11 GBDT

April 3, 2022

[1]:
```
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")

import pandas as pd
import numpy as np
import math as m
import matplotlib.pyplot as plt
import seaborn as sns
import nltk
import re

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.preprocessing import Normalizer
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.metrics import roc_curve, auc
from tqdm import tqdm
import nltk

nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer
sid = SentimentIntensityAnalyzer()
import pickle
```

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

1 1. Loading Data set

[2]:
```
data  = pd.read_csv('preprocessed_data.csv',nrows=35000)
data.head(3)
```

[2]:
```
  school_state teacher_prefix project_grade_category  \
0           ca            mrs          grades_prek_2
1           ut             ms             grades_3_5
2           ca            mrs          grades_prek_2

   teacher_number_of_previously_posted_projects  project_is_approved  \
0                                            53                   1
1                                             4                   1
2                                            10                   1

    clean_categories                clean_subcategories  \
0       math_science  appliedsciences health_lifescience
1       specialneeds                       specialneeds
2  literacy_language                           literacy

                                               essay   price
0  i fortunate enough use fairy tale stem kits cl…  725.05
1  imagine 8 9 years old you third grade classroo…  213.03
2  having class 24 students comes diverse learner…  329.00
```

[4]:
```
X=data.drop(['project_is_approved'],axis=1)
Y=data['project_is_approved'].values

X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.
↪33,stratify=Y,random_state=42)

X_train.shape,'train-x',Y_train.shape,'train-y',X_test.shape,'test-x',Y_test.
↪shape,'test-y'
```

[4]:
```
((23450, 8),
 'train-x',
 (23450,),
 'train-y',
 (11550, 8),
 'test-x',
 (11550,),
 'test-y')
```

2 2. Categorical data set handling

2.0.1 2.1 Response coding

[5]:
```
# https://stackoverflow.com/questions/66122577/
↪response-coding-for-categorical-data
def response_coding(xtrain, ytrain, feature):

    from prettytable import PrettyTable
```

```
    dictionary = dict()
    x = PrettyTable()
    x = PrettyTable([feature, 'class 1', 'class 0'])

    unique_cat_labels = xtrain[feature].unique()

    for i in tqdm(range(len(unique_cat_labels))):
        total_count = xtrain.loc[:,feature][(xtrain[feature] ==␣
↪unique_cat_labels[i])].count()
        p_0 = xtrain.loc[:, feature][((xtrain[feature] == unique_cat_labels[i])␣
↪& (ytrain==0))].count()
        p_1 = xtrain.loc[:, feature][((xtrain[feature] == unique_cat_labels[i])␣
↪& (ytrain==1))].count()

        dictionary[unique_cat_labels[i]] = [p_1/total_count, p_0/total_count]

        row = []
        row.append(unique_cat_labels[i])
        row.append(p_1/total_count)
        row.append(p_0/total_count)
        x.add_row(row)
    #print()
    print(x)#[![enter image description here][1]][1]
    return dictionary

grade=response_coding(X_train,Y_train,'project_grade_category')
school_state=response_coding(X_train,Y_train,'school_state')
teacher_pre=response_coding(X_train,Y_train,'teacher_prefix')
cate=response_coding(X_train,Y_train,'clean_categories')
subcate=response_coding(X_train,Y_train,'clean_subcategories')
```

```
100%|
     | 4/4 [00:00<00:00, 58.16it/s]
 39%|
| 20/51 [00:00<00:00, 194.61it/s]
```

```
+------------------------+--------------------+---------------------+
| project_grade_category |      class 1       |       class 0       |
+------------------------+--------------------+---------------------+
|       grades_3_5       | 0.8530190344132106 | 0.14698096558678936 |
|       grades_6_8       | 0.8352322524101665 | 0.16476774758983348 |
|      grades_9_12       | 0.8461538461538461 | 0.15384615384615385 |
|     grades_prek_2      | 0.8453110865413609 | 0.15468891345863905 |
+------------------------+--------------------+---------------------+
```

```
100%|
   | 51/51 [00:00<00:00, 193.66it/s]
100%|
```

```
    | 5/5 [00:00<00:00, 172.59it/s]
 49%|
| 21/43 [00:00<00:00, 200.45it/s]
```

school_state	class 1	class 0
in	0.8613861386138614	0.13861386138613863
ny	0.8614130434782609	0.13858695652173914
sc	0.8513824884792627	0.14861751152073732
tn	0.8489702517162472	0.15102974828375287
nc	0.8397129186602871	0.16028708133971292
ca	0.8696081659532433	0.13039183404675667
la	0.800531914893617	0.19946808510638298
ma	0.833648393194707	0.166351606805293
tx	0.8045454545454546	0.19545454545454546
co	0.8207171314741036	0.17928286852589642
mi	0.8421828908554573	0.15781710914454278
ut	0.849624060150376	0.15037593984962405
mo	0.8690095846645367	0.13099041533546327
fl	0.84363894811656	0.15636105188343993
wi	0.8296089385474861	0.17039106145251395
al	0.8481012658227848	0.1518987341772152
ms	0.8120300751879699	0.18796992481203006
az	0.8459958932238193	0.1540041067761807
ok	0.8212180746561886	0.1787819253438114
il	0.8480749219562955	0.15192507804370448
ar	0.8646288209606987	0.13537117903930132
ga	0.8329545454545455	0.16704545454545455
md	0.8099173553719008	0.19008264462809918
ct	0.8461538461538461	0.15384615384615385
va	0.84375	0.15625
nv	0.8403908794788274	0.15960912052117263
oh	0.864321608040201	0.135678391959799
or	0.8847926267281107	0.1152073732718894
wa	0.8694817658349329	0.13051823416506717
ia	0.8475609756097561	0.1524390243902439
nj	0.835	0.165
id	0.9313725490196079	0.06862745098039216
ne	0.835820895522388	0.16417910447761194
de	0.8878504672897196	0.11214953271028037
ky	0.8593155893536122	0.14068441064638784
dc	0.84375	0.15625
pa	0.851528384279476	0.14847161572052403
wv	0.8098591549295775	0.19014084507042253
hi	0.8780487804878049	0.12195121951219512
sd	0.8405797101449275	0.15942028985507245
vt	0.782608695652174	0.21739130434782608

```
|      mn      | 0.8306451612903226 |  0.1693548387096774  |
|      mt      |        0.84        |         0.16         |
|      ks      | 0.8652482269503546 |  0.1347517730496454  |
|      me      | 0.8584905660377359 | 0.14150943396226415  |
|      ri      | 0.8771929824561403 | 0.12280701754385964  |
|      nd      | 0.9411764705882353 | 0.058823529411764705 |
|      nm      | 0.8503937007874016 | 0.14960629921259844  |
|      ak      | 0.8809523809523809 | 0.11904761904761904  |
|      nh      | 0.9016393442622951 | 0.09836065573770492  |
|      wy      |        0.75        |         0.25         |
+--------------+--------------------+----------------------+
+----------------+--------------------+----------------------+
| teacher_prefix |      class 1       |       class 0        |
+----------------+--------------------+----------------------+
|       ms       | 0.8404434803567125 | 0.15955651964328754  |
|      mrs       | 0.8525732383214568 | 0.14742676167854316  |
|       mr       | 0.8542589857213195 | 0.14574101427868044  |
|    teacher     | 0.7622950819672131 | 0.23770491803278687  |
|       dr       | 0.6666666666666666 |  0.3333333333333333  |
+----------------+--------------------+----------------------+

100%|
   | 43/43 [00:00<00:00, 199.61it/s]
  7%|
| 22/318 [00:00<00:01, 214.16it/s]

+----------------------------------+--------------------+----------------------
+
|          clean_categories        |      class 1       |       class 0
|
+----------------------------------+--------------------+----------------------
+
|           health_sports          | 0.8555090655509066 | 0.14449093444909344
|
|           math_science           | 0.817046518888542  |  0.182953481111458
|
|            music_arts            | 0.8656084656084656 |  0.1343915343915344
|
|         literacy_language        | 0.8590004063388866 | 0.14099959366111336
|
|   math_science literacy_language | 0.8620689655172413 | 0.13793103448275862
|
|    math_science history_civics   |       0.875        |        0.125
|
|   literacy_language math_science | 0.8617807778849245 | 0.13821922211507554
|
|    appliedlearning specialneeds  | 0.8051948051948052 | 0.19480519480519481
|
|    appliedlearning math_science  | 0.8010204081632653 |  0.1989795918367347
```

history_civics	0.8376811594202899	0.16231884057971013
appliedlearning	0.8058124174372523	0.19418758256274768
literacy_language music_arts	0.8345498783454988	0.1654501216545012
appliedlearning health_sports	0.8655913978494624	0.13440860215053763
appliedlearning literacy_language	0.884990253411306	0.11500974658869395
specialneeds	0.8181818181818182	0.18181818181818182
math_science appliedlearning	0.8177339901477833	0.18226600985221675
literacy_language history_civics	0.8409090909090909	0.1590909090909091
appliedlearning music_arts	0.8273381294964028	0.17266187050359713
history_civics literacy_language	0.8552631578947368	0.14473684210526316
literacy_language specialneeds	0.8617571059431525	0.13824289405684753
math_science music_arts	0.792	0.208
math_science health_sports	0.7931034482758621	0.20689655172413793
math_science specialneeds	0.8371428571428572	0.16285714285714287
literacy_language appliedlearning	0.8581560283687943	0.14184397163120568
health_sports specialneeds	0.8882521489971347	0.11174785100286533
health_sports literacy_language	0.8554216867469879	0.14457831325301204
history_civics music_arts	0.8983050847457628	0.1016949152542373
appliedlearning history_civics	0.8181818181818182	0.18181818181818182
health_sports math_science	0.7530864197530864	0.24691358024691357
history_civics math_science	0.7777777777777778	0.2222222222222222
health_sports music_arts	0.8	0.2
music_arts specialneeds	0.9565217391304348	0.043478260869565216
health_sports appliedlearning	0.8297872340425532	0.1702127659574468

```
|
|   history_civics appliedlearning  | 0.8333333333333334 | 0.16666666666666666
|
|    history_civics specialneeds    | 0.7058823529411765 | 0.29411764705882354
|
|      specialneeds music_arts      | 0.864406779661017  | 0.13559322033898305
|
|    health_sports history_civics   |        0.75        |         0.25
|
|  literacy_language health_sports  | 0.7058823529411765 | 0.29411764705882354
|
|     specialneeds health_sports    |       0.875        |        0.125
|
|     music_arts history_civics     |        0.8         |         0.2
|
|      music_arts health_sports     | 0.5555555555555556 |  0.4444444444444444
|
|     music_arts appliedlearning    | 0.6666666666666666 |  0.3333333333333333
|
|    history_civics health_sports   |        1.0         |         0.0
|
+-----------------------------------+--------------------+----------------------
+

100%|
  | 318/318 [00:01<00:00, 199.16it/s]

+----------------------------------------+--------------------+----------------
------+
|           clean_subcategories          |      class 1       |       class 0
|
+----------------------------------------+--------------------+----------------
------+
|             health_wellness            | 0.8756944444444444 |
0.12430555555555556  |
| environmentalscience health_lifescience |        0.79        |         0.21
|
|                visualarts              | 0.8449197860962567 |
0.15508021390374332  |
|               esl literacy             | 0.8509615384615384 |
0.14903846153846154  |
|                literacy                | 0.8862433862433863 |
0.11375661375661375  |
|        performingarts visualarts       | 0.7391304347826086 |
0.2608695652173913  |
|           literature_writing           | 0.8474923234390993 |
0.15250767656090072  |
|  health_lifescience literature_writing  | 0.8775510204081632 |
0.12244897959183673  |
```

health_lifescience history_geography	0.7647058823529411	0.23529411764705882
mathematics	0.814638783269962	0.18536121673003803
literature_writing mathematics	0.8639097744360902	0.13609022556390976
earlydevelopment specialneeds	0.8222222222222222	0.17777777777777778
earlydevelopment mathematics	0.85	0.15
history_geography	0.8080808080808081	0.1919191919191919
gym_fitness	0.867816091954023	0.13218390804597702
environmentalscience literature_writing	0.8	0.2
literacy mathematics	0.8614035087719298	0.13859649122807016
charactereducation earlydevelopment	0.8292682926829268	0.17073170731707318
gym_fitness health_wellness	0.8672055427251733	0.13279445727482678
appliedsciences environmentalscience	0.7647058823529411	0.23529411764705882
literature_writing visualarts	0.8141025641025641	0.1858974358974359
nutritioneducation	0.7884615384615384	0.21153846153846154
earlydevelopment health_wellness	0.875	0.125
appliedsciences	0.8222222222222222	0.17777777777777778
esl	0.7746478873239436	0.22535211267605634
literacy literature_writing	0.8455538221528861	0.1544461778471139
literacy visualarts	0.8345323741007195	0.16546762589928057
music	0.8641114982578397	0.13588850174216027
charactereducation literacy	0.9	0.1
specialneeds	0.8181818181818182	0.18181818181818182
appliedsciences college_careerprep	0.7887323943661971	0.2112676056338028
esl literature_writing	0.7861635220125787	0.2138364779874214

earlydevelopment literacy	0.8922155688622755	0.10778443113772455
earlydevelopment other	0.7714285714285715	0.22857142857142856
environmentalscience socialsciences	0.95	0.05
appliedsciences literacy	0.8455284552845529	0.15447154471544716
literacy socialsciences	0.8620689655172413	0.13793103448275862
appliedsciences other	0.8260869565217391	0.17391304347826086
history_geography socialsciences	0.9027777777777778	0.09722222222222222
earlydevelopment performingarts	1.0	0.0
health_wellness nutritioneducation	0.8352601156069365	0.16473988439306358
appliedsciences mathematics	0.8236173393124065	0.17638266068759342
civics_government literature_writing	0.8571428571428571	0.14285714285714285
literature_writing specialneeds	0.825925925925926	0.17407407407407408
music visualarts	0.8888888888888888	0.1111111111111111
mathematics visualarts	0.7402597402597403	0.2597402597402597
environmentalscience gym_fitness	1.0	0.0
gym_fitness teamsports	0.7246376811594203	0.2753623188405797
literacy specialneeds	0.8763326226012793	0.12366737739872068
health_wellness teamsports	0.8181818181818182	0.18181818181818182
environmentalscience	0.8523489932885906	0.1476510067114094
mathematics specialneeds	0.8127853881278538	0.1872146118721461
environmentalscience history_geography	0.9166666666666666	0.08333333333333333
esl earlydevelopment	0.9047619047619048	0.09523809523809523
history_geography literature_writing	0.8145161290322581	0.18548387096774194
health_wellness specialneeds	0.8775510204081632	0.12244897959183673

| college_careerprep mathematics | 0.7708333333333334 |
0.22916666666666666 |
| teamsports | 0.8252427184466019 |
0.17475728155339806 |
| socialsciences | 0.8222222222222222 |
0.17777777777777778 |
| other | 0.7967914438502673 |
0.20320855614973263 |
| health_lifescience mathematics | 0.83 | 0.17
|
| health_wellness literature_writing | 0.9120879120879121 |
0.08791208791208792 |
| literature_writing socialsciences | 0.8636363636363636 |
0.13636363636363635 |
| health_lifescience | 0.8958333333333334 |
0.10416666666666667 |
| health_lifescience health_wellness | 0.7804878048780488 |
0.21951219512195122 |
| earlydevelopment | 0.7906976744186046 |
0.20930232558139536 |
| charactereducation | 0.8157894736842105 |
0.18421052631578946 |
| college_careerprep visualarts | 0.92 | 0.08
|
| esl health_lifescience | 0.8888888888888888 |
0.1111111111111111 |
| appliedsciences specialneeds | 0.9295774647887324 |
0.07042253521126761 |
| charactereducation literature_writing | 0.8 | 0.2
|
| civics_government history_geography | 0.7777777777777778 |
0.2222222222222222 |
| foreignlanguages literacy | 0.8333333333333334 |
0.16666666666666666 |
| foreignlanguages specialneeds | 1.0 | 0.0
|
| environmentalscience mathematics | 0.7702702702702703 |
0.22972972972972974 |
| appliedsciences esl | 0.9285714285714286 |
0.07142857142857142 |
| mathematics socialsciences | 0.8823529411764706 |
0.11764705882352941 |
| esl mathematics | 0.8775510204081632 |
0.12244897959183673 |
| civics_government visualarts | 1.0 | 0.0
|
| college_careerprep socialsciences | 0.7142857142857143 |
0.2857142857142857 |

esl environmentalscience	0.75	0.25
health_wellness mathematics	0.75	0.25
health_lifescience specialneeds	0.7777777777777778	0.2222222222222222
music performingarts	0.9144385026737968	0.0855614973262032
communityservice environmentalscience	1.0	0.0
earlydevelopment literature_writing	0.8222222222222222	0.17777777777777778
other specialneeds	0.8484848484848485	0.15151515151515152
health_lifescience literacy	0.9487179487179487	0.05128205128205128
history_geography mathematics	0.8333333333333334	0.16666666666666666
college_careerprep literacy	0.9272727272727272	0.07272727272727272
health_wellness visualarts	0.8181818181818182	0.18181818181818182
appliedsciences music	0.7777777777777778	0.2222222222222222
appliedsciences health_wellness	1.0	0.0
performingarts specialneeds	1.0	0.0
appliedsciences health_lifescience	0.8257575757575758	0.17424242424242425
charactereducation esl	0.4	0.6
charactereducation specialneeds	0.75	0.25
health_wellness literacy	0.831081081081081	0.16891891891891891
financialliteracy mathematics	0.625	0.375
health_lifescience visualarts	0.8125	0.1875
health_wellness other	0.813953488372093	0.18604651162790697
civics_government environmentalscience	0.6666666666666666	0.3333333333333333
literature_writing parentinvolvement	0.8333333333333334	0.16666666666666666
communityservice specialneeds	0.8	0.2

economics other	1.0	0.0
college_careerprep specialneeds	0.6666666666666666	0.3333333333333333
environmentalscience literacy	0.8482142857142857	0.15178571428571427
history_geography literacy	0.8653846153846154	0.1346153846153846
civics_government socialsciences	0.8421052631578947	0.15789473684210525
history_geography specialneeds	0.7619047619047619	0.23809523809523808
college_careerprep health_wellness	0.8	0.2
literature_writing performingarts	0.875	0.125
civics_government literacy	0.918918918918919	0.08108108108108109
charactereducation health_wellness	0.8108108108108109	0.1891891891891892
appliedsciences literature_writing	0.8470588235294118	0.15294117647058825
environmentalscience health_wellness	0.8235294117647058	0.17647058823529413
appliedsciences socialsciences	1.0	0.0
environmentalscience nutritioneducation	0.6470588235294118	0.35294117647058826
college_careerprep literature_writing	0.9672131147540983	0.03278688524590164
extracurricular	0.75	0.25
appliedsciences parentinvolvement	0.75	0.25
esl specialneeds	0.9354838709677419	0.06451612903225806
specialneeds visualarts	0.864406779661017	0.13559322033898305
parentinvolvement	1.0	0.0
appliedsciences visualarts	0.7931034482758621	0.20689655172413793
college_careerprep	0.873015873015873	0.12698412698412698
history_geography visualarts	0.9210526315789473	0.07894736842105263
gym_fitness performingarts	0.6666666666666666	0.3333333333333333

characteducation environmentalscience	0.75	0.25
literacy other	0.8958333333333334	0.10416666666666667
literacy performingarts	0.8235294117647058	0.17647058823529413
extracurricular visualarts	0.9166666666666666	0.08333333333333333
communityservice health_lifescience	0.6666666666666666	0.3333333333333333
characteducation college_careerprep	0.7435897435897436	0.2564102564102564
college_careerprep earlydevelopment	0.8333333333333334	0.16666666666666666
characteducation history_geography	0.5	0.5
mathematics other	0.7619047619047619	0.23809523809523808
environmentalscience other	1.0	0.0
college_careerprep other	0.7058823529411765	0.29411764705882354
extracurricular specialneeds	1.0	0.0
environmentalscience visualarts	0.7692307692307693	0.23076923076923078
performingarts	0.8923076923076924	0.1076923076923077
communityservice visualarts	0.875	0.125
characteducation communityservice	0.9090909090909091	0.09090909090909091
appliedsciences earlydevelopment	0.8333333333333334	0.16666666666666666
gym_fitness specialneeds	0.9583333333333334	0.041666666666666664
communityservice literature_writing	0.75	0.25
civics_government	0.9	0.1
appliedsciences extracurricular	1.0	0.0
literature_writing other	0.7435897435897436	0.2564102564102564
earlydevelopment visualarts	0.6842105263157895	0.3157894736842105
health_wellness music	0.75	0.25

earlydevelopment extracurricular	0.8	0.2
communityservice	0.7777777777777778	0.2222222222222222
civics_government economics	0.8	0.2
socialsciences visualarts	0.7	0.3
health_wellness performingarts	1.0	0.0
history_geography other	0.75	0.25
gym_fitness mathematics	0.7272727272727273	0.2727272727272727
extracurricular literature_writing	1.0	0.0
mathematics music	1.0	0.0
esl history_geography	0.6666666666666666	0.3333333333333333
appliedsciences civics_government	0.6666666666666666	0.3333333333333333
gym_fitness music	0.6666666666666666	0.3333333333333333
music specialneeds	0.95	0.05
appliedsciences charactereducation	0.8571428571428571	0.14285714285714285
charactereducation other	0.875	0.125
extracurricular mathematics	0.875	0.125
appliedsciences history_geography	0.7894736842105263	0.21052631578947367
esl visualarts	0.8571428571428571	0.14285714285714285
gym_fitness nutritioneducation	0.8636363636363636	0.13636363636363635
appliedsciences nutritioneducation	1.0	0.0
extracurricular teamsports	1.0	0.0
health_wellness socialsciences	1.0	0.0
communityservice health_wellness	0.6666666666666666	0.3333333333333333
foreignlanguages	0.819672131147541	0.18032786885245902

extracurricular literacy	0.8888888888888888	0.1111111111111111
financialliteracy specialneeds	0.0	1.0
foreignlanguages health_wellness	0.5	0.5
charactereducation mathematics	0.782608695652174	0.21739130434782608
earlydevelopment health_lifescience	0.3333333333333333	0.6666666666666666
appliedsciences communityservice	1.0	0.0
foreignlanguages literature_writing	0.9285714285714286	0.07142857142857142
charactereducation teamsports	0.8	0.2
charactereducation visualarts	0.8235294117647058	0.17647058823529413
literacy parentinvolvement	1.0	0.0
nutritioneducation visualarts	1.0	0.0
earlydevelopment environmentalscience	0.7	0.3
gym_fitness literacy	0.7142857142857143	0.2857142857142857
teamsports visualarts	1.0	0.0
literature_writing music	0.8888888888888888	0.1111111111111111
charactereducation health_lifescience	1.0	0.0
charactereducation socialsciences	1.0	0.0
earlydevelopment socialsciences	1.0	0.0
college_careerprep communityservice	0.875	0.125
charactereducation financialliteracy	1.0	0.0
gym_fitness literature_writing	0.6666666666666666	0.3333333333333333
earlydevelopment music	0.875	0.125
nutritioneducation specialneeds	0.8571428571428571	0.14285714285714285
other parentinvolvement	1.0	0.0

environmentalscience specialneeds	0.8484848484848485	0.15151515151515152
esl foreignlanguages	1.0	0.0
environmentalscience foreignlanguages	1.0	0.0
parentinvolvement visualarts	1.0	0.0
college_careerprep environmentalscience	0.8571428571428571	0.14285714285714285
foreignlanguages mathematics	0.4	0.6
literacy music	0.9318181818181818	0.06818181818181818
health_lifescience socialsciences	0.8421052631578947	0.15789473684210525
college_careerprep nutritioneducation	1.0	0.0
extracurricular health_wellness	1.0	0.0
civics_government mathematics	0.0	1.0
environmentalscience extracurricular	0.3333333333333333	0.6666666666666666
health_lifescience nutritioneducation	0.8571428571428571	0.14285714285714285
earlydevelopment teamsports	1.0	0.0
communityservice earlydevelopment	1.0	0.0
specialneeds teamsports	0.875	0.125
history_geography performingarts	1.0	0.0
extracurricular music	1.0	0.0
environmentalscience parentinvolvement	1.0	0.0
college_careerprep foreignlanguages	1.0	0.0
charactereducation music	1.0	0.0
charactereducation parentinvolvement	0.8333333333333334	0.16666666666666666
esl performingarts	0.5	0.5
esl health_wellness	0.8333333333333334	0.16666666666666666

esl other	0.8	0.2
esl socialsciences	0.5714285714285714	0.42857142857142855
environmentalscience performingarts	1.0	0.0
charactereducation extracurricular	0.7272727272727273	0.2727272727272727
financialliteracy	0.75	0.25
college_careerprep extracurricular	0.75	0.25
charactereducation performingarts	0.6666666666666666	0.3333333333333333
mathematics performingarts	1.0	0.0
economics financialliteracy	1.0	0.0
college_careerprep performingarts	0.6666666666666666	0.3333333333333333
communityservice literacy	0.6666666666666666	0.3333333333333333
performingarts socialsciences	1.0	0.0
communityservice performingarts	1.0	0.0
music teamsports	0.5	0.5
college_careerprep financialliteracy	0.6666666666666666	0.3333333333333333
civics_government health_lifescience	0.75	0.25
earlydevelopment gym_fitness	0.8461538461538461	0.15384615384615385
music socialsciences	0.6666666666666666	0.3333333333333333
gym_fitness other	1.0	0.0
economics literature_writing	1.0	0.0
college_careerprep history_geography	1.0	0.0
performingarts teamsports	0.6	0.4
earlydevelopment nutritioneducation	1.0	0.0
mathematics parentinvolvement	0.75	0.25

health_wellness history_geography	0.5714285714285714	0.42857142857142855
extracurricular health_lifescience	0.5	0.5
extracurricular socialsciences	1.0	0.0
economics history_geography	1.0	0.0
communityservice mathematics	0.6666666666666666	0.3333333333333333
college_careerprep parentinvolvement	1.0	0.0
appliedsciences teamsports	1.0	0.0
communityservice other	0.5	0.5
financialliteracy literacy	1.0	0.0
nutritioneducation teamsports	0.6	0.4
gym_fitness visualarts	0.6666666666666666	0.3333333333333333
other visualarts	0.6666666666666666	0.3333333333333333
health_lifescience other	0.0	1.0
communityservice socialsciences	0.0	1.0
history_geography music	1.0	0.0
college_careerprep health_lifescience	0.75	0.25
literacy teamsports	0.8	0.2
college_careerprep economics	1.0	0.0
economics mathematics	1.0	0.0
civics_government esl	1.0	0.0
college_careerprep esl	1.0	0.0
economics literacy	1.0	0.0
music other	0.6666666666666666	0.3333333333333333
socialsciences specialneeds	0.75	0.25

extracurricular other	0.6666666666666666	0.3333333333333333
parentinvolvement specialneeds	0.3333333333333333	0.6666666666666666
gym_fitness health_lifescience	1.0	0.0
civics_government specialneeds	1.0	0.0
communityservice nutritioneducation	1.0	0.0
esl financialliteracy	0.0	1.0
mathematics teamsports	1.0	0.0
civics_government health_wellness	1.0	0.0
civics_government financialliteracy	1.0	0.0
earlydevelopment parentinvolvement	1.0	0.0
extracurricular performingarts	1.0	0.0
communityservice esl	1.0	0.0
economics	0.75	0.25
literature_writing teamsports	0.0	1.0
foreignlanguages performingarts	0.0	1.0
foreignlanguages other	1.0	0.0
earlydevelopment history_geography	1.0	0.0
appliedsciences economics	1.0	0.0
health_lifescience music	1.0	0.0
mathematics nutritioneducation	0.5	0.5
communityservice financialliteracy	1.0	0.0
communityservice history_geography	0.6666666666666666	0.3333333333333333
college_careerprep teamsports	1.0	0.0
financialliteracy health_wellness	1.0	0.0

appliedsciences gym_fitness	0.5	0.5
nutritioneducation socialsciences	1.0	0.0
nutritioneducation other	1.0	0.0
parentinvolvement socialsciences	1.0	0.0
civics_government communityservice	1.0	0.0
foreignlanguages history_geography	0.0	1.0
extracurricular gym_fitness	1.0	0.0
communityservice parentinvolvement	1.0	0.0
esl nutritioneducation	1.0	0.0
economics health_lifescience	1.0	0.0
environmentalscience music	1.0	0.0
foreignlanguages visualarts	0.5	0.5
charactereducation civics_government	1.0	0.0
extracurricular nutritioneducation	1.0	0.0
literacy nutritioneducation	0.5	0.5
esl music	1.0	0.0
college_careerprep gym_fitness	1.0	0.0
extracurricular history_geography	1.0	0.0
health_lifescience teamsports	1.0	0.0
economics socialsciences	1.0	0.0
foreignlanguages socialsciences	1.0	0.0
college_careerprep music	1.0	0.0

2.0.2 2.2 Coding for replacment of the words to values

[6]:
```
def replace(data_frame,feature,dic):
    #creating two feature 0 & 1
    df=data_frame.copy()
    u=list(df[feature].unique())# for test data set unique category
    df[feature+'_0']=df[feature]
    df=df.rename(columns={feature:feature+'_1'})
    col=np.array(df.columns)
    ind_col=np.where(col == feature+'_1')[0][0]
    cols=list(df.columns)
    df=df[cols[0:ind_col]+[cols[-1]]+cols[ind_col:len(cols)-1]]

    # Using the dict for replacing the values
    for i in tqdm(dic):
        temp=dic[i]
        df[feature+'_0']=df[feature+'_0'].replace(i,temp[1])
        df[feature+'_1']=df[feature+'_1'].replace(i,temp[0])

    # Test data set new category
    keys=list(dic.keys())
    for i in keys:
        try:
            u.remove(i)
        except:
            print("This element is not present in train data set=",i)
            for i in u:
                df[feature+'_0']=df[feature+'_0'].replace(i,0.5)
                df[feature+'_1']=df[feature+'_1'].replace(i,0.5)

    return df


# Train data set
d=X_train.
 ↪drop(['teacher_number_of_previously_posted_projects','essay','price'],axis=1)
df=replace(d,'project_grade_category',grade)
df1=replace(df,'school_state',school_state)
df2=replace(df1,'teacher_prefix',teacher_pre)
df3=replace(df2,'clean_categories',cate)
X_cat_train=replace(df3,'clean_subcategories',subcate)

# Test data set
d=X_test.
 ↪drop(['teacher_number_of_previously_posted_projects','essay','price'],axis=1)
df=replace(d,'project_grade_category',grade)
df1=replace(df,'school_state',school_state)
```

```
df2=replace(df1,'teacher_prefix',teacher_pre)
df3=replace(df2,'clean_categories',cate)
X_cat_test=replace(df3,'clean_subcategories',subcate)


#replace(X_test,'clean_categories',cate)
```

```
100%|
     | 4/4 [00:00<00:00, 18.15it/s]
100%|
    | 51/51 [00:05<00:00,  9.61it/s]
100%|
     | 5/5 [00:00<00:00,  8.86it/s]
100%|
    | 43/43 [00:06<00:00,  7.16it/s]
100%|
   | 318/318 [00:51<00:00,  6.17it/s]
100%|
     | 4/4 [00:00<00:00, 29.06it/s]
100%|
    | 51/51 [00:03<00:00, 15.25it/s]
100%|
     | 5/5 [00:00<00:00, 20.98it/s]
 12%|
| 5/43 [00:00<00:01, 35.08it/s]
```

```
This element is not present in train data set= dr
```

```
100%|
    | 43/43 [00:02<00:00, 15.19it/s]
  3%|
| 9/318 [00:00<00:03, 77.79it/s]
```

```
This element is not present in train data set= music_arts appliedlearning
This element is not present in train data set= history_civics health_sports
```

```
100%|
   | 318/318 [00:23<00:00, 13.67it/s]
```

```
This element is not present in train data set= environmentalscience gym_fitness
This element is not present in train data set= performingarts specialneeds
This element is not present in train data set= economics other
This element is not present in train data set= appliedsciences parentinvolvement
This element is not present in train data set= environmentalscience other
This element is not present in train data set= history_geography other
This element is not present in train data set= gym_fitness music
This element is not present in train data set= nutritioneducation visualarts
This element is not present in train data set= charactereducation
financialliteracy
This element is not present in train data set= environmentalscience
```

```
foreignlanguages
This element is not present in train data set= college_careerprep
nutritioneducation
This element is not present in train data set= earlydevelopment teamsports
This element is not present in train data set= history_geography performingarts
This element is not present in train data set= environmentalscience
parentinvolvement
This element is not present in train data set= environmentalscience
performingarts
This element is not present in train data set= charactereducation performingarts
This element is not present in train data set= communityservice literacy
This element is not present in train data set= communityservice performingarts
This element is not present in train data set= music socialsciences
This element is not present in train data set= economics literature_writing
This element is not present in train data set= earlydevelopment
nutritioneducation
This element is not present in train data set= extracurricular
health_lifescience
This element is not present in train data set= extracurricular socialsciences
This element is not present in train data set= appliedsciences teamsports
This element is not present in train data set= communityservice socialsciences
This element is not present in train data set= college_careerprep economics
This element is not present in train data set= civics_government esl
This element is not present in train data set= economics literacy
This element is not present in train data set= music other
This element is not present in train data set= civics_government specialneeds
This element is not present in train data set= communityservice
nutritioneducation
This element is not present in train data set= esl financialliteracy
This element is not present in train data set= civics_government health_wellness
This element is not present in train data set= civics_government
financialliteracy
This element is not present in train data set= communityservice esl
This element is not present in train data set= foreignlanguages performingarts
This element is not present in train data set= foreignlanguages other
This element is not present in train data set= earlydevelopment
history_geography
This element is not present in train data set= appliedsciences economics
This element is not present in train data set= health_lifescience music
This element is not present in train data set= communityservice
financialliteracy
This element is not present in train data set= college_careerprep teamsports
This element is not present in train data set= financialliteracy health_wellness
This element is not present in train data set= nutritioneducation socialsciences
This element is not present in train data set= extracurricular gym_fitness
This element is not present in train data set= esl nutritioneducation
This element is not present in train data set= economics health_lifescience
This element is not present in train data set= environmentalscience music
```

```
This element is not present in train data set= extracurricular
nutritioneducation
This element is not present in train data set= college_careerprep gym_fitness
This element is not present in train data set= extracurricular history_geography
```

2.0.3 2.3 Words vectorization

[7]:
```
train_essays = X_train['essay'].values
test_essays = X_test['essay'].values
```

[8]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer_tfidf = TfidfVectorizer(min_df=10)      #taking words which has been␣
 ↪appeared in at least 10 doc (min_df=10)
vectorizer_tfidf=vectorizer_tfidf.fit(train_essays)
train_text_tfidf = vectorizer_tfidf.transform(train_essays)
test_text_tfidf = vectorizer_tfidf.transform(test_essays)
words_dict = vectorizer_tfidf.vocabulary_
print("Shape of matrix after one hot encodig train and test",train_text_tfidf.
 ↪shape,test_text_tfidf.shape)
```

```
Shape of matrix after one hot encodig train and test (23450, 8943) (11550, 8943)
```

2.3.1 W2V

[9]:
```
with open('glove_vectors', 'rb') as f:
    model = pickle.load(f)
    glove_words =  set(model.keys())
```

[11]:
```
# average Word2Vec
# compute average word2vec for each review.

def avg_w2v_vectors(data):
    avg_w2v_vectors = []; # the avg-w2v for each sentence/review is stored in␣
 ↪this list
    for sentence in tqdm(data): # 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)
    return avg_w2v_vectors

train_w2v=np.array(avg_w2v_vectors(train_essays))
test_w2v=np.array(avg_w2v_vectors(test_essays))
```

```
train_w2v.shape,test_w2v.shape
```

```
100%|                                                |
23450/23450 [00:06<00:00, 3653.19it/s]
100%|                                                |
11550/11550 [00:03<00:00, 3739.16it/s]
```

[11]:
```
((23450, 300), (11550, 300))
```

2.3.2 Sentiment

[12]:
```
sid = SentimentIntensityAnalyzer()

def analyzer(data):
    neg=[]
    neu=[]
    pos=[]
    compound=[]
    for i in tqdm(data):
        s=sid.polarity_scores(i)
        neg.append(s['neg'])
        neu.append(s['neu'])
        pos.append(s['pos'])
        compound.append(s['compound'])

    return neg,neu,pos,compound

neg,neu,pos,compound=analyzer(train_essays)
neg_test,neu_test,pos_test,compound_test=analyzer(test_essays)
```

```
100%|
| 23450/23450 [00:49<00:00, 474.69it/s]
100%|
| 11550/11550 [00:25<00:00, 452.34it/s]
```

[13]:
```
neg=np.reshape(neg,(-1,1))
neu=np.reshape(neu,(-1,1))
pos=np.reshape(pos,(-1,1))
compound=np.reshape(compound,(-1,1))


neg_test=np.reshape(neg_test,(-1,1))
neu_test=np.reshape(neu_test,(-1,1))
pos_test=np.reshape(pos_test,(-1,1))
compound_test=np.reshape(compound_test,(-1,1))
```

2.0.4 2.4 Numerical features

[14]:
```
normalizer = Normalizer()
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))

normalizer.fit(X_train['teacher_number_of_previously_posted_projects'].values.
 ↪reshape(-1,1))
X_train_num_norm = normalizer.
 ↪transform(X_train['teacher_number_of_previously_posted_projects'].values.
 ↪reshape(-1,1))
X_test_num_norm = normalizer.
 ↪transform(X_test['teacher_number_of_previously_posted_projects'].values.
 ↪reshape(-1,1))
```

3 3. Data set 1

[15]:
```
from scipy.sparse import hstack
X_tr = hstack((train_text_tfidf ,  X_cat_train , X_train_price_norm ,␣
 ↪X_train_num_norm , neg , neu , pos , compound)).tocsr()
X_te = hstack((test_text_tfidf , X_cat_test , X_test_price_norm ,␣
 ↪X_test_num_norm , neg_test , neu_test , pos_test , compound_test)).tocsr()

print(X_tr.get_shape(),"shape of training data set")
print(X_te.get_shape(),"shape of testing data set")
```

```
(23450, 8959) shape of training data set
(11550, 8959) shape of testing data set
```

3.0.1 3.1 Finding the best hyperparameters

[16]:
```
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import GridSearchCV
# https://stackoverflow.com/questions/37689942/
 ↪grid-search-finding-parameters-for-auc

DB = GradientBoostingClassifier(random_state=0)

parameters={'max_depth':[1, 2, 3, 4],
            "n_estimators":[5,10,15,20] }

clf = GridSearchCV(DB, parameters, cv=5, scoring='roc_auc', n_jobs=-1,␣
 ↪return_train_score=True)

%time clf.fit(X_tr,Y_train)
```

```
Wall time: 3min 51s
```

[16]:
```
GridSearchCV(cv=5, estimator=GradientBoostingClassifier(random_state=0),
             n_jobs=-1,
             param_grid={'max_depth': [1, 2, 3, 4],
                         'n_estimators': [5, 10, 15, 20]},
             return_train_score=True, scoring='roc_auc')
```

[17]:
```
from sklearn.metrics import roc_auc_score
print(clf.best_params_)
print(clf.best_score_)

Y_pred=clf.predict_proba(X_te)[:,1]
roc_auc_score(Y_test,Y_pred)
```

```
{'max_depth': 4, 'n_estimators': 20}
0.6693683391153492
```

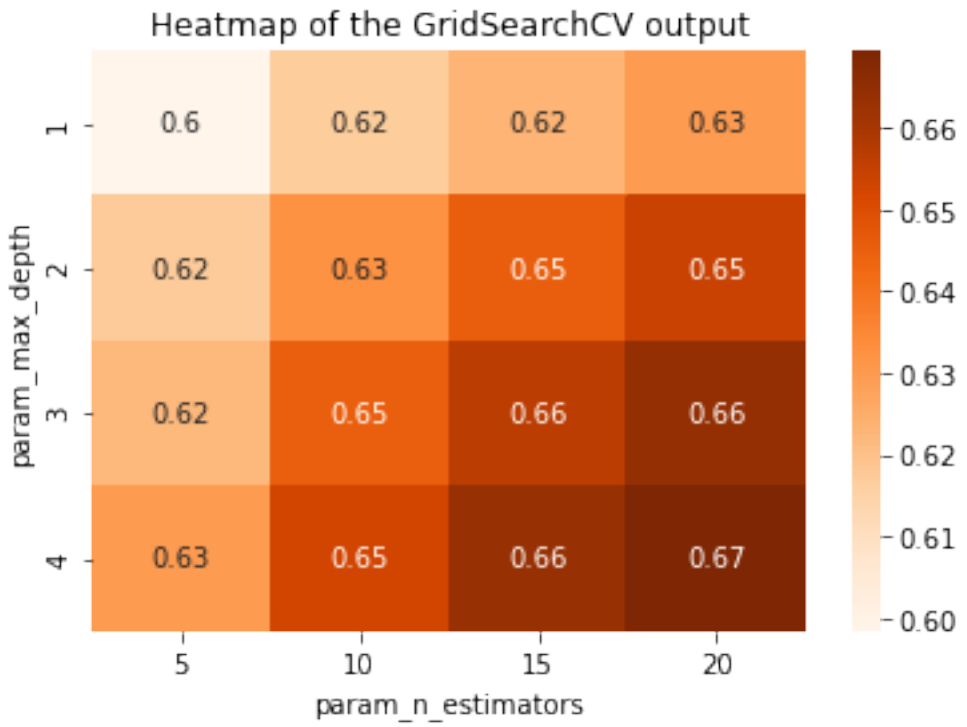
[17]:
```
0.6551416041078073
```

3.0.2 3.2 Plotting heat map of parameters and there outputs

[18]:
```
import seaborn as sns
x=clf.cv_results_['param_n_estimators'].data
y=clf.cv_results_['param_max_depth'].data
z=clf.cv_results_['mean_test_score']

mat=np.zeros((4,4))
for i in range(len(z)):
    mat.itemset(i, z[i])


ax= plt.subplot();
sns.heatmap(mat, annot=True,cmap='Oranges',ax=ax);
# labels, title and ticks
ax.set_xlabel('param_n_estimators');ax.set_ylabel('param_max_depth');
ax.set_title('Heatmap of the GridSearchCV output');
ax.xaxis.set_ticklabels(['5','10','15','20']);
ax.yaxis.set_ticklabels(['1','2','3','4']);
```



3.0.3 3.3 Training the best model

[19]:
```
from sklearn.metrics import confusion_matrix

DB_best = GradientBoostingClassifier(random_state=0,max_depth= 4,n_estimators=␣
↪20)
DB_best.fit(X_tr,Y_train)
```

[19]:
```
GradientBoostingClassifier(max_depth=4, n_estimators=20, random_state=0)
```

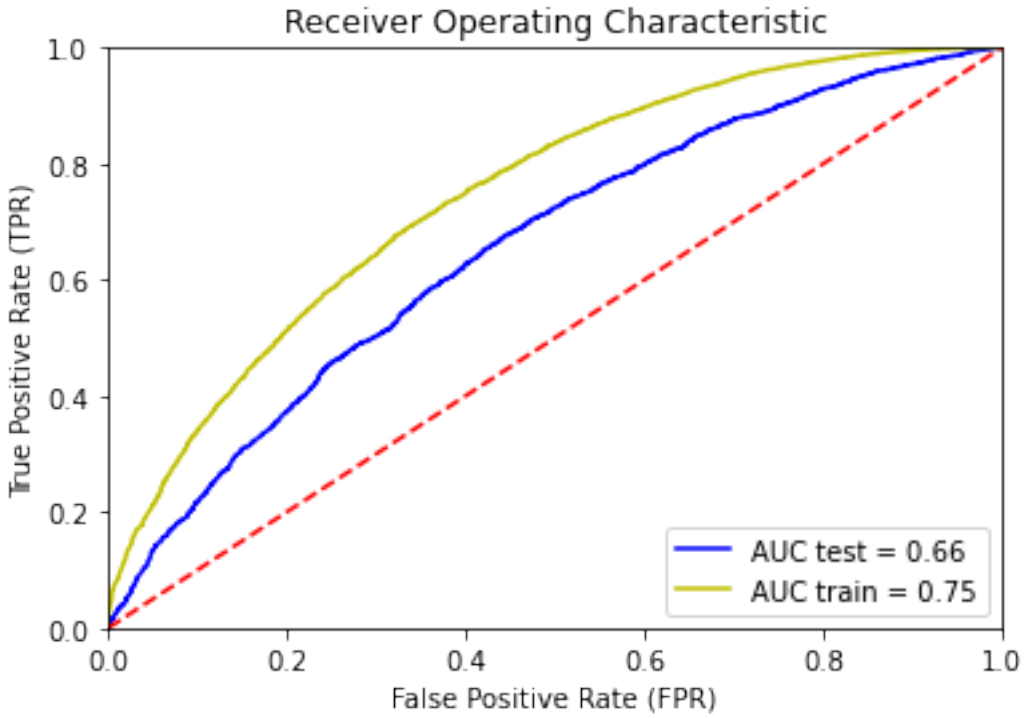
3.0.4 3.4 AUC

[22]:
```
y_pred=DB_best.predict_proba(X_te)[:,1]     # predict the probabitity of class 1␣
↪= (approved)
print("test= ",metrics.roc_auc_score(Y_test, y_pred))
fpr, tpr, threshold = metrics.roc_curve(Y_test, y_pred)
roc_auc_test = metrics.auc(fpr, tpr)

y_pred_class_train = DB_best.predict_proba(X_tr)[:,1]    # predict the␣
↪probabitity of class 1 = (approved)
print("train= ",metrics.roc_auc_score(Y_train, y_pred_class_train))
fpr_1, tpr_1, threshold_1 = metrics.roc_curve(Y_train, y_pred_class_train)
roc_auc_train = metrics.auc(fpr_1, tpr_1)
```

```
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b', label = 'AUC test = %0.2f' % roc_auc_test)
plt.plot(fpr_1, tpr_1, 'y', label = 'AUC train = %0.2f' % roc_auc_train)
plt.legend(loc = 'lower right')
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
plt.ylim([0, 1])
plt.ylabel('True Positive Rate (TPR)')
plt.xlabel('False Positive Rate (FPR)')
plt.show()
```

```
test=  0.6551416041078073
train=  0.745515594373414
```



3.0.5 3.5 confusion matrix

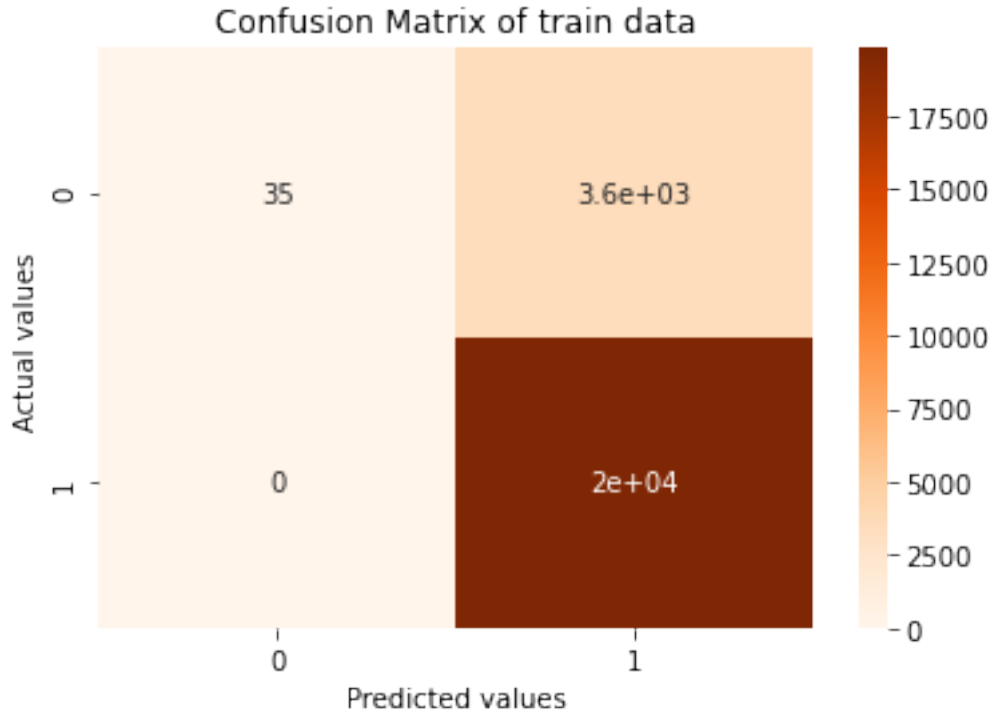
Train data set

[24]:
```
import seaborn as sns

y_pred_bi=DB_best.predict(X_tr)
cf_matrix=confusion_matrix(Y_train,y_pred_bi)
```

```
ax= plt.subplot();
sns.heatmap(cf_matrix, annot=True,cmap='Oranges',ax=ax);
# labels, title and ticks
ax.set_xlabel('Predicted values');ax.set_ylabel('Actual values');
ax.set_ylim(2.0, 0)
ax.set_title('Confusion Matrix of train data');
ax.xaxis.set_ticklabels(['0','1']);
ax.yaxis.set_ticklabels(['0','1']);
print(cf_matrix)
```

```
[[   35  3564]
 [    0 19851]]
```



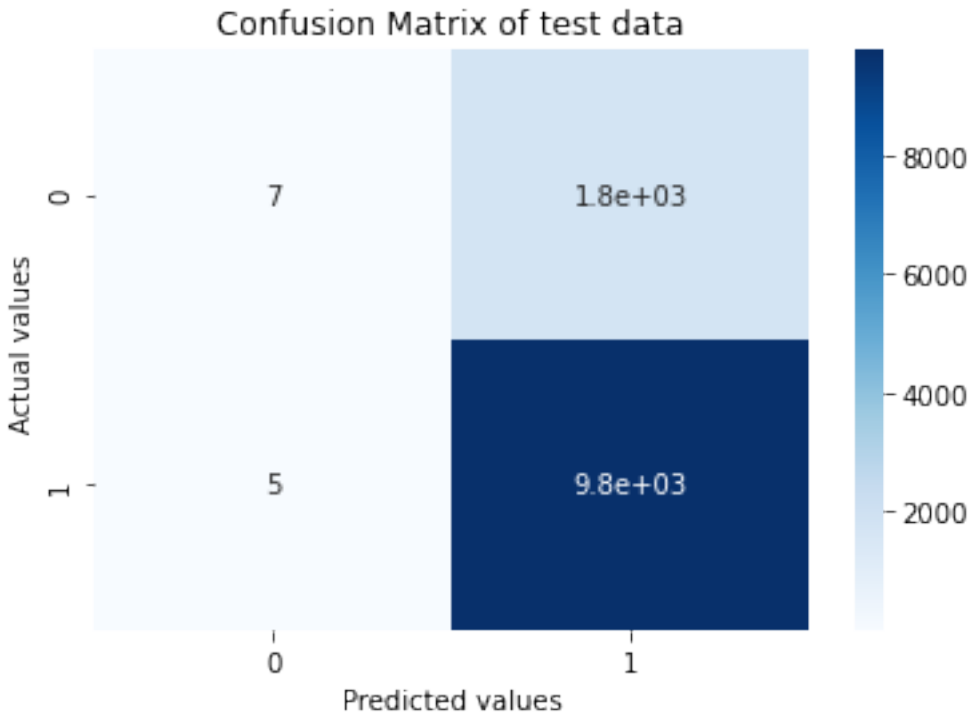
Test data set

[25]:
```
y_pred_bi=DB_best.predict(X_te)
cf_matrix=confusion_matrix(Y_test,y_pred_bi)

ax= plt.subplot();
sns.heatmap(cf_matrix, annot=True,cmap='Blues',ax=ax);
# labels, title and ticks
ax.set_xlabel('Predicted values');ax.set_ylabel('Actual values');
ax.set_ylim(2.0, 0)
ax.set_title('Confusion Matrix of test data');
```

```
ax.xaxis.set_ticklabels(['0','1']);
ax.yaxis.set_ticklabels(['0','1']);
print(cf_matrix)
```

```
[[   7 1765]
 [   5 9773]]
```



4 4. Data set 2

[35]:
```
from scipy.sparse import hstack


X_tr_2 = np.column_stack((train_w2v ,X_cat_train , X_train_price_norm ,␣
 ↪X_train_num_norm , neg , neu , pos , compound))
X_te_2 = np.column_stack((test_w2v , X_cat_test , X_test_price_norm ,␣
 ↪X_test_num_norm , neg_test , neu_test , pos_test , compound_test))

print(X_tr_2.shape,"shape of training data set")
print(X_te_2.shape,"shape of testing data set")
```

```
(23450, 316) shape of training data set
(11550, 316) shape of testing data set
```

4.0.1 4.1 Finding the best hyperparameter

[37]:
```
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import GridSearchCV
# https://stackoverflow.com/questions/37689942/
 ↪grid-search-finding-parameters-for-auc

DB_2 = GradientBoostingClassifier(random_state=0)

parameters={'max_depth':[1, 2, 3, 4],
            "n_estimators":[5,10,15,20] }

clf = GridSearchCV(DB_2, parameters, cv=5, scoring='roc_auc', n_jobs=-1,␣
 ↪return_train_score=True)

%time clf.fit(X_tr_2,Y_train)
```

```
Wall time: 8min 1s
```

[37]:
```
GridSearchCV(cv=5, estimator=GradientBoostingClassifier(random_state=0),
             n_jobs=-1,
             param_grid={'max_depth': [1, 2, 3, 4],
                         'n_estimators': [5, 10, 15, 20]},
             return_train_score=True, scoring='roc_auc')
```

[39]:
```
from sklearn.metrics import roc_auc_score
print(clf.best_params_)
print(clf.best_score_)

Y_pred=clf.predict_proba(X_te_2)[:,1]
roc_auc_score(Y_test,Y_pred)
```

```
{'max_depth': 4, 'n_estimators': 20}
0.6680806998246132
```

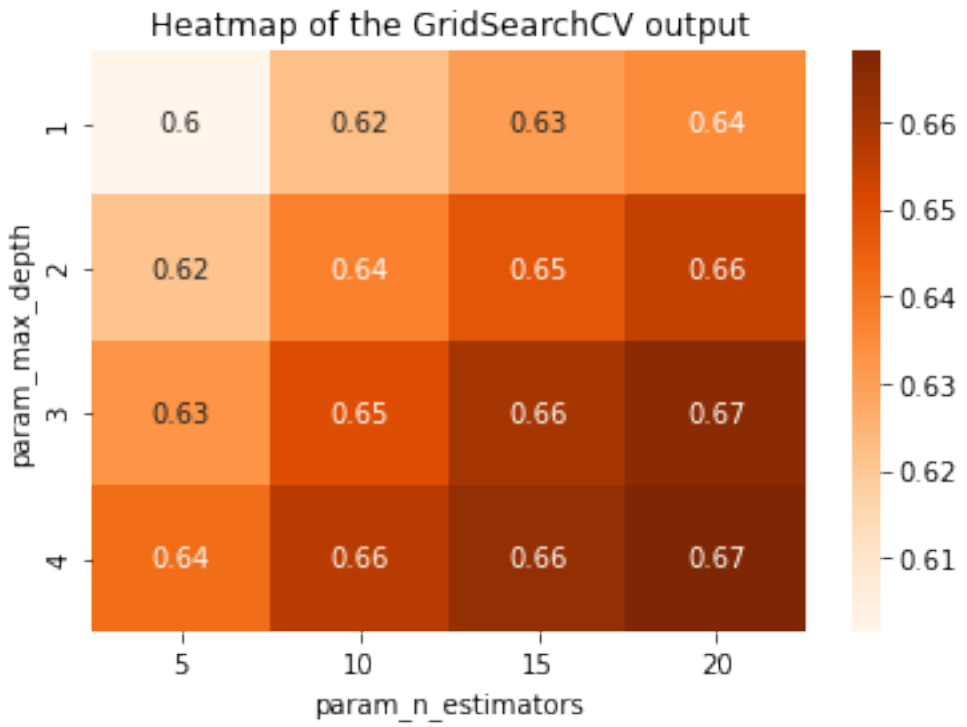
[39]:
```
0.6642239315513196
```

4.0.2 4.2 Heatmap of hyperparameter

[40]:
```
import seaborn as sns
x=clf.cv_results_['param_n_estimators'].data
y=clf.cv_results_['param_max_depth'].data
z=clf.cv_results_['mean_test_score']

mat=np.zeros((4,4))
for i in range(len(z)):
    mat.itemset(i, z[i])
```

```
ax= plt.subplot();
sns.heatmap(mat, annot=True,cmap='Oranges',ax=ax);
# labels, title and ticks
ax.set_xlabel('param_n_estimators');ax.set_ylabel('param_max_depth');
ax.set_title('Heatmap of the GridSearchCV output');
ax.xaxis.set_ticklabels(['5','10','15','20']);
ax.yaxis.set_ticklabels(['1','2','3','4']);
```



4.0.3 4.3 Confusion matrix

[42]:
```
from sklearn.metrics import confusion_matrix

DB_best = GradientBoostingClassifier(random_state=0,max_depth= 4,n_estimators=␣
↪20)
DB_best.fit(X_tr_2,Y_train)
```

[42]:
```
GradientBoostingClassifier(max_depth=4, n_estimators=20, random_state=0)
```

4.0.4 4.4 AUC

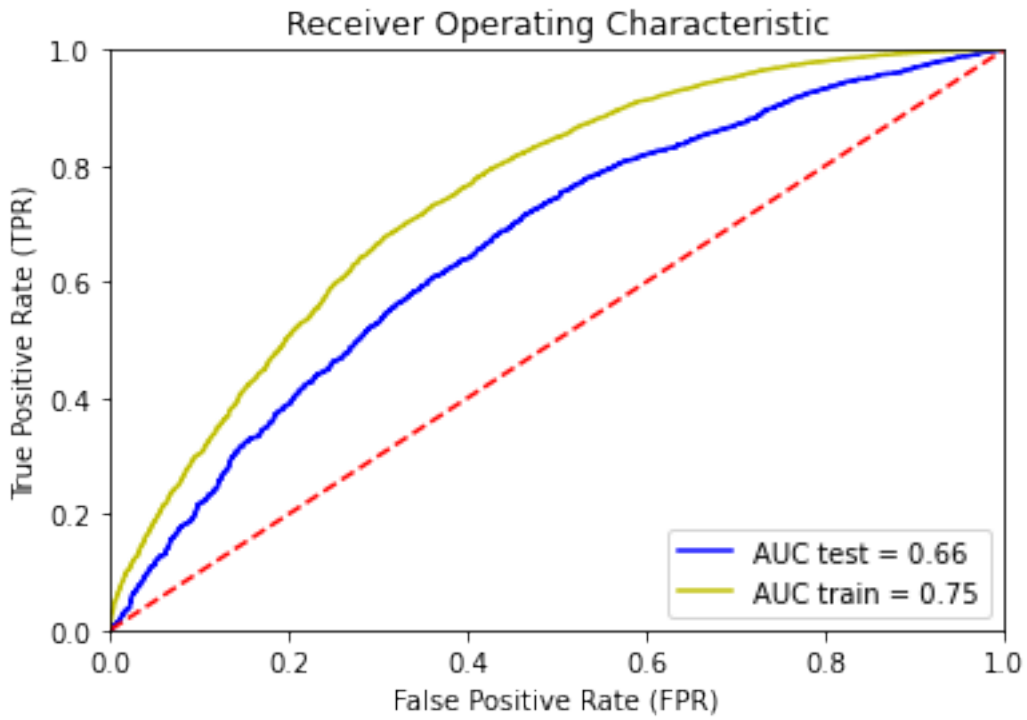
[43]:
```
y_pred=DB_best.predict_proba(X_te_2)[:,1]     # predict the probabitity of class␣
↪1 = (approved)
print("test= ",metrics.roc_auc_score(Y_test, y_pred))
```

```
fpr, tpr, threshold = metrics.roc_curve(Y_test, y_pred)
roc_auc_test = metrics.auc(fpr, tpr)

y_pred_class_train = DB_best.predict_proba(X_tr_2)[:,1]   # predict the␣
 ↪probabitity of class 1 = (approved)
print("train= ",metrics.roc_auc_score(Y_train, y_pred_class_train))
fpr_1, tpr_1, threshold_1 = metrics.roc_curve(Y_train, y_pred_class_train)
roc_auc_train = metrics.auc(fpr_1, tpr_1)

plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b', label = 'AUC test = %0.2f' % roc_auc_test)
plt.plot(fpr_1, tpr_1, 'y', label = 'AUC train = %0.2f' % roc_auc_train)
plt.legend(loc = 'lower right')
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
plt.ylim([0, 1])
plt.ylabel('True Positive Rate (TPR)')
plt.xlabel('False Positive Rate (FPR)')
plt.show()
```

```
test=  0.6642239315513196
train=  0.7479904225070833
```



4.0.5 4.5 Confusion matrix

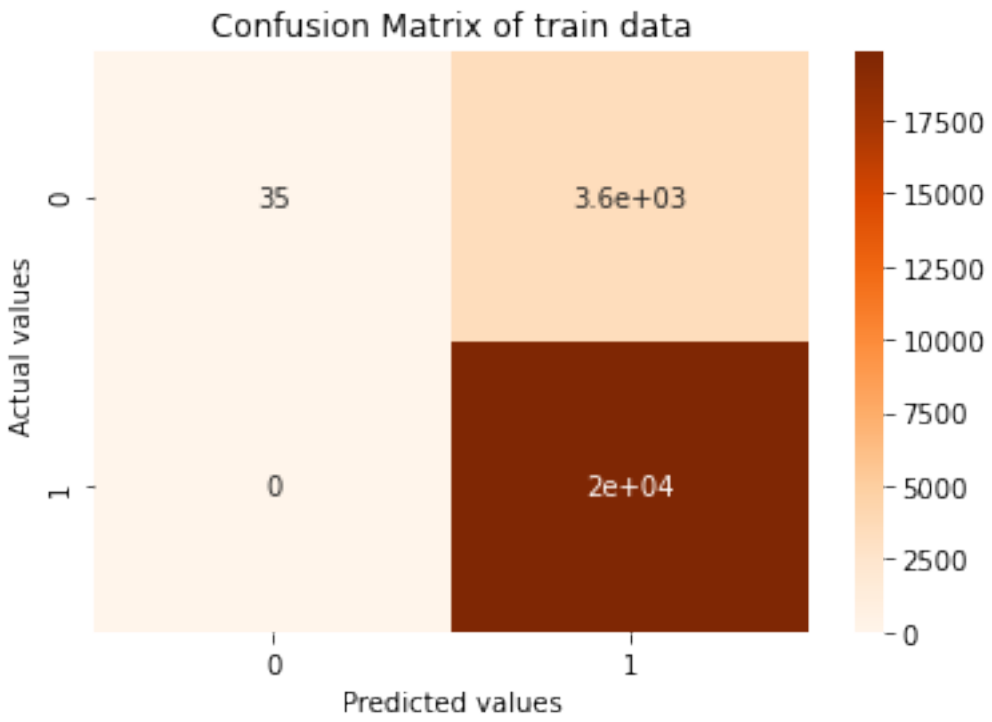
Train data set

[44]:
```
import seaborn as sns

y_pred_bi=DB_best.predict(X_tr_2)
cf_matrix=confusion_matrix(Y_train,y_pred_bi)

ax= plt.subplot();
sns.heatmap(cf_matrix, annot=True,cmap='Oranges',ax=ax);
# labels, title and ticks
ax.set_xlabel('Predicted values');ax.set_ylabel('Actual values');
ax.set_ylim(2.0, 0)
ax.set_title('Confusion Matrix of train data');
ax.xaxis.set_ticklabels(['0','1']);
ax.yaxis.set_ticklabels(['0','1']);
print(cf_matrix)
```

```
[[   35  3564]
 [    0 19851]]
```

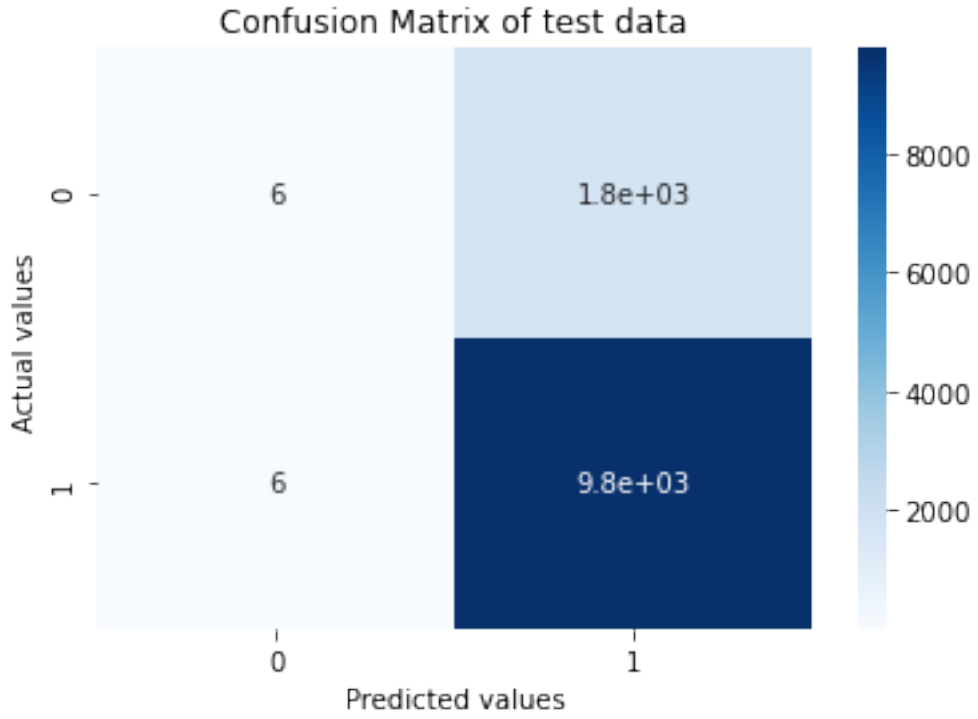


Test data set

[45]:
```
y_pred_bi=DB_best.predict(X_te_2)
cf_matrix=confusion_matrix(Y_test,y_pred_bi)

ax= plt.subplot();
sns.heatmap(cf_matrix, annot=True,cmap='Blues',ax=ax);
# labels, title and ticks
ax.set_xlabel('Predicted values');ax.set_ylabel('Actual values');
ax.set_ylim(2.0, 0)
ax.set_title('Confusion Matrix of test data');
ax.xaxis.set_ticklabels(['0','1']);
ax.yaxis.set_ticklabels(['0','1']);
print(cf_matrix)
```

```
[[   6 1766]
 [   6 9772]]
```



5 5. Outputs

[54]:
```
from prettytable import PrettyTable
x = PrettyTable()
x.field_names = ["Vectorizer", "Model", "Hyper parameter(max depth)","hyper␣
 ↪parameter(n_estimators) ","AUC"]
x.add_row(["TFIDF", "GBDT", 4,20, 0.6551416041078073])
```

```
x.add_row(["TFIDF W2V", "GBDT", 4,20, 0.6642239315513196])
print(x)
```

```
+------------+-------+-----------------------------+-----------------------------
---+--------------------+
| Vectorizer | Model | Hyper parameter(max depth) | hyper
parameter(n_estimators)  |        AUC         |
+------------+-------+-----------------------------+-----------------------------
---+--------------------+
|   TFIDF    |  GBDT |             4               |               20
| 0.6551416041078073 |
| TFIDF W2V  |  GBDT |             4               |               20
| 0.6642239315513196 |
+------------+-------+-----------------------------+-----------------------------
---+--------------------+
```

[ ]: