Imports

In [42]:
```
import itertools

from sklearn.model_selection import train_test_split, KFold, cross_val
_score
from sklearn.ensemble import RandomForestClassifier, RandomForestRegre
ssor
from sklearn.metrics import classification_report, confusion_matrix, r
oc_auc_score, roc_curve
from sklearn.linear_model import LogisticRegression
from imblearn import under_sampling
from imblearn.under_sampling import RandomUnderSampler
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score
from sklearn.svm import SVC
import numpy as np
import pandas as pd
import ast

import matplotlib.pyplot as plt
%matplotlib inline
```

Data pre-processing

In [5]:
```
# load file and get long term and short term features
def load(file):

    df = pd.read_csv(file)

    # convert the column values from literal string to dictionary
    df['ltiFeatures'] = df['ltiFeatures'].apply(ast.literal_eval)
    df['stiFeatures'] = df['stiFeatures'].apply(ast.literal_eval)

    return df

# load all the data
training = load("training.csv")
validation = load("validation.csv")
interest_topics = pd.read_csv("interest_topics.csv")
```

In [7]:
```
# inspect the data
print(interest_topics.head())

print(training.head())

print(validation.head())
```

```
   topic_id                topic_name
0         3     /Arts & Entertainment
1         5  /Computers & Electronics
2         7                  /Finance
3         8                    /Games
4        11            /Home & Garden
   userID  inAudience                                              ltiFeat
ures  \
0       1        True  {'45': 0.020536141517834786, '47': 0.00311752
9...
1       2        True  {'45': 0.001158253110658664, '592': 0.0154638
0...
2       3        True  {'908': 0.002470851264264668, '590': 0.002140
2...
3       4        True  {'1187': 0.001127974558171163, '1780': 0.0011
7...
4       5        True  {'907': 0.025339209040149392, '1187': 0.00602
0...

                                        stiFeatures
0                                                {}
1                                                {}
2                                                {}
3                                                {}
4  {'907': 0.10445132121076425, '908': 0.05651522...
   userID  inAudience                                              ltiFeat
ures  \
0       0        True  {'89': 0.0027281240558934, '1264': 0.00186295
8...
1       1        True  {'47': 0.0019292939671486482, '1187': 0.01226
1...
2       2        True  {'45': 0.001961152113619305, '47': 0.00158412
6...
3       3        True  {'1253': 0.006566573072362829, '1164': 0.0032
7...
4       4        True  {'78': 0.013096540307802428, '1198': 0.002554
6...

  stiFeatures
0          {}
1          {}
2          {}
3          {}
4          {}
```

In [8]:
```
# Use topic names for each topic ID to craete new IDs that group every
topic with the same general category together

Id_to_topic= {}
topic_to_newID= {}
index= 0

print(len(interest_topics))
for row in interest_topics.iterrows():
    Id= row[1][0]
    Id=str(Id)
    topic= row[1][1]

    topic= topic.split('/')[1]
    Id_to_topic[Id]= topic
    if topic not in topic_to_newID.keys():
        topic_to_newID[topic]= index
        index+=1

print(Id_to_topic)
print(topic_to_newID)
```

```
1411
{'3': 'Arts & Entertainment', '5': 'Computers & Electronics', '7': '
Finance', '8': 'Games', '11': 'Home & Garden', '12': 'Business & Ind
ustrial', '13': 'Internet & Telecom', '14': 'People & Society', '16'
: 'News', '18': 'Shopping', '19': 'Law & Government', '20': 'Sports'
, '22': 'Books & Literature', '23': 'Arts & Entertainment', '24': 'A
rts & Entertainment', '25': 'Business & Industrial', '28': 'Business
& Industrial', '29': 'Real Estate', '30': 'Computers & Electronics',
'31': 'Computers & Electronics', '32': 'Computers & Electronics', '3
3': 'Arts & Entertainment', '34': 'Arts & Entertainment', '35': 'Art
s & Entertainment', '36': 'Arts & Entertainment', '37': 'Finance', '
38': 'Finance', '39': 'Games', '41': 'Games', '42': 'Arts & Entertai
nment', '43': 'Online Communities', '44': 'Beauty & Fitness', '45':
'Health', '46': 'Business & Industrial', '47': 'Autos & Vehicles', '
48': 'Business & Industrial', '49': 'Business & Industrial', '50': '
Business & Industrial', '53': 'Internet & Telecom', '54': 'People &
Society', '55': 'Online Communities', '56': 'People & Society', '57'
: 'People & Society', '58': 'People & Society', '59': 'People & Soci
ety', '60': 'Jobs & Education', '61': 'Shopping', '63': 'News', '64'
: 'Shopping', '65': 'Hobbies & Leisure', '66': 'Pets & Animals', '67
': 'Travel', '68': 'Shopping', '69': 'Shopping', '70': 'Shopping', '
71': 'Food & Drink', '73': 'Shopping', '74': 'Jobs & Education', '75
': 'Law & Government', '76': 'Law & Government', '77': 'Computers &
Electronics', '78': 'Computers & Electronics', '82': 'People & Socie
ty', '83': 'Business & Industrial', '84': 'Internet & Telecom', '89'
: 'Autos & Vehicles', '91': 'Computers & Electronics', '93': 'Beauty
```

& Fitness', '94': 'Beauty & Fitness', '95': 'Business & Industrial', '96': 'Real Estate', '97': 'Shopping', '98': 'Beauty & Fitness', '99': 'Shopping', '100': 'Shopping', '101': 'People & Society', '102': 'Online Communities', '104': 'Internet & Telecom', '105': 'Games', '107': 'Finance', '108': 'Reference', '112': 'News', '113': 'People & Society', '115': 'People & Society', '118': 'Sports', '119': 'Pets & Animals', '120': 'Home & Garden', '121': 'Food & Drink', '122': 'Food & Drink', '124': 'Shopping', '137': 'Home & Garden', '138': 'Autos & Vehicles', '143': 'Beauty & Fitness', '144': 'Beauty & Fitness', '145': 'Beauty & Fitness', '146': 'Beauty & Fitness', '147': 'Beauty & Fitness', '148': 'Autos & Vehicles', '157': 'Business & Industrial', '158': 'Home & Garden', '166': 'Law & Government', '168': 'Law & Government', '170': 'Autos & Vehicles', '171': 'People & Society', '174': 'Science', '179': 'Travel', '180': 'Sports', '182': 'Arts & Entertainment', '184': 'Arts & Entertainment', '185': 'Beauty & Fitness', '188': 'Arts & Entertainment', '189': 'Hobbies & Leisure', '191': 'Online Communities', '195': 'Health', '198': 'Health', '202': 'Health', '203': 'Travel', '205': 'Travel', '206': 'Travel', '208': 'Travel', '210': 'Arts & Entertainment', '211': 'Arts & Entertainment', '213': 'Arts & Entertainment', '215': 'Arts & Entertainment', '216': 'Arts & Entertainment', '217': 'Arts & Entertainment', '218': 'Arts & Entertainment', '220': 'Arts & Entertainment', '224': 'Computers & Electronics', '225': 'Computers & Electronics', '226': 'Computers & Electronics', '227': 'Computers & Electronics', '228': 'Computers & Electronics', '229': 'Computers & Electronics', '230': 'Computers & Electronics', '231': 'Science', '233': 'Business & Industrial', '234': 'Beauty & Fitness', '235': 'Beauty & Fitness', '236': 'Beauty & Fitness', '237': 'Health', '238': 'Health', '239': 'Beauty & Fitness', '241': 'Beauty & Fitness', '242': 'Beauty & Fitness', '244': 'Beauty & Fitness', '245': 'Health', '246': 'Health', '248': 'Health', '249': 'Finance', '250': 'Health', '251': 'Health', '252': 'Health', '253': 'Health', '254': 'Health', '255': 'Business & Industrial', '256': 'Health', '257': 'Health', '258': 'Sports', '259': 'Sports', '260': 'Sports', '261': 'Sports', '262': 'Sports', '263': 'Sports', '264': 'Sports', '265': 'Sports', '268': 'Reference', '269': 'Home & Garden', '270': 'Home & Garden', '271': 'Home & Garden', '272': 'Home & Garden', '273': 'Autos & Vehicles', '275': 'Online Communities', '276': 'Food & Drink', '278': 'Finance', '279': 'Finance', '280': 'Business & Industrial', '284': 'Hobbies & Leisure', '287': 'Business & Industrial', '288': 'Business & Industrial', '289': 'Business & Industrial', '290': 'Business & Industrial', '291': 'Business & Industrial', '292': 'Shopping', '293': 'Hobbies & Leisure', '294': 'Sports', '296': 'Sports', '297': 'Food & Drink', '298': 'People & Society', '299': 'Online Communities', '301': 'Internet & Telecom', '302': 'Internet & Telecom', '303': 'Computers & Electronics', '304': 'Computers & Electronics', '305': 'Computers & Electronics', '306': 'Computers & Electronics', '307': 'Computers & Electronics', '308': 'Computers & Electronics', '309': 'Computers & Electronics', '310': 'Computers & Electronics', '311': 'Computers & Electronics', '312': 'Computers & Electronics', '313': 'Computers & Electronics', '314': 'Computer

s & Electronics', '315': 'Computers & Electronics', '316': 'Arts & E
ntertainment', '317': 'Arts & Entertainment', '318': 'Arts & Enterta
inment', '319': 'Arts & Entertainment', '320': 'Online Communities',
'321': 'Online Communities', '323': 'Shopping', '324': 'Shopping', '
325': 'People & Society', '326': 'Internet & Telecom', '327': 'Busin
ess & Industrial', '328': 'Business & Industrial', '329': 'Business
& Industrial', '330': 'Business & Industrial', '331': 'Business & In
dustrial', '332': 'Business & Industrial', '333': 'Business & Indust
rial', '334': 'Business & Industrial', '335': 'Business & Industrial
', '336': 'Business & Industrial', '337': 'Business & Industrial', '
338': 'Business & Industrial', '340': 'Business & Industrial', '341'
: 'Computers & Electronics', '342': 'Computers & Electronics', '343'
: 'Computers & Electronics', '344': 'Computers & Electronics', '346'
: 'Computers & Electronics', '347': 'Computers & Electronics', '350'
: 'Shopping', '352': 'Shopping', '353': 'Shopping', '354': 'Business
& Industrial', '355': 'Books & Literature', '356': 'Business & Indus
trial', '357': 'Arts & Entertainment', '358': 'Arts & Entertainment'
, '359': 'Arts & Entertainment', '360': 'Arts & Entertainment', '361
': 'Computers & Electronics', '362': 'Computers & Electronics', '365
': 'Shopping', '366': 'Law & Government', '367': 'Jobs & Education',
'369': 'Jobs & Education', '371': 'Jobs & Education', '372': 'Jobs &
Education', '373': 'Jobs & Education', '374': 'Reference', '375': 'R
eference', '377': 'Reference', '378': 'Real Estate', '379': 'Pets &
Animals', '380': 'Pets & Animals', '381': 'Games', '382': 'Internet
& Telecom', '383': 'Internet & Telecom', '384': 'Internet & Telecom'
, '385': 'Internet & Telecom', '386': 'Internet & Telecom', '389': '
Internet & Telecom', '390': 'Internet & Telecom', '392': 'Internet &
Telecom', '394': 'Internet & Telecom', '396': 'News', '398': 'News',
'400': 'People & Society', '401': 'People & Society', '402': 'People
& Society', '403': 'People & Society', '408': 'News', '409': 'News',
'410': 'News', '412': 'Books & Literature', '418': 'Health', '419':
'Health', '420': 'Health', '421': 'Health', '422': 'Internet & Telec
om', '423': 'Law & Government', '424': 'Law & Government', '425': 'R
eal Estate', '426': 'Law & Government', '427': 'Law & Government', '
428': 'Shopping', '429': 'Health', '432': 'Shopping', '433': 'Refere
nce', '434': 'Computers & Electronics', '435': 'Science', '436': 'Sc
ience', '437': 'Health', '438': 'Autos & Vehicles', '439': 'Arts & E
ntertainment', '440': 'Science', '441': 'Science', '442': 'Science',
'443': 'Science', '444': 'Science', '445': 'Science', '446': 'Scienc
e', '447': 'Arts & Entertainment', '448': 'People & Society', '449':
'Arts & Entertainment', '450': 'Shopping', '451': 'Shopping', '456':
'Health', '457': 'Health', '458': 'Sports', '459': 'Hobbies & Leisur
e', '461': 'Hobbies & Leisure', '462': 'Hobbies & Leisure', '463': '
Real Estate', '465': 'Finance', '466': 'Finance', '467': 'Finance',
'468': 'Finance', '471': 'Home & Garden', '472': 'Home & Garden', '4
73': 'Autos & Vehicles', '474': 'Reference', '477': 'Arts & Entertai
nment', '485': 'Internet & Telecom', '486': 'Computers & Electronics
', '487': 'Computers & Electronics', '488': 'Computers & Electronics
', '492': 'Computers & Electronics', '493': 'Computers & Electronics
', '494': 'Computers & Electronics', '495': 'Computers & Electronics

', '496': 'Computers & Electronics', '497': 'Computers & Electronics
', '498': 'Computers & Electronics', '499': 'Health', '500': 'Health
', '501': 'Internet & Telecom', '502': 'People & Society', '503': 'P
eople & Society', '504': 'Online Communities', '505': 'Science', '50
7': 'News', '508': 'Law & Government', '509': 'People & Society', '5
10': 'People & Society', '511': 'Health', '512': 'Sports', '513': 'S
ports', '514': 'Sports', '515': 'Sports', '516': 'Sports', '517': 'S
ports', '518': 'Sports', '519': 'Sports', '520': 'People & Society',
'521': 'People & Society', '522': 'Law & Government', '525': 'People
& Society', '527': 'Reference', '528': 'People & Society', '529': 'O
nline Communities', '530': 'Shopping', '531': 'Shopping', '532': 'In
ternet & Telecom', '533': 'Reference', '534': 'Sports', '535': 'Law
& Government', '536': 'Health', '538': 'Arts & Entertainment', '539'
: 'Arts & Entertainment', '540': 'Books & Literature', '541': 'Sport
s', '542': 'Hobbies & Leisure', '543': 'People & Society', '546': 'O
nline Communities', '547': 'People & Society', '548': 'People & Soci
ety', '549': 'People & Society', '550': 'People & Society', '551': '
Business & Industrial', '552': 'Business & Industrial', '554': 'Spor
ts', '555': 'Law & Government', '556': 'People & Society', '557': 'B
eauty & Fitness', '558': 'Health', '559': 'Health', '560': 'Food & D
rink', '563': 'Pets & Animals', '565': 'Books & Literature', '566':
'Business & Industrial', '567': 'Shopping', '568': 'Sports', '569':
'Arts & Entertainment', '571': 'Health', '572': 'News', '573': 'Comp
uters & Electronics', '574': 'Shopping', '575': 'Arts & Entertainmen
t', '576': 'Shopping', '577': 'Computers & Electronics', '578': 'Onl
ine Communities', '579': 'People & Society', '580': 'People & Societ
y', '581': 'Arts & Entertainment', '582': 'Online Communities', '585
': 'Arts & Entertainment', '586': 'Arts & Entertainment', '587': 'Ar
ts & Entertainment', '588': 'Arts & Entertainment', '589': 'Arts & E
ntertainment', '590': 'Arts & Entertainment', '592': 'Arts & Enterta
inment', '594': 'People & Society', '606': 'Business & Industrial',
'607': 'Pets & Animals', '608': 'Books & Literature', '609': 'Refere
nce', '610': 'Autos & Vehicles', '611': 'Beauty & Fitness', '612': '
Arts & Entertainment', '613': 'Arts & Entertainment', '614': 'Arts &
Entertainment', '615': 'Arts & Entertainment', '616': 'Arts & Entert
ainment', '617': 'Arts & Entertainment', '618': 'Arts & Entertainmen
t', '619': 'Finance', '620': 'Business & Industrial', '621': 'Busine
ss & Industrial', '622': 'Games', '623': 'Health', '624': 'Health',
'625': 'Health', '626': 'Health', '627': 'Health', '628': 'Health',
'629': 'Health', '630': 'Health', '631': 'Health', '632': 'Health',
'633': 'Health', '634': 'Health', '635': 'Health', '636': 'Health',
'638': 'Health', '639': 'Health', '640': 'Health', '641': 'Health',
'642': 'Health', '643': 'Health', '644': 'Health', '645': 'Health',
'646': 'Health', '647': 'Health', '648': 'Health', '649': 'Health',
'650': 'Business & Industrial', '651': 'Business & Industrial', '652
': 'Business & Industrial', '653': 'Arts & Entertainment', '654': 'A
rts & Entertainment', '655': 'Arts & Entertainment', '656': 'Arts &
Entertainment', '657': 'Business & Industrial', '658': 'Business & I
ndustrial', '659': 'Business & Industrial', '660': 'Business & Indus
trial', '661': 'Business & Industrial', '662': 'Business & Industria

l', '663': 'Business & Industrial', '664': 'Business & Industrial', '665': 'Business & Industrial', '666': 'Business & Industrial', '667': 'Business & Industrial', '668': 'Business & Industrial', '669': 'Business & Industrial', '670': 'Business & Industrial', '671': 'Business & Industrial', '672': 'Business & Industrial', '673': 'Business & Industrial', '674': 'Business & Industrial', '675': 'Internet & Telecom', '676': 'People & Society', '677': 'People & Society', '678': 'Hobbies & Leisure', '681': 'People & Society', '682': 'People & Society', '683': 'People & Society', '685': 'Reference', '686': 'Business & Industrial', '687': 'Real Estate', '688': 'Hobbies & Leisure', '689': 'Hobbies & Leisure', '690': 'Reference', '691': 'Reference', '692': 'Reference', '693': 'Reference', '694': 'Reference', '695': 'Reference', '696': 'Shopping', '697': 'Shopping', '699': 'Sports', '700': 'Jobs & Education', '701': 'Law & Government', '702': 'Law & Government', '703': 'People & Society', '704': 'Law & Government', '705': 'Law & Government', '706': 'Law & Government', '707': 'Travel', '708': 'Travel', '717': 'Computers & Electronics', '718': 'Business & Industrial', '719': 'Business & Industrial', '720': 'Business & Industrial', '721': 'Business & Industrial', '722': 'Business & Industrial', '723': 'Business & Industrial', '724': 'Business & Industrial', '725': 'Business & Industrial', '726': 'Business & Industrial', '727': 'Business & Industrial', '728': 'Computers & Electronics', '729': 'Computers & Electronics', '730': 'Computers & Electronics', '731': 'Computers & Electronics', '732': 'Computers & Electronics', '733': 'Computers & Electronics', '734': 'Computers & Electronics', '735': 'Computers & Electronics', '736': 'Computers & Electronics', '737': 'Computers & Electronics', '739': 'Computers & Electronics', '740': 'Computers & Electronics', '741': 'Computers & Electronics', '742': 'Computers & Electronics', '743': 'Computers & Electronics', '744': 'Computers & Electronics', '745': 'Computers & Electronics', '746': 'Computers & Electronics', '747': 'Business & Industrial', '748': 'Business & Industrial', '749': 'Business & Industrial', '750': 'Business & Industrial', '751': 'Business & Industrial', '752': 'Business & Industrial', '784': 'News', '785': 'News', '786': 'Hobbies & Leisure', '787': 'Hobbies & Leisure', '788': 'Science', '791': 'Jobs & Education', '792': 'Law & Government', '793': 'Law & Government', '794': 'Computers & Electronics', '799': 'Business & Industrial', '800': 'Business & Industrial', '801': 'Business & Industrial', '802': 'Jobs & Education', '803': 'Autos & Vehicles', '804': 'Computers & Electronics', '805': 'Computers & Electronics', '806': 'Autos & Vehicles', '807': 'Computers & Electronics', '808': 'Computers & Electronics', '809': 'Arts & Entertainment', '810': 'Autos & Vehicles', '811': 'Finance', '812': 'Finance', '813': 'Finance', '814': 'Finance', '815': 'Autos & Vehicles', '816': 'Business & Industrial', '817': 'Health', '818': 'Health', '819': 'Health', '820': 'Autos & Vehicles', '821': 'Autos & Vehicles', '822': 'Autos & Vehicles', '823': 'Autos & Vehicles', '824': 'Health', '825': 'Food & Drink', '826': 'Autos & Vehicles', '827': 'Home & Garden', '828': 'Home & Garden', '829': 'Business & Industrial', '830': 'Business & Industrial', '831': 'Business & Industrial', '832': 'Home & Garden', '833': 'Autos & Vehicles'

, '834': 'Autos & Vehicles', '835': 'Business & Industrial', '836': 'Autos & Vehicles', '837': 'Business & Industrial', '838': 'Autos & Vehicles', '839': 'Business & Industrial', '840': 'Autos & Vehicles', '841': 'Business & Industrial', '842': 'Autos & Vehicles', '843': 'Autos & Vehicles', '844': 'Business & Industrial', '845': 'Autos & Vehicles', '846': 'Autos & Vehicles', '847': 'Online Communities', '848': 'Autos & Vehicles', '849': 'Autos & Vehicles', '850': 'Autos & Vehicles', '851': 'Autos & Vehicles', '852': 'Autos & Vehicles', '853': 'Autos & Vehicles', '854': 'Autos & Vehicles', '855': 'Autos & Vehicles', '856': 'Autos & Vehicles', '857': 'Autos & Vehicles', '858': 'Autos & Vehicles', '859': 'Autos & Vehicles', '860': 'Autos & Vehicles', '861': 'Autos & Vehicles', '862': 'People & Society', '863': 'Autos & Vehicles', '864': 'People & Society', '865': 'Autos & Vehicles', '866': 'People & Society', '867': 'Autos & Vehicles', '868': 'People & Society', '869': 'People & Society', '870': 'People & Society', '882': 'Pets & Animals', '883': 'Pets & Animals', '884': 'Pets & Animals', '885': 'Pets & Animals', '886': 'Pets & Animals', '887': 'Pets & Animals', '888': 'Pets & Animals', '889': 'Pets & Animals', '890': 'Pets & Animals', '891': 'Arts & Entertainment', '892': 'Arts & Entertainment', '893': 'Arts & Entertainment', '894': 'Arts & Entertainment', '895': 'Arts & Entertainment', '896': 'Autos & Vehicles', '897': 'Autos & Vehicles', '898': 'Autos & Vehicles', '899': 'Computers & Electronics', '900': 'Computers & Electronics', '901': 'Computers & Electronics', '902': 'Computers & Electronics', '903': 'Finance', '904': 'Finance', '905': 'Business & Industrial', '906': 'Food & Drink', '907': 'Food & Drink', '908': 'Food & Drink', '909': 'Food & Drink', '910': 'Food & Drink', '911': 'Food & Drink', '912': 'Food & Drink', '913': 'Food & Drink', '914': 'Food & Drink', '915': 'Food & Drink', '916': 'Food & Drink', '917': 'Food & Drink', '918': 'Food & Drink', '919': 'Games', '920': 'Games', '921': 'Games', '922': 'Games', '923': 'Games', '925': 'Games', '926': 'Games', '927': 'Games', '928': 'Games', '929': 'Games', '930': 'Games', '931': 'Games', '932': 'Games', '933': 'Games', '935': 'Games', '936': 'Games', '937': 'Games', '938': 'Games', '939': 'Games', '940': 'Games', '941': 'Health', '942': 'Health', '943': 'Health', '944': 'Health', '945': 'Health', '946': 'Health', '947': 'Health', '948': 'Home & Garden', '949': 'Home & Garden', '950': 'Home & Garden', '951': 'Home & Garden', '952': 'Home & Garden', '953': 'Home & Garden', '954': 'Business & Industrial', '955': 'Business & Industrial', '956': 'Business & Industrial', '957': 'Business & Industrial', '958': 'Jobs & Education', '959': 'Jobs & Education', '960': 'Jobs & Education', '961': 'Jobs & Education', '962': 'Law & Government', '963': 'Law & Government', '964': 'Law & Government', '965': 'Law & Government', '966': 'Law & Government', '967': 'Law & Government', '968': 'Autos & Vehicles', '969': 'Law & Government', '970': 'Law & Government', '972': 'Online Communities', '973': 'People & Society', '974': 'People & Society', '975': 'People & Society', '976': 'People & Society', '977': 'Hobbies & Leisure', '978': 'Online Communities', '979': 'Online Communities', '980': 'Reference', '981': 'Science', '982': 'Science', '983': 'Shopping', '984': 'Shopping', '985': 'Shopping', '986': 'Shopping', '9

87': 'Shopping', '988': 'Shopping', '989': 'Shopping', '990': 'Shopp
ing', '991': 'Shopping', '992': 'Shopping', '993': 'Shopping', '994'
: 'Shopping', '995': 'Shopping', '996': 'Shopping', '997': 'Shopping
', '998': 'Sports', '999': 'Hobbies & Leisure', '1000': 'Sports', '1
001': 'Sports', '1002': 'Hobbies & Leisure', '1003': 'Travel', '1004
': 'Travel', '1005': 'Travel', '1006': 'Travel', '1007': 'Travel', '
1008': 'Travel', '1009': 'Travel', '1010': 'Travel', '1011': 'Travel
', '1012': 'Jobs & Education', '1013': 'Autos & Vehicles', '1014': '
Reference', '1015': 'Jobs & Education', '1016': 'Sports', '1017': 'S
ports', '1019': 'Travel', '1020': 'Arts & Entertainment', '1021': 'A
rts & Entertainment', '1022': 'Arts & Entertainment', '1023': 'Arts
& Entertainment', '1024': 'Arts & Entertainment', '1025': 'Arts & En
tertainment', '1026': 'Arts & Entertainment', '1027': 'Arts & Entert
ainment', '1028': 'Arts & Entertainment', '1030': 'Arts & Entertainm
ent', '1035': 'Arts & Entertainment', '1036': 'Arts & Entertainment'
, '1037': 'Arts & Entertainment', '1038': 'Arts & Entertainment', '1
039': 'Arts & Entertainment', '1040': 'Arts & Entertainment', '1041'
: 'Arts & Entertainment', '1043': 'Computers & Electronics', '1044':
'Computers & Electronics', '1045': 'Computers & Electronics', '1046'
: 'Computers & Electronics', '1047': 'Arts & Entertainment', '1048':
'Arts & Entertainment', '1049': 'Arts & Entertainment', '1050': 'Art
s & Entertainment', '1055': 'Arts & Entertainment', '1056': 'Autos &
Vehicles', '1057': 'Autos & Vehicles', '1058': 'Autos & Vehicles', '
1059': 'Autos & Vehicles', '1060': 'Autos & Vehicles', '1061': 'Auto
s & Vehicles', '1062': 'Autos & Vehicles', '1063': 'Autos & Vehicles
', '1064': 'Autos & Vehicles', '1065': 'Autos & Vehicles', '1066': '
Autos & Vehicles', '1067': 'Autos & Vehicles', '1068': 'Autos & Vehi
cles', '1069': 'Autos & Vehicles', '1070': 'Autos & Vehicles', '1071
': 'Internet & Telecom', '1072': 'Arts & Entertainment', '1073': 'Sp
orts', '1074': 'Travel', '1075': 'Law & Government', '1076': 'Busine
ss & Industrial', '1077': 'News', '1078': 'Hobbies & Leisure', '1079
': 'Hobbies & Leisure', '1080': 'Real Estate', '1081': 'Real Estate'
, '1082': 'Sports', '1083': 'Sports', '1084': 'Reference', '1085': '
Arts & Entertainment', '1086': 'Arts & Entertainment', '1087': 'Arts
& Entertainment', '1088': 'Computers & Electronics', '1089': 'Comput
ers & Electronics', '1090': 'Computers & Electronics', '1091': 'Arts
& Entertainment', '1092': 'Computers & Electronics', '1093': 'Refere
nce', '1094': 'Arts & Entertainment', '1095': 'Arts & Entertainment'
, '1096': 'Arts & Entertainment', '1097': 'Arts & Entertainment', '1
098': 'Arts & Entertainment', '1099': 'Arts & Entertainment', '1100'
: 'Arts & Entertainment', '1101': 'Arts & Entertainment', '1102': 'A
rts & Entertainment', '1103': 'Arts & Entertainment', '1104': 'Arts
& Entertainment', '1105': 'Arts & Entertainment', '1106': 'Arts & En
tertainment', '1107': 'Arts & Entertainment', '1108': 'Arts & Entert
ainment', '1109': 'Internet & Telecom', '1110': 'Arts & Entertainmen
t', '1111': 'Arts & Entertainment', '1112': 'Arts & Entertainment',
'1113': 'Arts & Entertainment', '1114': 'Arts & Entertainment', '111
5': 'Arts & Entertainment', '1116': 'Arts & Entertainment', '1117':
'Arts & Entertainment', '1118': 'Jobs & Education', '1119': 'Travel'
, '1120': 'Travel', '1121': 'People & Society', '1122': 'Hobbies & L

eisure', '1123': 'Hobbies & Leisure', '1124': 'Hobbies & Leisure', '
1125': 'Hobbies & Leisure', '1126': 'Sports', '1127': 'People & Soci
ety', '1131': 'People & Society', '1132': 'People & Society', '1133'
: 'People & Society', '1134': 'People & Society', '1135': 'People &
Society', '1136': 'Reference', '1137': 'Reference', '1138': 'Busines
s & Industrial', '1139': 'Business & Industrial', '1140': 'Autos & V
ehicles', '1141': 'Science', '1142': 'Internet & Telecom', '1143': '
Shopping', '1144': 'Shopping', '1145': 'Arts & Entertainment', '1146
': 'Games', '1147': 'Autos & Vehicles', '1148': 'Sports', '1149': 'S
ports', '1150': 'Business & Industrial', '1152': 'Business & Industr
ial', '1153': 'Home & Garden', '1155': 'Beauty & Fitness', '1157': '
Computers & Electronics', '1158': 'Computers & Electronics', '1159':
'Business & Industrial', '1160': 'Business & Industrial', '1161': 'L
aw & Government', '1162': 'Business & Industrial', '1163': 'News', '
1164': 'News', '1165': 'News', '1166': 'People & Society', '1167': '
Arts & Entertainment', '1168': 'Science', '1169': 'Science', '1170':
'Internet & Telecom', '1171': 'Internet & Telecom', '1173': 'Games',
'1174': 'Arts & Entertainment', '1175': 'Home & Garden', '1176': 'Bu
siness & Industrial', '1177': 'Books & Literature', '1178': 'Real Es
tate', '1179': 'News', '1180': 'Arts & Entertainment', '1181': 'Law
& Government', '1182': 'Internet & Telecom', '1183': 'Books & Litera
ture', '1184': 'Books & Literature', '1185': 'Arts & Entertainment',
'1186': 'Arts & Entertainment', '1187': 'Arts & Entertainment', '118
8': 'Computers & Electronics', '1189': 'Computers & Electronics', '1
190': 'Business & Industrial', '1191': 'Autos & Vehicles', '1192': '
Computers & Electronics', '1193': 'Arts & Entertainment', '1194': 'A
rts & Entertainment', '1195': 'Arts & Entertainment', '1198': 'Sport
s', '1199': 'Business & Industrial', '1200': 'Business & Industrial'
, '1201': 'News', '1202': 'News', '1203': 'News', '1204': 'News', '1
205': 'People & Society', '1210': 'Shopping', '1211': 'Health', '121
2': 'Autos & Vehicles', '1213': 'Autos & Vehicles', '1214': 'Autos &
Vehicles', '1215': 'Autos & Vehicles', '1216': 'Autos & Vehicles', '
1217': 'Autos & Vehicles', '1218': 'Autos & Vehicles', '1219': 'Beau
ty & Fitness', '1220': 'Beauty & Fitness', '1221': 'Law & Government
', '1222': 'Arts & Entertainment', '1223': 'Online Communities', '12
24': 'Health', '1225': 'Shopping', '1226': 'Science', '1227': 'Scien
ce', '1228': 'Shopping', '1229': 'Jobs & Education', '1230': 'Hobbie
s & Leisure', '1231': 'People & Society', '1232': 'Home & Garden', '
1233': 'Reference', '1234': 'Internet & Telecom', '1235': 'Health',
'1236': 'Health', '1237': 'Health', '1238': 'Health', '1239': 'Healt
h', '1240': 'News', '1241': 'News', '1242': 'Arts & Entertainment',
'1243': 'Arts & Entertainment', '1244': 'Arts & Entertainment', '124
5': 'Travel', '1246': 'Hobbies & Leisure', '1247': 'Law & Government
', '1248': 'Law & Government', '1249': 'Law & Government', '1250': '
Law & Government', '1251': 'People & Society', '1252': 'Science', '1
253': 'News', '1254': 'Science', '1255': 'Science', '1256': 'Health'
, '1257': 'People & Society', '1258': 'People & Society', '1259': 'N
ews', '1260': 'People & Society', '1261': 'People & Society', '1262'
: 'Health', '1263': 'Health', '1264': 'Reference', '1265': 'Referenc
e', '1266': 'Reference', '1267': 'Autos & Vehicles', '1268': 'Autos

& Vehicles', '1269': 'Autos & Vehicles', '1270': 'Hobbies & Leisure'
, '1271': 'Hobbies & Leisure', '1272': 'Law & Government', '1273': '
Arts & Entertainment', '1274': 'Hobbies & Leisure', '1275': 'Hobbies
& Leisure', '1276': 'Hobbies & Leisure', '1277': 'Computers & Electr
onics', '1278': 'Science', '1279': 'Computers & Electronics', '1280'
: 'People & Society', '1281': 'People & Society', '1282': 'Finance',
'1283': 'Finance', '1284': 'Law & Government', '1288': 'Reference',
'1289': 'Jobs & Education', '1290': 'Games', '1291': 'Arts & Enterta
inment', '1292': 'Home & Garden', '1293': 'Home & Garden', '1294': '
Autos & Vehicles', '1296': 'People & Society', '1298': 'Computers &
Electronics', '1299': 'Science', '1300': 'Computers & Electronics',
'1301': 'People & Society', '1302': 'People & Society', '1303': 'Peo
ple & Society', '1304': 'People & Society', '1305': 'Hobbies & Leisu
re', '1306': 'Business & Industrial', '1307': 'Business & Industrial
', '1308': 'Jobs & Education', '1309': 'Shopping', '1310': 'Arts & E
ntertainment', '1311': 'Games', '1312': 'Law & Government', '1313':
'People & Society', '1314': 'People & Society', '1315': 'Computers &
Electronics', '1316': 'Law & Government', '1317': 'Autos & Vehicles'
, '1318': 'Computers & Electronics', '1319': 'Computers & Electronic
s', '1320': 'Computers & Electronics', '1321': 'Computers & Electron
ics', '1322': 'Computers & Electronics', '1323': 'Computers & Electr
onics', '1324': 'Computers & Electronics', '1325': 'Arts & Entertain
ment', '1326': 'Arts & Entertainment', '1327': 'Arts & Entertainment
', '1328': 'Health', '1329': 'Health', '1330': 'Computers & Electron
ics', '1331': 'Computers & Electronics', '1332': 'Computers & Electr
onics', '1333': 'Computers & Electronics', '1334': 'Computers & Elec
tronics', '1339': 'Travel', '1340': 'People & Society', '1341': 'Com
puters & Electronics', '1342': 'Games', '1343': 'Games', '1344': 'Co
mputers & Electronics', '1345': 'Computers & Electronics', '1346': '
Computers & Electronics', '1347': 'Business & Industrial', '1348': '
Home & Garden', '1349': 'Business & Industrial', '1350': 'Health', '
1351': 'Health', '1352': 'Health', '1353': 'Health', '1354': 'Comput
ers & Electronics', '1355': 'Computers & Electronics', '1356': 'Comp
uters & Electronics', '1357': 'Computers & Electronics', '1358': 'Co
mputers & Electronics', '1359': 'Computers & Electronics', '1360': '
Business & Industrial', '1361': 'Hobbies & Leisure', '1362': 'Home &
Garden', '1363': 'Home & Garden', '1364': 'Home & Garden', '1365': '
Home & Garden', '1366': 'Home & Garden', '1367': 'Home & Garden', '1
368': 'Home & Garden', '1369': 'Home & Garden', '1370': 'Home & Gard
en', '1371': 'Home & Garden', '1372': 'Home & Garden', '1373': 'Home
& Garden', '1374': 'People & Society', '1375': 'Business & Industria
l', '1376': 'Sports', '1377': 'Autos & Vehicles', '1378': 'Autos & V
ehicles', '1379': 'Internet & Telecom', '1380': 'Autos & Vehicles',
'1381': 'Online Communities', '1382': 'Computers & Electronics', '13
83': 'Computers & Electronics', '1384': 'Computers & Electronics', '
1385': 'Law & Government', '1386': 'Law & Government', '1387': 'Law
& Government', '1388': 'Jobs & Education', '1389': 'Travel', '1390':
'Travel', '1391': 'Travel', '1392': 'Travel', '1393': 'Computers & E
lectronics', '1394': 'Computers & Electronics', '1395': 'Computers &
Electronics', '1396': 'Computers & Electronics', '1397': 'Games', '1

398': 'Autos & Vehicles', '1399': 'Autos & Vehicles', '1400': 'Autos & Vehicles', '1401': 'Autos & Vehicles', '1402': 'Autos & Vehicles', '1403': 'Autos & Vehicles', '1404': 'Autos & Vehicles', '1405': 'Autos & Vehicles', '1406': 'Autos & Vehicles', '1407': 'Shopping', '1408': 'Arts & Entertainment', '1409': 'Arts & Entertainment', '1410': 'Arts & Entertainment', '1411': 'Arts & Entertainment', '1412': 'Arts & Entertainment', '1413': 'Autos & Vehicles', '1414': 'Autos & Vehicles', '1415': 'Autos & Vehicles', '1416': 'Autos & Vehicles', '1417': 'Beauty & Fitness', '1418': 'Beauty & Fitness', '1419': 'Beauty & Fitness', '1420': 'Business & Industrial', '1421': 'Home & Garden', '1422': 'Business & Industrial', '1423': 'Business & Industrial', '1424': 'Business & Industrial', '1425': 'Business & Industrial', '1426': 'Business & Industrial', '1427': 'Finance', '1428': 'Finance', '1429': 'Finance', '1430': 'Finance', '1431': 'Finance', '1432': 'Finance', '1433': 'Finance', '1434': 'Finance', '1435': 'Finance', '1436': 'Finance', '1437': 'Finance', '1438': 'Finance', '1439': 'Finance', '1440': 'Finance', '1441': 'Finance', '1442': 'Finance', '1443': 'Finance', '1444': 'Finance', '1445': 'Finance', '1446': 'Finance', '1447': 'Finance', '1448': 'Finance', '1449': 'Finance', '1450': 'Finance', '1451': 'Finance', '1452': 'Finance', '1453': 'Finance', '1454': 'Finance', '1455': 'Finance', '1456': 'Finance', '1457': 'Finance', '1458': 'Finance', '1459': 'Finance', '1460': 'Real Estate', '1461': 'Internet & Telecom', '1462': 'Internet & Telecom', '1463': 'Internet & Telecom', '1464': 'Internet & Telecom', '1465': 'Jobs & Education', '1466': 'Jobs & Education', '1467': 'Jobs & Education', '1468': 'Hobbies & Leisure', '1469': 'Hobbies & Leisure', '1470': 'Hobbies & Leisure', '1471': 'Jobs & Education', '1472': 'Jobs & Education', '1473': 'Jobs & Education', '1474': 'Jobs & Education', '1475': 'Jobs & Education', '1476': 'Jobs & Education', '1477': 'Jobs & Education', '1478': 'Jobs & Education', '1479': 'Jobs & Education', '1480': 'Jobs & Education', '1481': 'Jobs & Education', '1482': 'Online Communities', '1483': 'People & Society', '1484': 'People & Society', '1485': 'People & Society', '1486': 'Shopping', '1487': 'Shopping', '1488': 'Shopping', '1489': 'Shopping', '1490': 'Shopping', '1491': 'Games', '1492': 'Games', '1493': 'Games', '1494': 'Games', '1495': 'Games', '1496': 'Games', '1497': 'Games', '1498': 'Games', '1499': 'Games', '1500': 'Computers & Electronics', '1501': 'Food & Drink', '1502': 'Health', '1503': 'Health', '1504': 'Shopping', '1505': 'Shopping', '1506': 'Shopping', '1507': 'Shopping', '1508': 'Sports', '1509': 'Food & Drink', '1510': 'Food & Drink', '1511': 'Food & Drink', '1512': 'Food & Drink', '1513': 'Food & Drink', '1514': 'Food & Drink', '1515': 'Food & Drink', '1516': 'Food & Drink', '1517': 'Food & Drink', '1518': 'Food & Drink', '1519': 'Reference', '1520': 'Reference', '1521': 'Arts & Entertainment', '1522': 'Internet & Telecom', '1523': 'Food & Drink', '1524': 'Food & Drink', '1525': 'Food & Drink', '1526': 'Food & Drink', '1527': 'Food & Drink', '1528': 'Food & Drink', '1529': 'Food & Drink', '1530': 'Arts & Entertainment', '1531': 'Arts & Entertainment', '1532': 'Food & Drink', '1533': 'Computers & Electronics', '1534': 'Computers & Electronics', '1535': 'Computers & Electronics', '1536': 'Computers & Electronics', '1537'

: 'Computers & Electronics', '1538': 'Food & Drink', '1539': 'Food & Drink', '1540': 'Food & Drink', '1541': 'Food & Drink', '1542': 'Food & Drink', '1543': 'Food & Drink', '1544': 'Food & Drink', '1545': 'Food & Drink', '1546': 'Food & Drink', '1547': 'Business & Industrial', '1548': 'Business & Industrial', '1549': 'Games', '1550': 'Food & Drink', '1551': 'Food & Drink', '1552': 'Food & Drink', '1553': 'Food & Drink', '1554': 'Food & Drink', '1555': 'Food & Drink', '1556': 'Food & Drink', '1557': 'Food & Drink', '1558': 'Food & Drink', '1559': 'Food & Drink', '1560': 'Food & Drink', '1561': 'Food & Drink', '1562': 'Food & Drink', '1563': 'Food & Drink', '1564': 'Food & Drink', '1565': 'Food & Drink', '1566': 'Food & Drink', '1567': 'Food & Drink', '1568': 'Food & Drink', '1569': 'Food & Drink', '1570': 'Health', '1571': 'Health', '1572': 'Health', '1573': 'Food & Drink', '1574': 'Food & Drink', '1575': 'Food & Drink', '1576': 'Food & Drink', '1577': 'Food & Drink', '1578': 'Business & Industrial', '1579': 'Jobs & Education', '1580': 'Shopping', '1581': 'Shopping', '1582': 'Shopping', '1583': 'Shopping', '1584': 'People & Society', '1585': 'People & Society', '1586': 'Shopping', '1587': 'Shopping', '1588': 'Shopping', '1589': 'Shopping', '1590': 'Hobbies & Leisure', '1591': 'Arts & Entertainment', '1592': 'Arts & Entertainment', '1593': 'Sports', '1594': 'Sports', '1595': 'Sports', '1596': 'Sports', '1597': 'Shopping', '1598': 'Shopping', '1599': 'Sports', '1600': 'Home & Garden', '1601': 'Home & Garden', '1602': 'Home & Garden', '1603': 'Home & Garden', '1604': 'Home & Garden', '1605': 'Home & Garden', '1606': 'Home & Garden', '1607': 'Home & Garden', '1608': 'Home & Garden', '1609': 'Home & Garden', '1611': 'Sports', '1612': 'Sports', '1613': 'Sports', '1614': 'Sports', '1615': 'Sports', '1616': 'Sports', '1617': 'Sports', '1618': 'Sports', '1619': 'Sports', '1620': 'Sports', '1621': 'Sports', '1622': 'Sports', '1623': 'Sports', '1624': 'Sports', '1625': 'Sports', '1626': 'Sports', '1627': 'Sports', '1628': 'Sports', '1629': 'Sports', '1630': 'Sports', '1631': 'Shopping', '1632': 'Shopping', '1633': 'Sports', '1634': 'Sports', '1635': 'Sports', '1636': 'Sports', '1641': 'People & Society', '1666': 'Sports', '1674': 'Sports', '1681': 'Sports', '1684': 'Autos & Vehicles', '1685': 'Autos & Vehicles', '1686': 'Autos & Vehicles', '1687': 'Autos & Vehicles', '1700': 'Autos & Vehicles', '1701': 'Autos & Vehicles', '1702': 'Autos & Vehicles', '1707': 'Real Estate', '1708': 'Real Estate', '1709': 'Real Estate', '1710': 'Real Estate', '1711': 'Travel', '1712': 'Real Estate', '1713': 'Real Estate', '1715': 'Real Estate', '1716': 'Real Estate', '1720': 'Home & Garden', '1721': 'Home & Garden', '1722': 'Home & Garden', '1723': 'Home & Garden', '1724': 'Home & Garden', '1725': 'Home & Garden', '1726': 'Home & Garden', '1727': 'Home & Garden', '1728': 'Home & Garden', '1732': 'Home & Garden', '1735': 'People & Society', '1738': 'People & Society', '1739': 'Computers & Electronics', '1740': 'Computers & Electronics', '1741': 'People & Society', '1747': 'Autos & Vehicles', '1748': 'Autos & Vehicles', '1750': 'Autos & Vehicles', '1751': 'Autos & Vehicles', '1757': 'Shopping', '1758': 'Shopping', '1763': 'Beauty & Fitness', '1779': 'Arts & Entertainment', '1780': 'Arts & Entertainment', '1783': 'Autos & Vehicles', '1784': 'Autos & Vehicles', '1785': 'Computers

```
& Electronics', '1786': 'Computers & Electronics', '1787': 'Computer
s & Electronics', '1788': 'Computers & Electronics', '1789': 'Comput
ers & Electronics', '1790': 'Computers & Electronics', '1791': 'Comp
uters & Electronics', '1795': 'Finance', '1799': 'Games', '1800': 'H
ome & Garden', '1801': 'Jobs & Education', '1802': 'Computers & Elec
tronics', '1804': 'Sports', '1820': 'Finance', '1826': 'Food & Drink
'}
{'Arts & Entertainment': 0, 'Computers & Electronics': 1, 'Finance':
2, 'Games': 3, 'Home & Garden': 4, 'Business & Industrial': 5, 'Inte
rnet & Telecom': 6, 'People & Society': 7, 'News': 8, 'Shopping': 9,
'Law & Government': 10, 'Sports': 11, 'Books & Literature': 12, 'Rea
l Estate': 13, 'Online Communities': 14, 'Beauty & Fitness': 15, 'He
alth': 16, 'Autos & Vehicles': 17, 'Jobs & Education': 18, 'Hobbies
& Leisure': 19, 'Pets & Animals': 20, 'Travel': 21, 'Food & Drink':
22, 'Reference': 23, 'Science': 24}
```

In [9]:
```
# create processed dataframe with only LTI features using the whole da
taset
def model_LTIOnly(file):

    columns= ['']*26

    for key, val in topic_to_newID.items():
        columns[val]= key+ ' LTI'

    columns[25]= 'Convert'

    new_user2= pd.DataFrame()
    data=[]
    i= 0

    for row in file.iterrows():
        LTI = np.zeros(25)

        for key, val in row[1][2].items():
            if key in Id_to_topic:
                topic= Id_to_topic[key]
                index= topic_to_newID[topic]

            else:
                continue
            LTI[index]+= val

        LTI=LTI/np.sum(LTI)


        new_entry= [entry for entry in LTI]
        new_entry.append(int(row[1][1]))
        data.append(new_entry)
        i+=1
        if i>1000:

            new_data= pd.DataFrame(data, columns = columns)
            new_user2= new_user2.append(new_data)
            data=[]
            i=0

    new_data= pd.DataFrame(data, columns = columns)
    new_user2= new_user2.append(new_data)

    return new_user2
```

In [10]:
```
# create processed dataframe with only LTI features only using data fo
r users without short term features
def model_noSTIData_LTIfeature(file):
    columns= ['']*26

    for key, val in topic_to_newID.items():
        columns[val]= key+ ' LTI'

    columns[25]= 'Convert'

    new_user2= pd.DataFrame()
    data=[]
    i= 0

    for row in file.iterrows():
        LTI = np.zeros(25)

        if not row[1][3]:
            for key, val in row[1][2].items():
                if key in Id_to_topic:
                    topic= Id_to_topic[key]
                    index= topic_to_newID[topic]

                else:
                    continue
                LTI[index]+= val

            LTI=LTI/np.sum(LTI)


        new_entry= [entry for entry in LTI]
        new_entry.append(int(row[1][1]))
        data.append(new_entry)
        i+=1
        if i>1000:

            new_data= pd.DataFrame(data, columns = columns)
            new_user2= new_user2.append(new_data)
            data=[]
            i=0

    new_data= pd.DataFrame(data, columns = columns)
    new_user2= new_user2.append(new_data)

    return new_user2
```

In [11]:
```
# create processed dataframe with both LTI and STI features only using
data from users with both types of features
def model_STIData_bothFeat(file):
```

```
columns_3= ['']*51

for key, val in topic_to_newID.items():
    columns_3[val]= key+ ' LTI'
    columns_3[val+25]= key+' STI'


columns_3[50]= 'Convert'

new_user3= pd.DataFrame()
data=[]
i= 0

for row in training.iterrows():
    LTI = np.zeros(25)
    STI = np.zeros(25)

    if row[1][3]:

        for key, val in row[1][2].items():
            if key in Id_to_topic:
                topic= Id_to_topic[key]
                index= topic_to_newID[topic]

            else:
                continue
            LTI[index]+= val

        LTI=LTI/np.sum(LTI)

        for key, val in row[1][3].items():
            if key in Id_to_topic:
                topic= Id_to_topic[key]
                index= topic_to_newID[topic]

            else:
                continue

            STI[index]+=val
        STI=STI/np.sum(STI)

    else:
        continue
    new_entry= [entry for entry in LTI]
    for entry in STI:
        new_entry.append(entry)
    new_entry.append(int(row[1][1]))
    data.append(new_entry)
    i+=1
    if i>1000:
```

```
            new_data= pd.DataFrame(data, columns = columns_3)
            new_user3= new_user3.append(new_data)
            data=[]
            i=0

    new_data= pd.DataFrame(data, columns = columns_3)
    new_user3= new_user3.append(new_data)

    return new_user3
```

In [12]:
```
def Split(train, test):
    X_train = train.iloc[:, :-1]
    y_train = train.iloc[:, -1]
    X_test = test.iloc[:, :-1]
    y_test = test.iloc[:, -1]
    #X_train, X_test, y_train, y_test = train_test_split(X, y, test_si
ze=0.2, random_state=0)
    cols = X_train.columns
    return X_train, y_train, X_test, y_test, cols
```

In [13]:
```
# randomly undersample training data to account for class imbalance
def underSamp(X_train, y_train):
    usamp = RandomUnderSampler()
    us_X_train, us_y_train=usamp.fit_sample(X_train, y_train)
    us_X_train = pd.DataFrame(data=us_X_train,columns=cols)
    return us_X_train, us_y_train
```

Run models

In [14]:
```
# run random forest algorithm
def RunAlgRF(us_X_train, us_y_train, X_test, y_test):
    clf = RandomForestClassifier(n_estimators=10)
    clf.fit(us_X_train, us_y_train)
    scores = cross_val_score(clf, us_X_train, us_y_train, cv=3, scorin
g='accuracy')
    print('Training accuracy= '+str(scores.mean()))
    print('Training standard deviation= '+ str(np.std(scores)))
    #clf.score(X_test, y_test)
    print('Testing accuracy= '+str(clf.score(X_test, y_test)))
    return clf
```

In [15]:
```
# run boosted decision trees
def RunAlgXG(us_X_train, us_y_train, X_test, y_test):
    clf = XGBClassifier()
    clf.fit(X_train, y_train)
    y_pred = clf.predict(X_test)
    predictions = [round(value) for value in y_pred]
    accuracy = accuracy_score(y_test, predictions)
    print("Accuracy: %.2f%%" % (accuracy * 100.0))
    return clf
```

In [16]:
```
# run support vector machine
def RunAlgSVM(us_X_train, us_y_train, X_test, y_test):
    clf = SVC(gamma = 'auto')
    clf.fit(X_train, y_train)
    y_pred = clf.predict(X_test)
    predictions = [round(value) for value in y_pred]
    accuracy = accuracy_score(y_test, predictions)
    print("Accuracy: %.2f%%" % (accuracy * 100.0))
    return clf
```

In [17]:
```
# isolate feature importance
def var_import(cols, clf):
    importance = pd.DataFrame({'feature': list(cols), 'feature_importa
nce': [round(i, 4) for i in list(clf.feature_importances_)]})
    importance = importance.sort_values(by=['feature_importance'], asc
ending=False)
    importance = importance.set_index('feature')
    return importance
```

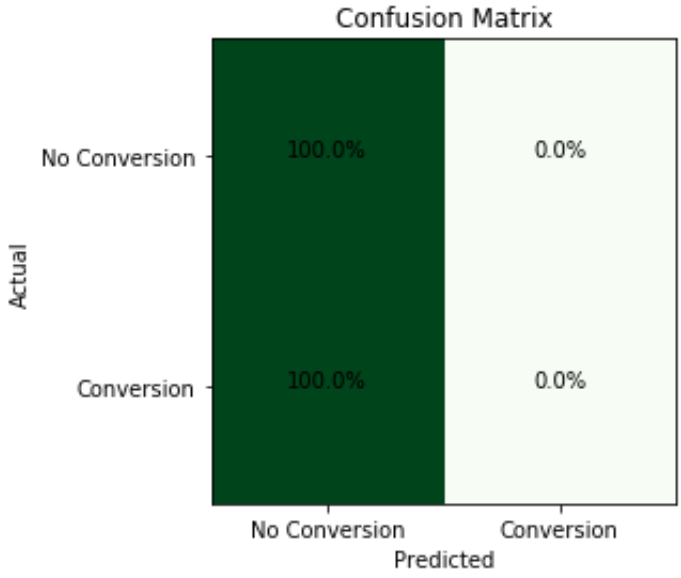
In [18]:
```
# create plot of most important features
def plot_importance(importance):
    plt.rcParams['figure.figsize']= [16, 10]
    ax = plt.bar(importance.index, importance['feature_importance'])
    plt.xticks( rotation='vertical')
    plt.xlabel('Variable')
    plt.ylabel('Feature Importance')
    plt.show()
```

In [19]:
```
# create easy to understand visualization of confusion matrix
def plot_confusion(x, y, clf):
    names = ['No Conversion', "Conversion"]
    conf = confusion_matrix(y, pd.DataFrame(clf.predict(x)))
    print(conf)
    conf = conf.astype('float')/ conf.sum(axis=1)[:, np.newaxis]
    plt.imshow(conf, interpolation='nearest', cmap=plt.get_cmap('Green
s'))
    marks = np.arange(len(names))
    for i, j in itertools.product(range(conf.shape[0]), range(conf.sha
pe[1])):
        plt.text(j, i, "{}%".format(round(conf[i, j]*100,2)), horizont
alalignment="center")
    plt.xticks(marks, names)
    plt.yticks(marks, names)
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.title('Confusion Matrix')
    plt.show()
```

In [20]:
```
# show ROC curve as ewll as AUC
def ROC_AUC(x, y, clf):
    AUC = roc_auc_score(y, clf.predict(x))
    FPR, TPR, thresh = roc_curve(y, clf.predict_proba(x)[:,1])
    plt.plot(FPR, TPR, label='Area Under Curve = %0.2f' % AUC)
    plt.xlim([-0.05, 1.05])
    plt.ylim([-0.05, 1.05])
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title('ROC Curve')
    plt.legend()
    plt.show()
```

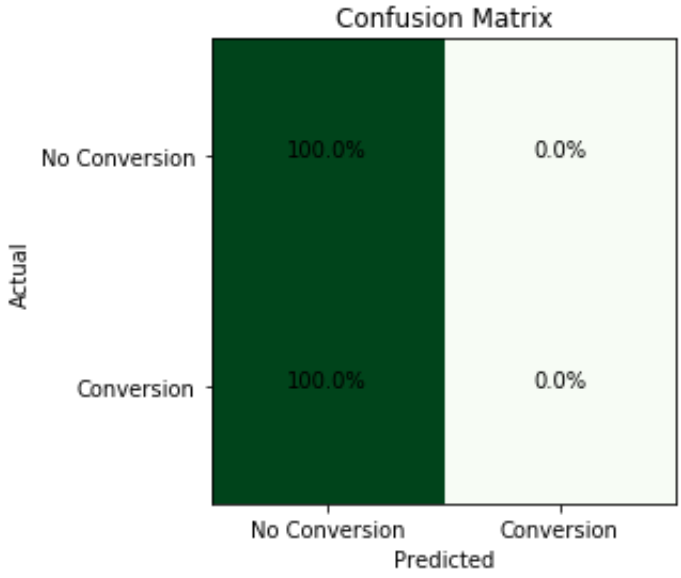
In [21]:
```
# train and analyze SVM model

train= model_LTIOnly(training)
test= model_LTIOnly(validation)
X_train, y_train, X_test, y_test, cols= Split(train, test)
us_X, us_y= underSamp(X_train, y_train)

clf= RunAlgSVM(us_X, us_y, X_test, y_test)
plot_confusion(X_test, y_test, clf)
```

```
Accuracy: 99.23%
[[79388     0]
 [  620     0]]
```



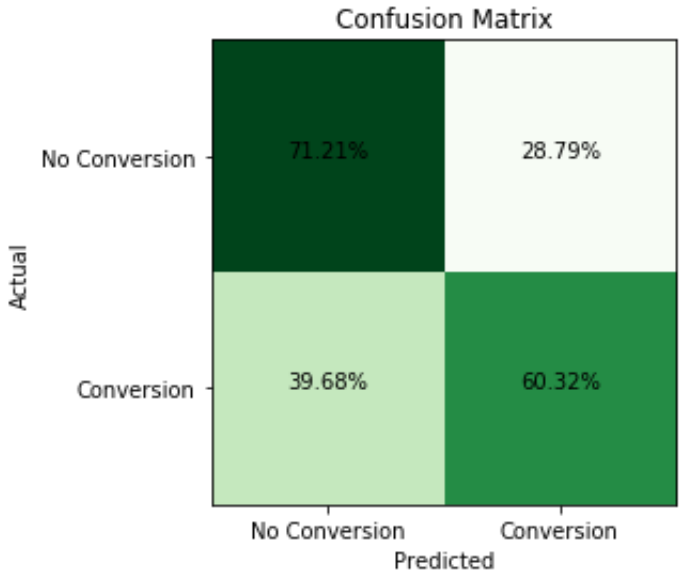
In [22]:
```
# train and analyze boosted decision trees model

train= model_LTIOnly(training)
test= model_LTIOnly(validation)
X_train, y_train, X_test, y_test, cols= Split(train, test)
us_X, us_y= underSamp(X_train, y_train)

clf= RunAlgXG(us_X, us_y, X_test, y_test)
plot_confusion(X_test, y_test, clf)
```

```
Accuracy: 99.23%
[[79388     0]
 [  620     0]]
```



In [23]:
```
# train and analyze random forest algorithm
train= model_LTIOnly(training)
test= model_LTIOnly(validation)

X_train, y_train, X_test, y_test, cols= Split(train, test)
us_X, us_y= underSamp(X_train, y_train)
clf= RunAlgRF(us_X, us_y, X_test, y_test)

plot_confusion(X_test, y_test, clf)
```

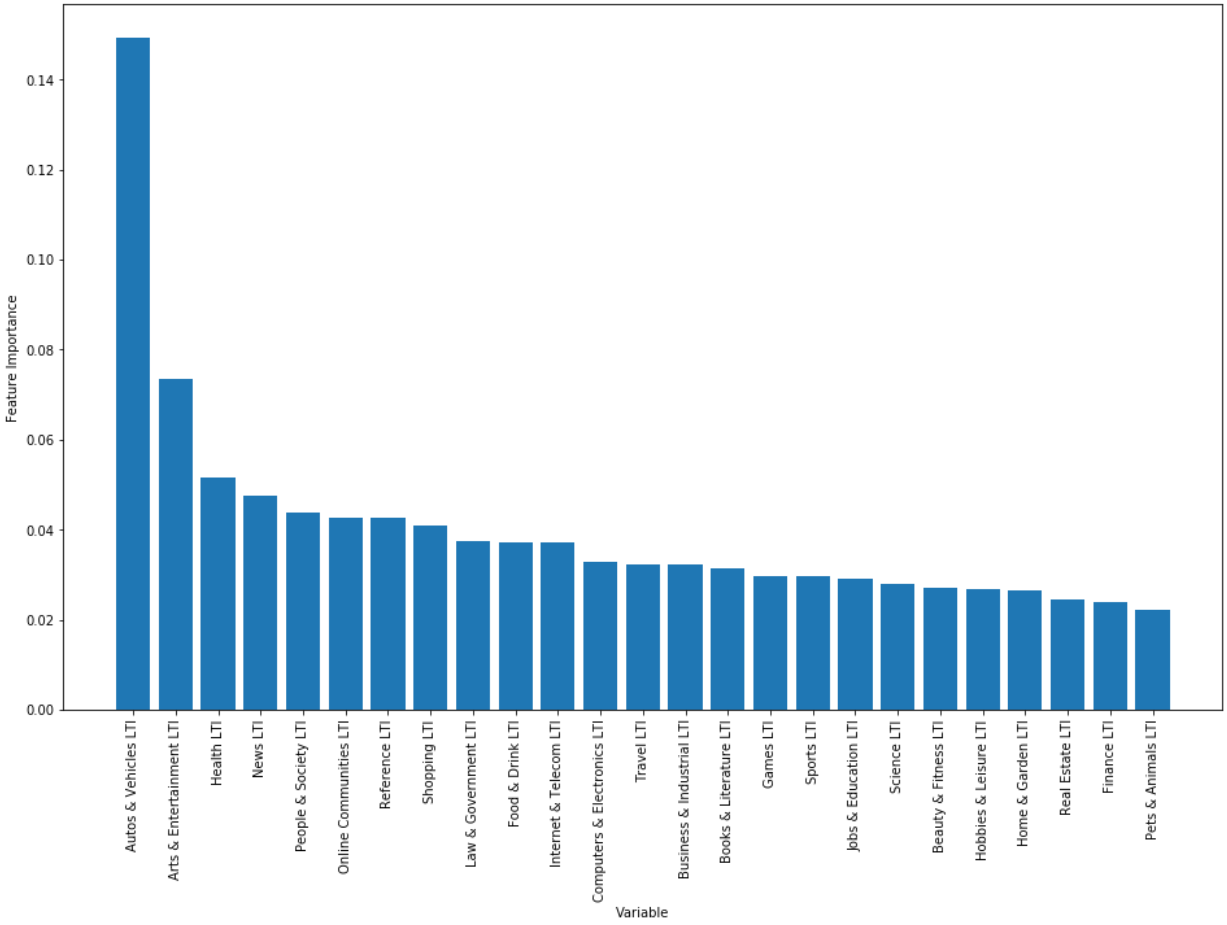
```
Training accuracy= 0.6474411087644015
Training standard deviation= 0.005506219057775745
Testing accuracy= 0.7112538746125388
[[56532 22856]
 [  246   374]]
```

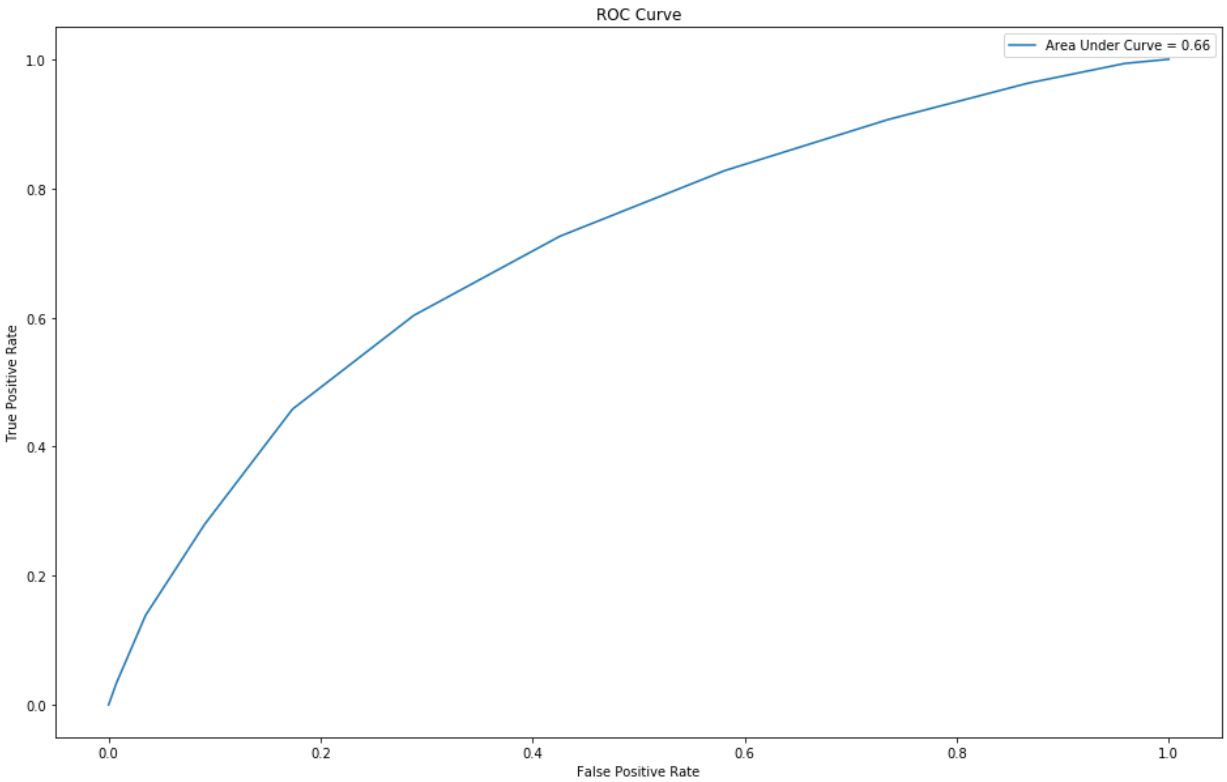


In [43]:
```
#calculate precision and recall for LTI only random forest

precision= 374/(22856+374)
recall= 374/(246+374)
print(precision)
print(recall)
```

```
0.01609987085665088
0.603225806451613
```

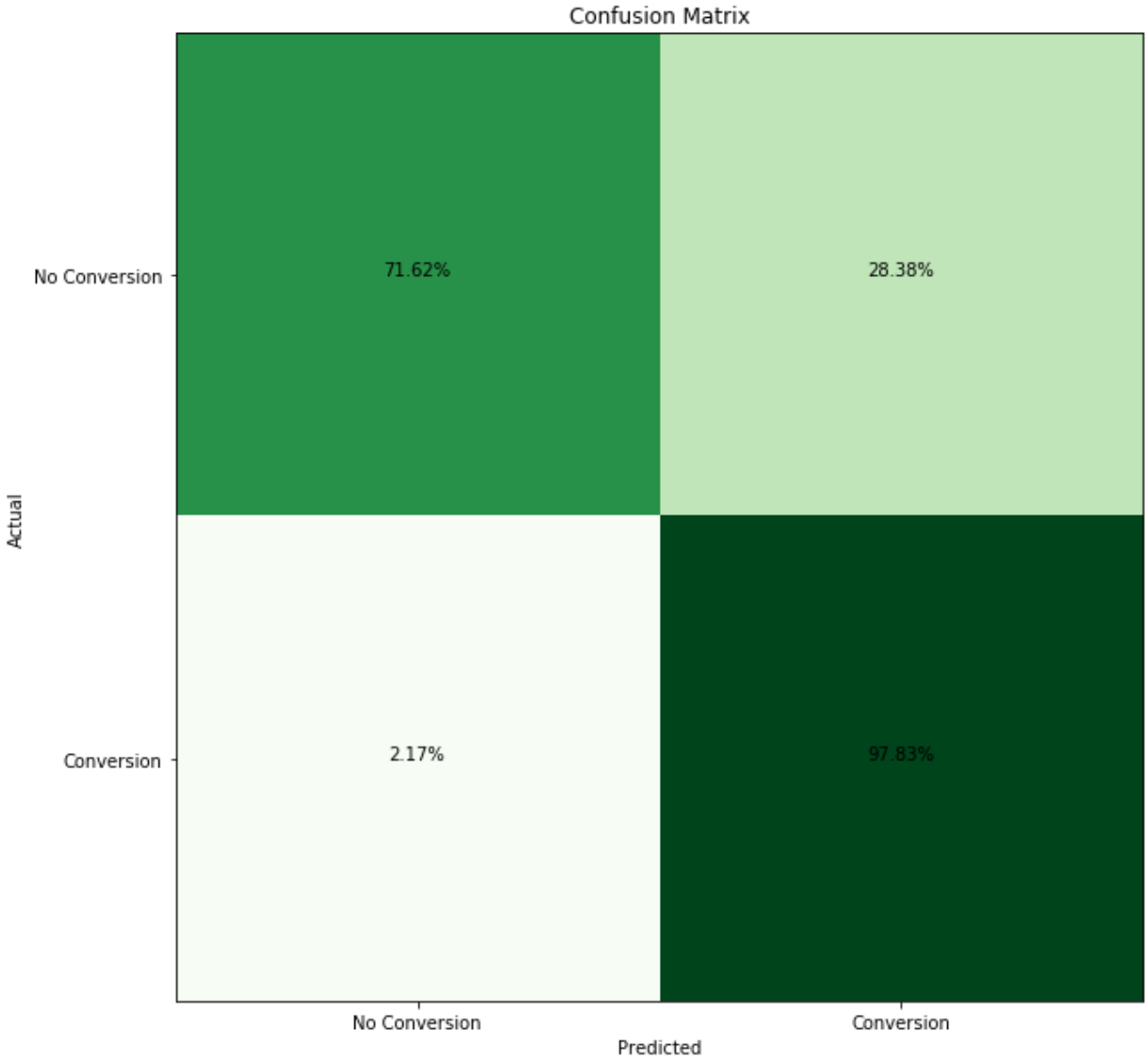
In [26]:
```
# show variable importance and AUROC for LTI only random forest
importance= var_import(cols, clf)
plot_importance(importance)
ROC_AUC(X_test, y_test, clf)
```

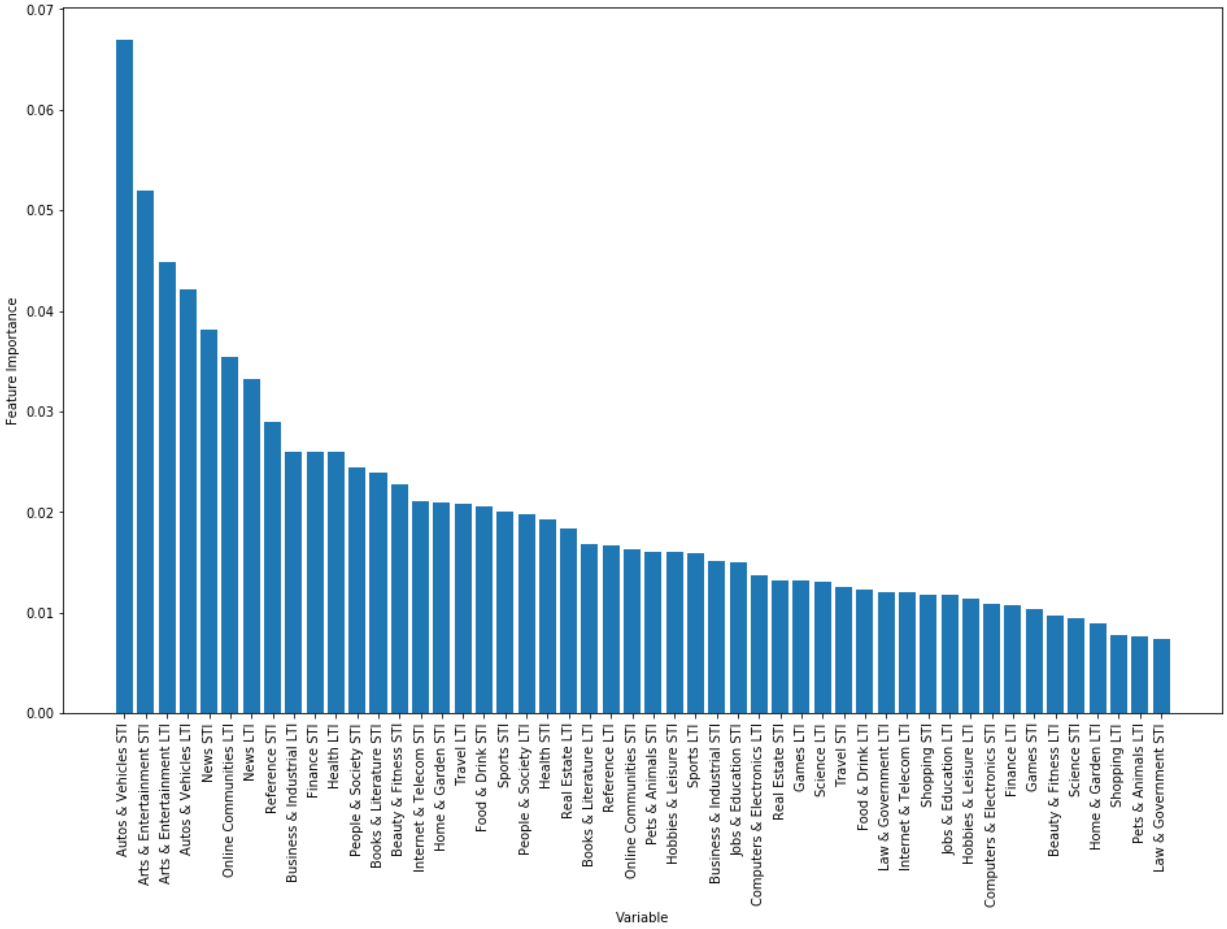

Feature Importance
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Autos & Vehicles LTI
Arts & Entertainment LTI
Health LTI
News LTI
People & Society LTI
Online Communities LTI
Reference LTI
Shopping LTI
Law & Government LTI
Food & Drink LTI
Internet & Telecom LTI
Computers & Electronics LTI
Travel LTI
Business & Industrial LTI
Books & Literature LTI
Games LTI
Sports LTI
Jobs & Education LTI
Science LTI
Beauty & Fitness LTI
Hobbies & Leisure LTI
Home & Garden LTI
Real Estate LTI
Finance LTI
Pets & Animals LTI
Variable

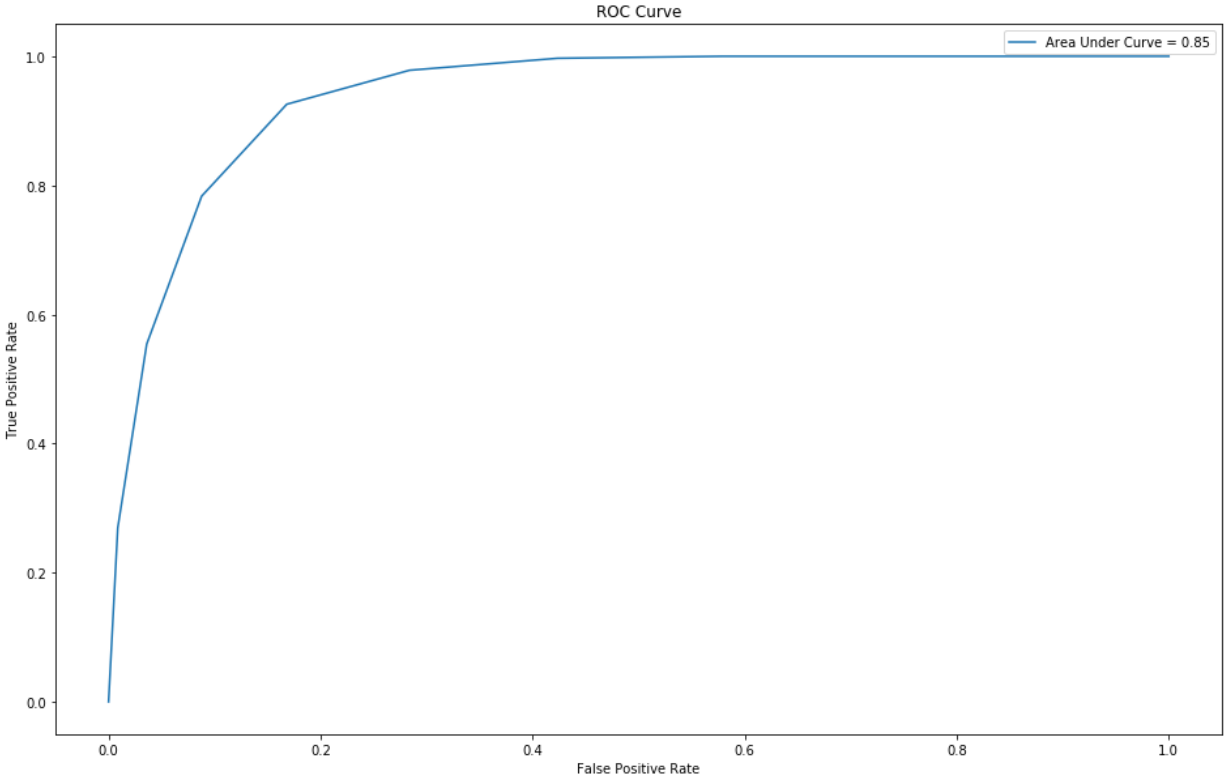

ROC Curve
Area Under Curve = 0.66
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

In [27]:
```
# train and analyze random forest model with STI data using both featu
res

train= model_STIData_bothFeat(training)
test= model_STIData_bothFeat(validation)

X_train, y_train, X_test, y_test, cols= Split(train, test)
us_X, us_y= underSamp(X_train, y_train)

clf= RunAlgRF(us_X, us_y, X_test, y_test)

plot_confusion(X_test, y_test, clf)
importance= var_import(cols, clf)
plot_importance(importance)
ROC_AUC(X_test, y_test, clf)
```

```
Training accuracy= 0.6361630321910696
Training standard deviation= 0.019028122805707457
Testing accuracy= 0.7194454153559362
[[18209  7217]
 [    7   316]]
```


Confusion Matrix
No Conversion
Conversion
Actual
71.62%
28.38%
2.17%
97.83%
No Conversion
Conversion
Predicted


Feature Importance
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Autos & Vehicles STI
Arts & Entertainment STI
Arts & Entertainment LTI
Autos & Vehicles LTI
News STI
Online Communities LTI
News LTI
Reference STI
Business & Industrial LTI
Finance STI
Health LTI
People & Society STI
Books & Literature STI
Beauty & Fitness STI
Internet & Telecom STI
Home & Garden STI
Travel LTI
Food & Drink STI
Sports STI
People & Society LTI
Health STI
Real Estate LTI
Books & Literature LTI
Reference LTI
Online Communities STI
Pets & Animals STI
Hobbies & Leisure STI
Sports LTI
Business & Industrial STI
Jobs & Education STI
Computers & Electronics LTI
Real Estate STI
Games LTI
Science LTI
Travel STI
Food & Drink LTI
Law & Government LTI
Internet & Telecom LTI
Shopping STI
Jobs & Education LTI
Hobbies & Leisure LTI
Computers & Electronics STI
Finance LTI
Games STI
Beauty & Fitness LTI
Science STI
Home & Garden LTI
Shopping LTI
Pets & Animals LTI
Law & Government STI
Variable


ROC Curve
Area Under Curve = 0.85
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

In [44]:
```
# precision and recall for random forest model using STI data only wit
h both features

precision= 316/(7217+316)
recall= 316/(7+316)
print(precision)
print(recall)
```

```
0.04194875879463693
0.978328173374613
```

Nested Cross Validation

In [30]:
```
# implement random selection nested cross validation
def RandomCV(X_train, y_train):
    from sklearn.model_selection import RandomizedSearchCV

    n_estimators = [int(x) for x in np.linspace(start=10, stop= 200, n
um=10)]
    max_depth = [int(x) for x in np.linspace(10, 110, num=5)]
    max_depth.append(None)
    min_samples_split= [2, 5, 10]
    min_samples_leaf= [1, 2, 4]
    bootstrap= [True, False]

    random_grid = {'n_estimators': n_estimators,
                'max_depth': max_depth,
                'min_samples_split': min_samples_split,
                'min_samples_leaf': min_samples_leaf,
                'bootstrap': bootstrap}

    clf= RandomForestClassifier()
    clf_random= RandomizedSearchCV(estimator = clf, param_distribution
s=random_grid,
                   n_iter=20, scoring='roc_auc', cv=3)

    clf_random.fit(X_train, y_train)

    print(clf_random.best_params_)

    return clf_random.best_estimator_
```

In [31]:
```
# retrain model using only data without STI features and only using LT
i features

train= model_noSTIData_LTIfeature(training)
test= model_noSTIData_LTIfeature(validation)

X_train, y_train, X_test, y_test, cols= Split(train, test)
us_X, us_y= underSamp(X_train, y_train)
```

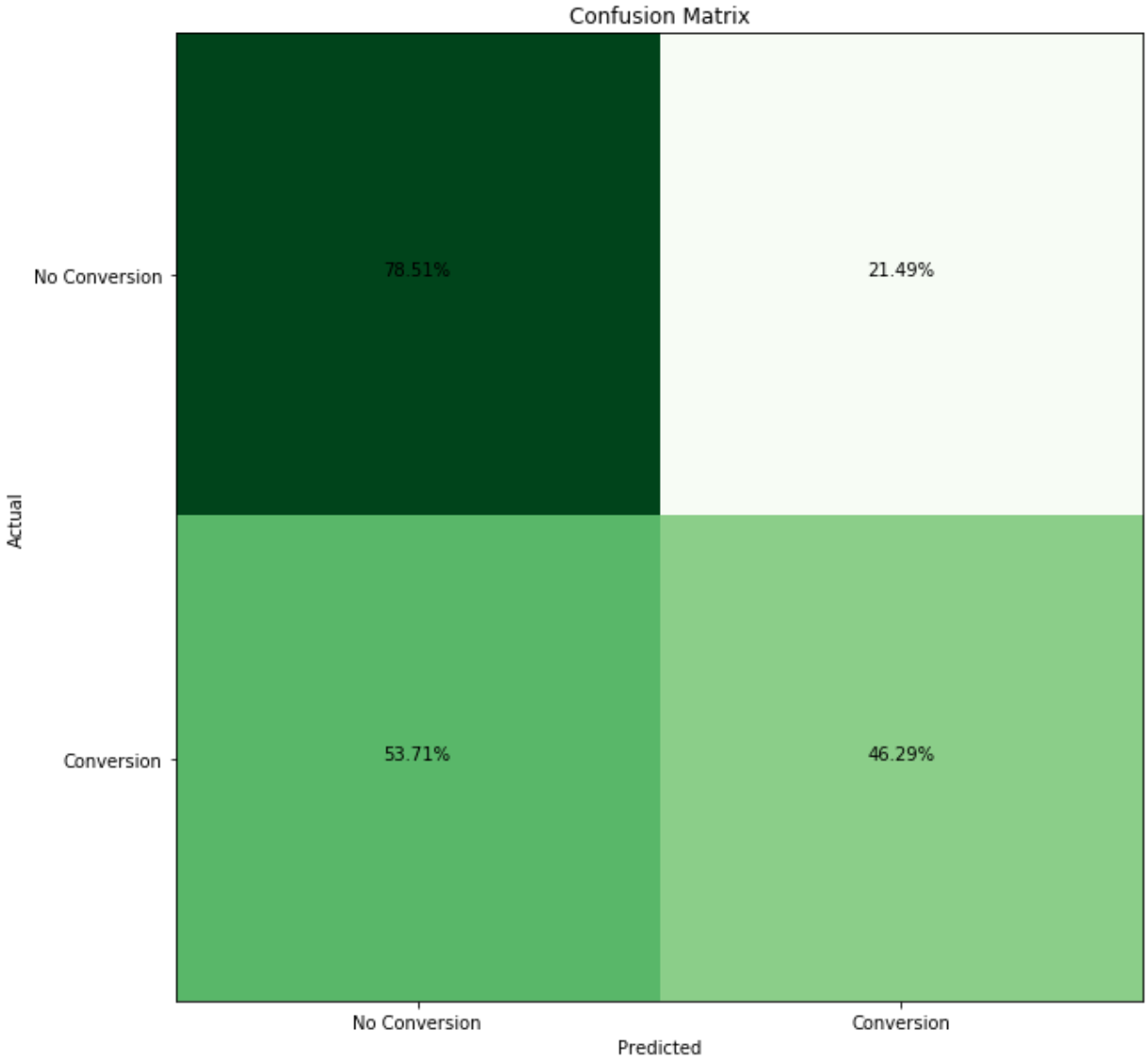
In [32]:
```
# run cross validation and compare results to the basic untuned random
forest algorithm
best_random= RandomCV(us_X, us_y)
scores = cross_val_score(best_random, us_X, us_y, cv=3, scoring='accur
acy')
print('Training accuracy= '+str(scores.mean()))
print('Training standard deviation= '+ str(np.std(scores)))

clf = RandomForestClassifier(n_estimators=10)
clf.fit(us_X, us_y)

scores = cross_val_score(clf, us_X, us_y, cv=3, scoring='accuracy')
print('Training accuracy= '+str(scores.mean()))
print('Training standard deviation= '+ str(np.std(scores)))
```

```
{'n_estimators': 200, 'min_samples_split': 10, 'min_samples_leaf': 1
, 'max_depth': 35, 'bootstrap': True}
Training accuracy= 0.6399358286678513
Training standard deviation= 0.006253376302251197
Training accuracy= 0.6092176796629678
Training standard deviation= 0.004666888917723448
```

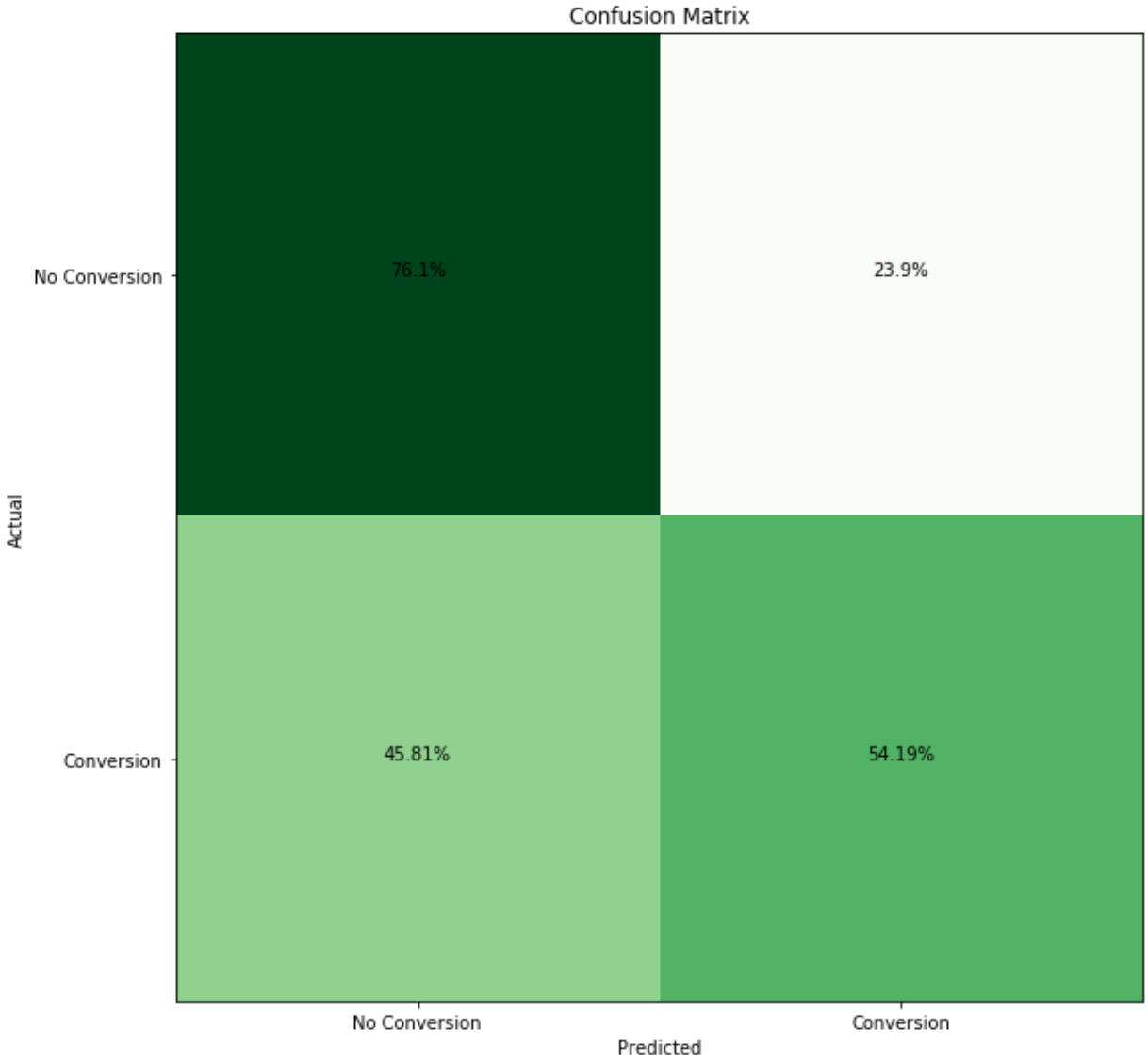
In [ ]:
```

```

In [33]:
```
# show confusion matrices for both
plot_confusion(X_test, y_test, clf)
plot_confusion(X_test, y_test, best_random)

# notice the sharp decrease in false negatives
```

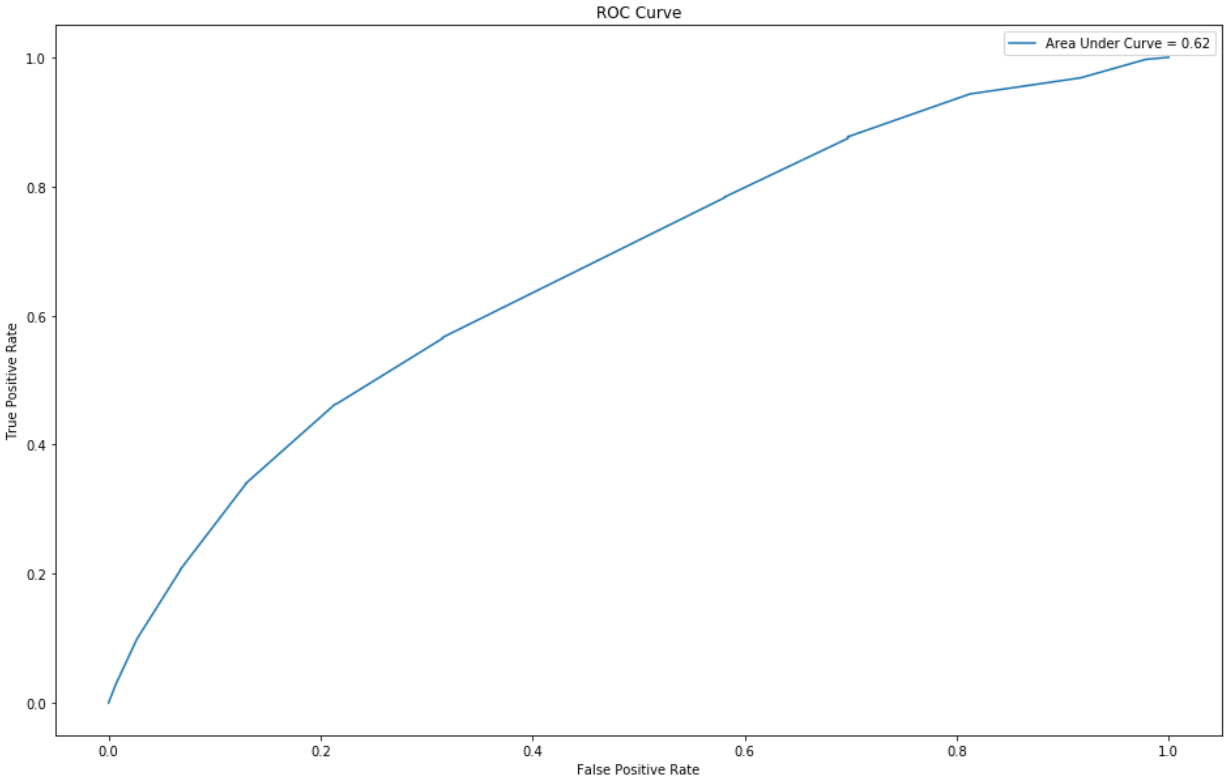
```
[[62330 17058]
 [  333   287]]
```

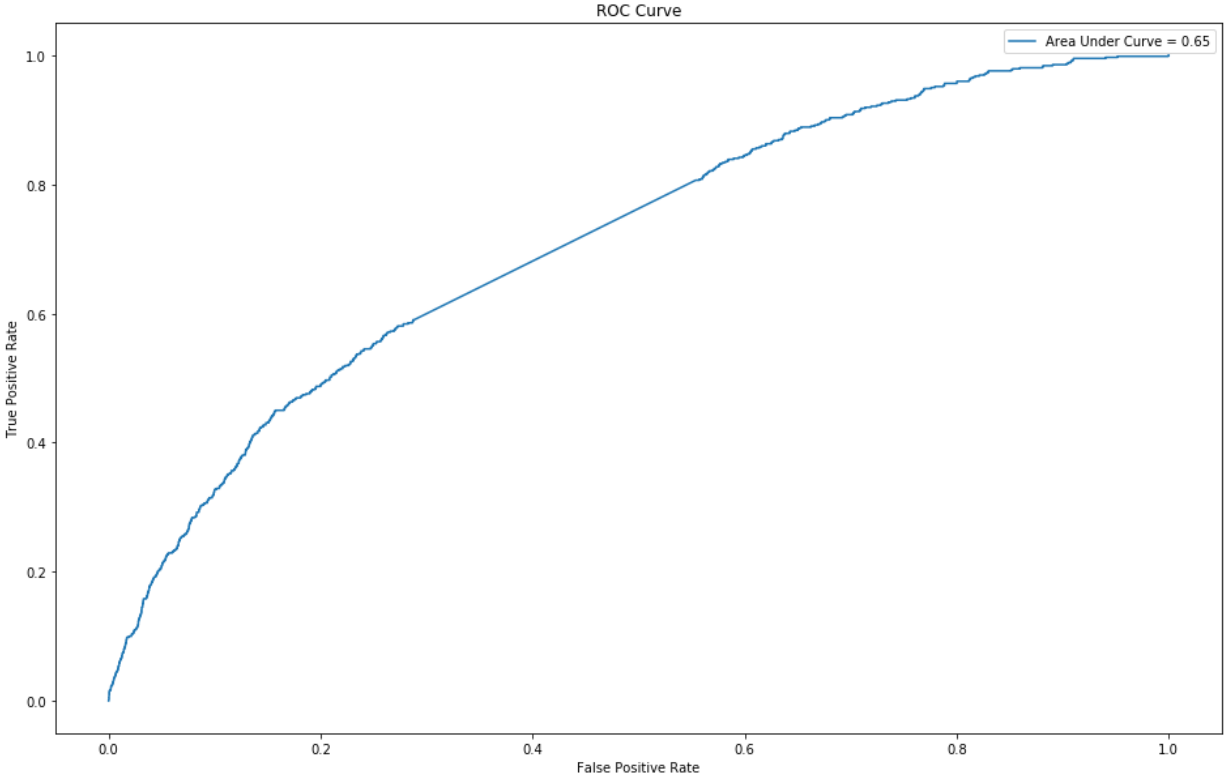



```
[[60415 18973]
 [  284   336]]
```



In [34]:
```
# show ROC for both basic algorithm and tuned
ROC_AUC(X_test, y_test, clf)
ROC_AUC(X_test, y_test, best_random)

# notice AUC improves by 0.3 with tuning
```


ROC Curve
Area Under Curve = 0.62
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve
Area Under Curve = 0.65
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

In [35]:
```
# retrain algorithm on STI data only with both features
train= model_STIData_bothFeat(training)
test= model_STIData_bothFeat(validation)

X_train, y_train, X_test, y_test, cols= Split(train, test)
us_X, us_y= underSamp(X_train, y_train)
```

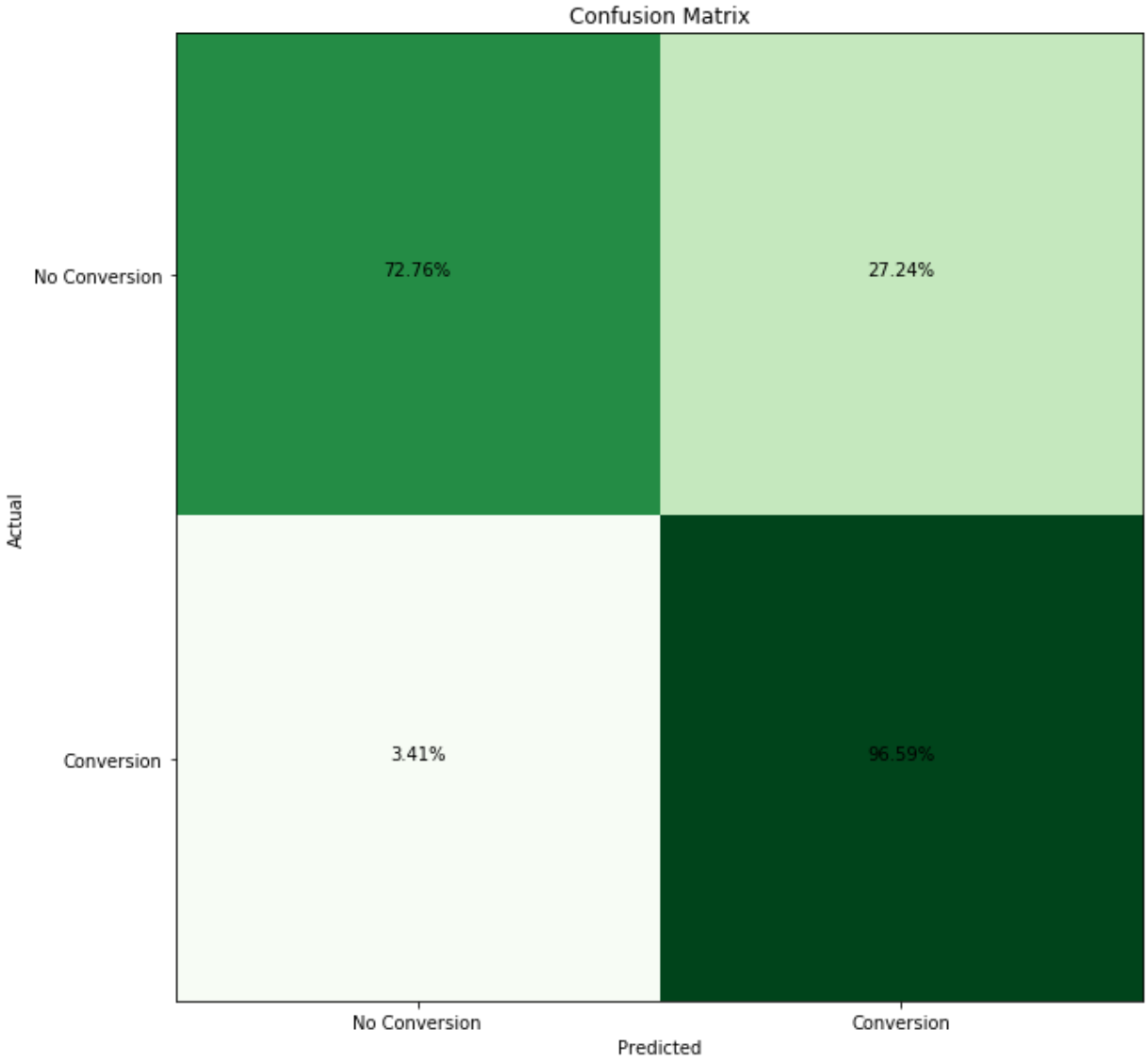
In [36]:
```
# repeat hyperparameter tuning with random selection nested cross vali
dation

best_random= RandomCV(us_X, us_y)
scores = cross_val_score(best_random, us_X, us_y, cv=3, scoring='accur
acy')
print('Training accuracy= '+str(scores.mean()))
print('Training standard deviation= '+ str(np.std(scores)))

clf = RandomForestClassifier(n_estimators=10)
clf.fit(us_X, us_y)
scores = cross_val_score(clf, us_X, us_y, cv=3, scoring='accuracy')
print('Training accuracy= '+str(scores.mean()))
print('Training standard deviation= '+ str(np.std(scores)))
```

```
{'n_estimators': 52, 'min_samples_split': 2, 'min_samples_leaf': 1,
'max_depth': None, 'bootstrap': True}
Training accuracy= 0.6671714549440405
Training standard deviation= 0.0031819235089912477
Training accuracy= 0.6392350294219452
Training standard deviation= 0.035795346173965695
```

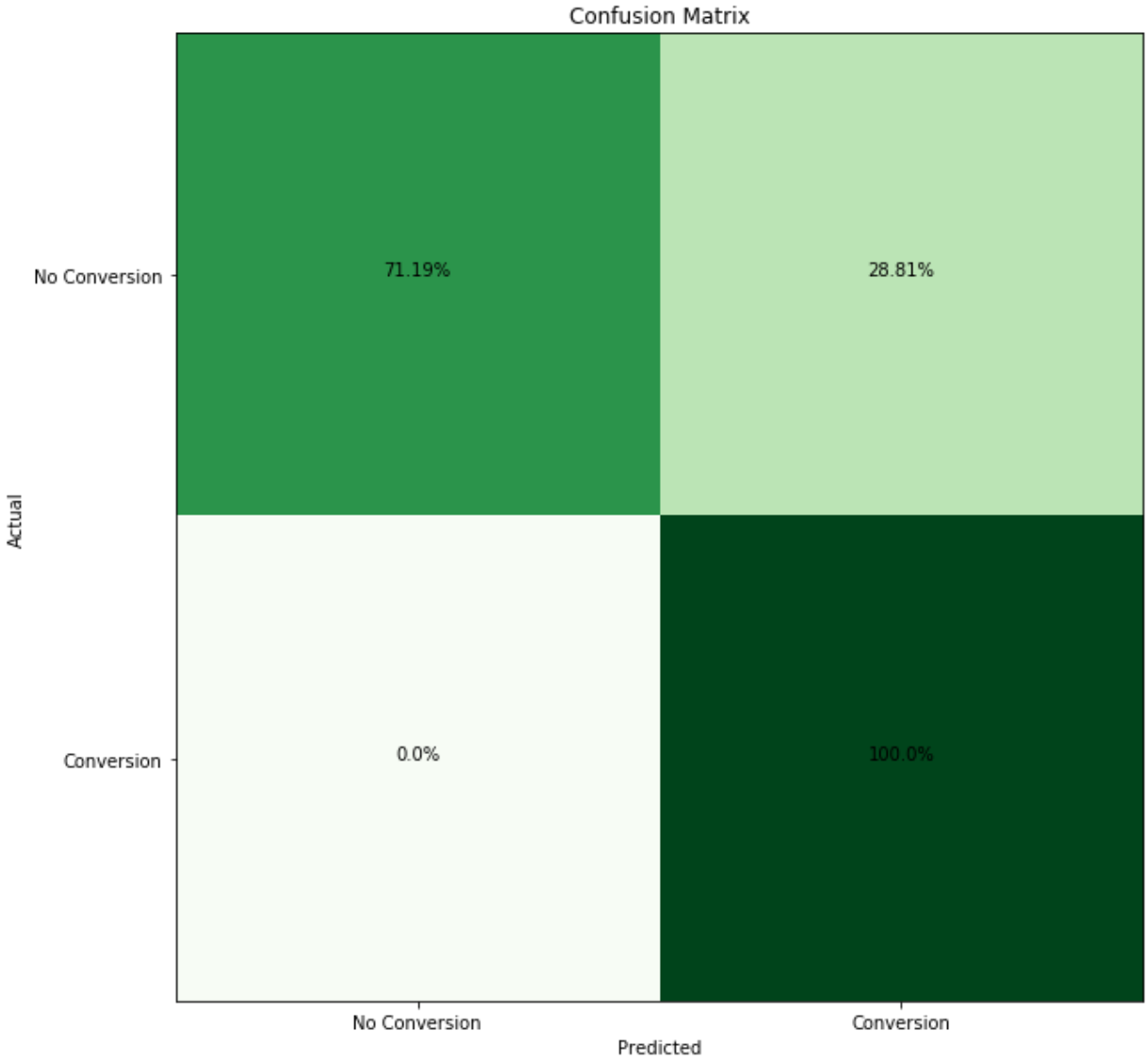
In [ ]:
```

```

In [37]:
```
# plot and compare confusion matrices
plot_confusion(X_test, y_test, clf)
plot_confusion(X_test, y_test, best_random)

# notice the sharp decrease in false negatives
```

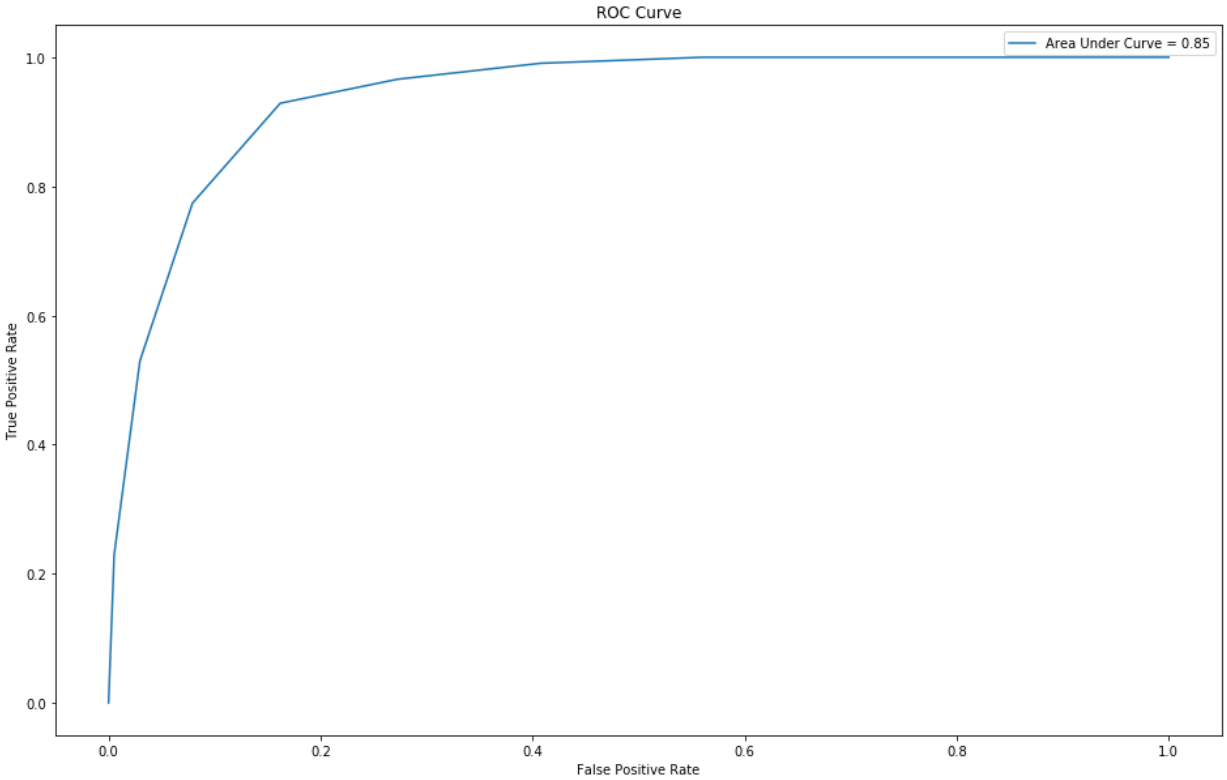
```
[[18499  6927]
 [   11   312]]
```

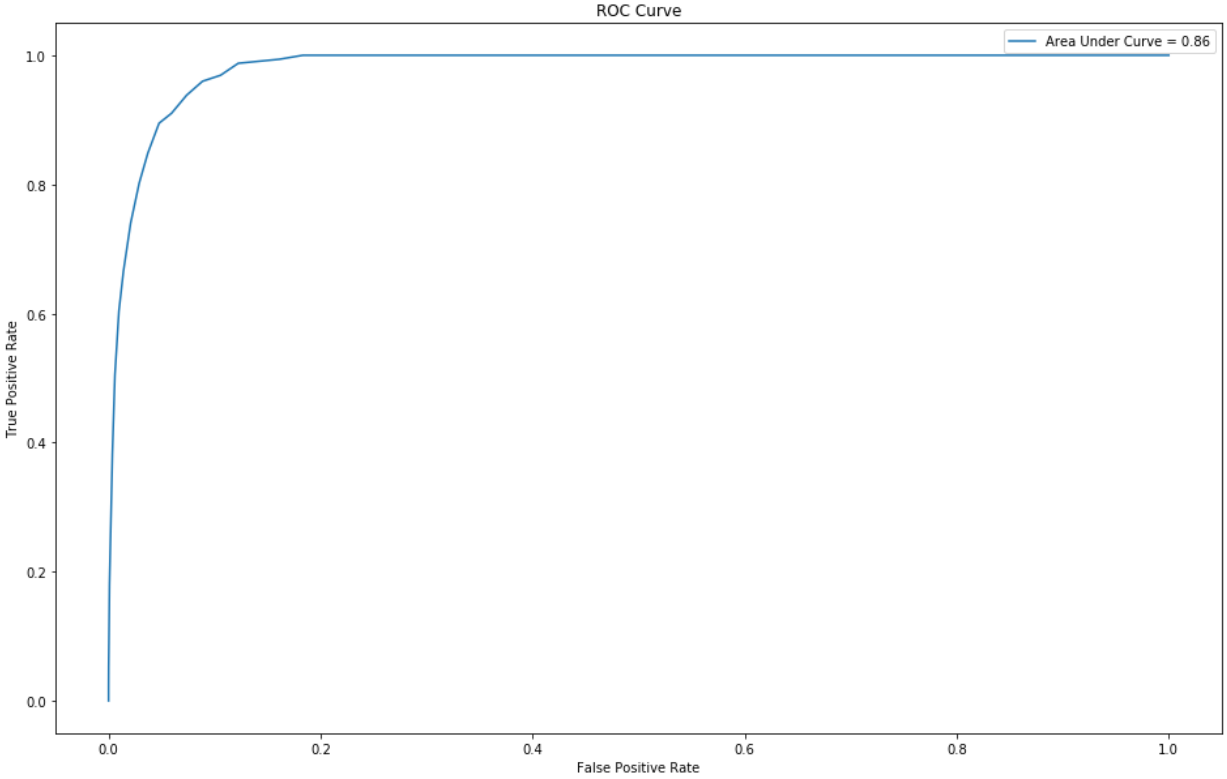



```
[[18102  7324]
 [    0   323]]
```

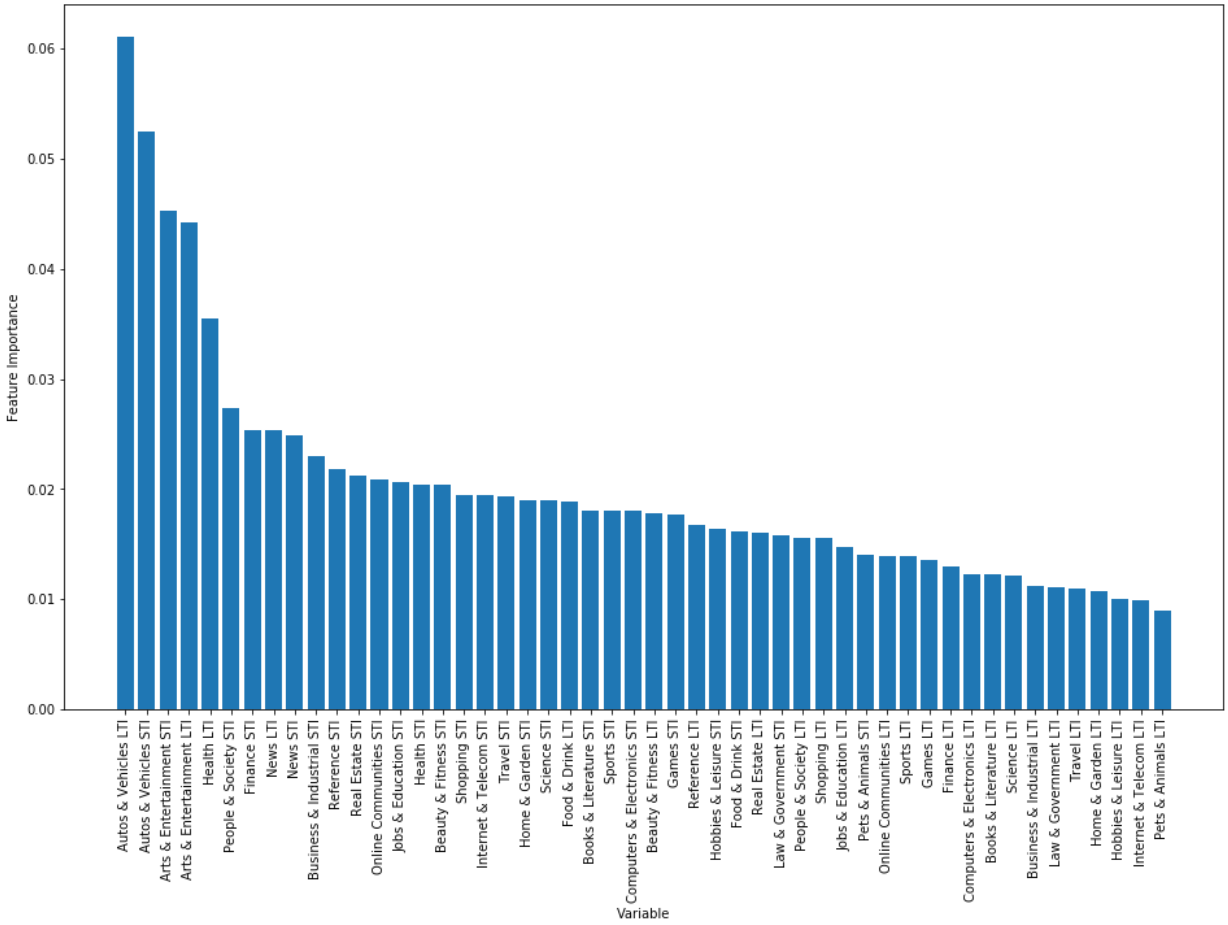


In [38]:
```
# compare ROC curves

ROC_AUC(X_test, y_test, clf)
ROC_AUC(X_test, y_test, best_random)
```


ROC Curve
Area Under Curve = 0.85
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve
Area Under Curve = 0.86
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

In [39]:
```
# show final overall variable importance

impor= var_import(cols, best_random)
plot_importance(impor)
impor
```



Out[39]:

feature	feature_importance
Autos & Vehicles LTI	0.0610
Autos & Vehicles STI	0.0525
Arts & Entertainment STI	0.0453
Arts & Entertainment LTI	0.0442
Health LTI	0.0355
People & Society STI	0.0274
Finance STI	0.0253
News LTI	0.0253

News STI	0.0249
Business & Industrial STI	0.0230
Reference STI	0.0218
Real Estate STI	0.0212
Online Communities STI	0.0209
Jobs & Education STI	0.0206
Health STI	0.0204
Beauty & Fitness STI	0.0204
Shopping STI	0.0195
Internet & Telecom STI	0.0195
Travel STI	0.0194
Home & Garden STI	0.0190
Science STI	0.0190
Food & Drink LTI	0.0189
Books & Literature STI	0.0181
Sports STI	0.0181
Computers & Electronics STI	0.0180
Beauty & Fitness LTI	0.0178
Games STI	0.0177
Reference LTI	0.0167
Hobbies & Leisure STI	0.0164
Food & Drink STI	0.0162
Real Estate LTI	0.0160
Law & Government STI	0.0158
People & Society LTI	0.0156
Shopping LTI	0.0156
Jobs & Education LTI	0.0148
Pets & Animals STI	0.0140
Online Communities LTI	0.0139
Sports LTI	0.0139
Games LTI	0.0136
Finance LTI	0.0130

Computers & Electronics LTI	0.0123
Books & Literature LTI	0.0123
Science LTI	0.0121
Business & Industrial LTI	0.0112
Law & Government LTI	0.0111
Travel LTI	0.0110
Home & Garden LTI	0.0107
Hobbies & Leisure LTI	0.0100
Internet & Telecom LTI	0.0099
Pets & Animals LTI	0.0090