Test confusion matrix")
print(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)))
```

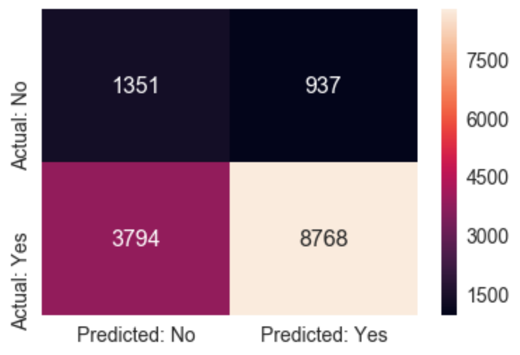
```
Test confusion matrix
[[1351  937]
 [3794 8768]]
```

In [144]:
```
print("Test data confusion matrix")

confusion_matrix_df_test = pd.DataFrame(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)), ['Actua
l: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_test, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Test data confusion matrix
```

Out[144]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c387733438>
```



In [ ]:

2.4.3 Applying Random Forests on AVG W2V, SET 3

In [ ]:
```
# Please write all the code with proper documentation
```

In [145]:
```
import warnings
warnings.filterwarnings('ignore')
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt
#from sklearn.grid_search import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import learning_curve, GridSearchCV

#n_estimators = [10, 50, 100, 150, 200, 300, 500, 1000], max_depth = [2, 3, 4, 5, 6, 7, 8, 9, 10]

clf = RandomForestClassifier(class_weight='balanced')
parameters ={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]}
set3 = GridSearchCV(clf, parameters, cv=5, scoring='roc_auc',return_train_score=True)
set3.fit(X_tr_AVG_W2V, y_train)
```

Out[145]:
```
GridSearchCV(cv=5, error_score='raise',
       estimator=RandomForestClassifier(bootstrap=True, class_weight='balanced',
            criterion='gini', max_depth=None, max_features='auto',
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=10, n_jobs=1, oob_score=False, random_state=None,
            verbose=0, warm_start=False),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
       scoring='roc_auc', verbose=0)
```

```
In [146]: print(set3.cv_results_)
```

```
{'mean_fit_time': array([  3.73062325,  13.32516446,  25.57221727,  33.69788375,
        41.92189856,  60.54669194,  99.69279761,   4.06591187,
        16.30799503,  31.07567439,  46.57105794,  61.64554644,
        92.21979003, 153.77058063,   5.51824231,  22.29996467,
        43.26291308,  65.49645438,  88.41017094, 128.48839798,
       213.09653106,   6.65020905,  29.62936335,  57.91731505,
        86.75618744, 115.32339101, 185.9726584 , 286.61373773,
         8.31237168,  38.32829809,  76.14577565, 114.00532498,
       152.73555007, 229.14879632, 381.00829735,  10.5434124 ,
        49.14317856,  97.81561079, 146.13300571, 196.76520405,
       291.58683152, 484.32141342,  12.91427064,  61.29288054,
       121.27448444, 180.93973322, 242.88846173, 364.30078444,
       606.01726732]), 'std_fit_time': array([ 0.46871538,  0.94828071,  1.6721416 ,  2.05765362,  1.95092412,
        1.23487479,  1.11740305,  0.29004007,  0.74801564,  0.07524828,
        0.78092064,  0.45758603,  0.8748903 ,  0.824161  ,  0.69650726,
        0.37874563,  0.19436051,  0.75199033,  2.63021382,  1.31256689,
        0.82624271,  0.13649627,  0.66789174,  0.7844865 ,  0.93759825,
        1.53875575, 25.72804289,  0.56307094,  0.05390959,  0.86349005,
        1.12045847,  2.65580884,  1.02227508,  1.05258202,  4.35531515,
        0.12423184,  0.91764252,  0.97585077,  1.13059805,  3.27768592,
        0.83730496,  2.10063469,  0.17608668,  0.9915423 ,  0.94019382,
        1.31081392,  0.97941753,  1.55535882,  2.50909234]), 'mean_score_time': array([0.43264499, 0.89720206, 1.4
7086325, 1.76807261, 2.16500397,
       3.10249586, 5.19471078, 0.31894779, 0.69912319, 1.17266507,
       1.64579749, 2.12611423, 3.10389242, 5.41352496, 0.31795282,
       0.73343873, 1.20457859, 1.66813121, 2.27929773, 3.12106166,
       5.26392646, 0.33011689, 0.70511503, 1.18901477, 1.68050752,
       2.16042938, 3.15616808, 5.10474176, 0.3165544 , 0.70391197,
       1.24286914, 1.61387749, 2.30143914, 3.23315587, 5.29284568,
       0.31295619, 0.72206225, 1.2428925 , 1.77187657, 2.33735733,
       3.2967833 , 5.47077117, 0.33051081, 0.73264065, 1.24945831,
       1.74273496, 2.30264359, 3.29320908, 5.32076492]), 'std_score_time': array([0.0963797 , 0.07375091, 0.129240
07, 0.10906159, 0.07975843,
       0.08687339, 0.42801299, 0.01892752, 0.01032365, 0.01000217,
       0.02256441, 0.0107869 , 0.02361238, 0.78216354, 0.01134418,
       0.07193042, 0.04774801, 0.01823479, 0.18517518, 0.03122215,
       0.30411509, 0.02837078, 0.00993368, 0.01550209, 0.01824914,
       0.01663906, 0.02906525, 0.06053538, 0.01355238, 0.01538633,
       0.10891573, 0.20514628, 0.20089182, 0.02100894, 0.18498651,
       0.01024336, 0.01236752, 0.02474342, 0.06167748, 0.16855486,
       0.08634432, 0.17034525, 0.01551989, 0.01122304, 0.02398037,
       0.00280214, 0.10672358, 0.0313729 , 0.05331423]), 'param_max_depth': masked_array(data=[2, 2, 2, 2, 2, 2,
2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4,
                   4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7,
                   7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'param_n_estimators': masked_array(data=[10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
150, 200,
                   300, 500, 10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
                   150, 200, 300, 500, 10, 50, 100, 150, 200, 300, 500,
                   10, 50, 100, 150, 200, 300, 500, 10, 50, 100, 150, 200,
                   300, 500],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'params': [{'max_depth': 2, 'n_estimators': 10}, {'max_depth': 2, 'n_estimators': 50},
{'max_depth': 2, 'n_estimators': 100}, {'max_depth': 2, 'n_estimators': 150}, {'max_depth': 2, 'n_estimators': 20
0}, {'max_depth': 2, 'n_estimators': 300}, {'max_depth': 2, 'n_estimators': 500}, {'max_depth': 3, 'n_estimators':
10}, {'max_depth': 3, 'n_estimators': 50}, {'max_depth': 3, 'n_estimators': 100}, {'max_depth': 3, 'n_estimators':
150}, {'max_depth': 3, 'n_estimators': 200}, {'max_depth': 3, 'n_estimators': 300}, {'max_depth': 3, 'n_estimator
s': 500}, {'max_depth': 4, 'n_estimators': 10}, {'max_depth': 4, 'n_estimators': 50}, {'max_depth': 4, 'n_estimato
rs': 100}, {'max_depth': 4, 'n_estimators': 150}, {'max_depth': 4, 'n_estimators': 200}, {'max_depth': 4, 'n_estim
ators': 300}, {'max_depth': 4, 'n_estimators': 500}, {'max_depth': 5, 'n_estimators': 10}, {'max_depth': 5, 'n_est
imators': 50}, {'max_depth': 5, 'n_estimators': 100}, {'max_depth': 5, 'n_estimators': 150}, {'max_depth': 5, 'n_e
stimators': 200}, {'max_depth': 5, 'n_estimators': 300}, {'max_depth': 5, 'n_estimators': 500}, {'max_depth': 6,
'n_estimators': 10}, {'max_depth': 6, 'n_estimators': 50}, {'max_depth': 6, 'n_estimators': 100}, {'max_depth': 6,
'n_estimators': 150}, {'max_depth': 6, 'n_estimators': 200}, {'max_depth': 6, 'n_estimators': 300}, {'max_depth':
6, 'n_estimators': 500}, {'max_depth': 7, 'n_estimators': 10}, {'max_depth': 7, 'n_estimators': 50}, {'max_depth':
7, 'n_estimators': 100}, {'max_depth': 7, 'n_estimators': 150}, {'max_depth': 7, 'n_estimators': 200}, {'max_dept
h': 7, 'n_estimators': 300}, {'max_depth': 7, 'n_estimators': 500}, {'max_depth': 8, 'n_estimators': 10}, {'max_de
pth': 8, 'n_estimators': 50}, {'max_depth': 8, 'n_estimators': 100}, {'max_depth': 8, 'n_estimators': 150}, {'max_
depth': 8, 'n_estimators': 200}, {'max_depth': 8, 'n_estimators': 300}, {'max_depth': 8, 'n_estimators': 500}], 's
plit0_test_score': array([0.63610115, 0.6869933 , 0.68486661, 0.69110068, 0.68178143,
       0.68837333, 0.68944469, 0.66675746, 0.69036288, 0.69557412,
       0.69425739, 0.68701524, 0.69769997, 0.69635454, 0.65939126,
       0.68491704, 0.69845866, 0.7016968 , 0.70094971, 0.69878272,
       0.70136345, 0.67315262, 0.69344342, 0.70289538, 0.6996929 ,
       0.70043492, 0.70206053, 0.7034165 , 0.6757345 , 0.7060563 ,
       0.69800357, 0.70335321, 0.70627298, 0.70510203, 0.70796083,
       0.65024206, 0.69314446, 0.69789069, 0.70463998, 0.70051361,
       0.70751102, 0.70918916, 0.64080487, 0.69017785, 0.69290753,
       0.70335891, 0.6990148 , 0.70544213, 0.70724602]), 'split1_test_score': array([0.62741861, 0.67582709, 0.667
89369, 0.67741039, 0.67344253,
       0.67613602, 0.67382955, 0.65175806, 0.67061525, 0.67845601,
       0.68567741, 0.68195223, 0.68358998, 0.6852368 , 0.65587015,
       0.68123201, 0.68064727, 0.68956846, 0.68268743, 0.68553054,
       0.68746224, 0.67216711, 0.67998339, 0.69213512, 0.69074301,
       0.69006774, 0.6911009 , 0.69280386, 0.64957482, 0.68397003,
       0.69190005, 0.69140055, 0.69455408, 0.69453614, 0.69747125,
       0.64226996, 0.67862651, 0.69222713, 0.69439644, 0.69682215,
       0.69691141, 0.69515697, 0.65355154, 0.68178721, 0.68634825,
       0.69720958, 0.69339008, 0.69586305, 0.69819   ]), 'split2_test_score': array([0.64858544, 0.69111462, 0.697
2003 , 0.68939626, 0.69700342,
       0.70081934, 0.69455598, 0.65857399, 0.71021512, 0.6955577 ,
       0.70418008, 0.70850394, 0.70663639, 0.70533818, 0.65842511,
       0.69734126, 0.70682546, 0.70615652, 0.70770289, 0.70921361,
       0.7058689 , 0.67066948, 0.70151297, 0.70963523, 0.71458582,
       0.71307553, 0.71541218, 0.71472805, 0.66269199, 0.6975506 ,
       0.70934317, 0.71686001, 0.71279297, 0.71531849, 0.71643585,
       0.67509378, 0.70499611, 0.71051097, 0.71370459, 0.71582452,
       0.71301644, 0.71894132, 0.64833823, 0.7053922 , 0.70908024,
       0.70512779, 0.71390506, 0.71732489, 0.71551305]), 'split3_test_score': array([0.63751646, 0.68057657, 0.686
86209, 0.68409432, 0.68668694,
       0.68930215, 0.68860957, 0.66063462, 0.69638533, 0.69059677,
       0.69381128, 0.69616502, 0.6976253 , 0.69661218, 0.6722511 ,
       0.6966278 , 0.68942644, 0.69907777, 0.70439786, 0.70081512,
       0.70498598, 0.6643713 , 0.69508227, 0.70740303, 0.70915557,
       0.70458672, 0.70666973, 0.71059749, 0.6671414 , 0.70229164,
       0.70943286, 0.70560427, 0.71095243, 0.70674549, 0.71162074,
       0.64778619, 0.70437612, 0.70475238, 0.70328134, 0.70588325,
       0.71116134, 0.71033857, 0.6447692 , 0.69117223, 0.70751086,
       0.71244098, 0.70816525, 0.70908446, 0.71027674]), 'split4_test_score': array([0.63168648, 0.67160344, 0.699
70666, 0.69202012, 0.69741079,
       0.68903959, 0.69308545, 0.66955748, 0.69021139, 0.69388609,
       0.6993826 , 0.70292463, 0.69723968, 0.70244213, 0.67112665,
       0.69297433, 0.70323453, 0.69844507, 0.70570869, 0.70659975,
       0.70919284, 0.65551929, 0.70369527, 0.70744855, 0.71075419,
       0.71062617, 0.71487969, 0.71123837, 0.67388895, 0.7007815 ,
       0.701974  , 0.71506559, 0.71717851, 0.71713647, 0.71462239,
       0.66618002, 0.70211343, 0.71191543, 0.71501658, 0.71332193,
       0.71660983, 0.71966006, 0.6490787 , 0.70092895, 0.71074384,
       0.71494792, 0.71519044, 0.71895998, 0.71818617]), 'mean_test_score': array([0.63626178, 0.68122351, 0.68730
538, 0.68680432, 0.6872645 ,
       0.68873406, 0.68790493, 0.66145623, 0.691558  , 0.69081419,
       0.69546158, 0.69531168, 0.69655828, 0.69719656, 0.66341246,
       0.69061822, 0.69571831, 0.69898903, 0.70028916, 0.70018809,
       0.70177442, 0.66717655, 0.69474312, 0.70390331, 0.70498593,
       0.70375788, 0.70602418, 0.70655659, 0.66580639, 0.69813019,
       0.7021306 , 0.70645634, 0.70834983, 0.70776732, 0.70962199,
       0.65631387, 0.69665103, 0.70345886, 0.70620744, 0.70647267,
       0.70904171, 0.71065687, 0.64730823, 0.69389133, 0.70131755,
       0.70661665, 0.70593259, 0.70933446, 0.70988203]), 'std_test_score': array([0.0071078 , 0.00711618, 0.011266
18, 0.00544004, 0.0091567 ,
       0.00781423, 0.00737637, 0.00627318, 0.01275329, 0.00644001,
       0.00618527, 0.00979825, 0.00738446, 0.00689759, 0.00686372,
       0.00644165, 0.00952983, 0.00543612, 0.00907159, 0.00824348,
       0.00757773, 0.00658125, 0.008314  , 0.00627974, 0.00864515,
       0.00816254, 0.00900417, 0.00779765, 0.00936852, 0.00758981,
       0.00673896, 0.00915835, 0.00773414, 0.00809938, 0.00672047,
       0.01230313, 0.00996325, 0.00748423, 0.00754075, 0.00725752,
       0.00673863, 0.0088564 , 0.00428769, 0.00836045, 0.00982117,
       0.00639759, 0.00847841, 0.00840448, 0.00699235]), 'rank_test_score': array([49, 42, 39, 41, 40, 37, 38, 46,
34, 35, 30, 31, 28, 26, 45, 36, 29,
       24, 22, 23, 20, 43, 32, 16, 15, 17, 13,  9, 44, 25, 19, 11,  6,  7,
        3, 47, 27, 18, 12, 10,  5,  1, 48, 33, 21,  8, 14,  4,  2]), 'split0_train_score': array([0.66364709, 0.71
38774 , 0.71283148, 0.72180735, 0.71310233,
       0.71667777, 0.72059565, 0.70149002, 0.72580386, 0.7375143 ,
       0.7361672 , 0.73008994, 0.74138487, 0.73812959, 0.71078282,
       0.75425496, 0.76272026, 0.76516625, 0.76647352, 0.76512323,
       0.76638417, 0.7514201 , 0.79158955, 0.79783071, 0.79733806,
       0.80207676, 0.80340369, 0.80698442, 0.7974557 , 0.844013  ,
       0.844691  , 0.85256955, 0.85243691, 0.85458251, 0.85483721,
       0.83345418, 0.89701565, 0.9020852 , 0.90246731, 0.9050249 ,
       0.90635939, 0.9088287 , 0.87541317, 0.9385451 , 0.94485469,
       0.94741036, 0.94972686, 0.95151787, 0.9530801 ]), 'split1_train_score': array([0.66047507, 0.70876092, 0.70
752867, 0.71586727, 0.71449483,
       0.71661421, 0.71292916, 0.70326151, 0.72248798, 0.73218205,
       0.73785337, 0.73384733, 0.7386269 , 0.73909411, 0.71830649,
       0.75227495, 0.75965912, 0.76386609, 0.76242682, 0.76467112,
       0.76495272, 0.75607755, 0.78553101, 0.80005765, 0.79979761,
       0.80037342, 0.80277553, 0.80463241, 0.79486654, 0.83734598,
       0.84306219, 0.84468481, 0.84998246, 0.85146176, 0.85244371,
       0.83169803, 0.88535032, 0.89821605, 0.90357681, 0.90363985,
       0.90444963, 0.9049147 , 0.88040459, 0.9373195 , 0.94437776,
       0.94819441, 0.9507175 , 0.95126645, 0.95208128]), 'split2_train_score': array([0.65882608, 0.71368526, 0.71
238546, 0.70482392, 0.71381609,
       0.71751177, 0.713851  , 0.68784827, 0.73515631, 0.72773324,
       0.73134683, 0.73808064, 0.73768763, 0.73851251, 0.70371042,
       0.75090839, 0.76091826, 0.76248484, 0.76538729, 0.765286  ,
       0.76242811, 0.7498334 , 0.78767221, 0.80200072, 0.80295762,
       0.80228947, 0.80350416, 0.80438628, 0.79408625, 0.83630793,
       0.84681669, 0.85192651, 0.85074253, 0.85280983, 0.85444084,
       0.83442446, 0.88877092, 0.9022733 , 0.90389166, 0.90412418,
       0.90552647, 0.90701497, 0.87404027, 0.93646803, 0.94741945,
       0.95185846, 0.95195233, 0.95185515, 0.95312406]), 'split3_train_score': array([0.65832276, 0.71036832, 0.71
142205, 0.71013214, 0.71253773,
       0.71398904, 0.7128005 , 0.70281393, 0.73058804, 0.72903862,
       0.73546226, 0.73403023, 0.73615265, 0.73665525, 0.71416587,
       0.7564818 , 0.75246933, 0.75955025, 0.76464216, 0.76192491,
       0.76561417, 0.75285278, 0.79172076, 0.7980515 , 0.80280585,
       0.7974574 , 0.80122337, 0.80327662, 0.78610772, 0.83699254,
       0.85006546, 0.85094227, 0.8500859 , 0.84940617, 0.85389873,
       0.83824803, 0.89092621, 0.90057769, 0.90380211, 0.90418237,
       0.90740474, 0.90796608, 0.89033023, 0.93631721, 0.94661412,
       0.94820036, 0.95027359, 0.95325439, 0.95378183]), 'split4_train_score': array([0.66011317, 0.68705572, 0.71
277743, 0.70823878, 0.71203725,
       0.70897882, 0.71102431, 0.6918687 , 0.71893966, 0.725421  ,
       0.72606004, 0.73004135, 0.72858542, 0.73335005, 0.71077676,
       0.74905214, 0.75941142, 0.75137612, 0.75843031, 0.76157965,
       0.76306724, 0.74501035, 0.78818519, 0.79434547, 0.79675803,
       0.79732779, 0.79939781, 0.80145671, 0.78110743, 0.83641192,
       0.84136515, 0.84499697, 0.84980276, 0.84886618, 0.84971257,
       0.84103537, 0.88720685, 0.89474628, 0.90041932, 0.9003604 ,
       0.90473653, 0.90598569, 0.88095832, 0.93581123, 0.94493436,
       0.94999632, 0.94909904, 0.95155438, 0.9518798 ]), 'mean_train_score': array([0.66027683, 0.70674952, 0.7113
8902, 0.71217389, 0.71319765,
       0.71475432, 0.71424012, 0.69745649, 0.72659517, 0.73037784,
       0.73337794, 0.7332179 , 0.73648749, 0.7371483 , 0.71154847,
       0.75259445, 0.75903568, 0.76048871, 0.76347202, 0.76371698,
       0.76448928, 0.75103884, 0.78893974, 0.79845721, 0.79993143,
       0.79990497, 0.80206091, 0.80414729, 0.79072473, 0.83821427,
       0.8452001 , 0.84902402, 0.85061011, 0.85142529, 0.85306661,
       0.83577201, 0.88985399, 0.8995797 , 0.90283144, 0.90346634,
       0.90569535, 0.90694203, 0.88022932, 0.93689221, 0.94564008,
       0.94913198, 0.95035387, 0.95188965, 0.95278941]), 'std_train_score': array([0.00186267, 0.01003891, 0.00199
523, 0.00600201, 0.00087821,
       0.00312104, 0.00330686, 0.00635942, 0.00575045, 0.0041843 ,
       0.00423865, 0.00298586, 0.00430248, 0.00206301, 0.00479872,
       0.00258308, 0.00348617, 0.00492322, 0.00284808, 0.00162046,
       0.00150608, 0.00365021, 0.00238961, 0.0025527 , 0.00261664,
       0.00215659, 0.00156143, 0.00180771, 0.00612486, 0.00292419,
       0.00302727, 0.00345603, 0.00096699, 0.0021219 , 0.00186298,
       0.00339468, 0.0040225 , 0.00281959, 0.00130941, 0.00161565,
       0.00108283, 0.00138839, 0.00572871, 0.00095851, 0.00116818,
       0.00160572, 0.00096555, 0.00070749, 0.00070855])}
```

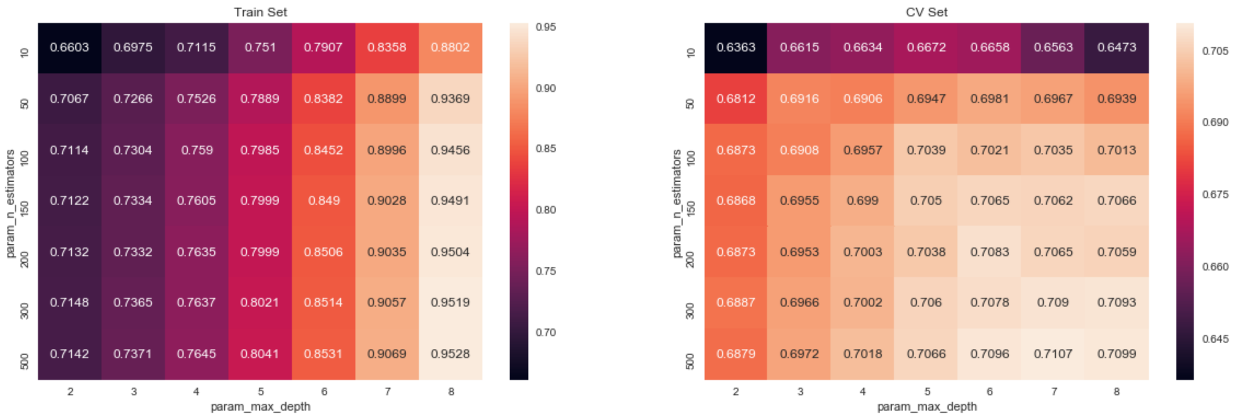
In [159]:
```python
import seaborn as sns; sns.set()
max_scores1 = pd.DataFrame(set3.cv_results_).groupby(['param_n_estimators', 'param_max_depth']).max().unstack()
max_scores1
```

Out[159]:

	mean_fit_time							mean_score_time			...	std
param_max_depth	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**2**	**3**	**4**	**...**	**6**
param_n_estimators												
10	3.730623	4.065912	5.518242	6.650209	8.312372	10.543412	12.914271	0.432645	0.318948	0.317953	...	0.0
50	13.325164	16.307995	22.299965	29.629363	38.328298	49.143179	61.292881	0.897202	0.699123	0.733439	...	0.0
100	25.572217	31.075674	43.262913	57.917315	76.145776	97.815611	121.274484	1.470863	1.172665	1.204579	...	0.0
150	33.697884	46.571058	65.496454	86.756187	114.005325	146.133006	180.939733	1.768073	1.645797	1.668131	...	0.0
200	41.921899	61.645546	88.410171	115.323391	152.735550	196.765204	242.888462	2.165004	2.126114	2.279298	...	0.0
300	60.546692	92.219790	128.488398	185.972658	229.148796	291.586832	364.300784	3.102496	3.103892	3.121062	...	0.0
500	99.692798	153.770581	213.096531	286.613738	381.008297	484.321413	606.017267	5.194711	5.413525	5.263926	...	0.0

7 rows × 140 columns

In [160]:
```python
fig, ax = plt.subplots(1,2, figsize=(20,6))
sns.heatmap(max_scores1.mean_train_score, annot = True, fmt='.4g', ax=ax[0])
sns.heatmap(max_scores1.mean_test_score, annot = True, fmt='.4g', ax=ax[1])
ax[0].set_title('Train Set')
ax[1].set_title('CV Set')
plt.show()
```



In [162]:
```python
print(set3.best_estimator_)
```

```
RandomForestClassifier(bootstrap=True, class_weight='balanced',
            criterion='gini', max_depth=7, max_features='auto',
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=500, n_jobs=1, oob_score=False, random_state=None,
            verbose=0, warm_start=False)
```

Training our model with best Hyperparameters

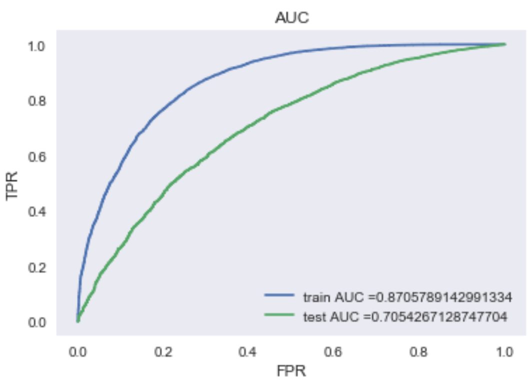
In [150]:
```python
# https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
from sklearn.metrics import roc_curve, auc
model = RandomForestClassifier(max_depth = 7, n_estimators = 500)

model.fit(X_tr_AVG_W2V, y_train)

y_train_pred = pred_prob(model,X_tr_AVG_W2V)
y_test_pred = pred_prob(model,X_te_AVG_W2V)

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

plt.close
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("FPR")
plt.ylabel("TPR")
plt.title("AUC")
plt.grid()
plt.show()
```



Confusion matrix

In [151]:
```
#our objective here is to make auc the maximum
#so we find  the best threshold that will give the least fpr
best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")
print(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)))
```

```
the maximum value of tpr*(1-fpr) 0.6217242934107549 for threshold 0.831
Train confusion matrix
[[ 3461  1183]
 [ 4228 21278]]
```

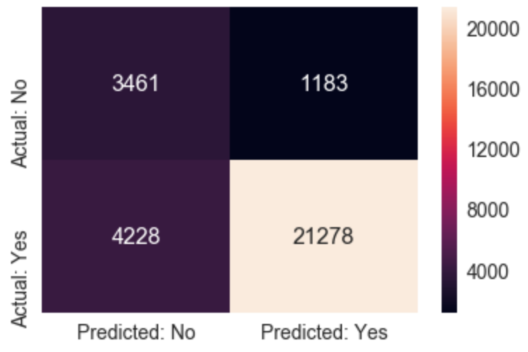
In [152]:
```
#plotting confusion matrix using seaborn's heatmap
# https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix

print("Train data confusion matrix")

confusion_matrix_df_train = pd.DataFrame(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)), ['Actual: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Train data confusion matrix
```

Out[152]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c387725630>
```



In [153]:
```
print("Test confusion matrix")
print(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)))
```

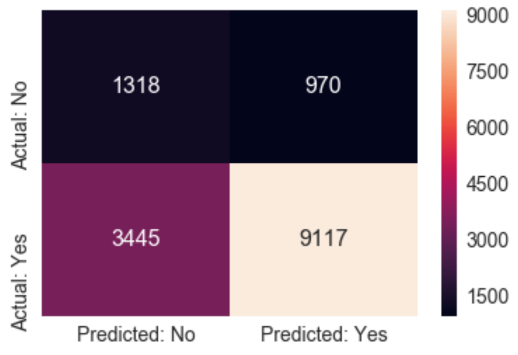
```
Test confusion matrix
[[1318  970]
 [3445 9117]]
```

In [154]:
```
print("Test data confusion matrix")

confusion_matrix_df_test = pd.DataFrame(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)), ['Actual: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_test, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Test data confusion matrix
```

Out[154]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c38752fd30>
```



In [ ]:

2.4.4 Applying Random Forests on TFIDF W2V, SET 4

In [155]:
```
# Please write all the code with proper documentation
```

In [156]:
```
import warnings
warnings.filterwarnings('ignore')
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt
#from sklearn.grid_search import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import learning_curve, GridSearchCV

#n_estimators = [10, 50, 100, 150, 200, 300, 500, 1000], max_depth = [2, 3, 4, 5, 6, 7, 8, 9, 10]

clf = RandomForestClassifier(class_weight='balanced')
parameters ={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]}
set4 = GridSearchCV(clf, parameters, cv=5, scoring='roc_auc',return_train_score=True)
set4.fit(X_tr_TFIDF_W2V, y_train)
```

Out[156]:
```
GridSearchCV(cv=5, error_score='raise',
       estimator=RandomForestClassifier(bootstrap=True, class_weight='balanced',
            criterion='gini', max_depth=None, max_features='auto',
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=10, n_jobs=1, oob_score=False, random_state=None,
            verbose=0, warm_start=False),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
       scoring='roc_auc', verbose=0)
```

In [157]:
```
print(set4.cv_results_)
```

```
{'mean_fit_time': array([  3.37656288,  10.6826333 ,  20.47025652,  30.79544525,
        40.52462788,  60.3326427 , 104.0308053 ,   3.88859434,
        16.65785251,  34.24960699,  54.64071875,  74.67776875,
       122.55498152, 206.21353583,   6.43122153,  29.2857357 ,
        60.86756506,  74.09678359,  90.94179354, 128.84166279,
       212.76262474,   6.85785995,  29.63774123,  58.11718302,
        87.6504045 , 114.88916621, 172.28625979, 284.34959168,
         8.3546576 ,  38.32970505,  75.07503328, 112.3206285 ,
       149.93225298, 224.23852587, 373.22491245,  10.34992142,
        48.8280076 ,  96.95551763, 144.45628934, 192.09070592,
       289.58498349, 488.22677097,  13.24418283,  61.58411808,
       122.79661183, 182.61325154, 246.40206442, 366.70414157,
       608.54800487]), 'std_fit_time': array([4.28545541e-01, 7.61170269e-02, 5.90249980e-02, 7.99001322e-01,
       8.90242182e-01, 8.71740600e-01, 7.91750712e+00, 1.16674431e-02,
       7.28747650e-01, 2.26865342e+00, 4.61617486e+00, 3.48773115e+00,
       7.14010361e+00, 1.19057430e+01, 2.21140056e-01, 1.69020244e+00,
       1.87773033e+00, 5.64744887e+00, 3.56865319e+00, 9.10843284e-01,
       6.51808558e-01, 4.90771691e-01, 1.00180000e+00, 6.42902104e-01,
       2.09860041e+00, 9.63375966e-01, 1.25906308e+00, 3.58010075e+00,
       8.66561950e-02, 7.37421187e-01, 8.33439082e-01, 8.93958631e-01,
       7.41448110e-01, 3.00439923e-01, 8.97358049e-01, 9.96862699e-02,
       9.58495140e-01, 1.77353620e+00, 1.52469492e+00, 1.03095950e+00,
       2.47810716e+00, 1.13793768e+00, 7.65855530e-01, 3.86180498e-01,
       1.53431913e+00, 1.53774178e+00, 5.12089202e+00, 4.70079336e+00,
       1.42538533e+00]), 'mean_score_time': array([0.34108763, 0.70411696, 1.17226534, 1.65158348, 2.12052798,
       3.07357922, 4.9950346 , 0.33211946, 0.74061933, 1.40205269,
       1.81175551, 2.3819171 , 4.02391467, 6.37122655, 0.39280996,
       0.95877771, 1.63521338, 1.80209165, 2.14700336, 3.12722216,
       5.06106601, 0.32493229, 0.71369162, 1.24388118, 1.68648815,
       2.18754344, 3.26906638, 5.12469568, 0.3187479 , 0.71309357,
       1.20658002, 1.70383019, 2.20170531, 3.25928626, 5.22523341,
       0.32075152, 0.72705679, 1.22872148, 1.72837906, 2.32199097,
       3.29917164, 5.41452093, 0.32015142, 0.77851014, 1.25683856,
       1.76468863, 2.34473109, 3.35881925, 5.37662187]), 'std_score_time': array([0.03874087, 0.00591467, 0.017914
64, 0.02724038, 0.01600061,
       0.02083997, 0.04578612, 0.01495054, 0.05211174, 0.18115562,
       0.10471114, 0.12188756, 0.52108004, 0.44387613, 0.04177432,
       0.0704989 , 0.1013927 , 0.06076347, 0.12855792, 0.02017844,
       0.04936749, 0.02439235, 0.0131341 , 0.08006261, 0.01698248,
       0.01031206, 0.2474431 , 0.0221704 , 0.01329948, 0.00986708,
       0.00363261, 0.00825937, 0.02207959, 0.15445154, 0.06385135,
       0.01039855, 0.00790376, 0.02043356, 0.01399096, 0.17366591,
       0.08822745, 0.2090541 , 0.00754415, 0.06737551, 0.01233435,
       0.01034942, 0.10759653, 0.10994987, 0.05945013]), 'param_max_depth': masked_array(data=[2, 2, 2, 2, 2, 2,
2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4,
                   4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7,
                   7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'param_n_estimators': masked_array(data=[10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
150, 200,
                   300, 500, 10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
                   150, 200, 300, 500, 10, 50, 100, 150, 200, 300, 500,
                   10, 50, 100, 150, 200, 300, 500, 10, 50, 100, 150, 200,
                   300, 500],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'params': [{'max_depth': 2, 'n_estimators': 10}, {'max_depth': 2, 'n_estimators': 50},
{'max_depth': 2, 'n_estimators': 100}, {'max_depth': 2, 'n_estimators': 150}, {'max_depth': 2, 'n_estimators': 20
0}, {'max_depth': 2, 'n_estimators': 300}, {'max_depth': 2, 'n_estimators': 500}, {'max_depth': 3, 'n_estimators':
10}, {'max_depth': 3, 'n_estimators': 50}, {'max_depth': 3, 'n_estimators': 100}, {'max_depth': 3, 'n_estimators':
150}, {'max_depth': 3, 'n_estimators': 200}, {'max_depth': 3, 'n_estimators': 300}, {'max_depth': 3, 'n_estimator
s': 500}, {'max_depth': 4, 'n_estimators': 10}, {'max_depth': 4, 'n_estimators': 50}, {'max_depth': 4, 'n_estimato
rs': 100}, {'max_depth': 4, 'n_estimators': 150}, {'max_depth': 4, 'n_estimators': 200}, {'max_depth': 4, 'n_estim
ators': 300}, {'max_depth': 4, 'n_estimators': 500}, {'max_depth': 5, 'n_estimators': 10}, {'max_depth': 5, 'n_est
imators': 50}, {'max_depth': 5, 'n_estimators': 100}, {'max_depth': 5, 'n_estimators': 150}, {'max_depth': 5, 'n_e
stimators': 200}, {'max_depth': 5, 'n_estimators': 300}, {'max_depth': 5, 'n_estimators': 500}, {'max_depth': 6,
'n_estimators': 10}, {'max_depth': 6, 'n_estimators': 50}, {'max_depth': 6, 'n_estimators': 100}, {'max_depth': 6,
'n_estimators': 150}, {'max_depth': 6, 'n_estimators': 200}, {'max_depth': 6, 'n_estimators': 300}, {'max_depth':
6, 'n_estimators': 500}, {'max_depth': 7, 'n_estimators': 10}, {'max_depth': 7, 'n_estimators': 50}, {'max_depth':
7, 'n_estimators': 100}, {'max_depth': 7, 'n_estimators': 150}, {'max_depth': 7, 'n_estimators': 200}, {'max_dept
h': 7, 'n_estimators': 300}, {'max_depth': 7, 'n_estimators': 500}, {'max_depth': 8, 'n_estimators': 10}, {'max_de
pth': 8, 'n_estimators': 50}, {'max_depth': 8, 'n_estimators': 100}, {'max_depth': 8, 'n_estimators': 150}, {'max_
depth': 8, 'n_estimators': 200}, {'max_depth': 8, 'n_estimators': 300}, {'max_depth': 8, 'n_estimators': 500}], 's
plit0_test_score': array([0.64216591, 0.67874014, 0.68367267, 0.67708394, 0.68837924,
        0.68504848, 0.68487505, 0.66158926, 0.67681704, 0.69147286,
        0.68521937, 0.68842987, 0.68843262, 0.68985231, 0.65958346,
        0.68055732, 0.68787077, 0.69124183, 0.69153404, 0.69227416,
        0.69570514, 0.66024584, 0.68966053, 0.69111313, 0.69630095,
        0.69340776, 0.69819619, 0.69484687, 0.66313196, 0.6882172 ,
        0.69670013, 0.69779428, 0.69404303, 0.69397341, 0.6977698 ,
        0.65150584, 0.68323741, 0.69285204, 0.69451162, 0.69306344,
        0.69922325, 0.69577877, 0.64541534, 0.67682253, 0.69299952,
        0.69418143, 0.69694908, 0.69301323, 0.69809492]), 'split1_test_score': array([0.65248377, 0.65829554, 0.660
40112, 0.66382328, 0.66712494,
        0.65977418, 0.6720922 , 0.65132863, 0.67430962, 0.66657797,
        0.67141334, 0.67213694, 0.67185037, 0.66998155, 0.66800691,
        0.6713245 , 0.67516891, 0.67596953, 0.6725594 , 0.68100664,
        0.68057446, 0.66210872, 0.67682565, 0.67773769, 0.68174564,
        0.68235254, 0.68107712, 0.6819383 , 0.64250345, 0.67104996,
        0.67795356, 0.68381725, 0.68434966, 0.68019082, 0.68428656,
        0.65318879, 0.66978277, 0.68361699, 0.68953955, 0.68172032,
        0.68221854, 0.68621868, 0.63866284, 0.67490766, 0.67764125,
        0.67960397, 0.68294804, 0.68549213, 0.68368198]), 'split2_test_score': array([0.65031657, 0.69834615, 0.692
86484, 0.69356016, 0.69440488,
        0.69777365, 0.69906827, 0.67379156, 0.6974014 , 0.7058537 ,
        0.70373672, 0.70409968, 0.70247523, 0.70490262, 0.67602703,
        0.70591933, 0.70570557, 0.70991146, 0.71285185, 0.70648888,
        0.70902347, 0.6796888 , 0.702375  , 0.70272234, 0.71038499,
        0.71158128, 0.71246947, 0.710703  , 0.66931451, 0.70616771,
        0.70711562, 0.71379153, 0.71466432, 0.71324583, 0.71343342,
        0.67130445, 0.7094797 , 0.7076991 , 0.71069456, 0.71180306,
        0.71290207, 0.71585744, 0.66456924, 0.69561869, 0.70895637,
        0.71055339, 0.71258005, 0.71068697, 0.71275583]), 'split3_test_score': array([0.65603802, 0.66939132, 0.687
47933, 0.68236119, 0.68615496,
        0.68996602, 0.68862814, 0.64253194, 0.6923548 , 0.69162508,
        0.68775408, 0.69501178, 0.69282074, 0.69430465, 0.6547683 ,
        0.6960536 , 0.69730349, 0.69750143, 0.6998959 , 0.69596392,
        0.69790617, 0.66941675, 0.69559653, 0.69925165, 0.70503283,
        0.7000586 , 0.69938924, 0.70330793, 0.65941375, 0.6945315 ,
        0.69751472, 0.70404798, 0.70291057, 0.7024708 , 0.70520333,
        0.66651846, 0.69713699, 0.69743622, 0.70366603, 0.7049193 ,
        0.70473592, 0.70394965, 0.66599386, 0.69345402, 0.70010292,
        0.69615785, 0.70188437, 0.70086365, 0.70302051]), 'split4_test_score': array([0.64902018, 0.67503688, 0.689
17986, 0.68977389, 0.6923146 ,
        0.69082951, 0.68922971, 0.65288584, 0.68354709, 0.68651579,
        0.69120363, 0.6945135 , 0.69406058, 0.69526999, 0.66559866,
        0.69672022, 0.70191084, 0.70734208, 0.69866583, 0.70343036,
        0.69802954, 0.66393971, 0.69455448, 0.70163981, 0.69830607,
        0.70473546, 0.70670178, 0.70764902, 0.66168229, 0.69849218,
        0.70355246, 0.7065575 , 0.70645377, 0.70821222, 0.70898666,
        0.66974803, 0.70129969, 0.7086664 , 0.70740165, 0.70933015,
        0.70856162, 0.70872872, 0.66313992, 0.6936516 , 0.7037992 ,
        0.70565482, 0.70618527, 0.71172214, 0.7110755 ]), 'mean_test_score': array([0.65000466, 0.67596213, 0.68271
938, 0.68132007, 0.68567559,
        0.68467818, 0.68677853, 0.65642574, 0.68488577, 0.68840924,
        0.68786523, 0.69083815, 0.68992772, 0.69086204, 0.66479667,
        0.69011446, 0.69359145, 0.69639273, 0.69510117, 0.69583242,
        0.69624768, 0.66707984, 0.69180227, 0.69449257, 0.69835403,
        0.69842675, 0.69956648, 0.6996886 , 0.65920924, 0.69169137,
        0.69656707, 0.70120142, 0.70048386, 0.69961814, 0.70193558,
        0.66245251, 0.69218672, 0.69805363, 0.70116225, 0.70016671,
        0.70152797, 0.70210622, 0.65555565, 0.68689034, 0.69669949,
        0.69722991, 0.70010906, 0.700355  , 0.70172532]), 'std_test_score': array([0.0045841 , 0.01315221, 0.011544
11, 0.01045413, 0.00971532,
        0.01309681, 0.00871525, 0.01058097, 0.00884522, 0.01268249,
        0.01039929, 0.010605  , 0.01011844, 0.0115363 , 0.00727764,
        0.01243473, 0.01097136, 0.01222699, 0.01320332, 0.00898625,
        0.00911142, 0.00701041, 0.00851633, 0.00931237, 0.00968924,
        0.00999072, 0.01059393, 0.01035658, 0.00897493, 0.01184917,
        0.0100741 , 0.01009075, 0.01043467, 0.01163603, 0.01021208,
        0.00841155, 0.01406148, 0.00939687, 0.00794208, 0.01124883,
        0.01065168, 0.01028398, 0.01127761, 0.00905467, 0.01085468,
        0.01067333, 0.01000079, 0.01010954, 0.01048337]), 'rank_test_score': array([49, 42, 40, 41, 37, 39, 36, 47,
38, 33, 34, 30, 32, 29, 44, 31, 25,
        20, 23, 22, 21, 43, 27, 24, 15, 14, 13, 11, 46, 28, 19,  5,  7, 12,
         2, 45, 26, 16,  6,  9,  4,  1, 48, 35, 18, 17, 10,  8,  3]), 'split0_train_score': array([0.67680809, 0.70
510608, 0.71476884, 0.70832104, 0.71593459,
        0.71082422, 0.71264009, 0.69110755, 0.72456599, 0.73299337,
        0.72895633, 0.72981627, 0.73285217, 0.7323845 , 0.71677938,
        0.74539533, 0.75276707, 0.75584443, 0.75635276, 0.75513614,
        0.75873966, 0.73941213, 0.78307697, 0.78785957, 0.79034149,
        0.79027098, 0.79088088, 0.7929642 , 0.79024024, 0.82227393,
        0.83626629, 0.83883337, 0.83709092, 0.83815196, 0.83975092,
        0.83079817, 0.87836695, 0.88645488, 0.88484418, 0.88704505,
        0.88937218, 0.8937839 , 0.85872241, 0.9306449 , 0.93320532,
        0.93472471, 0.93687206, 0.9387218 , 0.94001493]), 'split1_train_score': array([0.68242793, 0.69712027, 0.70
495431, 0.70660405, 0.70960032,
        0.70624489, 0.71377799, 0.69588581, 0.72007043, 0.72401875,
        0.72227014, 0.72629077, 0.72854231, 0.72743375, 0.7262144 ,
        0.74179556, 0.7511875 , 0.75384292, 0.75054962, 0.75597264,
        0.7554835 , 0.7507691 , 0.77488194, 0.78589458, 0.78684703,
       0.78873709, 0.78917651, 0.79182465, 0.77722909, 0.82665945,
       0.83402367, 0.83468557, 0.83607344, 0.83801115, 0.83880279,
       0.83695976, 0.87800286, 0.88126901, 0.88704211, 0.88625737,
       0.8866281 , 0.89036201, 0.86188917, 0.92569637, 0.93187126,
       0.93489264, 0.93394814, 0.93591907, 0.93777366]), 'split2_train_score': array([0.66680359, 0.7058891 , 0.70
192592, 0.70647298, 0.70808766,
       0.70707178, 0.70982968, 0.69464582, 0.71938802, 0.73007025,
       0.72774948, 0.72520243, 0.72852166, 0.72891788, 0.710762  ,
       0.7454056 , 0.74894973, 0.75284713, 0.75748633, 0.75222856,
       0.75384511, 0.74578748, 0.77987898, 0.78415743, 0.79087195,
       0.78845584, 0.79020799, 0.79131388, 0.77944   , 0.82615291,
       0.83168668, 0.83341903, 0.83815231, 0.83764324, 0.83865244,
       0.82737574, 0.87844686, 0.88481289, 0.88749982, 0.88963228,
       0.89280932, 0.89306127, 0.866116  , 0.92423195, 0.93374649,
       0.93676496, 0.94067677, 0.94006914, 0.94226045]), 'split3_train_score': array([0.67702321, 0.69670492, 0.70
924384, 0.70355341, 0.7086803 ,
       0.70958166, 0.71021617, 0.67329221, 0.72037252, 0.72311903,
       0.72251394, 0.72944448, 0.72992372, 0.73052327, 0.70822748,
       0.75048448, 0.75159763, 0.75243523, 0.75664843, 0.75348583,
       0.75519609, 0.74937739, 0.78329305, 0.786155  , 0.78991606,
       0.79023207, 0.79085018, 0.79286969, 0.77679072, 0.82744854,
       0.83271117, 0.83518555, 0.83396944, 0.83830264, 0.83824098,
       0.83059174, 0.87953934, 0.8865623 , 0.88607784, 0.89022103,
       0.89040928, 0.89084456, 0.87467705, 0.92611811, 0.93411461,
       0.93432032, 0.93873454, 0.93863088, 0.93937345]), 'split4_train_score': array([0.65561299, 0.69620168, 0.70
506971, 0.70949607, 0.70610422,
       0.70913535, 0.70645907, 0.68305044, 0.71423573, 0.71539159,
       0.72109855, 0.72270973, 0.72223133, 0.72504094, 0.71150929,
       0.74487785, 0.75046462, 0.7536886 , 0.75055892, 0.75214007,
       0.75071352, 0.74613728, 0.77928449, 0.78504957, 0.78216638,
       0.78733216, 0.78753671, 0.78825559, 0.77821811, 0.82558859,
       0.82917479, 0.83592857, 0.83434377, 0.83628873, 0.83663199,
       0.83205744, 0.87351428, 0.88403063, 0.88744837, 0.88668058,
       0.88963048, 0.89031503, 0.86627813, 0.92104818, 0.92986195,
       0.93373531, 0.93647775, 0.93696506, 0.93806028]), 'mean_train_score': array([0.67173516, 0.70020441, 0.7071
9253, 0.70688951, 0.70968142,
       0.70857158, 0.7105846 , 0.68759636, 0.71972654, 0.7251186 ,
       0.72451769, 0.72669274, 0.72841424, 0.72886007, 0.71469851,
       0.74559177, 0.75099331, 0.75373166, 0.75431921, 0.75379265,
       0.75479558, 0.74629668, 0.78008309, 0.78582323, 0.78802858,
       0.78900563, 0.78973045, 0.7914456 , 0.78038363, 0.82562468,
       0.83277252, 0.83561042, 0.83592598, 0.83767954, 0.83841582,
       0.83155657, 0.87757406, 0.88462594, 0.88658247, 0.88796726,
       0.88976987, 0.89167335, 0.86553655, 0.9255479 , 0.93255993,
       0.93488759, 0.93734185, 0.93806119, 0.93949655]), 'std_train_score': array([0.00951082, 0.00433872, 0.00444
661, 0.00201137, 0.00333009,
       0.00167782, 0.00253568, 0.00843816, 0.003293  , 0.00610497,
       0.00319071, 0.00266741, 0.00347074, 0.00252235, 0.00639511,
       0.00279044, 0.00126529, 0.00117823, 0.00309648, 0.00153811,
       0.00259831, 0.00392874, 0.0030659 , 0.00123574, 0.00324833,
       0.00112073, 0.00125894, 0.00171297, 0.00501184, 0.00178354,
       0.00236213, 0.00180703, 0.00159166, 0.00072898, 0.00101989,
       0.00311226, 0.00209373, 0.00193664, 0.00100786, 0.00162982,
       0.00198471, 0.00145824, 0.00536974, 0.00310824, 0.00154886,
       0.00101987, 0.00226035, 0.00145486, 0.00160961])}
```

In [158]:

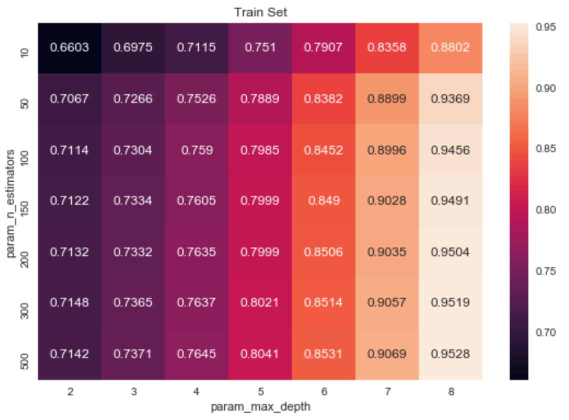
```python
import seaborn as sns; sns.set()
max_scores1 = pd.DataFrame(set4.cv_results_).groupby(['param_n_estimators', 'param_max_depth']).max().unstack()
max_scores1
```

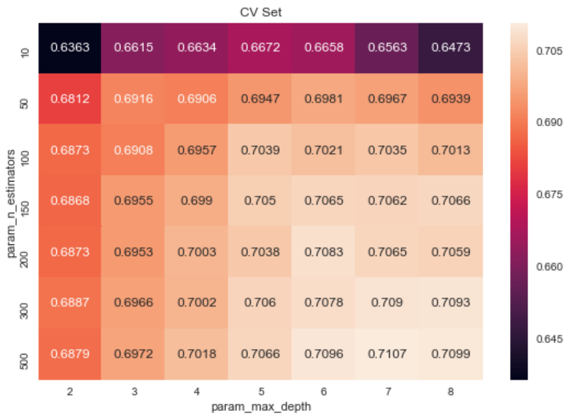
Out[158]:

	mean_fit_time							mean_score_time			...	st
param_max_depth	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**2**	**3**	**4**	**...**	**6**
param_n_estimators												
10	3.376563	3.888594	6.431222	6.857860	8.354658	10.349921	13.244183	0.341088	0.332119	0.392810	...	0.
50	10.682633	16.657853	29.285736	29.637741	38.329705	48.828008	61.584118	0.704117	0.740619	0.958778	...	0
100	20.470257	34.249607	60.867565	58.117183	75.075033	96.955518	122.796612	1.172265	1.402053	1.635213	...	0.
150	30.795445	54.640719	74.096784	87.650405	112.320628	144.456289	182.613252	1.651583	1.811756	1.802092	...	0.
200	40.524628	74.677769	90.941794	114.889166	149.932253	192.090706	246.402064	2.120528	2.381917	2.147003	...	0.
300	60.332643	122.554982	128.841663	172.286260	224.238526	289.584983	366.704142	3.073579	4.023915	3.127222	...	0
500	104.030805	206.213536	212.762625	284.349592	373.224912	488.226771	608.548005	4.995035	6.371227	5.061066	...	0.

7 rows × 140 columns

In [161]:
```
fig, ax = plt.subplots(1,2, figsize=(20,6))
sns.heatmap(max_scores1.mean_train_score, annot = True, fmt='.4g', ax=ax[0])
sns.heatmap(max_scores1.mean_test_score, annot = True, fmt='.4g', ax=ax[1])
ax[0].set_title('Train Set')
ax[1].set_title('CV Set')
plt.show()
```





In [163]:
```
print(set4.best_estimator_)
```

```
RandomForestClassifier(bootstrap=True, class_weight='balanced',
            criterion='gini', max_depth=7, max_features='auto',
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=500, n_jobs=1, oob_score=False, random_state=None,
            verbose=0, warm_start=False)
```

Training our model with best Hyperparameters

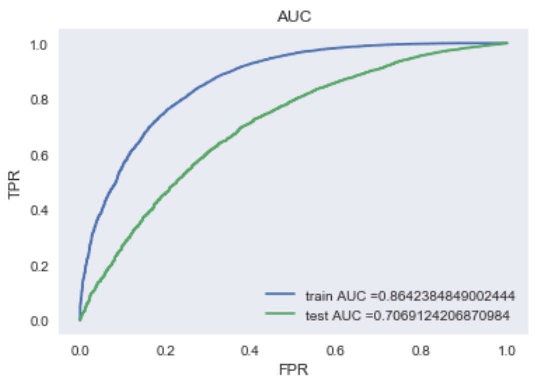
In [164]:
```
# https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
from sklearn.metrics import roc_curve, auc
model = RandomForestClassifier(max_depth = 7, n_estimators = 500)

model.fit(X_tr_TFIDF_W2V, y_train)

y_train_pred = pred_prob(model,X_tr_TFIDF_W2V)
y_test_pred = pred_prob(model,X_te_TFIDF_W2V)

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

plt.close
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("FPR")
plt.ylabel("TPR")
plt.title("AUC")
plt.grid()
plt.show()
```



Confusion matrix

```
In [165]: #our objective here is to make auc the maximum
          #so we find  the best threshold that will give the least fpr
          best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
          print("Train confusion matrix")
          print(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)))
```

```
the maximum value of tpr*(1-fpr) 0.6081086931429486 for threshold 0.829
Train confusion matrix
[[ 3467  1177]
 [ 4730 20776]]
```

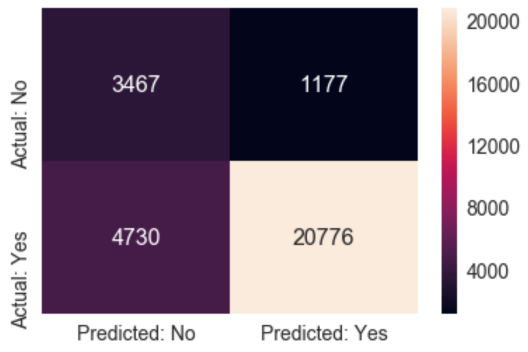
```
In [166]: #plotting confusion matrix using seaborn's heatmap
          # https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix

          print("Train data confusion matrix")

          confusion_matrix_df_train = pd.DataFrame(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)), ['Ac
          tual: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
          sns.set(font_scale=1.4)#for label size
          sns.heatmap(confusion_matrix_df_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Train data confusion matrix
```

```
Out[166]: <matplotlib.axes._subplots.AxesSubplot at 0x2c38803d4a8>
```



```
In [167]: print("Test confusion matrix")
          print(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)))
```

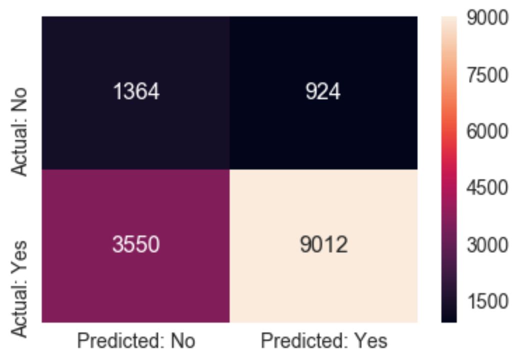
```
Test confusion matrix
[[1364  924]
 [3550 9012]]
```

```
In [168]: print("Test data confusion matrix")

          confusion_matrix_df_test = pd.DataFrame(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)), ['Actua
          l: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
          sns.set(font_scale=1.4)#for label size
          sns.heatmap(confusion_matrix_df_test, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Test data confusion matrix
```

```
Out[168]: <matplotlib.axes._subplots.AxesSubplot at 0x2c3877257b8>
```



```
In [ ]:
```

2.5 Applying GBDT

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

2.5.1 Applying XGBOOST on BOW, SET 1

In [ ]:
```
# Please write all the code with proper documentation
```

In [169]:
```
import warnings
warnings.filterwarnings('ignore')
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt
#from sklearn.grid_search import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import learning_curve, GridSearchCV
from xgboost import XGBClassifier

#n_estimators = [10, 50, 100, 150, 200, 300, 500, 1000], max_depth = [2, 3, 4, 5, 6, 7, 8, 9, 10]

clf = XGBClassifier(class_weight='balanced')
parameters ={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]}
set1 =GridSearchCV(clf, parameters, cv=5, scoring='roc_auc',return_train_score=True)
set1.fit(X_tr_BOW, y_train)
```

Out[169]:
```
GridSearchCV(cv=5, error_score='raise',
       estimator=XGBClassifier(base_score=None, booster=None, class_weight='balanced',
       colsample_bylevel=None, colsample_bynode=None,
       colsample_bytree=None, gamma=None, gpu_id=None,
       importance_type='gain', interaction_constraints=None,
       learning_rate=None, max_delta_step=None, ma...pos_weight=None, subsample=None,
       tree_method=None, validate_parameters=False, verbosity=None),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
       scoring='roc_auc', verbose=0)
```

In [170]:
```
print(set1.cv_results_)
```

```
{'mean_fit_time': array([  3.84232435,   9.4050508 ,  15.82866988,  21.06067882,
        26.64912529,  34.36430125,  56.76997375,   5.53260489,
        12.50954556,  20.48456411,  27.36482487,  35.95615177,
        48.30932741,  76.96787162,   5.36046486,  12.80814676,
        22.99769754,  36.17536006,  46.33272276,  57.97606778,
        92.10818729,   5.76398554,  14.74277344,  25.4407649 ,
        35.54414577,  48.21226854,  67.55414386, 110.30043049,
         6.45992465,  16.66363635,  29.25775747,  42.64794827,
        54.5114234 ,  79.05179667, 134.42591381,   7.2611825 ,
        19.01634479,  33.01072145,  47.89072747,  61.83483906,
        91.06886082, 148.09894938,   7.66250825,  21.12650194,
        37.28249769,  53.64194818,  71.07831941, 102.64570122,
       170.23475337]), 'std_fit_time': array([0.17780663, 1.32991613, 1.06304874, 0.36050813, 0.67260309,
       1.53422994, 3.24083598, 0.44798178, 0.37321193, 0.52752653,
       0.70008059, 2.10946   , 0.83758651, 3.19180344, 0.31027575,
       0.2308999 , 1.47656834, 1.266332  , 1.7334592 , 1.52799013,
       4.60531698, 0.04582247, 0.10286461, 0.5862431 , 0.08336988,
       2.51786896, 0.58374843, 0.81783406, 0.02642007, 0.06874055,
       0.20818553, 2.43370309, 0.68152492, 0.53617639, 6.91093466,
       0.32190937, 0.68561591, 0.12532967, 0.59572382, 0.53990805,
       0.73856385, 0.67431298, 0.04050711, 0.45971951, 0.03439483,
       0.41141913, 2.21945913, 1.55042674, 5.96975473]), 'mean_score_time': array([0.53716273, 0.36542006, 0.38636
713, 0.40372071, 0.41588788,
       0.41768327, 0.45378752, 0.68297405, 0.41489062, 0.4105022 ,
       0.39713826, 0.38098149, 0.39235063, 0.4553843 , 0.6626286 ,
       0.33191242, 0.35485239, 0.42745709, 0.44181876, 0.37759113,
       0.42626023, 0.5176157 , 0.34348192, 0.35425315, 0.35784345,
       0.35784378, 0.40272341, 0.40152626, 0.52280216, 0.36063576,
       0.36342888, 0.34787006, 0.37779026, 0.37779007, 0.41568875,
       0.46395898, 0.3448781 , 0.35006428, 0.36642065, 0.36382709,
       0.40072918, 0.44899955, 0.43902607, 0.37579598, 0.35604782,
       0.36622157, 0.38596854, 0.43324242, 0.49547501]), 'std_score_time': array([0.06289507, 0.0411293 , 0.030931
01, 0.04550041, 0.03832209,
       0.11275861, 0.04649716, 0.08806942, 0.03624585, 0.03850083,
       0.034614  , 0.04403102, 0.04812245, 0.08912063, 0.16417198,
       0.00965704, 0.02562567, 0.02898556, 0.05780059, 0.03402475,
       0.04481413, 0.02473647, 0.02149801, 0.03215918, 0.02570295,
       0.04121549, 0.01172309, 0.03514652, 0.01640487, 0.02714934,
       0.05277592, 0.01314641, 0.02828196, 0.02034593, 0.05883918,
       0.00938615, 0.01882201, 0.01542562, 0.01558718, 0.00562139,
       0.02381505, 0.05132373, 0.0173822 , 0.0565113 , 0.02003667,
       0.01185837, 0.04013749, 0.02923177, 0.05533344]), 'param_max_depth': masked_array(data=[2, 2, 2, 2, 2, 2,
2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4,
                   4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7,
                   7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'param_n_estimators': masked_array(data=[10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
150, 200,
                   300, 500, 10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
                   150, 200, 300, 500, 10, 50, 100, 150, 200, 300, 500,
                   10, 50, 100, 150, 200, 300, 500, 10, 50, 100, 150, 200,
                   300, 500],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'params': [{'max_depth': 2, 'n_estimators': 10}, {'max_depth': 2, 'n_estimators': 50},
{'max_depth': 2, 'n_estimators': 100}, {'max_depth': 2, 'n_estimators': 150}, {'max_depth': 2, 'n_estimators': 20
0}, {'max_depth': 2, 'n_estimators': 300}, {'max_depth': 2, 'n_estimators': 500}, {'max_depth': 3, 'n_estimators':
10}, {'max_depth': 3, 'n_estimators': 50}, {'max_depth': 3, 'n_estimators': 100}, {'max_depth': 3, 'n_estimators':
150}, {'max_depth': 3, 'n_estimators': 200}, {'max_depth': 3, 'n_estimators': 300}, {'max_depth': 3, 'n_estimator
s': 500}, {'max_depth': 4, 'n_estimators': 10}, {'max_depth': 4, 'n_estimators': 50}, {'max_depth': 4, 'n_estimato
rs': 100}, {'max_depth': 4, 'n_estimators': 150}, {'max_depth': 4, 'n_estimators': 200}, {'max_depth': 4, 'n_estim
ators': 300}, {'max_depth': 4, 'n_estimators': 500}, {'max_depth': 5, 'n_estimators': 10}, {'max_depth': 5, 'n_est
imators': 50}, {'max_depth': 5, 'n_estimators': 100}, {'max_depth': 5, 'n_estimators': 150}, {'max_depth': 5, 'n_e
stimators': 200}, {'max_depth': 5, 'n_estimators': 300}, {'max_depth': 5, 'n_estimators': 500}, {'max_depth': 6,
'n_estimators': 10}, {'max_depth': 6, 'n_estimators': 50}, {'max_depth': 6, 'n_estimators': 100}, {'max_depth': 6,
'n_estimators': 150}, {'max_depth': 6, 'n_estimators': 200}, {'max_depth': 6, 'n_estimators': 300}, {'max_depth':
6, 'n_estimators': 500}, {'max_depth': 7, 'n_estimators': 10}, {'max_depth': 7, 'n_estimators': 50}, {'max_depth':
7, 'n_estimators': 100}, {'max_depth': 7, 'n_estimators': 150}, {'max_depth': 7, 'n_estimators': 200}, {'max_dept
h': 7, 'n_estimators': 300}, {'max_depth': 7, 'n_estimators': 500}, {'max_depth': 8, 'n_estimators': 10}, {'max_de
pth': 8, 'n_estimators': 50}, {'max_depth': 8, 'n_estimators': 100}, {'max_depth': 8, 'n_estimators': 150}, {'max_
depth': 8, 'n_estimators': 200}, {'max_depth': 8, 'n_estimators': 300}, {'max_depth': 8, 'n_estimators': 500}], 's
plit0_test_score': array([0.68224433, 0.72594012, 0.73572395, 0.7345739 , 0.7364068 ,
       0.73710061, 0.72972291, 0.69293991, 0.73007694, 0.73187671,
       0.7325451 , 0.73073214, 0.72914577, 0.72473152, 0.70620905,
       0.72811249, 0.73124904, 0.72428286, 0.72319515, 0.72249195,
       0.72099419, 0.70543275, 0.73310905, 0.73085757, 0.72676116,
       0.72260039, 0.71984192, 0.71764835, 0.71217718, 0.7209252 ,
       0.71483301, 0.7121699 , 0.71350774, 0.71441031, 0.71557978,
       0.69805009, 0.71442593, 0.71664482, 0.71749613, 0.71360553,
       0.71572536, 0.70932989, 0.71080148, 0.71168412, 0.71223088,
       0.70933685, 0.7103292 , 0.71559286, 0.71775954]), 'split1_test_score': array([0.68737467, 0.72928702, 0.736
94408, 0.73877017, 0.73840151,
       0.73833852, 0.73572712, 0.69277495, 0.73256684, 0.73044986,
       0.73138258, 0.72955112, 0.72981194, 0.72987483, 0.70458862,
       0.73477899, 0.73066532, 0.7309387 , 0.72890581, 0.72558453,
       0.72122723, 0.7074561 , 0.72558801, 0.73056772, 0.72953825,
       0.72690722, 0.72519646, 0.72472661, 0.71083204, 0.72543249,
       0.72398951, 0.72294242, 0.7209299 , 0.71750384, 0.71940895,
       0.71492704, 0.71932507, 0.71355919, 0.71217657, 0.7108656 ,
       0.71184369, 0.71423162, 0.71152831, 0.72150451, 0.71895937,
       0.71324751, 0.71154583, 0.71245333, 0.7171776 ]), 'split2_test_score': array([0.69613369, 0.74384927, 0.756
31132, 0.75819047, 0.75771641,
       0.75269281, 0.75068735, 0.71215083, 0.74408361, 0.75359725,
       0.75337757, 0.7503513 , 0.74805907, 0.74157002, 0.71184274,
       0.74672266, 0.75035151, 0.74804282, 0.74596804, 0.74514411,
       0.73672736, 0.71378931, 0.74414323, 0.74160125, 0.74424694,
       0.7415601 , 0.73985145, 0.73462653, 0.71286134, 0.73810998,
       0.73565674, 0.73354472, 0.73048299, 0.72579618, 0.72424686,
       0.71914517, 0.73363209, 0.73175092, 0.73235054, 0.7364505 ,
       0.73392625, 0.73890797, 0.71278464, 0.7353363 , 0.73190708,
       0.72751538, 0.73023937, 0.73171811, 0.73019938]), 'split3_test_score': array([0.67508081, 0.72142379, 0.732
32997, 0.7342412 , 0.73580488,
       0.73479324, 0.73891831, 0.69204597, 0.73311487, 0.73491341,
       0.73406204, 0.7325315 , 0.73592189, 0.73316678, 0.6958659 ,
       0.73244519, 0.73582324, 0.73762442, 0.73803623, 0.73611392,
       0.72855541, 0.69739782, 0.73015454, 0.73060792, 0.72706453,
       0.72440554, 0.72396155, 0.72361284, 0.69972308, 0.72253747,
       0.7205394 , 0.72249663, 0.72376319, 0.72161709, 0.71744813,
       0.70642568, 0.73037959, 0.73075142, 0.72753047, 0.72637755,
       0.72588344, 0.73108968, 0.70075255, 0.72916031, 0.72518401,
       0.73096666, 0.73021668, 0.72724042, 0.72876105]), 'split4_test_score': array([0.68101896, 0.73959245, 0.747
08971, 0.7518383 , 0.754103  ,
       0.75458761, 0.75063195, 0.70703365, 0.75017111, 0.75193135,
       0.75281005, 0.75263291, 0.74543309, 0.74356828, 0.70787918,
       0.73919023, 0.74246218, 0.73904774, 0.73844241, 0.73629769,
       0.73538879, 0.71314321, 0.74579275, 0.74843464, 0.74628644,
       0.74235169, 0.73465691, 0.729335  , 0.71435558, 0.73439148,
       0.73351236, 0.73438366, 0.73019003, 0.72311643, 0.72176781,
       0.71259375, 0.73726405, 0.73836139, 0.73789263, 0.73774803,
       0.74011625, 0.74041823, 0.71301161, 0.73142648, 0.73521102,
       0.73624192, 0.73408833, 0.7340407 , 0.73803596]), 'mean_test_score': array([0.68437053, 0.73201808, 0.74167
943, 0.74352223, 0.74448593,
       0.74350198, 0.74113683, 0.69938859, 0.73800201, 0.74055305,
       0.7408348 , 0.73915907, 0.73767381, 0.73458166, 0.70527704,
       0.73624955, 0.73810989, 0.73598682, 0.73490902, 0.73312598,
       0.72857812, 0.70744358, 0.73575709, 0.73641324, 0.73477882,
       0.73156433, 0.72870117, 0.72598948, 0.70998977, 0.72827888,
       0.72570559, 0.72510673, 0.72377422, 0.72048848, 0.7196901 ,
       0.71022787, 0.72700459, 0.72621283, 0.72548859, 0.72500864,
       0.72549819, 0.72679445, 0.70977564, 0.72582169, 0.72469771,
       0.72346077, 0.72328309, 0.72420847, 0.72638603]), 'std_test_score': array([0.00706326, 0.00841408, 0.008816
66, 0.00972737, 0.00943591,
       0.00837651, 0.00831674, 0.00849186, 0.00776261, 0.01008729,
       0.01004638, 0.01013946, 0.00781879, 0.0070821 , 0.00528755,
       0.00634335, 0.00743675, 0.00799857, 0.00797296, 0.00816573,
       0.0066985 , 0.00596172, 0.00790996, 0.00735048, 0.00864079,
       0.00859728, 0.00738972, 0.00570296, 0.00525743, 0.00676993,
       0.00784637, 0.00819918, 0.00631937, 0.00405484, 0.00306865,
       0.00734815, 0.00869227, 0.0094922 , 0.00944659, 0.01118153,
       0.01064709, 0.01275656, 0.00458398, 0.008386  , 0.00837973,
       0.01039171, 0.0101862 , 0.00865642, 0.00793857]), 'rank_test_score': array([49, 20,  4,  2,  1,  3,  5, 48,
10,  7,  6,  8, 11, 18, 47, 13,  9,
       14, 16, 19, 23, 46, 15, 12, 17, 21, 22, 29, 44, 24, 31, 34, 38, 41,
       42, 43, 25, 28, 33, 35, 32, 26, 45, 30, 36, 39, 40, 37, 27]), 'split0_train_score': array([0.6952915 , 0.76
848328, 0.8007226 , 0.82277427, 0.84025178,
       0.86610881, 0.90380725, 0.72503745, 0.81028364, 0.85306492,
       0.88295366, 0.90589098, 0.93558437, 0.96924016, 0.75317304,
       0.85063053, 0.90123313, 0.93512771, 0.95347176, 0.97856737,
       0.99535522, 0.78718526, 0.89865411, 0.94476757, 0.96980433,
       0.98301726, 0.99485591, 0.99962516, 0.81525661, 0.93350242,
       0.97093778, 0.98629808, 0.99483937, 0.99919062, 0.99999377,
       0.8526611 , 0.9582356 , 0.98722312, 0.99673897, 0.99919882,
       0.99996494, 0.99999996, 0.87764278, 0.97964276, 0.99648345,
       0.99944958, 0.99992461, 0.99999979, 1.        ]), 'split1_train_score': array([0.70333386, 0.76801851, 0.79
947479, 0.82183499, 0.8382788 ,
       0.86460117, 0.9036328 , 0.72893612, 0.80789175, 0.8521552 ,
       0.87994222, 0.9015261 , 0.93408276, 0.96936957, 0.74952382,
       0.85230331, 0.9015659 , 0.93110019, 0.9528519 , 0.97681781,
       0.99460117, 0.7825595 , 0.89078275, 0.94071287, 0.96422949,
       0.98086865, 0.99424337, 0.999646  , 0.82033869, 0.92977012,
       0.97073545, 0.9859366 , 0.99326384, 0.9989732 , 0.99999256,
          0.84522344, 0.95780251, 0.98621527, 0.99653519, 0.9989301 ,
          0.99996339, 1.        , 0.87602816, 0.97489431, 0.99391762,
          0.99885524, 0.9998386 , 0.99999887, 1.        ]), 'split2_train_score': array([0.69882108, 0.7655466 , 0.79
571904, 0.81758143, 0.83429443,
          0.86354541, 0.90321559, 0.72452771, 0.80372866, 0.84604136,
          0.87376803, 0.89660997, 0.93056613, 0.96665871, 0.74770436,
          0.84743744, 0.89684929, 0.9287299 , 0.94805392, 0.97349859,
          0.99342639, 0.77409955, 0.88579297, 0.93575045, 0.96103401,
          0.97830758, 0.99298197, 0.9995203 , 0.81225371, 0.92604573,
          0.96686211, 0.98603791, 0.99377937, 0.99922508, 0.99998315,
          0.84385097, 0.9537116 , 0.98312868, 0.99459078, 0.99835045,
          0.99986428, 0.99999996, 0.87420579, 0.97393198, 0.99378933,
          0.99899471, 0.99979387, 0.99999728, 0.99999999]), 'split3_train_score': array([0.69845139, 0.76630707, 0.79
748871, 0.81867267, 0.83555787,
          0.86370662, 0.90520288, 0.72737804, 0.80772099, 0.85051675,
          0.88030789, 0.90308476, 0.93424323, 0.9709756 , 0.7507415 ,
          0.84892891, 0.9017088 , 0.93257922, 0.95064348, 0.97517321,
          0.99461815, 0.78085523, 0.89083164, 0.94092988, 0.96510852,
          0.98034148, 0.99384513, 0.99966588, 0.81281291, 0.92475088,
          0.97006018, 0.98645373, 0.99467423, 0.99930298, 0.99999861,
          0.84749234, 0.9586267 , 0.98731812, 0.9957925 , 0.99897678,
          0.99996558, 1.        , 0.87394488, 0.97574026, 0.99516033,
          0.99917777, 0.99989386, 0.99999962, 1.        ]), 'split4_train_score': array([0.69570386, 0.76360829, 0.79
515657, 0.81758652, 0.83538085,
          0.86214852, 0.90045136, 0.72388023, 0.80446913, 0.84862357,
          0.88084412, 0.9034609 , 0.93332626, 0.96895315, 0.75339334,
          0.85097895, 0.90030565, 0.93162656, 0.95350551, 0.9764327 ,
          0.99461071, 0.77913905, 0.89513627, 0.9416217 , 0.96793027,
          0.98141765, 0.99399772, 0.99943658, 0.81509546, 0.93198542,
          0.97196221, 0.98941093, 0.99579439, 0.99941099, 0.99999385,
          0.84833888, 0.96012859, 0.98684732, 0.99658119, 0.99917166,
          0.99996955, 1.        , 0.88374482, 0.97907312, 0.99454301,
          0.99894172, 0.99984348, 0.99999876, 1.        ]), 'mean_train_score': array([0.69832034, 0.76639275, 0.7977
1234, 0.81968998, 0.83675275,
          0.86402211, 0.90326197, 0.72595191, 0.80681883, 0.85008036,
          0.87956318, 0.90211454, 0.93356055, 0.96903944, 0.75090721,
          0.85005583, 0.90033256, 0.93183272, 0.95170531, 0.97609794,
          0.99452233, 0.78076772, 0.89223955, 0.9407565 , 0.96562132,
          0.98079053, 0.99398482, 0.99957878, 0.81515148, 0.92921092,
          0.97011155, 0.98682745, 0.99447024, 0.99922057, 0.99999239,
          0.84751335, 0.957701  , 0.9861465 , 0.99604773, 0.99892556,
          0.99994555, 0.99999998, 0.87711329, 0.97665649, 0.99477875,
          0.9990838 , 0.99985888, 0.99999886, 1.        ]), 'std_train_score': array([2.87830350e-03, 1.76010832e-03,
2.13181315e-03, 2.19177683e-03,
          2.18747474e-03, 1.30588901e-03, 1.55587079e-03, 1.90240777e-03,
          2.40998633e-03, 2.52343852e-03, 3.07955029e-03, 3.08746186e-03,
          1.66469406e-03, 1.38371292e-03, 2.16860951e-03, 1.69466979e-03,
          1.80887127e-03, 2.07946281e-03, 2.10408405e-03, 1.69352382e-03,
          6.19352493e-04, 4.27831858e-03, 4.36349398e-03, 2.89584982e-03,
          3.03553450e-03, 1.53129845e-03, 6.08446928e-04, 8.70583155e-05,
          2.85612733e-03, 3.35672304e-03, 1.73541458e-03, 1.30470377e-03,
          8.79087154e-04, 1.45022554e-04, 5.06074296e-06, 3.02780360e-03,
          2.14296608e-03, 1.55787961e-03, 7.98483620e-04, 3.06147975e-04,
          4.06867508e-05, 1.93884058e-08, 3.57568140e-03, 2.28585374e-03,
          9.82617977e-04, 2.11182664e-04, 4.56462228e-05, 8.89490424e-07,
          5.27672639e-09])}
```

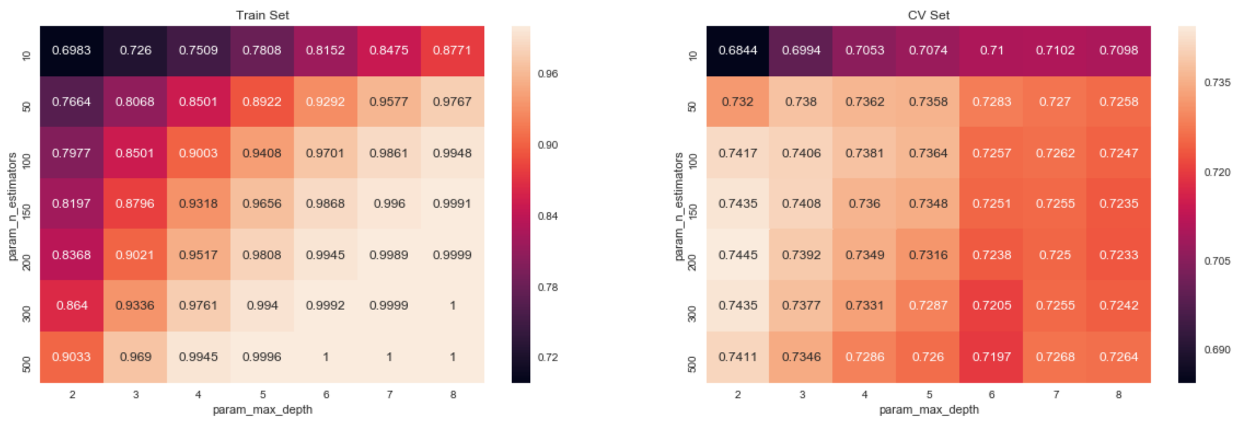
In [171]:
```
import seaborn as sns; sns.set()
max_scores1 = pd.DataFrame(set1.cv_results_).groupby(['param_n_estimators', 'param_max_depth']).max().unstack()
max_scores1
```

Out[171]:

	mean_fit_time							mean_score_time			...	std_te
param_max_depth	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**2**	**3**	**4**	**...**	**6**
param_n_estimators												
10	3.842324	5.532605	5.360465	5.763986	6.459925	7.261182	7.662508	0.537163	0.682974	0.662629	...	0.005
50	9.405051	12.509546	12.808147	14.742773	16.663636	19.016345	21.126502	0.365420	0.414891	0.331912	...	0.006
100	15.828670	20.484564	22.997698	25.440765	29.257757	33.010721	37.282498	0.386367	0.410502	0.354852	...	0.007
150	21.060679	27.364825	36.175360	35.544146	42.647948	47.890727	53.641948	0.403721	0.397138	0.427457	...	0.008
200	26.649125	35.956152	46.332723	48.212269	54.511423	61.834839	71.078319	0.415888	0.380981	0.441819	...	0.006
300	34.364301	48.309327	57.976068	67.554144	79.051797	91.068861	102.645701	0.417683	0.392351	0.377591	...	0.004
500	56.769974	76.967872	92.108187	110.300430	134.425914	148.098949	170.234753	0.453788	0.455384	0.426260	...	0.003

7 rows × 140 columns

```
In [172]: fig, ax = plt.subplots(1,2, figsize=(20,6))
          sns.heatmap(max_scores1.mean_train_score, annot = True, fmt='.4g', ax=ax[0])
          sns.heatmap(max_scores1.mean_test_score, annot = True, fmt='.4g', ax=ax[1])
          ax[0].set_title('Train Set')
          ax[1].set_title('CV Set')
          plt.show()
```



```
In [173]: print(set1.best_estimator_)

          XGBClassifier(base_score=0.5, booster=None, class_weight='balanced',
                 colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
                 gamma=0, gpu_id=-1, importance_type='gain',
                 interaction_constraints=None, learning_rate=0.300000012,
                 max_delta_step=0, max_depth=2, min_child_weight=1, missing=nan,
                 monotone_constraints=None, n_estimators=200, n_jobs=0,
                 num_parallel_tree=1, objective='binary:logistic', random_state=0,
                 reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
                 tree_method=None, validate_parameters=False, verbosity=None)
```

```
In [174]: max_d = set1.best_params_['max_depth']
          n_est = set1.best_params_['n_estimators']
```

Training our model with best Hyperparameters

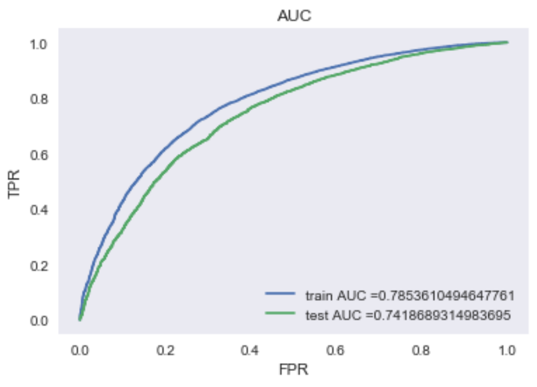
```
In [176]: # https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
          from sklearn.metrics import roc_curve, auc
          from sklearn.ensemble import GradientBoostingClassifier
          model = GradientBoostingClassifier(max_depth = max_d , n_estimators = n_est)

          model.fit(X_tr_BOW, y_train)

          y_train_pred = pred_prob(model,X_tr_BOW)
          y_test_pred = pred_prob(model,X_te_BOW)

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

          plt.close
          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("FPR")
          plt.ylabel("TPR")
          plt.title("AUC")
          plt.grid()
          plt.show()
```



Confusion matrix

In [177]:
```
#our objective here is to make auc the maximum
#so we find  the best threshold that will give the least fpr
best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")
print(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)))
```

```
the maximum value of tpr*(1-fpr) 0.5175654064068829 for threshold 0.835
Train confusion matrix
[[ 3344  1300]
 [ 7173 18333]]
```

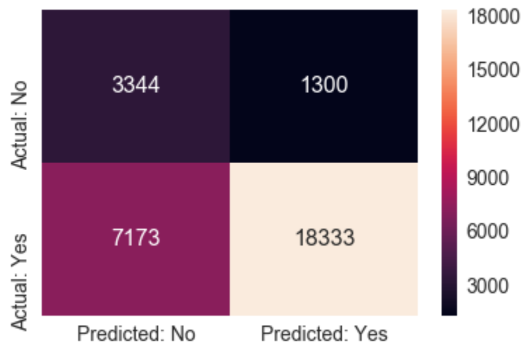
In [178]:
```
#plotting confusion matrix using seaborn's heatmap
# https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix

print("Train data confusion matrix")

confusion_matrix_df_train = pd.DataFrame(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)), ['Ac
tual: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Train data confusion matrix
```

Out[178]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c39885ff98>
```



In [179]:
```
print("Test confusion matrix")
print(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)))
```

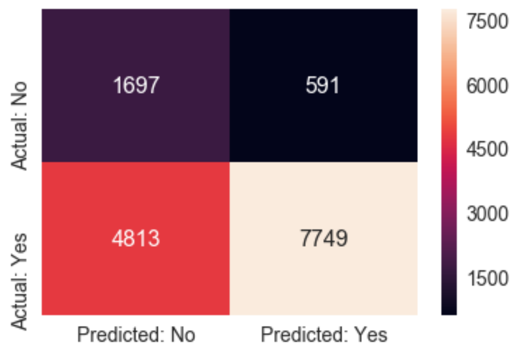
```
Test confusion matrix
[[1697  591]
 [4813 7749]]
```

In [180]:
```
print("Test data confusion matrix")

confusion_matrix_df_test = pd.DataFrame(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)), ['Actua
l: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_test, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Test data confusion matrix
```

Out[180]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c398915908>
```



2.5.2 Applying XGBOOST on TFIDF, SET 2

In [181]:
```
# Please write all the code with proper documentation
```

In [182]:
```
import warnings
warnings.filterwarnings('ignore')
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt
#from sklearn.grid_search import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import learning_curve, GridSearchCV
from xgboost import XGBClassifier

#n_estimators = [10, 50, 100, 150, 200, 300, 500, 1000], max_depth = [2, 3, 4, 5, 6, 7, 8, 9, 10]

clf = XGBClassifier(class_weight='balanced')
parameters ={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]}
set2 =GridSearchCV(clf, parameters, cv=5, scoring='roc_auc',return_train_score=True)
set2.fit(X_tr_TFIDF, y_train)
```

Out[182]:
```
GridSearchCV(cv=5, error_score='raise',
       estimator=XGBClassifier(base_score=None, booster=None, class_weight='balanced',
       colsample_bylevel=None, colsample_bynode=None,
       colsample_bytree=None, gamma=None, gpu_id=None,
       importance_type='gain', interaction_constraints=None,
       learning_rate=None, max_delta_step=None, ma...pos_weight=None, subsample=None,
       tree_method=None, validate_parameters=False, verbosity=None),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
       scoring='roc_auc', verbose=0)
```

In [183]:
```
print(set2.cv_results_)
```

```
{'mean_fit_time': array([  6.20540586,  13.83320036,  22.98772478,  32.25374517,
        41.3508182 ,  60.7429595 ,  97.63709579,   6.74635816,
        17.86642041,  31.69723334,  44.96255884,  58.67149854,
        85.64436684, 140.72287459,   7.88531289,  22.53992205,
        40.19630489,  57.82875199,  75.5485651 , 111.54570079,
       182.36112399,   8.8519269 ,  26.83663163,  49.40268517,
        70.68616858,  93.78479805, 137.17835402, 224.12025127,
         9.82173362,  30.78866339,  57.03647065,  83.33414445,
       109.27597022, 160.59712596, 266.09041381,  10.80829554,
        34.94415107,  65.49804277,  95.55885291, 125.56760244,
       186.67339268, 301.24061341,  11.49027114,  39.60847726,
        73.93428206, 108.14539371, 141.31130099, 210.38079228,
       339.7578095 ]), 'std_fit_time': array([0.63752742, 0.77116663, 0.36997205, 0.87610681, 0.62571369,
       0.82293503, 0.61692052, 0.06869409, 0.61078798, 0.79114296,
       0.56766296, 0.7120891 , 0.61699251, 1.34817852, 0.03359467,
       0.30024141, 0.39802436, 0.55285794, 0.61131142, 0.97704933,
       0.73735067, 0.05242002, 0.64925244, 0.69701516, 0.37649544,
       2.46646631, 0.87172225, 2.09130183, 0.16485065, 0.0460857 ,
       0.1899844 , 1.17133361, 1.34127467, 0.60288647, 5.09876577,
       0.31970744, 0.21346953, 0.83639106, 0.66562088, 0.64371018,
       1.25134913, 1.43857427, 0.07018854, 0.58008512, 1.63351462,
       0.57791883, 0.61454973, 1.4633432 , 2.13381024]), 'mean_score_time': array([0.59002094, 0.32532997, 0.35185
995, 0.34447904, 0.36223173,
       0.34088783, 0.35425239, 0.44580832, 0.32792373, 0.33051596,
       0.35305624, 0.33929286, 0.34627433, 0.36701894, 0.39115438,
       0.36422625, 0.36402726, 0.35485172, 0.36482458, 0.3456758 ,
       0.38955855, 0.35525012, 0.33570228, 0.36442566, 0.34767108,
       0.35345459, 0.40292244, 0.45298858, 0.36163301, 0.33590164,
       0.35923939, 0.35744438, 0.36562223, 0.39933319, 0.45298896,
       0.35983777, 0.33330832, 0.36801615, 0.35006375, 0.42386751,
       0.40930591, 0.43344102, 0.33550296, 0.38118067, 0.37040954,
       0.37360091, 0.40970483, 0.41848164, 0.45139341]), 'std_score_time': array([0.0795876 , 0.05722745, 0.016693
23, 0.02978408, 0.03834948,
       0.01470399, 0.03791634, 0.02789018, 0.009128  , 0.01284364,
       0.02268198, 0.00888461, 0.02931198, 0.01013171, 0.02384768,
       0.03792209, 0.04371961, 0.02066594, 0.05199286, 0.01905538,
       0.04323755, 0.02324789, 0.01521179, 0.02799362, 0.01127641,
       0.00835823, 0.01528324, 0.06751092, 0.05387684, 0.02023757,
       0.02466005, 0.01764575, 0.02317263, 0.03625511, 0.03329637,
       0.07641519, 0.01060008, 0.02696178, 0.01355792, 0.04227539,
       0.03384248, 0.04057439, 0.017892  , 0.04458161, 0.03073691,
       0.03891718, 0.03517284, 0.03160522, 0.03500966]), 'param_max_depth': masked_array(data=[2, 2, 2, 2, 2, 2,
2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4,
                   4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7,
                   7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'param_n_estimators': masked_array(data=[10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
150, 200,
                   300, 500, 10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
                   150, 200, 300, 500, 10, 50, 100, 150, 200, 300, 500,
                   10, 50, 100, 150, 200, 300, 500, 10, 50, 100, 150, 200,
                   300, 500],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'params': [{'max_depth': 2, 'n_estimators': 10}, {'max_depth': 2, 'n_estimators': 50},
{'max_depth': 2, 'n_estimators': 100}, {'max_depth': 2, 'n_estimators': 150}, {'max_depth': 2, 'n_estimators': 20
0}, {'max_depth': 2, 'n_estimators': 300}, {'max_depth': 2, 'n_estimators': 500}, {'max_depth': 3, 'n_estimators':
10}, {'max_depth': 3, 'n_estimators': 50}, {'max_depth': 3, 'n_estimators': 100}, {'max_depth': 3, 'n_estimators':
150}, {'max_depth': 3, 'n_estimators': 200}, {'max_depth': 3, 'n_estimators': 300}, {'max_depth': 3, 'n_estimator
s': 500}, {'max_depth': 4, 'n_estimators': 10}, {'max_depth': 4, 'n_estimators': 50}, {'max_depth': 4, 'n_estimato
rs': 100}, {'max_depth': 4, 'n_estimators': 150}, {'max_depth': 4, 'n_estimators': 200}, {'max_depth': 4, 'n_estim
ators': 300}, {'max_depth': 4, 'n_estimators': 500}, {'max_depth': 5, 'n_estimators': 10}, {'max_depth': 5, 'n_est
imators': 50}, {'max_depth': 5, 'n_estimators': 100}, {'max_depth': 5, 'n_estimators': 150}, {'max_depth': 5, 'n_e
stimators': 200}, {'max_depth': 5, 'n_estimators': 300}, {'max_depth': 5, 'n_estimators': 500}, {'max_depth': 6,
'n_estimators': 10}, {'max_depth': 6, 'n_estimators': 50}, {'max_depth': 6, 'n_estimators': 100}, {'max_depth': 6,
'n_estimators': 150}, {'max_depth': 6, 'n_estimators': 200}, {'max_depth': 6, 'n_estimators': 300}, {'max_depth':
6, 'n_estimators': 500}, {'max_depth': 7, 'n_estimators': 10}, {'max_depth': 7, 'n_estimators': 50}, {'max_depth':
7, 'n_estimators': 100}, {'max_depth': 7, 'n_estimators': 150}, {'max_depth': 7, 'n_estimators': 200}, {'max_dept
h': 7, 'n_estimators': 300}, {'max_depth': 7, 'n_estimators': 500}, {'max_depth': 8, 'n_estimators': 10}, {'max_de
pth': 8, 'n_estimators': 50}, {'max_depth': 8, 'n_estimators': 100}, {'max_depth': 8, 'n_estimators': 150}, {'max_
depth': 8, 'n_estimators': 200}, {'max_depth': 8, 'n_estimators': 300}, {'max_depth': 8, 'n_estimators': 500}], 's
plit0_test_score': array([0.68794135, 0.72667856, 0.7294019 , 0.7319129 , 0.73198452,
       0.73058519, 0.72675767, 0.70464146, 0.72974032, 0.7319956 ,
       0.73086136, 0.72876062, 0.72757491, 0.7211653 , 0.71096088,
          0.72336763, 0.72656389, 0.72390025, 0.72199794, 0.71838014,
          0.71507374, 0.71320909, 0.72721825, 0.72282119, 0.7196648 ,
          0.71658901, 0.71307617, 0.70650601, 0.71194447, 0.73040575,
          0.72514441, 0.72061095, 0.71842286, 0.71679778, 0.71308809,
          0.71620745, 0.73173356, 0.72359496, 0.72414393, 0.72037444,
          0.71479398, 0.71086625, 0.70491046, 0.7172747 , 0.71390744,
          0.71059271, 0.70710235, 0.70571092, 0.70696932]), 'split1_test_score': array([0.6860429 , 0.73048816, 0.738
86671, 0.73772107, 0.73645989,
          0.72920968, 0.73079868, 0.70354638, 0.73170735, 0.73157493,
          0.72996293, 0.72841698, 0.72436197, 0.72140913, 0.70090459,
          0.72623268, 0.72162321, 0.71694473, 0.71959644, 0.71696425,
          0.71477331, 0.70474225, 0.72071856, 0.7209491 , 0.71768279,
          0.7109538 , 0.70916707, 0.7041497 , 0.70403859, 0.71450616,
          0.70736726, 0.70821484, 0.70801869, 0.70547734, 0.70896091,
          0.70923608, 0.71757474, 0.71180855, 0.71078752, 0.70932671,
          0.70998721, 0.71074394, 0.70428633, 0.71001127, 0.70768264,
          0.70549011, 0.71050876, 0.71248931, 0.71616553]), 'split2_test_score': array([0.6956036 , 0.74306796, 0.748
58272, 0.74812469, 0.74880218,
          0.74595507, 0.74644474, 0.70882131, 0.74652472, 0.74707159,
          0.74433895, 0.74079525, 0.73896304, 0.73232491, 0.71880047,
          0.74016017, 0.73927293, 0.74100954, 0.7418033 , 0.74064489,
          0.73699726, 0.7157457 , 0.74008558, 0.73744389, 0.73617079,
          0.73510386, 0.73187901, 0.72982745, 0.70993214, 0.73457768,
          0.73068115, 0.73209922, 0.73243822, 0.73132898, 0.7288252 ,
          0.72135068, 0.73838685, 0.73617786, 0.73627029, 0.73252168,
          0.73196131, 0.73103018, 0.71590355, 0.72956178, 0.72663479,
          0.72454661, 0.72558337, 0.73101013, 0.73346949]), 'split3_test_score': array([0.675967  , 0.72777832, 0.736
99589, 0.7392705 , 0.7404478 ,
          0.74282708, 0.73711544, 0.69224559, 0.72621612, 0.73015707,
          0.72794756, 0.72786653, 0.72967731, 0.72176692, 0.69785183,
          0.72714145, 0.72850941, 0.72852302, 0.72999933, 0.72549569,
          0.72191073, 0.69991669, 0.72739468, 0.71930988, 0.71778355,
          0.71813976, 0.71551506, 0.71356774, 0.69872568, 0.71901307,
          0.72009066, 0.71726095, 0.71757232, 0.71259387, 0.72028564,
          0.70557737, 0.72015745, 0.72225786, 0.71945094, 0.71853078,
          0.71490752, 0.72008201, 0.69672824, 0.7153628 , 0.71506178,
          0.71211042, 0.71621312, 0.71624182, 0.71777448]), 'split4_test_score': array([0.68451895, 0.74019737, 0.744
02247, 0.74495915, 0.74402712,
          0.74658778, 0.74525702, 0.70179201, 0.74385917, 0.74354388,
          0.74160324, 0.74100582, 0.74256115, 0.74125689, 0.71263167,
          0.73929237, 0.7301538 , 0.73121491, 0.73213036, 0.72990928,
          0.72745836, 0.71713193, 0.73781161, 0.73844357, 0.7366283 ,
          0.73604715, 0.73444218, 0.73515039, 0.71132572, 0.72706512,
          0.72685672, 0.72297458, 0.72384651, 0.71979801, 0.72026413,
          0.71566343, 0.72503522, 0.72157145, 0.72376645, 0.72172229,
          0.72209778, 0.72277273, 0.70962632, 0.72349056, 0.72439111,
          0.72778421, 0.72568608, 0.72685376, 0.73227887]), 'mean_test_score': array([0.68601487, 0.73364162, 0.73957
345, 0.74039723, 0.7403439 ,
          0.73903243, 0.7372741 , 0.70220945, 0.73560907, 0.73686823,
          0.73494245, 0.73336863, 0.73262718, 0.72758396, 0.70822983,
          0.73123833, 0.72922453, 0.72831825, 0.72910514, 0.72627847,
          0.72324227, 0.710149  , 0.73064538, 0.72779301, 0.72558548,
          0.72336607, 0.72081519, 0.71783931, 0.70719334, 0.72511367,
          0.72202798, 0.72023203, 0.72005954, 0.7171991 , 0.71828456,
          0.71360702, 0.72657778, 0.7230822 , 0.72288384, 0.72049514,
          0.71874932, 0.71909863, 0.70629083, 0.71914002, 0.7175352 ,
          0.71610424, 0.71701812, 0.71846049, 0.7213307 ]), 'std_test_score': array([0.0063064 , 0.00670286, 0.006505
05, 0.0056736 , 0.00583255,
          0.00757958, 0.00775096, 0.00549313, 0.00806354, 0.00700672,
          0.00667885, 0.00615631, 0.00694774, 0.00803203, 0.0077451 ,
          0.00704635, 0.00578283, 0.00797971, 0.007901  , 0.0085946 ,
          0.00833312, 0.00668394, 0.00722938, 0.00836757, 0.00885839,
          0.01025547, 0.01031256, 0.01246961, 0.00507554, 0.00736373,
          0.008082  , 0.00776832, 0.00801771, 0.00854606, 0.00682592,
          0.00555943, 0.00762489, 0.00776938, 0.00824221, 0.00741283,
          0.00765439, 0.00767346, 0.00633869, 0.00676193, 0.00701644,
          0.00856347, 0.00761371, 0.00928531, 0.01012721]), 'rank_test_score': array([49,  9,  3,  1,  2,  4,  5, 48,
7,  6,  8, 10, 11, 18, 45, 12, 14,
          16, 15, 20, 24, 44, 13, 17, 21, 23, 29, 38, 46, 22, 27, 31, 32, 40,
          37, 43, 19, 25, 26, 30, 35, 34, 47, 33, 39, 42, 41, 36, 28]), 'split0_train_score': array([0.7060963 , 0.78
227148, 0.81847065, 0.84662878, 0.86782471,
          0.90066443, 0.94100651, 0.73911529, 0.83056136, 0.88203216,
          0.91346902, 0.93661882, 0.9644383 , 0.99003457, 0.76526132,
          0.87621169, 0.93064287, 0.95910481, 0.97626371, 0.99281292,
          0.9995276 , 0.79976739, 0.91841608, 0.96060373, 0.98220792,
          0.99198657, 0.99892231, 0.99999549, 0.838716  , 0.95042045,
          0.9841931 , 0.99524042, 0.99850851, 0.99995511, 1.        ,
          0.87328542, 0.97551903, 0.99483036, 0.99911642, 0.99991028,
          0.99999996, 1.        , 0.90747754, 0.98761944, 0.99802707,
          0.99979855, 0.99999474, 1.        , 1.        ]), 'split1_train_score': array([0.70798278, 0.78025762, 0.81
982527, 0.84902834, 0.86958923,
          0.90062873, 0.94110865, 0.73449807, 0.82771215, 0.88217257,
          0.91051342, 0.93299889, 0.96147723, 0.98775055, 0.76345271,
          0.87498055, 0.92860643, 0.95842141, 0.97461333, 0.99088433,
          0.99927392, 0.80430593, 0.91856824, 0.96446864, 0.98479684,
          0.99324788, 0.99893435, 0.99999406, 0.83002097, 0.95059865,
          0.98551487, 0.99603265, 0.99891859, 0.99997145, 1.        ,
       0.87241914, 0.97430127, 0.99478722, 0.99904857, 0.99990519,
       0.99999865, 1.        , 0.90212386, 0.98490836, 0.99807237,
       0.99982138, 0.99999388, 1.        , 1.        ]), 'split2_train_score': array([0.70331516, 0.77692227, 0.81
774215, 0.84558403, 0.86581346,
       0.89771491, 0.9400615 , 0.73316219, 0.82372001, 0.87797127,
       0.90892708, 0.93160935, 0.96075091, 0.98816745, 0.76444192,
       0.87491855, 0.92768964, 0.95578525, 0.97436686, 0.99173795,
       0.99931696, 0.80304041, 0.9180666 , 0.96126936, 0.98229092,
       0.99236867, 0.99854602, 0.99999098, 0.83757024, 0.95139257,
       0.9840311 , 0.99458904, 0.99843302, 0.99991636, 1.        ,
       0.86870556, 0.97333662, 0.99402132, 0.99866877, 0.99984762,
       0.9999993 , 1.        , 0.89906165, 0.98762029, 0.99826688,
       0.99986844, 0.99999367, 0.99999999, 1.        ]), 'split3_train_score': array([0.70365233, 0.78257753, 0.82
023565, 0.8487298 , 0.87077524,
       0.90224831, 0.94289419, 0.73773967, 0.83032315, 0.87814938,
       0.91404717, 0.93636252, 0.96402458, 0.98969228, 0.76915647,
       0.87631206, 0.92932494, 0.95638871, 0.97372495, 0.99077105,
       0.99942194, 0.79835918, 0.91220569, 0.95997427, 0.9826133 ,
       0.99181053, 0.99866354, 0.9999947 , 0.83729613, 0.95157323,
       0.98396089, 0.99484768, 0.99860315, 0.99994959, 0.99999999,
       0.86978496, 0.97088783, 0.99386166, 0.99879263, 0.99988258,
       0.999999  , 1.        , 0.90759684, 0.98666028, 0.99855825,
       0.99989115, 0.99999441, 1.        , 1.        ]), 'split4_train_score': array([0.70457988, 0.77862603, 0.81
796943, 0.84554649, 0.86795125,
       0.90030409, 0.93904221, 0.73085821, 0.82697347, 0.87668126,
       0.90874073, 0.93055839, 0.96142873, 0.98903115, 0.76235436,
       0.87630558, 0.93033683, 0.95949649, 0.97567608, 0.99170949,
       0.99928237, 0.79436468, 0.91374688, 0.96369077, 0.98287211,
       0.9920378 , 0.99858486, 0.99998797, 0.83414824, 0.95067242,
       0.98394148, 0.99478487, 0.99818742, 0.9998881 , 1.        ,
       0.8704519 , 0.97109398, 0.9947288 , 0.99880409, 0.999759  ,
       0.99999722, 1.        , 0.89813617, 0.98411346, 0.99752203,
       0.99977887, 0.99999029, 1.        , 1.        ]), 'mean_train_score': array([0.70512529, 0.78013099, 0.8188
4863, 0.84710349, 0.86839078,
       0.90031209, 0.94082261, 0.73507469, 0.82785803, 0.87940133,
       0.91113948, 0.93362959, 0.96242395, 0.9889352 , 0.76493336,
       0.87574569, 0.92932014, 0.95783934, 0.97492899, 0.99158315,
       0.99936456, 0.79996752, 0.9162007 , 0.96200135, 0.98295622,
       0.99229029, 0.99873022, 0.99999264, 0.83555031, 0.95093146,
       0.98432829, 0.99509893, 0.99853014, 0.99993612, 1.        ,
       0.87092939, 0.97302775, 0.99444587, 0.9988861 , 0.99986093,
       0.99999883, 1.        , 0.90287921, 0.98618437, 0.99808932,
       0.99983168, 0.9999934 , 1.        , 1.        ]), 'std_train_score': array([1.72347394e-03, 2.15146852e-03,
1.00177731e-03, 1.50388189e-03,
       1.68986147e-03, 1.46415617e-03, 1.27720130e-03, 3.00657289e-03,
       2.50246792e-03, 2.26328540e-03, 2.23258118e-03, 2.46234984e-03,
       1.50368908e-03, 8.69921903e-04, 2.32496529e-03, 6.51307581e-04,
       1.09098228e-03, 1.48391261e-03, 9.17056174e-04, 7.34861227e-04,
       9.70685240e-05, 3.52939224e-03, 2.68239077e-03, 1.76294461e-03,
       9.50277553e-04, 5.11722039e-04, 1.66176640e-04, 2.79107903e-06,
       3.15242866e-03, 4.61189969e-04, 5.99863699e-04, 5.12564098e-04,
       2.38154066e-04, 2.99565572e-05, 2.63836317e-09, 1.68839264e-03,
       1.80235679e-03, 4.16159506e-04, 1.68603750e-04, 5.55522314e-05,
       9.12052095e-07, 0.00000000e+00, 4.02595042e-03, 1.43280607e-03,
       3.39823365e-04, 4.21432144e-05, 1.59797973e-06, 2.63836317e-09,
       0.00000000e+00])}
```

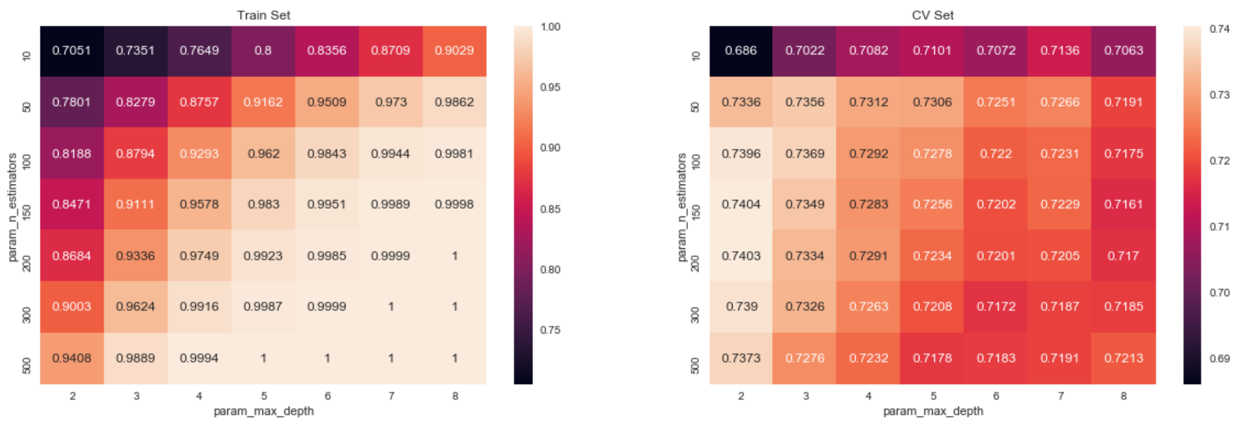
In [184]:
```
import seaborn as sns; sns.set()
max_scores2 = pd.DataFrame(set2.cv_results_).groupby(['param_n_estimators', 'param_max_depth']).max().unstack()
max_scores2
```

Out[184]:

	mean_fit_time							mean_score_time			...	std
param_max_depth	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**2**	**3**	**4**	...	**6**
param_n_estimators												
10	6.205406	6.746358	7.885313	8.851927	9.821734	10.808296	11.490271	0.590021	0.445808	0.391154	...	0.0
50	13.833200	17.866420	22.539922	26.836632	30.788663	34.944151	39.608477	0.325330	0.327924	0.364226	...	0.0
100	22.987725	31.697233	40.196305	49.402685	57.036471	65.498043	73.934282	0.351860	0.330516	0.364027	...	0.0
150	32.253745	44.962559	57.828752	70.686169	83.334144	95.558853	108.145394	0.344479	0.353056	0.354852	...	0.0
200	41.350818	58.671499	75.548565	93.784798	109.275970	125.567602	141.311301	0.362232	0.339293	0.364825	...	0.0
300	60.742959	85.644367	111.545701	137.178354	160.597126	186.673393	210.380792	0.340888	0.346274	0.345676	...	0.0
500	97.637096	140.722875	182.361124	224.120251	266.090414	301.240613	339.757809	0.354252	0.367019	0.389559	...	0.0

7 rows × 140 columns

In [185]:
```
fig, ax = plt.subplots(1,2, figsize=(20,6))
sns.heatmap(max_scores2.mean_train_score, annot = True, fmt='.4g', ax=ax[0])
sns.heatmap(max_scores2.mean_test_score, annot = True, fmt='.4g', ax=ax[1])
ax[0].set_title('Train Set')
ax[1].set_title('CV Set')
plt.show()
```



In [186]:
```
print(set2.best_estimator_)
```

```
XGBClassifier(base_score=0.5, booster=None, class_weight='balanced',
       colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
       gamma=0, gpu_id=-1, importance_type='gain',
       interaction_constraints=None, learning_rate=0.300000012,
       max_delta_step=0, max_depth=2, min_child_weight=1, missing=nan,
       monotone_constraints=None, n_estimators=150, n_jobs=0,
       num_parallel_tree=1, objective='binary:logistic', random_state=0,
       reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
       tree_method=None, validate_parameters=False, verbosity=None)
```

In [187]:
```
max_d = set2.best_params_['max_depth']
n_est = set2.best_params_['n_estimators']
```

Training our model with best Hyperparameters

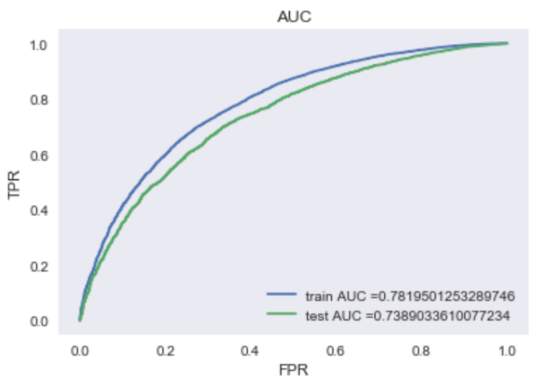
In [188]:
```
# https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
from sklearn.metrics import roc_curve, auc
from sklearn.ensemble import GradientBoostingClassifier
model = GradientBoostingClassifier(max_depth = max_d , n_estimators = n_est)

model.fit(X_tr_TFIDF, y_train)

y_train_pred = pred_prob(model,X_tr_TFIDF)
y_test_pred = pred_prob(model,X_te_TFIDF)

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

plt.close
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("FPR")
plt.ylabel("TPR")
plt.title("AUC")
plt.grid()
plt.show()
```



Confusion matrix

In [189]:
```
#our objective here is to make auc the maximum
#so we find  the best threshold that will give the least fpr
best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")
print(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)))
```

```
the maximum value of tpr*(1-fpr) 0.5068760568606479 for threshold 0.842
Train confusion matrix
[[ 3373  1271]
 [ 7706 17800]]
```

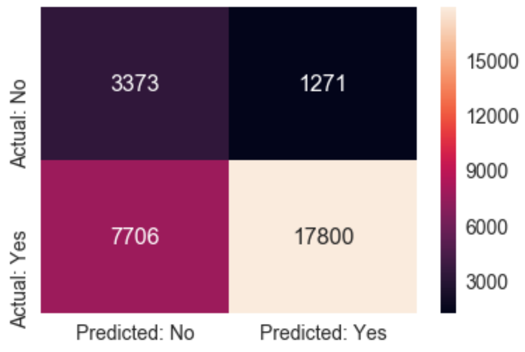
In [190]:
```
#plotting confusion matrix using seaborn's heatmap
# https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix

print("Train data confusion matrix")

confusion_matrix_df_train = pd.DataFrame(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)), ['Ac
tual: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Train data confusion matrix
```

Out[190]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c3880fe3c8>
```



In [191]:
```
print("Test confusion matrix")
print(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)))
```

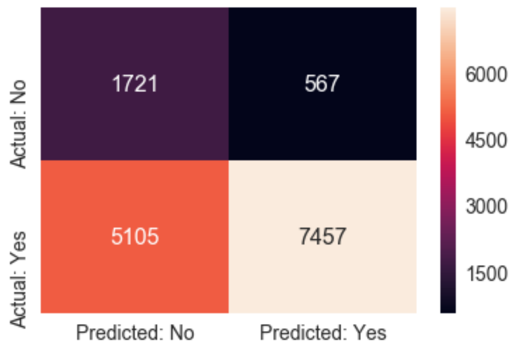
```
Test confusion matrix
[[1721  567]
 [5105 7457]]
```

In [192]:
```
print("Test data confusion matrix")

confusion_matrix_df_test = pd.DataFrame(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)), ['Actua
l: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_test, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Test data confusion matrix
```

Out[192]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c398aa9fd0>
```



In [ ]:

2.5.3 Applying XGBOOST on AVG W2V, SET 3

In [193]:
```
# Please write all the code with proper documentation
```

In [194]:
```
import warnings
warnings.filterwarnings('ignore')
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt
#from sklearn.grid_search import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import learning_curve, GridSearchCV
from xgboost import XGBClassifier

#n_estimators = [10, 50, 100, 150, 200, 300, 500, 1000], max_depth = [2, 3, 4, 5, 6, 7, 8, 9, 10]

clf = XGBClassifier(class_weight='balanced')
parameters ={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]}
set3 =GridSearchCV(clf, parameters, cv=5, scoring='roc_auc',return_train_score=True)
set3.fit(X_tr_AVG_W2V, y_train)
```

Out[194]:
```
GridSearchCV(cv=5, error_score='raise',
       estimator=XGBClassifier(base_score=None, booster=None, class_weight='balanced',
       colsample_bylevel=None, colsample_bynode=None,
       colsample_bytree=None, gamma=None, gpu_id=None,
       importance_type='gain', interaction_constraints=None,
       learning_rate=None, max_delta_step=None, ma...pos_weight=None, subsample=None,
       tree_method=None, validate_parameters=False, verbosity=None),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
       scoring='roc_auc', verbose=0)
```

In [195]:

```
print(set3.cv_results_)
```

```
{'mean_fit_time': array([  19.27864342,   47.58255305,   83.20867219,  117.41878643,
        152.62663283,  222.51114612,  359.90273104,   22.05920777,
         62.44948101,  114.03205285,  188.03859687, 2292.41762762,
        328.57044487,  552.16559982,   26.01861939,   80.92837677,
        149.83191419,  218.33452291,  283.53077412,  422.03069911,
        708.13122778,   29.21307554,   96.95731335,  184.58955021,
        270.58519268,  354.59493318,  529.50576563,  858.58774056,
         32.81963134,  114.82910514,  217.20094638,  319.71600556,
        422.60186582,  625.10689721, 1000.86305742,   35.77313313,
        132.08736744,  253.39576006,  368.35871034,  487.1687921 ,
        700.49232454, 1083.74441366,   38.56446152,  148.96623073,
        283.55569863,  412.61178131,  533.26013932,  758.98310766,
       1132.78866611]), 'std_fit_time': array([1.97872227e-01, 6.46769251e-01, 1.08047054e+00, 1.88429111e+00,
       1.29623262e+00, 1.81974236e+00, 1.93491423e+00, 1.44497880e-01,
       3.27315822e-01, 2.08912435e+00, 1.13704845e+01, 4.09543143e+03,
       7.29658578e+00, 1.80173758e+01, 7.30870471e-01, 5.59679249e-01,
       1.06440083e+00, 2.28994560e+00, 2.51637348e+00, 2.88975666e+00,
       1.71159745e+01, 4.49685894e-01, 8.29444065e-01, 8.86773500e-01,
       1.93835503e+00, 2.54843187e+00, 4.76653260e+00, 3.04810683e+00,
       7.73980331e-01, 1.45693686e+00, 1.73386781e+00, 9.02395369e-01,
       4.17519980e+00, 6.26485960e+00, 9.32291971e+00, 6.23596215e-01,
       1.88852042e+00, 1.64934077e+00, 1.12023914e+00, 4.55921952e+00,
       5.54393471e+00, 8.41964608e+00, 9.96083838e-01, 1.94495407e+00,
       2.55663404e+00, 5.24741476e+00, 3.16193724e+00, 8.92680471e+00,
       9.22849998e+00]), 'mean_score_time': array([1.9695334 , 2.03475981, 2.26434517, 2.11354818, 2.40676503,
       2.05390744, 2.09659333, 2.016608  , 2.14985151, 2.17059617,
       2.16401324, 2.36557651, 2.25098119, 2.48150959, 2.04832268,
       2.09041095, 2.12292352, 2.049719  , 2.31002336, 2.39320059,
       2.27990389, 2.20510364, 2.1841599 , 2.08841591, 2.21527696,
       2.25457101, 2.15563622, 2.42431746, 2.12611485, 2.10237947,
       2.25058298, 2.12192669, 2.09739223, 2.28489079, 2.28528986,
       2.12930727, 2.0772459 , 2.17877431, 2.43089981, 2.13489161,
       2.39240332, 2.62956944, 2.09619517, 2.3648756 , 2.0393476 ,
       2.18396053, 2.41873293, 2.32278862, 2.39778833]), 'std_score_time': array([0.07982026, 0.04019878, 0.240198
82, 0.14410187, 0.39084014,
       0.04384032, 0.20765519, 0.05098698, 0.28268421, 0.466258  ,
       0.05344305, 0.22315496, 0.39782543, 0.16326633, 0.06709399,
       0.1026334 , 0.12327277, 0.13679359, 0.50270965, 0.29202876,
       0.06528082, 0.18500394, 0.08352419, 0.07818048, 0.05193957,
       0.11335287, 0.12678698, 0.2451503 , 0.10397387, 0.06223898,
       0.37535307, 0.1148579 , 0.09207958, 0.24997949, 0.11908448,
       0.11737175, 0.15967505, 0.16501259, 0.33578261, 0.09353571,
       0.15258358, 0.41985474, 0.0331398 , 0.26136399, 0.04420457,
       0.05311776, 0.1633771 , 0.22297649, 0.18761129]), 'param_max_depth': masked_array(data=[2, 2, 2, 2, 2, 2,
2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4,
                   4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7,
                   7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'param_n_estimators': masked_array(data=[10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
150, 200,
                   300, 500, 10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
                   150, 200, 300, 500, 10, 50, 100, 150, 200, 300, 500,
                   10, 50, 100, 150, 200, 300, 500, 10, 50, 100, 150, 200,
                   300, 500],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'params': [{'max_depth': 2, 'n_estimators': 10}, {'max_depth': 2, 'n_estimators': 50},
{'max_depth': 2, 'n_estimators': 100}, {'max_depth': 2, 'n_estimators': 150}, {'max_depth': 2, 'n_estimators': 20
0}, {'max_depth': 2, 'n_estimators': 300}, {'max_depth': 2, 'n_estimators': 500}, {'max_depth': 3, 'n_estimators':
10}, {'max_depth': 3, 'n_estimators': 50}, {'max_depth': 3, 'n_estimators': 100}, {'max_depth': 3, 'n_estimators':
150}, {'max_depth': 3, 'n_estimators': 200}, {'max_depth': 3, 'n_estimators': 300}, {'max_depth': 3, 'n_estimator
s': 500}, {'max_depth': 4, 'n_estimators': 10}, {'max_depth': 4, 'n_estimators': 50}, {'max_depth': 4, 'n_estimato
rs': 100}, {'max_depth': 4, 'n_estimators': 150}, {'max_depth': 4, 'n_estimators': 200}, {'max_depth': 4, 'n_estim
ators': 300}, {'max_depth': 4, 'n_estimators': 500}, {'max_depth': 5, 'n_estimators': 10}, {'max_depth': 5, 'n_est
imators': 50}, {'max_depth': 5, 'n_estimators': 100}, {'max_depth': 5, 'n_estimators': 150}, {'max_depth': 5, 'n_e
stimators': 200}, {'max_depth': 5, 'n_estimators': 300}, {'max_depth': 5, 'n_estimators': 500}, {'max_depth': 6,
'n_estimators': 10}, {'max_depth': 6, 'n_estimators': 50}, {'max_depth': 6, 'n_estimators': 100}, {'max_depth': 6,
'n_estimators': 150}, {'max_depth': 6, 'n_estimators': 200}, {'max_depth': 6, 'n_estimators': 300}, {'max_depth':
6, 'n_estimators': 500}, {'max_depth': 7, 'n_estimators': 10}, {'max_depth': 7, 'n_estimators': 50}, {'max_depth':
7, 'n_estimators': 100}, {'max_depth': 7, 'n_estimators': 150}, {'max_depth': 7, 'n_estimators': 200}, {'max_dept
h': 7, 'n_estimators': 300}, {'max_depth': 7, 'n_estimators': 500}, {'max_depth': 8, 'n_estimators': 10}, {'max_de
pth': 8, 'n_estimators': 50}, {'max_depth': 8, 'n_estimators': 100}, {'max_depth': 8, 'n_estimators': 150}, {'max_
depth': 8, 'n_estimators': 200}, {'max_depth': 8, 'n_estimators': 300}, {'max_depth': 8, 'n_estimators': 500}], 's
plit0_test_score': array([0.6795694 , 0.71585068, 0.71734738, 0.71737397, 0.71652688,
       0.71292754, 0.70954899, 0.69644051, 0.72075051, 0.71828752,
       0.71664629, 0.71671064, 0.71092026, 0.70559045, 0.70483757,
       0.71584752, 0.70868217, 0.69804334, 0.69471743, 0.69146969,
       0.69492208, 0.70192655, 0.70337716, 0.70440295, 0.69762707,
       0.69474855, 0.69430064, 0.69798448, 0.69171622, 0.69176085,
       0.6906169 , 0.69394724, 0.69710553, 0.69993289, 0.70060687,
       0.68843768, 0.68802036, 0.68909257, 0.69225391, 0.69465382,
       0.69978056, 0.70158276, 0.68707938, 0.69231984, 0.69856583,
       0.70218722, 0.70536133, 0.70777654, 0.70953834]), 'split1_test_score': array([0.69207614, 0.72158491, 0.721
27133, 0.72040603, 0.7189906 ,
       0.71807814, 0.71412959, 0.70019956, 0.71021586, 0.70867107,
       0.70372354, 0.70252018, 0.69676106, 0.6927339 , 0.70657297,
       0.70875991, 0.70356274, 0.70124771, 0.6997695 , 0.69949401,
       0.70149292, 0.704301  , 0.69859938, 0.69375958, 0.68824703,
       0.68943498, 0.68938497, 0.69993536, 0.70952824, 0.69676886,
       0.69405807, 0.69478873, 0.699961  , 0.70614681, 0.71310159,
       0.70072469, 0.68939225, 0.69252119, 0.6982188 , 0.69890716,
       0.70140165, 0.70293104, 0.68863996, 0.68009386, 0.6856944 ,
       0.69543235, 0.69815159, 0.70237236, 0.70377967]), 'split2_test_score': array([0.70037672, 0.73850185, 0.741
21043, 0.7404172 , 0.73848476,
       0.7295201 , 0.72370885, 0.70546162, 0.73260346, 0.72856649,
       0.72935392, 0.72652917, 0.72017939, 0.70819574, 0.71231068,
       0.73154623, 0.73263173, 0.72076351, 0.71925301, 0.71700224,
       0.71832925, 0.71465451, 0.70798841, 0.70224079, 0.70028471,
       0.70397402, 0.70538587, 0.71045189, 0.71134504, 0.70615536,
       0.70837342, 0.70652634, 0.70907982, 0.71608619, 0.71919582,
       0.70458356, 0.71925353, 0.71426053, 0.71651836, 0.72050353,
       0.72226894, 0.72424274, 0.6957186 , 0.70295763, 0.71085283,
       0.71485603, 0.72071824, 0.7234609 , 0.72495273]), 'split3_test_score': array([0.69039809, 0.72729381, 0.730
48911, 0.73311613, 0.73188798,
       0.73092171, 0.72632754, 0.70220354, 0.72533014, 0.72487486,
       0.72049635, 0.71922979, 0.71337233, 0.70721037, 0.70162555,
       0.7207788 , 0.71869612, 0.72033196, 0.71385906, 0.71753876,
       0.71008365, 0.7056852 , 0.72623311, 0.71455638, 0.70754136,
       0.70472632, 0.70908087, 0.71709076, 0.70354353, 0.71874887,
       0.7131631 , 0.71296654, 0.71427076, 0.71745351, 0.72917054,
       0.6947708 , 0.69568083, 0.69502434, 0.69986784, 0.70074189,
       0.7072971 , 0.71491354, 0.70532878, 0.70273479, 0.70970065,
       0.71479779, 0.71999686, 0.72530355, 0.72678461]), 'split4_test_score': array([0.69968501, 0.74233659, 0.747
96239, 0.74201   , 0.740705  ,
       0.73291748, 0.72490128, 0.71235937, 0.73311289, 0.73177472,
       0.73080097, 0.7269258 , 0.71868536, 0.71273223, 0.71608086,
       0.73276496, 0.73074446, 0.72175651, 0.71962774, 0.71539408,
       0.71269917, 0.72148769, 0.72016759, 0.70723973, 0.70210709,
       0.70449791, 0.70976469, 0.71833954, 0.70385381, 0.70319915,
       0.69881899, 0.70217797, 0.70535907, 0.71219703, 0.71882499,
       0.70396134, 0.71159993, 0.71087249, 0.71428724, 0.72075867,
       0.72471845, 0.72963814, 0.69834156, 0.70647363, 0.71502165,
       0.72158128, 0.72441773, 0.72846982, 0.72988742]), 'mean_test_score': array([0.6924204 , 0.72911269, 0.73165
511, 0.73066385, 0.72931824,
       0.72487233, 0.71972274, 0.70333239, 0.72440216, 0.72243448,
       0.72020374, 0.71838278, 0.71198342, 0.7052923 , 0.70828515,
       0.72193892, 0.71886271, 0.71242782, 0.70944452, 0.70817896,
       0.70750483, 0.70961034, 0.71127257, 0.70443979, 0.6991613 ,
       0.69947603, 0.7015829 , 0.70875973, 0.70399697, 0.70332624,
       0.70100582, 0.70208109, 0.70515496, 0.71036288, 0.71617936,
       0.6984951 , 0.7007886 , 0.7003535 , 0.7042285 , 0.70711215,
       0.71109251, 0.71466071, 0.69502128, 0.69691548, 0.70396653,
       0.70977029, 0.71372852, 0.71747595, 0.71898788]), 'std_test_score': array([0.00755583, 0.00998884, 0.011585
46, 0.0101166 , 0.00990527,
       0.0078965 , 0.00665028, 0.00537775, 0.00846847, 0.0082148 ,
       0.00980152, 0.00888224, 0.00832901, 0.006712  , 0.005218  ,
       0.00918289, 0.01156301, 0.01049678, 0.0102947 , 0.01069747,
       0.00831011, 0.00733673, 0.01036021, 0.00676887, 0.00635042,
       0.00626399, 0.00823945, 0.00846194, 0.00686523, 0.0091941 ,
       0.00852152, 0.0071779 , 0.00616969, 0.00652648, 0.00934755,
       0.0061148 , 0.01246467, 0.01020403, 0.00949472, 0.01121306,
       0.01045798, 0.01117866, 0.00665671, 0.009652  , 0.01063522,
       0.00952582, 0.01014939, 0.01039374, 0.0103513 ]), 'rank_test_score': array([49,  4,  1,  2,  3,  5, 10, 37,
6,  7,  9, 13, 19, 31, 27,  8, 12,
       18, 25, 28, 29, 24, 20, 33, 45, 44, 40, 26, 35, 38, 41, 39, 32, 22,
       15, 46, 42, 43, 34, 30, 21, 16, 48, 47, 36, 23, 17, 14, 11]), 'split0_train_score': array([0.71356962, 0.78
04419 , 0.81851405, 0.84754631, 0.87012102,
       0.90505219, 0.94762625, 0.74623377, 0.83603628, 0.8947691 ,
       0.93280584, 0.95621876, 0.98305736, 0.99855858, 0.77300033,
       0.90437659, 0.96260347, 0.98831131, 0.99689382, 0.99993495,
       1.        , 0.81721606, 0.96016409, 0.99547828, 0.99979643,
       0.9999993 , 1.        , 1.        , 0.86491425, 0.99278072,
       0.9999891 , 1.        , 1.        , 1.        , 1.        ,
       0.91202874, 0.99978876, 1.        , 1.        , 1.        ,
       1.        , 1.        , 0.95578076, 0.99999985, 1.        ,
       1.        , 1.        , 1.        , 1.        ]), 'split1_train_score': array([0.71655368, 0.78266239, 0.82
133619, 0.8504111 , 0.87188366,
       0.90492285, 0.94603028, 0.7459786 , 0.83599312, 0.896279  ,
       0.93234719, 0.95681972, 0.98380043, 0.99841615, 0.77368419,
       0.89966596, 0.95957545, 0.98639355, 0.99632292, 0.99981117,
       1.        , 0.81782144, 0.96123579, 0.99587706, 0.99976905,
       0.99999719, 1.        , 1.        , 0.86301543, 0.99104334,
       0.99998154, 1.        , 1.        , 1.        , 1.        ,
       0.90717215, 0.99976755, 1.        , 1.        , 1.        ,
       1.        , 1.        , 0.95747986, 0.99999923, 1.        ,
       1.        , 1.        , 1.        , 1.        ]), 'split2_train_score': array([0.71846418, 0.77853479, 0.81
69577 , 0.84507854, 0.86626327,
       0.9006545 , 0.9441765 , 0.74296381, 0.83393052, 0.89416409,
       0.93290884, 0.95689001, 0.98288567, 0.99844545, 0.77702397,
       0.8988115 , 0.96113296, 0.98720573, 0.99663809, 0.99990627,
       1.        , 0.81589748, 0.95648891, 0.99553818, 0.99979466,
       0.99999723, 1.        , 1.        , 0.86762137, 0.99260471,
       0.99998578, 1.        , 1.        , 1.        , 1.        ,
       0.92638046, 0.99966788, 1.        , 1.        , 1.        ,
       1.        , 1.        , 0.95141994, 0.99999999, 1.        ,
       1.        , 1.        , 1.        , 1.        ]), 'split3_train_score': array([0.71480899, 0.77906585, 0.81
691528, 0.84377444, 0.86580349,
       0.90010471, 0.94352013, 0.74155755, 0.83131132, 0.89077731,
       0.92841528, 0.95247855, 0.9808927 , 0.99824272, 0.77172829,
       0.89642006, 0.95931877, 0.98550919, 0.99562008, 0.99984278,
       1.        , 0.81258282, 0.95931757, 0.99486323, 0.9997993 ,
       0.9999926 , 1.        , 1.        , 0.86643612, 0.99152833,
       0.99997877, 1.        , 1.        , 1.        , 1.        ,
       0.91322202, 0.9996217 , 1.        , 1.        , 1.        ,
       1.        , 1.        , 0.95223647, 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        ]), 'split4_train_score': array([0.71011708, 0.7766665 , 0.81
62679 , 0.84433997, 0.86628778,
       0.90031945, 0.94303836, 0.74106027, 0.83354831, 0.89230807,
       0.93003088, 0.95329151, 0.97949836, 0.99786983, 0.77495211,
       0.89994562, 0.95912765, 0.9848093 , 0.9958365 , 0.99980102,
       1.        , 0.81286918, 0.95746336, 0.99468734, 0.99965559,
       0.99999873, 1.        , 1.        , 0.86777079, 0.99172442,
       0.99993886, 0.99999999, 1.        , 1.        , 1.        ,
       0.91657845, 0.99962156, 1.        , 1.        , 1.        ,
       1.        , 1.        , 0.95421494, 0.99999993, 1.        ,
       1.        , 1.        , 1.        , 1.        ]), 'mean_train_score': array([0.71470271, 0.77947428, 0.8179
9822, 0.84623007, 0.86807184,
       0.90221074, 0.9448783 , 0.7435588 , 0.83416391, 0.89365951,
       0.93130161, 0.95513971, 0.9820269 , 0.99830655, 0.77407778,
       0.89984395, 0.96035166, 0.98644582, 0.99626228, 0.99985924,
       1.        , 0.81527739, 0.95893394, 0.99528882, 0.999763  ,
       0.99999701, 1.        , 1.        , 0.86595159, 0.9919363 ,
       0.99997481, 1.        , 1.        , 1.        , 1.        ,
       0.91507636, 0.99969349, 1.        , 1.        , 1.        ,
       1.        , 1.        , 0.95422639, 0.9999998 , 1.        ,
       1.        , 1.        , 1.        , 1.        ]), 'std_train_score': array([2.82481606e-03, 2.00166785e-03,
1.82524124e-03, 2.45528480e-03,
       2.46285779e-03, 2.27435993e-03, 1.70884966e-03, 2.17313363e-03,
       1.75627548e-03, 1.92254920e-03, 1.78240415e-03, 1.87338995e-03,
       1.58900496e-03, 2.40674378e-04, 1.80446074e-03, 2.58363508e-03,
       1.33030894e-03, 1.23415839e-03, 4.76961860e-04, 5.27546228e-05,
       0.00000000e+00, 2.17602257e-03, 1.73780000e-03, 4.44296954e-04,
       5.47923585e-05, 2.35600844e-06, 0.00000000e+00, 0.00000000e+00,
       1.79050408e-03, 6.58561415e-04, 1.83201867e-05, 5.27530637e-09,
       0.00000000e+00, 0.00000000e+00, 4.96506831e-17, 6.40690439e-03,
       7.14779822e-05, 0.00000000e+00, 0.00000000e+00, 4.96506831e-17,
       4.96506831e-17, 0.00000000e+00, 2.22877304e-03, 2.88175368e-07,
       0.00000000e+00, 0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
       0.00000000e+00])}
```

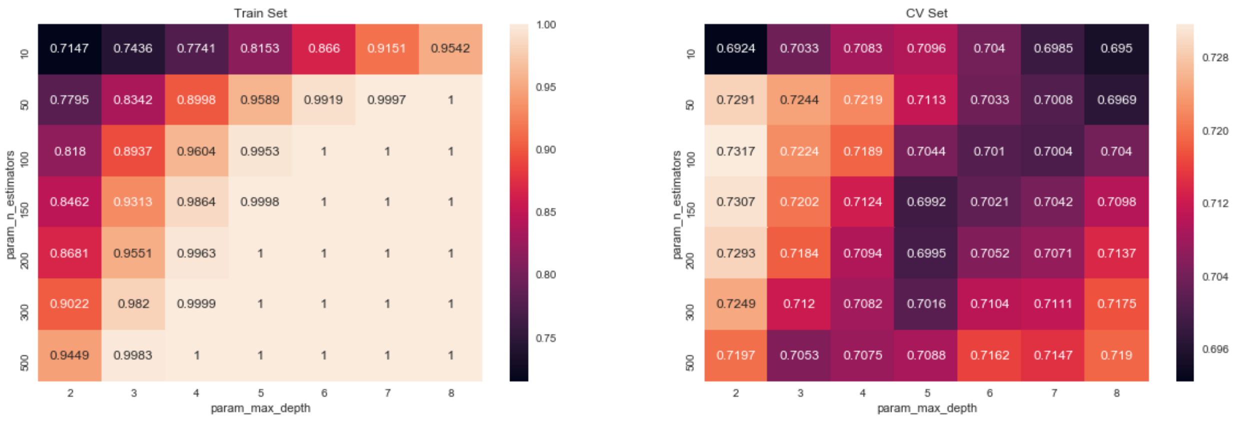
In [196]:
```python
import seaborn as sns; sns.set()
max_scores3 = pd.DataFrame(set3.cv_results_).groupby(['param_n_estimators', 'param_max_depth']).max().unstack()
max_scores3
```

Out[196]:

	mean_fit_time							**mean_score_time**			..
param_max_depth	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**2**	**3**	**4**	..
param_n_estimators											
10	19.278643	22.059208	26.018619	29.213076	32.819631	35.773133	38.564462	1.969533	2.016608	2.048323	.
50	47.582553	62.449481	80.928377	96.957313	114.829105	132.087367	148.966231	2.034760	2.149852	2.090411	.
100	83.208672	114.032053	149.831914	184.589550	217.200946	253.395760	283.555699	2.264345	2.170596	2.122924	.
150	117.418786	188.038597	218.334523	270.585193	319.716006	368.358710	412.611781	2.113548	2.164013	2.049719	.
200	152.626633	2292.417628	283.530774	354.594933	422.601866	487.168792	533.260139	2.406765	2.365577	2.310023	.
300	222.511146	328.570445	422.030699	529.505766	625.106897	700.492325	758.983108	2.053907	2.250981	2.393201	.
500	359.902731	552.165600	708.131228	858.587741	1000.863057	1083.744414	1132.788666	2.096593	2.481510	2.279904	.

7 rows × 140 columns

In [197]:
```python
fig, ax = plt.subplots(1,2, figsize=(20,6))
sns.heatmap(max_scores3.mean_train_score, annot = True, fmt='.4g', ax=ax[0])
sns.heatmap(max_scores3.mean_test_score, annot = True, fmt='.4g', ax=ax[1])
ax[0].set_title('Train Set')
ax[1].set_title('CV Set')
plt.show()
```



In [198]:
```python
print(set3.best_estimator_)
```

```
XGBClassifier(base_score=0.5, booster=None, class_weight='balanced',
       colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
       gamma=0, gpu_id=-1, importance_type='gain',
       interaction_constraints=None, learning_rate=0.300000012,
       max_delta_step=0, max_depth=2, min_child_weight=1, missing=nan,
       monotone_constraints=None, n_estimators=100, n_jobs=0,
       num_parallel_tree=1, objective='binary:logistic', random_state=0,
       reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
       tree_method=None, validate_parameters=False, verbosity=None)
```

In [199]:
```python
max_d = set3.best_params_['max_depth']
n_est = set3.best_params_['n_estimators']
```

Training our model with best Hyperparameters

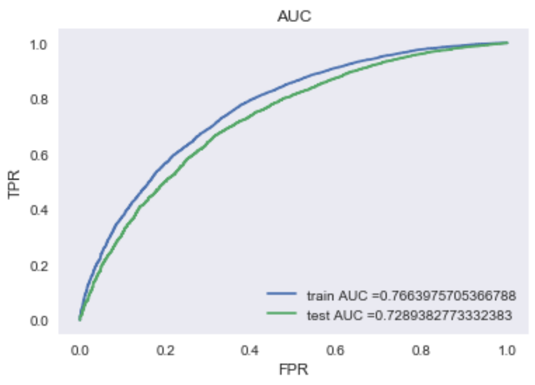
In [200]:
```python
# https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
from sklearn.metrics import roc_curve, auc
from sklearn.ensemble import GradientBoostingClassifier
model = GradientBoostingClassifier(max_depth = max_d , n_estimators = n_est)

model.fit(X_tr_AVG_W2V, y_train)

y_train_pred = pred_prob(model,X_tr_AVG_W2V)
y_test_pred = pred_prob(model,X_te_AVG_W2V)

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

plt.close
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("FPR")
plt.ylabel("TPR")
plt.title("AUC")
plt.grid()
plt.show()
```



Confusion Matrix

In [201]:
```
#our objective here is to make auc the maximum
#so we find  the best threshold that will give the least fpr
best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")
print(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)))
```

```
the maximum value of tpr*(1-fpr) 0.4875473727855019 for threshold 0.833
Train confusion matrix
[[ 3104  1540]
 [ 6901 18605]]
```

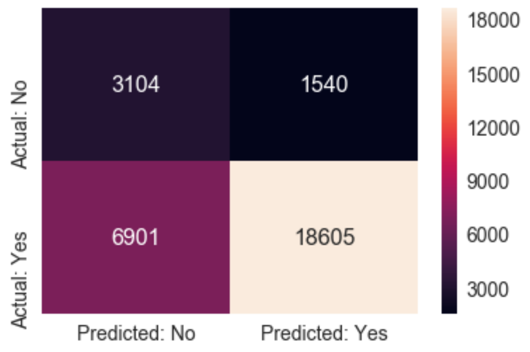
In [202]:
```
#plotting confusion matrix using seaborn's heatmap
# https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix

print("Train data confusion matrix")

confusion_matrix_df_train = pd.DataFrame(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)), ['Ac
tual: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Train data confusion matrix
```

Out[202]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c387776908>
```



In [203]:
```
print("Test confusion matrix")
print(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)))
```

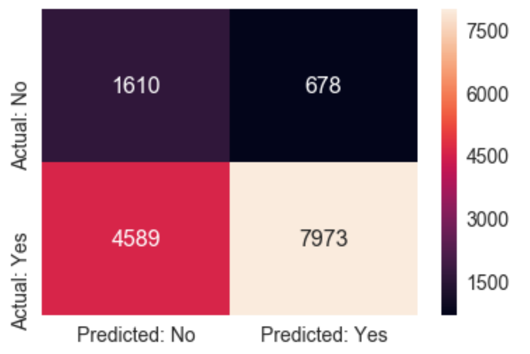
```
Test confusion matrix
[[1610  678]
 [4589 7973]]
```

In [204]:
```
print("Test data confusion matrix")

confusion_matrix_df_test = pd.DataFrame(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)), ['Actua
l: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_test, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Test data confusion matrix
```

Out[204]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c399700208>
```



In [ ]:

2.5.4 Applying XGBOOST on TFIDF W2V, SET 4

In [205]:
```
# Please write all the code with proper documentation
```

```
In [206]: import warnings
          warnings.filterwarnings('ignore')
          from sklearn.metrics import roc_auc_score
          import matplotlib.pyplot as plt
          #from sklearn.grid_search import GridSearchCV
          from sklearn.linear_model import LogisticRegression
          from sklearn.ensemble import RandomForestClassifier
          from sklearn.model_selection import learning_curve, GridSearchCV
          from xgboost import XGBClassifier

          #n_estimators = [10, 50, 100, 150, 200, 300, 500, 1000], max_depth = [2, 3, 4, 5, 6, 7, 8, 9, 10]

          clf = XGBClassifier(class_weight='balanced')
          parameters ={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]}
          set4 =GridSearchCV(clf, parameters, cv=5, scoring='roc_auc',return_train_score=True)
          set4.fit(X_tr_TFIDF_W2V, y_train)
```

```
Out[206]: GridSearchCV(cv=5, error_score='raise',
                 estimator=XGBClassifier(base_score=None, booster=None, class_weight='balanced',
                 colsample_bylevel=None, colsample_bynode=None,
                 colsample_bytree=None, gamma=None, gpu_id=None,
                 importance_type='gain', interaction_constraints=None,
                 learning_rate=None, max_delta_step=None, ma...pos_weight=None, subsample=None,
                 tree_method=None, validate_parameters=False, verbosity=None),
                 fit_params=None, iid=True, n_jobs=1,
                 param_grid={'n_estimators': [10, 50, 100, 150, 200, 300, 500], 'max_depth': [2, 3, 4, 5, 6, 7, 8]},
                 pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
                 scoring='roc_auc', verbose=0)
```

In [207]:
```
print(set4.cv_results_)
```

```
{'mean_fit_time': array([  19.48469605,   47.37709379,   82.38368616,  117.45549631,
        152.82570796,  223.17537723,  364.70648417,   21.87948871,
         63.26101732,  113.89102201,  165.12381191,  215.84577818,
        317.34195538,  521.85424166,   25.21297317,   87.6785192 ,
        148.81962237,  213.88183031,  284.80376167,  417.31221271,
        687.81263161,   28.15589647,   96.4504694 ,  179.76785836,
        263.8196857 ,  350.34808912,  518.70107446,  851.94749875,
         31.75726542,  112.16903315,  213.44559708,  315.74820976,
        416.35995812,  616.97108469, 1080.4634304 ,   37.59964519,
        135.89665718,  265.12946591,  427.3624784 ,  501.19556241,
        701.85629606, 1079.95056667,   38.20084052,  146.43100157,
        280.42188787,  404.2944171 ,  520.81182227,  737.96790586,
       1112.16820283]), 'std_fit_time': array([ 0.87990155,  0.67554766,  0.87800321,  0.68479648,  1.26395711,
        1.255631  ,  2.12955455,  0.15179577,  1.14723398,  1.11731203,
        0.89650286,  1.54894116,  0.35618541,  0.99193942,  0.18782266,
        3.57858682,  3.99376175,  1.40019698,  7.69641303,  1.13376646,
        9.76110331,  0.09394966,  1.55599129,  1.01599922,  0.89939265,
        2.32307986,  3.68629812,  4.92156663,  0.78434713,  0.99538231,
        1.24833553,  1.40370846,  2.71878782,  4.47600035, 58.46508408,
        1.24910096,  1.8463043 , 11.48124605,  5.36493279, 17.01518503,
        2.74721682,  7.27209956,  0.72766792,  1.1738526 ,  0.62927718,
        1.27655574,  1.6668454 ,  4.42273192,  7.39657824]), 'mean_score_time': array([2.04872212, 1.97711358, 2.1
8974447, 2.09200687, 2.07405438,
       2.09360118, 2.14546332, 1.83589101, 2.13150129, 2.14107461,
       2.06487918, 2.13449254, 2.18834839, 2.23023748, 2.08681974,
       2.76739936, 2.1307024 , 2.19433198, 2.57730827, 2.16600804,
       2.24659257, 2.05769768, 2.12711234, 2.06288438, 2.16800265,
       2.11654115, 2.26494389, 2.16620822, 2.04253879, 2.16042371,
       2.04553089, 2.09061055, 2.05969229, 2.33495684, 2.78358364,
       2.31979885, 2.40053988, 2.42577863, 2.54554138, 2.55592947,
       2.32418566, 2.29865351, 2.16840172, 1.88057146, 2.04652767,
       2.05231204, 2.18455863, 2.20370755, 2.34453106]), 'std_score_time': array([0.08071707, 0.0668331 , 0.234292
98, 0.17621296, 0.07333855,
       0.05722508, 0.08834063, 0.21270177, 0.09781107, 0.19752583,
       0.07607414, 0.02414632, 0.11023049, 0.06722093, 0.10226557,
       0.55446561, 0.19742174, 0.25718024, 0.28679319, 0.19852832,
       0.14978657, 0.09389204, 0.12809204, 0.09041742, 0.20595451,
       0.07930981, 0.09726354, 0.06714613, 0.02709347, 0.17884472,
       0.07908531, 0.06105609, 0.12998758, 0.42577661, 0.23597305,
       0.20978276, 0.24312423, 0.36606399, 0.26080117, 0.18435587,
       0.08700867, 0.04693785, 0.19227844, 0.2297595 , 0.30073633,
       0.29791673, 0.0860337 , 0.19366605, 0.06578584]), 'param_max_depth': masked_array(data=[2, 2, 2, 2, 2, 2,
2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4,
                   4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7,
                   7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'param_n_estimators': masked_array(data=[10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
150, 200,
                   300, 500, 10, 50, 100, 150, 200, 300, 500, 10, 50, 100,
                   150, 200, 300, 500, 10, 50, 100, 150, 200, 300, 500,
                   10, 50, 100, 150, 200, 300, 500, 10, 50, 100, 150, 200,
                   300, 500],
             mask=[False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False, False, False, False, False, False, False, False,
                   False],
       fill_value='?',
            dtype=object), 'params': [{'max_depth': 2, 'n_estimators': 10}, {'max_depth': 2, 'n_estimators': 50},
{'max_depth': 2, 'n_estimators': 100}, {'max_depth': 2, 'n_estimators': 150}, {'max_depth': 2, 'n_estimators': 20
0}, {'max_depth': 2, 'n_estimators': 300}, {'max_depth': 2, 'n_estimators': 500}, {'max_depth': 3, 'n_estimators':
10}, {'max_depth': 3, 'n_estimators': 50}, {'max_depth': 3, 'n_estimators': 100}, {'max_depth': 3, 'n_estimators':
150}, {'max_depth': 3, 'n_estimators': 200}, {'max_depth': 3, 'n_estimators': 300}, {'max_depth': 3, 'n_estimator
s': 500}, {'max_depth': 4, 'n_estimators': 10}, {'max_depth': 4, 'n_estimators': 50}, {'max_depth': 4, 'n_estimato
rs': 100}, {'max_depth': 4, 'n_estimators': 150}, {'max_depth': 4, 'n_estimators': 200}, {'max_depth': 4, 'n_estim
ators': 300}, {'max_depth': 4, 'n_estimators': 500}, {'max_depth': 5, 'n_estimators': 10}, {'max_depth': 5, 'n_est
imators': 50}, {'max_depth': 5, 'n_estimators': 100}, {'max_depth': 5, 'n_estimators': 150}, {'max_depth': 5, 'n_e
stimators': 200}, {'max_depth': 5, 'n_estimators': 300}, {'max_depth': 5, 'n_estimators': 500}, {'max_depth': 6,
'n_estimators': 10}, {'max_depth': 6, 'n_estimators': 50}, {'max_depth': 6, 'n_estimators': 100}, {'max_depth': 6,
'n_estimators': 150}, {'max_depth': 6, 'n_estimators': 200}, {'max_depth': 6, 'n_estimators': 300}, {'max_depth':
6, 'n_estimators': 500}, {'max_depth': 7, 'n_estimators': 10}, {'max_depth': 7, 'n_estimators': 50}, {'max_depth':
7, 'n_estimators': 100}, {'max_depth': 7, 'n_estimators': 150}, {'max_depth': 7, 'n_estimators': 200}, {'max_dept
h': 7, 'n_estimators': 300}, {'max_depth': 7, 'n_estimators': 500}, {'max_depth': 8, 'n_estimators': 10}, {'max_de
pth': 8, 'n_estimators': 50}, {'max_depth': 8, 'n_estimators': 100}, {'max_depth': 8, 'n_estimators': 150}, {'max_
depth': 8, 'n_estimators': 200}, {'max_depth': 8, 'n_estimators': 300}, {'max_depth': 8, 'n_estimators': 500}], 's
plit0_test_score': array([0.69395832, 0.71810755, 0.71346892, 0.71132535, 0.71106531,
       0.70598562, 0.69861679, 0.69949078, 0.71196145, 0.70531101,
       0.70213543, 0.69872829, 0.69140038, 0.68162246, 0.69914266,
       0.71090676, 0.6959426 , 0.68633261, 0.67948501, 0.68001679,
       0.67896937, 0.70077966, 0.69362908, 0.68208925, 0.67661503,
       0.67820541, 0.67775064, 0.68027956, 0.69605315, 0.68768353,
       0.67998451, 0.67792206, 0.68120187, 0.68653885, 0.69068305,
       0.69250687, 0.68206911, 0.67481736, 0.67909237, 0.68199484,
       0.68219643, 0.6831277 , 0.67953712, 0.67352996, 0.68099363,
       0.68067979, 0.68481935, 0.68744818, 0.6902748 ]), 'split1_test_score': array([0.68805943, 0.72035708, 0.720
4092 , 0.71932518, 0.71550271,
       0.71005917, 0.70722852, 0.69638765, 0.70968039, 0.70791096,
       0.70601514, 0.70160993, 0.70194409, 0.69621641, 0.70162523,
       0.7071536 , 0.69926315, 0.70017825, 0.69415229, 0.69213238,
       0.6950351 , 0.70124655, 0.69368308, 0.68795857, 0.68445675,
       0.68353279, 0.68861812, 0.69452211, 0.68918007, 0.68084468,
       0.68093014, 0.67713258, 0.68183406, 0.68812253, 0.69422573,
       0.69175507, 0.68623293, 0.68493366, 0.68822445, 0.69304526,
       0.69881969, 0.70065231, 0.67611144, 0.68652308, 0.69476658,
       0.7012378 , 0.7032941 , 0.70401422, 0.70465636]), 'split2_test_score': array([0.70455064, 0.73889752, 0.736
96983, 0.73806219, 0.73704464,
       0.73395157, 0.72675117, 0.7113787 , 0.73749791, 0.73415827,
       0.73049988, 0.72887479, 0.71849354, 0.70594814, 0.71999454,
       0.73184293, 0.72515436, 0.71895926, 0.70773687, 0.69506581,
       0.69799195, 0.71445731, 0.71372158, 0.71101331, 0.7107623 ,
       0.70243366, 0.70698373, 0.71326186, 0.70746908, 0.70675551,
       0.7010349 , 0.6984256 , 0.6985424 , 0.7038377 , 0.70755856,
       0.69847519, 0.70390058, 0.69968963, 0.70894814, 0.71036167,
       0.71530541, 0.71873959, 0.68989058, 0.69765621, 0.70244421,
       0.70580084, 0.70820365, 0.71098166, 0.71205186]), 'split3_test_score': array([0.68635912, 0.72576474, 0.726
46291, 0.72627615, 0.7251878 ,
       0.72488224, 0.71729989, 0.69853048, 0.71897963, 0.71393703,
       0.70664261, 0.70154895, 0.70141516, 0.69090951, 0.7016917 ,
       0.71820053, 0.71769218, 0.71053039, 0.70803916, 0.70439712,
       0.70123305, 0.69602163, 0.70923566, 0.7013356 , 0.69778947,
       0.69458615, 0.69824602, 0.7012398 , 0.69785436, 0.68946938,
       0.68885446, 0.69628721, 0.69738473, 0.70056875, 0.70371552,
       0.68871814, 0.6853211 , 0.693224  , 0.69784434, 0.70212747,
       0.70022204, 0.7036521 , 0.69109869, 0.69016227, 0.70064556,
       0.70209275, 0.70345279, 0.7071937 , 0.71023221]), 'split4_test_score': array([0.69532227, 0.73925815, 0.738
87674, 0.73606004, 0.73595673,
       0.72823766, 0.72260362, 0.707549  , 0.73548491, 0.73478673,
       0.72864104, 0.72109836, 0.71498373, 0.70535654, 0.71526754,
       0.72847711, 0.71819029, 0.71578648, 0.71399117, 0.70758649,
       0.70387451, 0.71575479, 0.7162376 , 0.7074289 , 0.70282101,
       0.70786154, 0.70339519, 0.70760614, 0.69583487, 0.70561553,
       0.69948506, 0.69892968, 0.70805061, 0.70756495, 0.7109103 ,
       0.70654049, 0.68909451, 0.68797014, 0.69423159, 0.70063468,
       0.70115045, 0.70499118, 0.69360238, 0.70444954, 0.70751393,
       0.71552442, 0.71630531, 0.71590679, 0.72050158]), 'mean_test_score': array([0.69364991, 0.72847631, 0.72723
668, 0.72620896, 0.72495061,
       0.72062252, 0.7144992 , 0.70266705, 0.72272008, 0.71921982,
       0.71478594, 0.71037132, 0.7056466 , 0.69600982, 0.7075438 ,
       0.7193156 , 0.71124778, 0.70635642, 0.70067975, 0.6958388 ,
       0.69541997, 0.70565149, 0.70530065, 0.69796429, 0.69448804,
       0.69332293, 0.69499789, 0.69938099, 0.69727831, 0.69407313,
       0.69005717, 0.68973873, 0.69340184, 0.69732586, 0.70141796,
       0.69559869, 0.68932341, 0.68812652, 0.69366768, 0.69763217,
       0.69953817, 0.70223185, 0.68604758, 0.69046319, 0.6972719 ,
       0.70106597, 0.703214  , 0.70510796, 0.70754236]), 'std_test_score': array([0.00642069, 0.00900704, 0.009664
35, 0.01006455, 0.01048363,
       0.01076706, 0.01028378, 0.0057681 , 0.01167125, 0.01276504,
       0.01218329, 0.0122279 , 0.00986725, 0.00915645, 0.00842062,
       0.00959928, 0.01149811, 0.0118689 , 0.01243405, 0.00975646,
       0.0087486 , 0.00794328, 0.00976944, 0.01116584, 0.01237176,
       0.01113878, 0.01061068, 0.01142199, 0.00588812, 0.010306  ,
       0.008895  , 0.01001375, 0.01038766, 0.00847139, 0.00774675,
       0.00631908, 0.00762561, 0.00833153, 0.00993336, 0.00955895,
       0.01051641, 0.01139876, 0.00690605, 0.01048544, 0.00910295,
       0.01138578, 0.0103415 , 0.00968051, 0.01002029]), 'rank_test_score': array([41,  1,  2,  3,  4,  6, 10, 21,
5,  8,  9, 12, 17, 33, 13,  7, 11,
       15, 25, 34, 36, 16, 18, 28, 38, 43, 37, 27, 31, 39, 45, 46, 42, 30,
       23, 35, 47, 48, 40, 29, 26, 22, 49, 44, 32, 24, 20, 19, 14]), 'split0_train_score': array([0.71451286, 0.77
904834, 0.81819727, 0.84466455, 0.86648696,
       0.89990607, 0.94453225, 0.74127991, 0.83226039, 0.89145469,
       0.9297588 , 0.95303642, 0.9822931 , 0.99855967, 0.77372333,
       0.89920479, 0.96112405, 0.98682374, 0.99655403, 0.99994475,
       1.        , 0.81616699, 0.95780666, 0.99606387, 0.99980673,
       0.9999998 , 1.        , 1.        , 0.86189666, 0.99281108,
       0.99997327, 1.        , 1.        , 1.        , 1.        ,
       0.91707326, 0.99961694, 1.        , 1.        , 1.        ,
       1.        , 1.        , 0.95900738, 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        ]), 'split1_train_score': array([0.71484358, 0.77963573, 0.81
558084, 0.84448199, 0.86579238,
       0.89906128, 0.94167678, 0.74078425, 0.83304802, 0.88927032,
       0.92623826, 0.9511994 , 0.98057051, 0.99810786, 0.77345671,
       0.89655948, 0.95933358, 0.98416974, 0.99556119, 0.99987168,
       1.        , 0.81721994, 0.95987974, 0.99486424, 0.99977276,
       0.99999963, 1.        , 1.        , 0.87268482, 0.9923069 ,
       0.99996023, 1.        , 1.        , 1.        , 1.        ,
       0.92064803, 0.9997864 , 1.        , 1.        , 1.        ,
       1.        , 1.        , 0.96230299, 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        ]), 'split2_train_score': array([0.71237402, 0.77475462, 0.80
980637, 0.8367689 , 0.85865921,
       0.893782  , 0.93902028, 0.73304956, 0.82664083, 0.88347827,
       0.92013324, 0.94827031, 0.97782929, 0.99792408, 0.76766856,
       0.89473109, 0.95963654, 0.9874216 , 0.99630616, 0.99990845,
       1.        , 0.81330602, 0.9578445 , 0.99529985, 0.99980444,
       0.99999446, 1.        , 1.        , 0.86954316, 0.99014862,
       0.99998434, 1.        , 1.        , 1.        , 1.        ,
       0.91506876, 0.99975867, 1.        , 1.        , 1.        ,
       1.        , 1.        , 0.95570732, 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        ]), 'split3_train_score': array([0.71110998, 0.77845069, 0.81
395312, 0.83985987, 0.86163775,
       0.8948691 , 0.93990326, 0.74184056, 0.82811902, 0.88524879,
       0.92350375, 0.95042936, 0.9807449 , 0.99824743, 0.7723717 ,
       0.89087095, 0.95854993, 0.98641276, 0.99630932, 0.99982632,
       1.        , 0.81022696, 0.95724189, 0.99505976, 0.9997571 ,
       0.99999938, 1.        , 1.        , 0.86075956, 0.9911833 ,
       0.99998839, 1.        , 1.        , 1.        , 1.        ,
       0.91783543, 0.99964286, 1.        , 1.        , 1.        ,
       1.        , 1.        , 0.95482579, 0.99999982, 1.        ,
       1.        , 1.        , 1.        , 1.        ]), 'split4_train_score': array([0.70906486, 0.77535367, 0.81
178455, 0.83969212, 0.8620944 ,
       0.8953129 , 0.93941314, 0.73490567, 0.82471515, 0.88309477,
       0.92162076, 0.94742159, 0.97869221, 0.99769044, 0.76884441,
       0.89384502, 0.9612552 , 0.98625601, 0.99599612, 0.99989068,
       1.        , 0.81260391, 0.95395875, 0.99424215, 0.9997804 ,
       0.99999949, 1.        , 1.        , 0.86463324, 0.99088261,
       0.99999054, 1.        , 1.        , 1.        , 1.        ,
       0.91263698, 0.99971163, 1.        , 1.        , 1.        ,
       1.        , 1.        , 0.95808246, 0.99999947, 1.        ,
       1.        , 1.        , 1.        , 1.        ]), 'mean_train_score': array([0.71238106, 0.77744861, 0.8138
6443, 0.84109348, 0.86293414,
       0.89658627, 0.94090914, 0.73837199, 0.82895668, 0.88650937,
       0.92425096, 0.95007142, 0.980026  , 0.9981059 , 0.77121294,
       0.89504227, 0.95997986, 0.98621677, 0.99614536, 0.99988838,
       1.        , 0.81390476, 0.95734631, 0.99510597, 0.99978429,
       0.99999855, 1.        , 1.        , 0.86590349, 0.9914665 ,
       0.99997936, 1.        , 1.        , 1.        , 1.        ,
       0.91665249, 0.9997033 , 1.        , 1.        , 1.        ,
       1.        , 1.        , 0.95798519, 0.99999986, 1.        ,
       1.        , 1.        , 1.        , 1.        ]), 'std_train_score': array([2.15506039e-03, 1.99965925e-03,
2.91620398e-03, 3.04703694e-03,
       2.87933555e-03, 2.43231617e-03, 2.02651154e-03, 3.65101486e-03,
       3.21586360e-03, 3.30197153e-03, 3.42687924e-03, 2.02273726e-03,
       1.58477584e-03, 2.93804298e-04, 2.48402284e-03, 2.77699953e-03,
       1.05032424e-03, 1.10002890e-03, 3.41548585e-04, 3.92843115e-05,
       0.00000000e+00, 2.51734265e-03, 1.91650103e-03, 5.93619086e-04,
       1.89590610e-05, 2.05097865e-06, 0.00000000e+00, 4.96506831e-17,
       4.54668998e-03, 9.66657656e-04, 1.12659233e-05, 0.00000000e+00,
       0.00000000e+00, 4.96506831e-17, 0.00000000e+00, 2.68972896e-03,
       6.50417130e-05, 0.00000000e+00, 0.00000000e+00, 4.96506831e-17,
       0.00000000e+00, 0.00000000e+00, 2.64076641e-03, 2.05400553e-07,
       0.00000000e+00, 0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
       0.00000000e+00])}
```

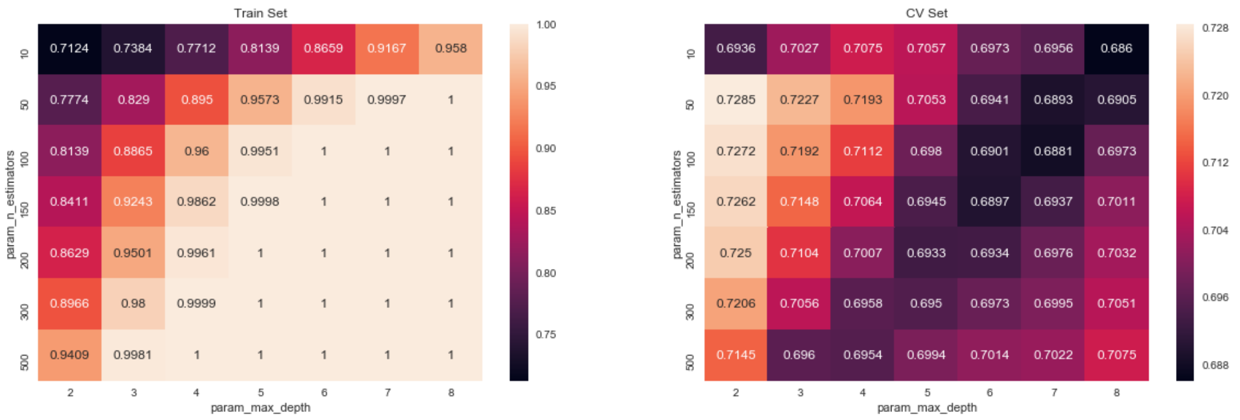
In [208]:
```python
import seaborn as sns; sns.set()
max_scores4 = pd.DataFrame(set4.cv_results_).groupby(['param_n_estimators', 'param_max_depth']).max().unstack()
max_scores4
```

Out[208]:

	mean_fit_time							mean_score_time			...
param_max_depth	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**2**	**3**	**4**	**...**
param_n_estimators											
10	19.484696	21.879489	25.212973	28.155896	31.757265	37.599645	38.200841	2.048722	1.835891	2.086820	...
50	47.377094	63.261017	87.678519	96.450469	112.169033	135.896657	146.431002	1.977114	2.131501	2.767399	...
100	82.383686	113.891022	148.819622	179.767858	213.445597	265.129466	280.421888	2.189744	2.141075	2.130702	...
150	117.455496	165.123812	213.881830	263.819686	315.748210	427.362478	404.294417	2.092007	2.064879	2.194332	...
200	152.825708	215.845778	284.803762	350.348089	416.359958	501.195562	520.811822	2.074054	2.134493	2.577308	...
300	223.175377	317.341955	417.312213	518.701074	616.971085	701.856296	737.967906	2.093601	2.188348	2.166008	...
500	364.706484	521.854242	687.812632	851.947499	1080.463430	1079.950567	1112.168203	2.145463	2.230237	2.246593	...

7 rows × 140 columns

In [209]:
```
fig, ax = plt.subplots(1,2, figsize=(20,6))
sns.heatmap(max_scores4.mean_train_score, annot = True, fmt='.4g', ax=ax[0])
sns.heatmap(max_scores4.mean_test_score, annot = True, fmt='.4g', ax=ax[1])
ax[0].set_title('Train Set')
ax[1].set_title('CV Set')
plt.show()
```



In [210]:
```
print(set4.best_estimator_)
```

```
XGBClassifier(base_score=0.5, booster=None, class_weight='balanced',
       colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
       gamma=0, gpu_id=-1, importance_type='gain',
       interaction_constraints=None, learning_rate=0.300000012,
       max_delta_step=0, max_depth=2, min_child_weight=1, missing=nan,
       monotone_constraints=None, n_estimators=50, n_jobs=0,
       num_parallel_tree=1, objective='binary:logistic', random_state=0,
       reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
       tree_method=None, validate_parameters=False, verbosity=None)
```

In [211]:
```
max_d = set4.best_params_['max_depth']
n_est = set4.best_params_['n_estimators']
```

Training our model with best Hyperparameters

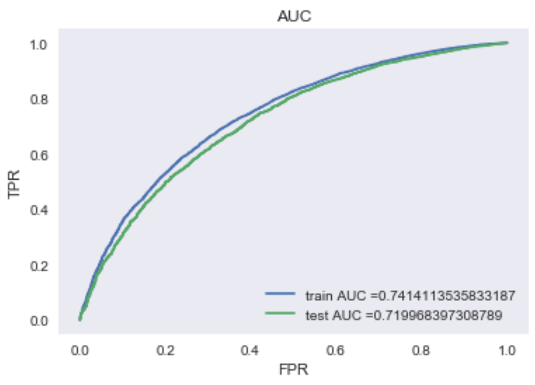
In [212]:
```
# https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html#sklearn.metrics.roc_curve
from sklearn.metrics import roc_curve, auc
from sklearn.ensemble import GradientBoostingClassifier
model = GradientBoostingClassifier(max_depth = max_d , n_estimators = n_est)

model.fit(X_tr_TFIDF_W2V, y_train)

y_train_pred = pred_prob(model,X_tr_TFIDF_W2V)
y_test_pred = pred_prob(model,X_te_TFIDF_W2V)

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

plt.close
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("FPR")
plt.ylabel("TPR")
plt.title("AUC")
plt.grid()
plt.show()
```



Confusion Matrix

In [213]:
```
#our objective here is to make auc the maximum
#so we find  the best threshold that will give the least fpr
best_t = find_best_threshold(tr_thresholds, train_fpr, train_tpr)
print("Train confusion matrix")
print(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)))
```

```
the maximum value of tpr*(1-fpr) 0.46242484499602293 for threshold 0.839
Train confusion matrix
[[ 3140  1504]
 [ 8062 17444]]
```

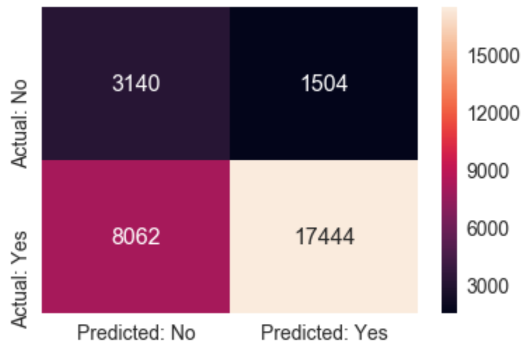
In [214]:
```
#plotting confusion matrix using seaborn's heatmap
# https://stackoverflow.com/questions/35572000/how-can-i-plot-a-confusion-matrix

print("Train data confusion matrix")

confusion_matrix_df_train = pd.DataFrame(confusion_matrix(y_train, predict_with_best_t(y_train_pred, best_t)), ['Ac
tual: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_train, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Train data confusion matrix
```

Out[214]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c399f1cba8>
```



In [215]:
```
print("Test confusion matrix")
print(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)))
```

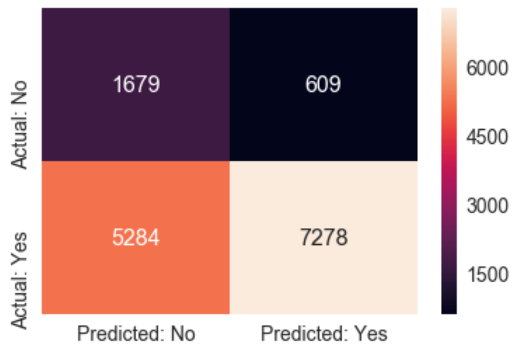
```
Test confusion matrix
[[1679  609]
 [5284 7278]]
```

In [216]:
```
print("Test data confusion matrix")

confusion_matrix_df_test = pd.DataFrame(confusion_matrix(y_test, predict_with_best_t(y_test_pred, best_t)), ['Actua
l: No','Actual: Yes'],['Predicted: No','Predicted: Yes'])
sns.set(font_scale=1.4)#for label size
sns.heatmap(confusion_matrix_df_test, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
Test data confusion matrix
```

Out[216]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2c399eaeb00>
```



In [ ]:

3. Conclusion

In [ ]:
```
# Please compare all your models using Prettytable library
```

In [217]:
```python
# Please compare all your models using Prettytable library

# 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 = ["Vectorizer", "Model", "Hyperparameters(n_estimators,max_depth)", "Test AUC"]

x.add_row(["BOW", "RF","(500, 8)", 0.698])
x.add_row(["TFIDF", "RF", "(500, 8)",  0.702])
x.add_row(["AVG W2V", "RF", "(500, 7)",  0.705])
x.add_row(["TFIDF W2V", "RF", "(500, 7)",  0.706])

x.add_row(["----- ---", "----", "-------------------------------", "---------"])

x.add_row(["BOW", "GBDT","(200, 2)",  0.741])
x.add_row(["TFIDF", "GBDT", "(150, 2)",  0.738])
x.add_row(["AVG W2V", "GBDT", "(100, 2)",  0.728])
x.add_row(["TFIDF W2V", "GBDT", "(50, 2)", 0.719])



print(x)
```

```
+------------+-------+-----------------------------------------+-----------+
| Vectorizer | Model | Hyperparameters(n_estimators,max_depth) |  Test AUC |
+------------+-------+-----------------------------------------+-----------+
|    BOW     |   RF  |                 (500, 8)                |   0.698   |
|   TFIDF    |   RF  |                 (500, 8)                |   0.702   |
|  AVG W2V   |   RF  |                 (500, 7)                |   0.705   |
| TFIDF W2V  |   RF  |                 (500, 7)                |   0.706   |
| ----- ---  |  ---- |      -------------------------------    | --------- |
|    BOW     |  GBDT |                 (200, 2)                |   0.741   |
|   TFIDF    |  GBDT |                 (150, 2)                |   0.738   |
|  AVG W2V   |  GBDT |                 (100, 2)                |   0.728   |
| TFIDF W2V  |  GBDT |                 (50, 2)                 |   0.719   |
+------------+-------+-----------------------------------------+-----------+
```

In [ ]: