

This jupyter notebook is prepared by "Patrick Ingram".

1. Load Data and perform basic EDA

1. import libraries: pandas, numpy, matplotlib (set %matplotlib inline), matplotlib's pyplot, seaborn, missingno, scipy's stats, sklearn (1 pt)
2. import the data to a dataframe and show the count of rows and columns (1 pt)
3. Show the top 5 and last 5 rows (1 pt)
4. Show how many columns have null values
5. Plot the count of target and discuss its imbalances and probably issues and solutions

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import scipy.stats as st
from sklearn import ensemble, tree, linear_model
import missingno as msno

from sklearn.preprocessing import MinMaxScaler
from imblearn.over_sampling import SMOTE, ADASYN

from sklearn.svm import SVC
from sklearn.datasets import make_classification
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedStratifiedKFold, train_test_split, GridSearchCV
from sklearn.pipeline import Pipeline, make_pipeline
from sklearn.decomposition import PCA
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.metrics import precision_score, recall_score
from sklearn.metrics import f1_score, make_scorer
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
from sklearn.metrics import precision_recall_curve
from sklearn.model_selection import cross_val_predict
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB, CategoricalNB
```

In [2]:

```
from google.colab import drive
drive.mount('/content/drive')
```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

In [3]:

```
hrdata = pd.read_csv('../content/drive/MyDrive/Colab Notebooks/hrdata2.csv')
```

In [4]:

```
hrdata.shape
```

Out[4]:

(8955, 15)

In [5]:

```
hrdata.describe()
```

Out[5]:

	Unnamed: 0	enrollee_id	city_development_index	experience	training_hours	target
count	8955.000000	8955.000000	8955.000000	8955.000000	8955.000000	8955.000000
mean	9527.135008	16869.638749	0.844570	11.635623	65.074930	0.165606
std	5554.606202	9963.804718	0.116178	6.545796	60.235087	0.371747
min	1.000000	2.000000	0.448000	0.000000	1.000000	0.000000
25%	4681.000000	8150.000000	0.794000	6.000000	23.000000	0.000000
50%	9488.000000	16924.000000	0.910000	10.000000	47.000000	0.000000
75%	14419.000000	25902.000000	0.920000	18.000000	88.000000	0.000000
max	19155.000000	33380.000000	0.949000	21.000000	336.000000	1.000000

In [6]:

```
hrdata.head()
```

Out[6]:

	Unnamed: 0	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level
0	1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate
1	4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters
2	7	402	city_46	0.762	Male	Has relevent experience	no_enrollment	Graduate
3	8	27107	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate
4	11	23853	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate

In [7]:

```
hrdata.tail()
```

Out[7]:

	Unnamed: 0	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level
8950	19147	21319	city_21	0.624	Male	No relevent experience	Full time course	Graduate
8951	19149	251	city_103	0.920	Male	Has relevent experience	no_enrollment	Masters
8952	19150	32313	city_160	0.920	Female	Has relevent experience	no_enrollment	Graduate
8953	19152	29754	city_103	0.920	Female	Has relevent experience	no_enrollment	Graduate
8954	19155	24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate

```
In [8]:
```

```
hrdata.isnull()
```

```
Out[8]:
```

	Unnamed: 0	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
...
8950	False	False	False	False	False	False	False	False
8951	False	False	False	False	False	False	False	False
8952	False	False	False	False	False	False	False	False
8953	False	False	False	False	False	False	False	False
8954	False	False	False	False	False	False	False	False

8955 rows x 15 columns



```
In [9]:
```

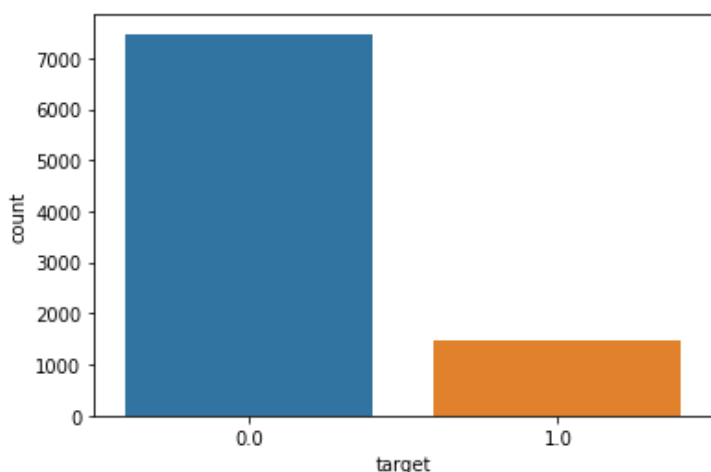
```
x_target = hrdata['target']  
sns.countplot(x_target)
```

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.

FutureWarning

```
Out[9]:
```

<matplotlib.axes._subplots.AxesSubplot at 0x7fe9d8caa250>



2. Feature Selection and Pre-Processing

```
In [10]:
```

```
x = hrdata['city']  
plt.figure(figsize = (20, 15))  
plt.xticks(rotation=90, horizontalalignment='right', fontweight='light', fontsize=8)  
sns.countplot(x, order=hrdata['city'].value_counts().index)
```

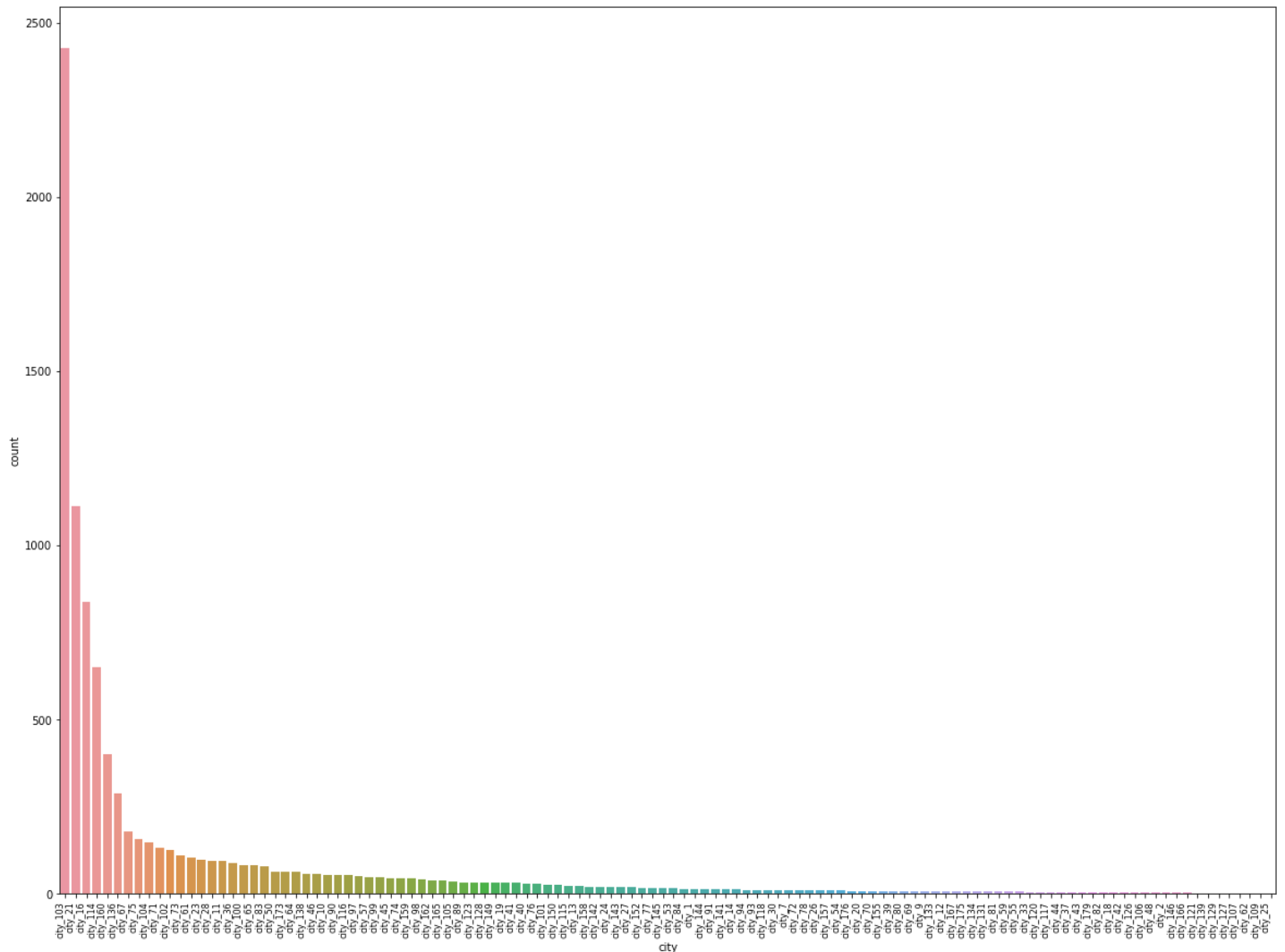
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the

following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.

FutureWarning

Out[10]:

<matplotlib.axes._subplots.AxesSubplot at 0x7fe9d8b44310>



In [11]:

```
#getting top 4 cities
```

```
city_top = hrdata[hrdata.city == 'city_103'].shape[0] + hrdata[hrdata.city == 'city_21']  
.shape[0] + hrdata[hrdata.city == 'city_16'].shape[0] + hrdata[hrdata.city == 'city_114']  
.shape[0]  
city_top
```

Out[11]:

5021

In [12]:

```
#Remaining rows
```

```
print(hrdata['city'].shape[0] - city_top)
```

3934

In [13]:

```
#Replacing with city_others
```

```
newSet = set(hrdata['city'])
```

```
newSet.remove('city_114')
newSet.remove('city_21')
newSet.remove('city_16')
newSet.remove('city_103')
```

```
hrdata.loc[hrdata['city'].isin(newSet), 'city'] = 'city_others'
newSet
```

Out[13]:

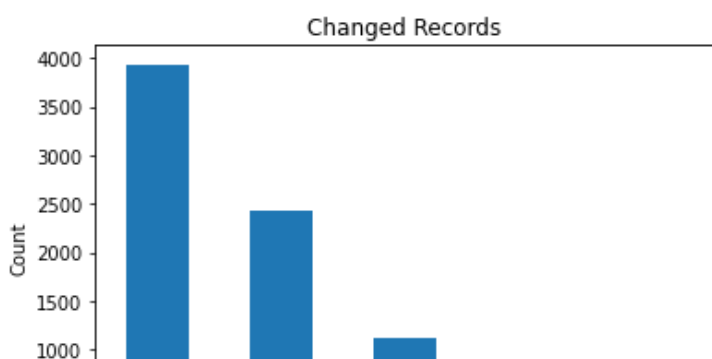
```
{'city_1',
 'city_10',
 'city_100',
 'city_101',
 'city_102',
 'city_104',
 'city_105',
 'city_106',
 'city_107',
 'city_109',
 'city_11',
 'city_115',
 'city_116',
 'city_117',
 'city_118',
 'city_12',
 'city_120',
 'city_121',
 'city_123',
 'city_126',
 'city_127',
 'city_128',
 'city_129',
 'city_13',
 'city_131',
 'city_133',
 'city_134',
 'city_136',
 'city_138',
 'city_139',
 'city_14',
 'city_141',
 'city_142',
 'city_143',
 'city_144',
 'city_145',
 'city_146',
 'city_149',
 'city_150',
 'city_152',
 'city_155',
 'city_157',
 'city_158',
 'city_159',
 'city_160',
 'city_162',
 'city_165',
 'city_166',
 'city_167',
 'city_173',
 'city_175',
 'city_176',
 'city_179',
 'city_18',
 'city_19',
 'city_2',
 'city_20',
 'city_23',
 'city_24',
 'city_25',
 'city_26',
 'city_27',
 'city_28'}
```

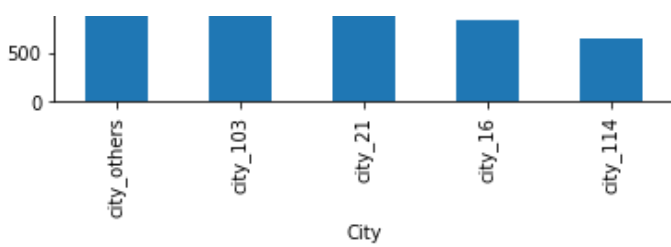
```
'city_20',  
'city_30',  
'city_33',  
'city_36',  
'city_37',  
'city_39',  
'city_40',  
'city_41',  
'city_42',  
'city_43',  
'city_44',  
'city_45',  
'city_46',  
'city_48',  
'city_50',  
'city_53',  
'city_54',  
'city_55',  
'city_57',  
'city_59',  
'city_61',  
'city_62',  
'city_64',  
'city_65',  
'city_67',  
'city_69',  
'city_7',  
'city_70',  
'city_71',  
'city_72',  
'city_73',  
'city_74',  
'city_75',  
'city_76',  
'city_77',  
'city_78',  
'city_80',  
'city_81',  
'city_82',  
'city_83',  
'city_84',  
'city_89',  
'city_9',  
'city_90',  
'city_91',  
'city_93',  
'city_94',  
'city_97',  
'city_98',  
'city_99'}
```

In [14]:

```
#plotting sample data for changed records
```

```
hrdata['city'].value_counts().plot(kind='bar')  
plt.title('Changed Records')  
plt.xlabel('City')  
plt.ylabel('Count')  
plt.show()
```





In [15]:

```
#Education Level

unq = set()

for i in hrdata['education_level']:
    unq.add(i)
unq
```

Out[15]:

```
{'Graduate', 'Masters', 'Phd'}
```

In [16]:

```
#Replacing edu level with Ordinal Values

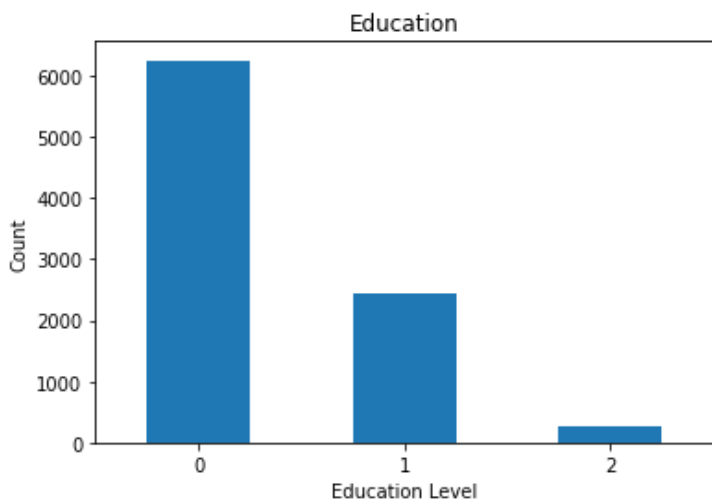
level = {
    'Graduate': 0,
    'Masters' : 1,
    'Phd' : 2
}

convEdu = hrdata.replace({'education_level': level})
hrdata = convEdu
```

In [17]:

```
#Show changes

hrdata['education_level'].value_counts().plot(kind='bar')
plt.xticks(rotation=0, fontweight='light')
plt.title('Education')
plt.xlabel('Education Level')
plt.ylabel('Count')
plt.show()
```



In [18]:

```
#Company Size

unq = set()

for i in hrdata['company_size']:
```

```
unq.add(i)
unq
```

Out[18]:

```
{'10/49',
 '100-500',
 '1000-4999',
 '10000+',
 '50-99',
 '500-999',
 '5000-9999',
 '<10'}
```

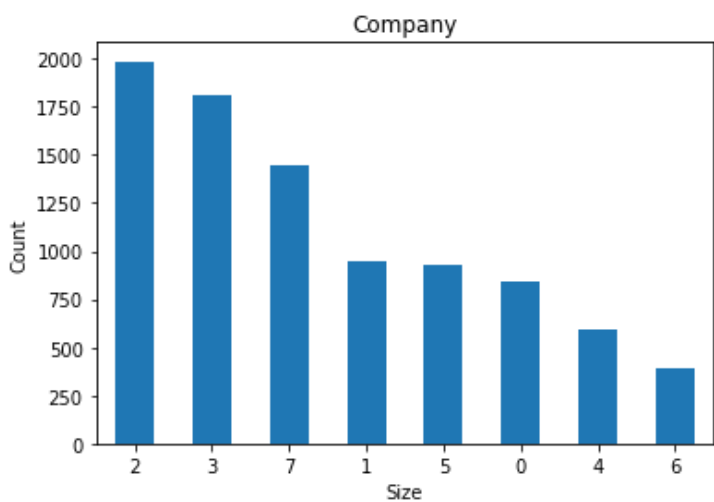
In [19]:

```
order = {'<10': 0, '10/49': 1, '50-99' : 2, '100-500':3, '500-999': 4, '1000-4999':5, '5000-9999':6, '10000+':7}
newSize = hrdata.replace({'company_size': order})
hrdata = newSize
```

In [20]:

#Show Changes

```
hrdata['company_size'].value_counts().plot(kind='bar')
plt.xticks(rotation=0, fontweight='light')
plt.title('Company')
plt.xlabel('Size')
plt.ylabel('Count')
plt.show()
```



In [21]:

```
unq = set()
for i in hrdata['last_new_job']:
    unq.add(i)
unq
```

Out[21]:

```
{'1', '2', '3', '4', '>4', 'never'}
```

In [22]:

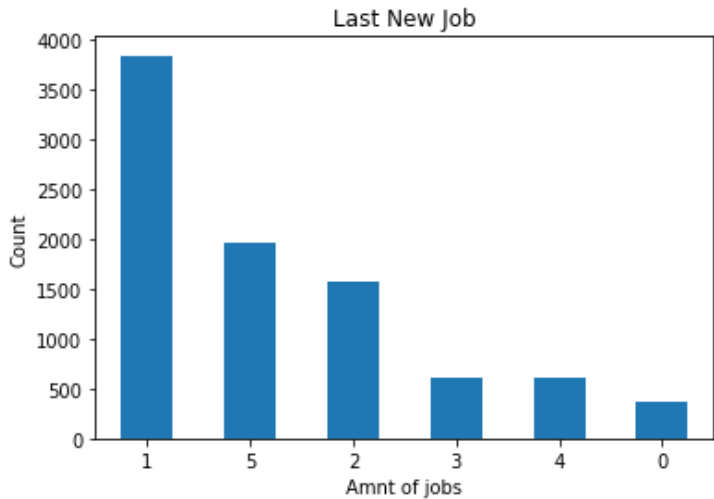
```
jobMap = {'never': 0, '1': 1, '2' : 2, '3':3, '4': 4, '>4':5}
lastNew = hrdata.replace({'last_new_job': jobMap})
hrdata = lastNew
```

In [23]:

```
hrdata['last_new_job'].value_counts().plot(kind='bar')
```



```
plt.xticks(rotation=0, fontweight='light')
plt.title('Last New Job')
plt.xlabel('Amnt of jobs')
plt.ylabel('Count')
plt.show()
```



In [24]:

```
#Rest of columns
```

```
comp = set()
gen = set()
ci = set()
relexp = set()
enruni = set()
mjdisc = set()

for i in hrdata['company_type']:
    comp.add(i)

for i in hrdata['gender']:
    gen.add(i)

for i in hrdata['city']:
    ci.add(i)

for i in hrdata['relevent_experience']:
    relexp.add(i)

for i in hrdata['enrolled_university']:
    enruni.add(i)

for i in hrdata['major_discipline']:
    mjdisc.add(i)

print(comp)
print(gen)
print(ci)
print(relexp)
print(enruni)
print(mjdisc)
```

```
{'Funded Startup', 'NGO', 'Other', 'Pvt Ltd', 'Public Sector', 'Early Stage Startup'}
{'Female', 'Male', 'Other'}
{'city_114', 'city_others', 'city_16', 'city_103', 'city_21'}
{'No relevent experience', 'Has relevent experience'}
{'Full time course', 'Part time course', 'no_enrollment'}
{'Other', 'Arts', 'STEM', 'Humanities', 'Business Degree', 'No Major'}
```

In [25]:

```
#get_dummies for binary columns
```

```
hrdata = pd.get_dummies(data=hrdata, columns=['company_type', 'gender', 'city', 'relevent_experience', 'enrolled_university', 'major_discipline'])
```

```
t_experience', 'enrolled_university', 'major_discipline']])
```

In [26]:

```
#Show top and last 5
hrdata.head()
```

Out[26]:

Unnamed: 0	enrollee_id	city_development_index	education_level	experience	company_size	last_new_job	training_hours	ta
0	1	29725	0.776	0	15.0	2	5	47
1	4	666	0.767	1	21.0	2	4	8
2	7	402	0.762	0	13.0	0	5	18
3	8	27107	0.920	0	7.0	2	1	46
4	11	23853	0.920	0	5.0	6	1	108

5 rows x 34 columns



In [27]:

```
hrdata.tail()
```

Out[27]:

Unnamed: 0	enrollee_id	city_development_index	education_level	experience	company_size	last_new_job	training_hours	
8950	19147	21319	0.624	0	1.0	3	1	52
8951	19149	251	0.920	1	9.0	2	1	36
8952	19150	32313	0.920	0	10.0	3	3	23
8953	19152	29754	0.920	0	7.0	1	1	25
8954	19155	24576	0.920	0	21.0	2	4	44

5 rows x 34 columns



In [28]:

```
#Show the shape of the new table
```

```
hrdata.shape
```

Out[28]:

```
(8955, 34)
```

In [29]:

```
#Drop Enrollee ID and duplicates
```

```
hrdata = hrdata.drop('enrollee_id', axis=1)
```

In [30]:

```
#Feature Scaling
scaler = MinMaxScaler()
col = hrdata.columns
data_scale = scaler.fit_transform(hrdata.to_numpy())
data_scale = pd.DataFrame(data_scale, columns = col)

data_scale.head()
```

Out[30]:

	Unnamed: 0	city_development_index	education_level	experience	company_size	last_new_job	training_hours	target	company_type
0	0.000000	0.654691	0.0	0.714286	0.285714	1.0	0.137313	0.0	company_type_Early Stage Startup
1	0.000157	0.636727	0.5	1.000000	0.285714	0.8	0.020896	0.0	company_type_Early Stage Startup
2	0.000313	0.626747	0.0	0.619048	0.000000	1.0	0.050746	1.0	company_type_Early Stage Startup
3	0.000365	0.942116	0.0	0.333333	0.285714	0.2	0.134328	1.0	company_type_Early Stage Startup
4	0.000522	0.942116	0.0	0.238095	0.857143	0.2	0.319403	0.0	company_type_Early Stage Startup

5 rows x 33 columns

In [31]:

```
#Move target to last column

data_scale = data_scale.loc[:, data_scale.columns != 'target']
data_scale['target'] = hrdata['target']
data_scale.head()
```

Out[31]:

	Unnamed: 0	city_development_index	education_level	experience	company_size	last_new_job	training_hours	company_type
0	0.000000	0.654691	0.0	0.714286	0.285714	1.0	0.137313	Stage S
1	0.000157	0.636727	0.5	1.000000	0.285714	0.8	0.020896	Stage S
2	0.000313	0.626747	0.0	0.619048	0.000000	1.0	0.050746	Stage S
3	0.000365	0.942116	0.0	0.333333	0.285714	0.2	0.134328	Stage S
4	0.000522	0.942116	0.0	0.238095	0.857143	0.2	0.319403	Stage S

5 rows x 33 columns

3. X/Y and Training/Test Split with stratified sampling and SMOTE

In [32]:

```
#Copy all features into X and target into Y

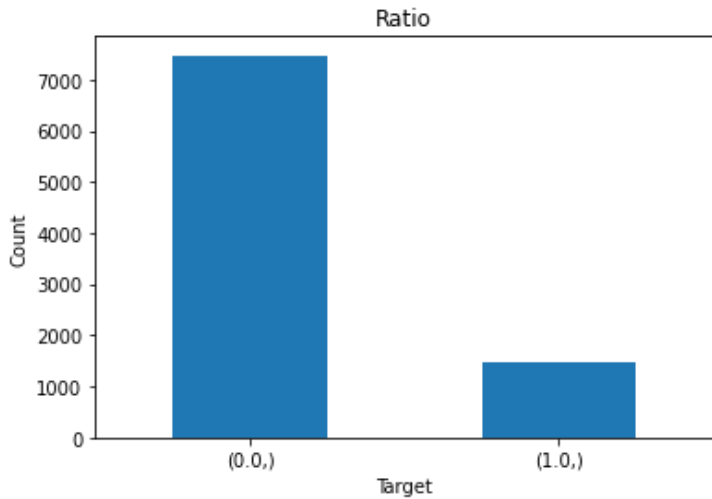
X = hrdata[['city_development_index', 'education_level', 'experience',
'company_size', 'last_new_job', 'training_hours',
'company_type_Early Stage Startup', 'company_type_Funded Startup',
'company_type_NGO', 'company_type_Other', 'company_type_Public Sector',
'company_type_Pvt Ltd', 'major_discipline_Arts',
'major_discipline_Business Degree', 'major_discipline_Humanities',
'major_discipline_No Major', 'major_discipline_Other',
'major_discipline_STEM', 'enrolled_university_Full time course',
'enrolled_university_Part time course',
'enrolled_university_no_enrollment',
'relevant_experience_Has relevant experience',
'relevant_experience_No relevant experience', 'gender_Female',
'gender_Male', 'gender_Other', 'city_city_103', 'city_city_114',
'city_city_16', 'city_city_21', 'city_city_others']]

Y = hrdata[['target']]
```

In [33]:

```
#Show the ratio of 1 and 0 in Y
```

```
Y.value_counts().plot(kind='bar')
plt.title('Ratio')
plt.xticks(rotation=0, fontweight='light')
plt.xlabel('Target')
plt.ylabel('Count')
plt.show()
```

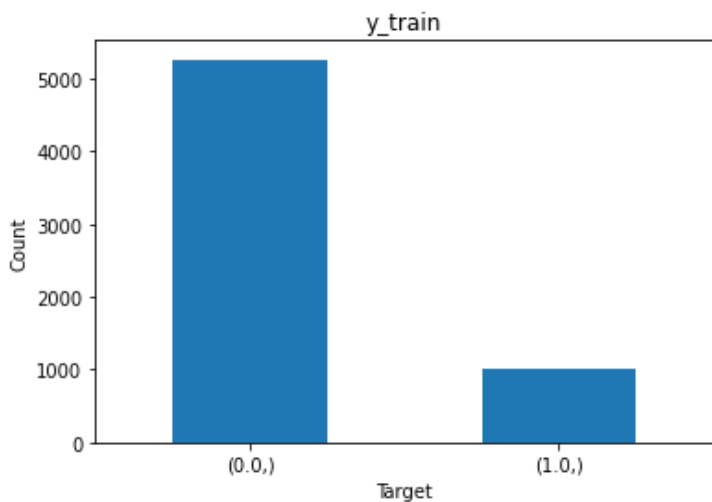


In [34]:

```
#Use sklearn's train_test_split to split the data set into training and test sets. There should be 30% records in the test set.
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = .30, random_state = 0)
```

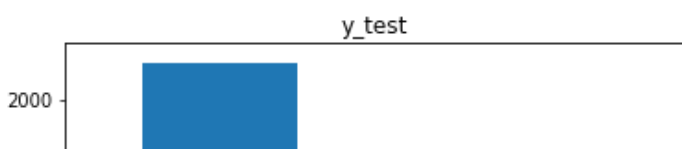
In [35]:

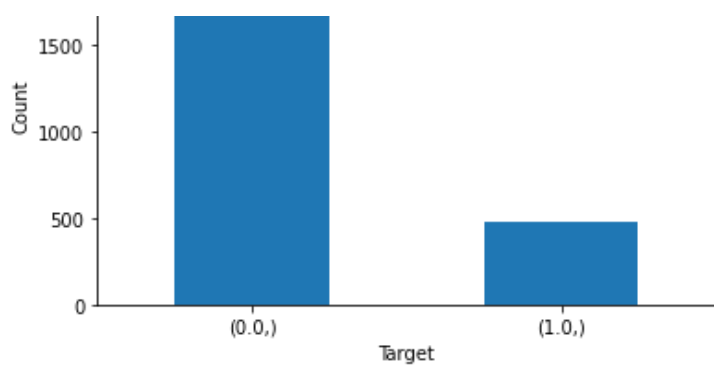
```
y_train.value_counts().plot(kind='bar')
plt.xticks(rotation=0, fontweight='light')
plt.title('y_train')
plt.xlabel('Target')
plt.ylabel('Count')
plt.show()
```



In [36]:

```
y_test.value_counts().plot(kind='bar')
plt.xticks(rotation=0, fontweight='light')
plt.title('y_test')
plt.xlabel('Target')
plt.ylabel('Count')
plt.show()
```



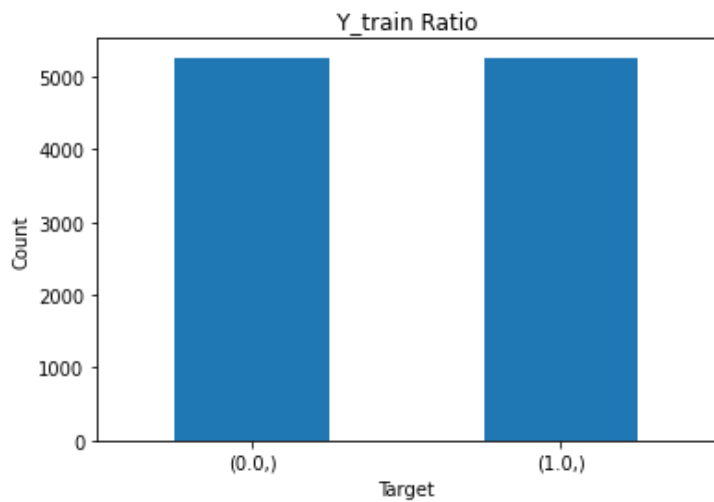


In [37]:

```
X_scaled, y_scaled = SMOTE().fit_resample(X_train, y_train)
```

In [38]:

```
y_scaled.value_counts().plot(kind='bar')
plt.xticks(rotation=0, fontweight='light')
plt.title('Y_train Ratio')
plt.xlabel('Target')
plt.ylabel('Count')
plt.show()
```



4. PCA and Logistic Regression

In [39]:

```
#Logistical Regression Model with Make_Pipeline

features = list(range(1,X.shape[1]+1))
model_s = []
for i in features:
    step = make_pipeline(PCA(n_components=i), LogisticRegression())
    RS = RepeatedStratifiedKFold(n_splits=10, n_repeats=3)
    scores = cross_val_score(step, X_scaled, y_scaled, scoring='accuracy', cv=RS, n_jobs=-1)
    model_s.append(scores)

plt.boxplot(model_s, showmeans=True)
```

Out[39]:

```
{'boxes': [<matplotlib.lines.Line2D at 0x7fe9d5803fd0>,
<matplotlib.lines.Line2D at 0x7fe9d57fccd0>,
<matplotlib.lines.Line2D at 0x7fe9d57e6250>,
<matplotlib.lines.Line2D at 0x7fe9d5773d50>,
<matplotlib.lines.Line2D at 0x7fe9d578a950>,
<matplotlib.lines.Line2D at 0x7fe9d57a7550>,
<matplotlib.lines.Line2D at 0x7fe9d5739fd0>,
<matplotlib.lines.Line2D at 0x7fe9d5753a90>,
```

```
<matplotlib.lines.Line2D at 0x7fe9d576c310>,  
<matplotlib.lines.Line2D at 0x7fe9d56fdc50>,  
<matplotlib.lines.Line2D at 0x7fe9d5718710>,  
<matplotlib.lines.Line2D at 0x7fe9d56b21d0>,  
<matplotlib.lines.Line2D at 0x7fe9d56c2c50>,  
<matplotlib.lines.Line2D at 0x7fe9d56dd750>,  
<matplotlib.lines.Line2D at 0x7fe9d56781d0>,  
<matplotlib.lines.Line2D at 0x7fe9d5687c50>,  
<matplotlib.lines.Line2D at 0x7fe9d569f750>,  
<matplotlib.lines.Line2D at 0x7fe9d563a1d0>,  
<matplotlib.lines.Line2D at 0x7fe9d564cc50>,  
<matplotlib.lines.Line2D at 0x7fe9d5666750>,  
<matplotlib.lines.Line2D at 0x7fe9d55ff210>,  
<matplotlib.lines.Line2D at 0x7fe9d5610c50>,  
<matplotlib.lines.Line2D at 0x7fe9d5629710>,  
<matplotlib.lines.Line2D at 0x7fe9d55c31d0>,  
<matplotlib.lines.Line2D at 0x7fe9d55d4c50>,  
<matplotlib.lines.Line2D at 0x7fe9d55ef750>,  
<matplotlib.lines.Line2D at 0x7fe9d55881d0>,  
<matplotlib.lines.Line2D at 0x7fe9d5599c50>,  
<matplotlib.lines.Line2D at 0x7fe9d5532750>,  
<matplotlib.lines.Line2D at 0x7fe9d554a210>,  
<matplotlib.lines.Line2D at 0x7fe9d555fc90>],  
'caps': [<matplotlib.lines.Line2D at 0x7fe9d5820d90>,  
<matplotlib.lines.Line2D at 0x7fe9d5820110>,  
<matplotlib.lines.Line2D at 0x7fe9d57c3290>,  
<matplotlib.lines.Line2D at 0x7fe9d57c3ad0>,  
<matplotlib.lines.Line2D at 0x7fe9d57ec2d0>,  
<matplotlib.lines.Line2D at 0x7fe9d57ec850>,  
<matplotlib.lines.Line2D at 0x7fe9d577bd90>,  
<matplotlib.lines.Line2D at 0x7fe9d5783310>,  
<matplotlib.lines.Line2D at 0x7fe9d5792990>,  
<matplotlib.lines.Line2D at 0x7fe9d5792e10>,  
<matplotlib.lines.Line2D at 0x7fe9d57af510>,  
<matplotlib.lines.Line2D at 0x7fe9d57afb50>,  
<matplotlib.lines.Line2D at 0x7fe9d5743fd0>,  
<matplotlib.lines.Line2D at 0x7fe9d574b550>,  
<matplotlib.lines.Line2D at 0x7fe9d575da50>,  
<matplotlib.lines.Line2D at 0x7fe9d575df90>,  
<matplotlib.lines.Line2D at 0x7fe9d56f42d0>,  
<matplotlib.lines.Line2D at 0x7fe9d56f4810>,  
<matplotlib.lines.Line2D at 0x7fe9d5708c90>,  
<matplotlib.lines.Line2D at 0x7fe9d5710210>,  
<matplotlib.lines.Line2D at 0x7fe9d5722750>,  
<matplotlib.lines.Line2D at 0x7fe9d5722c90>,  
<matplotlib.lines.Line2D at 0x7fe9d56ba210>,  
<matplotlib.lines.Line2D at 0x7fe9d56ba750>,  
<matplotlib.lines.Line2D at 0x7fe9d56ccc90>,  
<matplotlib.lines.Line2D at 0x7fe9d56d3210>,  
<matplotlib.lines.Line2D at 0x7fe9d56e6750>,  
<matplotlib.lines.Line2D at 0x7fe9d56e6c90>,  
<matplotlib.lines.Line2D at 0x7fe9d567f210>,  
<matplotlib.lines.Line2D at 0x7fe9d567f750>,  
<matplotlib.lines.Line2D at 0x7fe9d5692c90>,  
<matplotlib.lines.Line2D at 0x7fe9d5698210>,  
<matplotlib.lines.Line2D at 0x7fe9d56aa750>,  
<matplotlib.lines.Line2D at 0x7fe9d56aac90>,  
<matplotlib.lines.Line2D at 0x7fe9d5643210>,  
<matplotlib.lines.Line2D at 0x7fe9d5643750>,  
<matplotlib.lines.Line2D at 0x7fe9d5653c90>,  
<matplotlib.lines.Line2D at 0x7fe9d565f210>,  
<matplotlib.lines.Line2D at 0x7fe9d566e750>,  
<matplotlib.lines.Line2D at 0x7fe9d566ec90>,  
<matplotlib.lines.Line2D at 0x7fe9d5607210>,  
<matplotlib.lines.Line2D at 0x7fe9d5607750>,  
<matplotlib.lines.Line2D at 0x7fe9d5618c90>,  
<matplotlib.lines.Line2D at 0x7fe9d5620210>,  
<matplotlib.lines.Line2D at 0x7fe9d55b2750>,  
<matplotlib.lines.Line2D at 0x7fe9d55b2c90>,  
<matplotlib.lines.Line2D at 0x7fe9d55cb210>,  
<matplotlib.lines.Line2D at 0x7fe9d55cb750>,  
<matplotlib.lines.Line2D at 0x7fe9d55dec90>],
```

```
<matplotlib.lines.Line2D at 0x7fe9d55e6210>,  
<matplotlib.lines.Line2D at 0x7fe9d5577750>,  
<matplotlib.lines.Line2D at 0x7fe9d5577c90>,  
<matplotlib.lines.Line2D at 0x7fe9d5591210>,  
<matplotlib.lines.Line2D at 0x7fe9d5591750>,  
<matplotlib.lines.Line2D at 0x7fe9d55a1c90>,  
<matplotlib.lines.Line2D at 0x7fe9d55aa210>,  
<matplotlib.lines.Line2D at 0x7fe9d553b750>,  
<matplotlib.lines.Line2D at 0x7fe9d553bc90>,  
<matplotlib.lines.Line2D at 0x7fe9d5555210>,  
<matplotlib.lines.Line2D at 0x7fe9d5555750>,  
<matplotlib.lines.Line2D at 0x7fe9d5567c90>,  
<matplotlib.lines.Line2D at 0x7fe9d556d210>],  
'fliers': [<matplotlib.lines.Line2D at 0x7fe9d5800210>,  
<matplotlib.lines.Line2D at 0x7fe9d57dfd10>,  
<matplotlib.lines.Line2D at 0x7fe9d5773850>,  
<matplotlib.lines.Line2D at 0x7fe9d578a310>,  
<matplotlib.lines.Line2D at 0x7fe9d579d590>,  
<matplotlib.lines.Line2D at 0x7fe9d5739a90>,  
<matplotlib.lines.Line2D at 0x7fe9d5753590>,  
<matplotlib.lines.Line2D at 0x7fe9d5764dd0>,  
<matplotlib.lines.Line2D at 0x7fe9d56fd790>,  
<matplotlib.lines.Line2D at 0x7fe9d5718210>,  
<matplotlib.lines.Line2D at 0x7fe9d5729c90>,  
<matplotlib.lines.Line2D at 0x7fe9d56c2750>,  
<matplotlib.lines.Line2D at 0x7fe9d56dd210>,  
<matplotlib.lines.Line2D at 0x7fe9d56eec90>,  
<matplotlib.lines.Line2D at 0x7fe9d5687750>,  
<matplotlib.lines.Line2D at 0x7fe9d569f210>,  
<matplotlib.lines.Line2D at 0x7fe9d5632c90>,  
<matplotlib.lines.Line2D at 0x7fe9d564c750>,  
<matplotlib.lines.Line2D at 0x7fe9d5666210>,  
<matplotlib.lines.Line2D at 0x7fe9d55f7c90>,  
<matplotlib.lines.Line2D at 0x7fe9d5610750>,  
<matplotlib.lines.Line2D at 0x7fe9d5629210>,  
<matplotlib.lines.Line2D at 0x7fe9d55bac90>,  
<matplotlib.lines.Line2D at 0x7fe9d55d4750>,  
<matplotlib.lines.Line2D at 0x7fe9d55ef210>,  
<matplotlib.lines.Line2D at 0x7fe9d557ec90>,  
<matplotlib.lines.Line2D at 0x7fe9d5599750>,  
<matplotlib.lines.Line2D at 0x7fe9d5532210>,  
<matplotlib.lines.Line2D at 0x7fe9d5544c90>,  
<matplotlib.lines.Line2D at 0x7fe9d555f750>,  
<matplotlib.lines.Line2D at 0x7fe9d54f7210>],  
'means': [<matplotlib.lines.Line2D at 0x7fe9d5800a90>,  
<matplotlib.lines.Line2D at 0x7fe9d57df7d0>,  
<matplotlib.lines.Line2D at 0x7fe9d5773310>,  
<matplotlib.lines.Line2D at 0x7fe9d5783d90>,  
<matplotlib.lines.Line2D at 0x7fe9d579d950>,  
<matplotlib.lines.Line2D at 0x7fe9d5739550>,  
<matplotlib.lines.Line2D at 0x7fe9d574bfd0>,  
<matplotlib.lines.Line2D at 0x7fe9d57648d0>,  
<matplotlib.lines.Line2D at 0x7fe9d56fd290>,  
<matplotlib.lines.Line2D at 0x7fe9d5710c90>,  
<matplotlib.lines.Line2D at 0x7fe9d5729750>,  
<matplotlib.lines.Line2D at 0x7fe9d56c2210>,  
<matplotlib.lines.Line2D at 0x7fe9d56d3c90>,  
<matplotlib.lines.Line2D at 0x7fe9d56ee750>,  
<matplotlib.lines.Line2D at 0x7fe9d5687210>,  
<matplotlib.lines.Line2D at 0x7fe9d5698c90>,  
<matplotlib.lines.Line2D at 0x7fe9d5632750>,  
<matplotlib.lines.Line2D at 0x7fe9d564c210>,  
<matplotlib.lines.Line2D at 0x7fe9d565fc90>,  
<matplotlib.lines.Line2D at 0x7fe9d55f7750>,  
<matplotlib.lines.Line2D at 0x7fe9d5610210>,  
<matplotlib.lines.Line2D at 0x7fe9d5620c90>,  
<matplotlib.lines.Line2D at 0x7fe9d55ba750>,  
<matplotlib.lines.Line2D at 0x7fe9d55d4210>,  
<matplotlib.lines.Line2D at 0x7fe9d55e6c90>,  
<matplotlib.lines.Line2D at 0x7fe9d557e750>,  
<matplotlib.lines.Line2D at 0x7fe9d5599210>,  
<matplotlib.lines.Line2D at 0x7fe9d55aac90>]
```

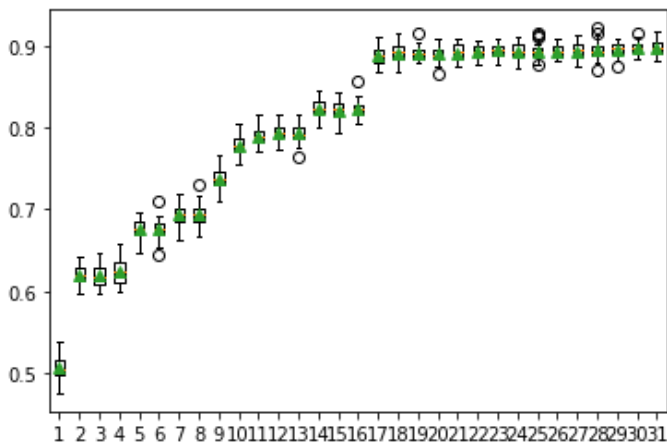
```
<matplotlib.lines.Line2D at 0x7fe9d5544750>,  
<matplotlib.lines.Line2D at 0x7fe9d555f210>,  
<matplotlib.lines.Line2D at 0x7fe9d556dc90>],  
'medians': [<matplotlib.lines.Line2D at 0x7fe9d58008d0>,  
<matplotlib.lines.Line2D at 0x7fe9d57df290>,  
<matplotlib.lines.Line2D at 0x7fe9d57ecd90>,  
<matplotlib.lines.Line2D at 0x7fe9d5783850>,  
<matplotlib.lines.Line2D at 0x7fe9d579d390>,  
<matplotlib.lines.Line2D at 0x7fe9d57affd0>,  
<matplotlib.lines.Line2D at 0x7fe9d574ba90>,  
<matplotlib.lines.Line2D at 0x7fe9d5764350>,  
<matplotlib.lines.Line2D at 0x7fe9d56f4d50>,  
<matplotlib.lines.Line2D at 0x7fe9d5710750>,  
<matplotlib.lines.Line2D at 0x7fe9d5729210>,  
<matplotlib.lines.Line2D at 0x7fe9d56bac90>,  
<matplotlib.lines.Line2D at 0x7fe9d56d3750>,  
<matplotlib.lines.Line2D at 0x7fe9d56ee210>,  
<matplotlib.lines.Line2D at 0x7fe9d567fc90>,  
<matplotlib.lines.Line2D at 0x7fe9d5698750>,  
<matplotlib.lines.Line2D at 0x7fe9d5632210>,  
<matplotlib.lines.Line2D at 0x7fe9d5643c90>,  
<matplotlib.lines.Line2D at 0x7fe9d565f750>,  
<matplotlib.lines.Line2D at 0x7fe9d55f7210>,  
<matplotlib.lines.Line2D at 0x7fe9d5607c90>,  
<matplotlib.lines.Line2D at 0x7fe9d5620750>,  
<matplotlib.lines.Line2D at 0x7fe9d55ba210>,  
<matplotlib.lines.Line2D at 0x7fe9d55cbc90>,  
<matplotlib.lines.Line2D at 0x7fe9d55e6750>,  
<matplotlib.lines.Line2D at 0x7fe9d557e210>,  
<matplotlib.lines.Line2D at 0x7fe9d5591c90>,  
<matplotlib.lines.Line2D at 0x7fe9d55aa750>,  
<matplotlib.lines.Line2D at 0x7fe9d5544210>,  
<matplotlib.lines.Line2D at 0x7fe9d5555c90>,  
<matplotlib.lines.Line2D at 0x7fe9d556d750>],  
'whiskers': [<matplotlib.lines.Line2D at 0x7fe9d810f990>,  
<matplotlib.lines.Line2D at 0x7fe9d588b8d0>,  
<matplotlib.lines.Line2D at 0x7fe9d57fc410>,  
<matplotlib.lines.Line2D at 0x7fe9d57c3710>,  
<matplotlib.lines.Line2D at 0x7fe9d57e67d0>,  
<matplotlib.lines.Line2D at 0x7fe9d57e6d50>,  
<matplotlib.lines.Line2D at 0x7fe9d577b310>,  
<matplotlib.lines.Line2D at 0x7fe9d577b850>,  
<matplotlib.lines.Line2D at 0x7fe9d578af10>,  
<matplotlib.lines.Line2D at 0x7fe9d5792450>,  
<matplotlib.lines.Line2D at 0x7fe9d57a7a90>,  
<matplotlib.lines.Line2D at 0x7fe9d57a7f90>,  
<matplotlib.lines.Line2D at 0x7fe9d5743550>,  
<matplotlib.lines.Line2D at 0x7fe9d5743a90>,  
<matplotlib.lines.Line2D at 0x7fe9d575d050>,  
<matplotlib.lines.Line2D at 0x7fe9d575d4d0>,  
<matplotlib.lines.Line2D at 0x7fe9d576c810>,  
<matplotlib.lines.Line2D at 0x7fe9d576cd50>,  
<matplotlib.lines.Line2D at 0x7fe9d5708210>,  
<matplotlib.lines.Line2D at 0x7fe9d5708750>,  
<matplotlib.lines.Line2D at 0x7fe9d5718c90>,  
<matplotlib.lines.Line2D at 0x7fe9d5722210>,  
<matplotlib.lines.Line2D at 0x7fe9d56b2750>,  
<matplotlib.lines.Line2D at 0x7fe9d56b2c90>,  
<matplotlib.lines.Line2D at 0x7fe9d56cc210>,  
<matplotlib.lines.Line2D at 0x7fe9d56cc750>,  
<matplotlib.lines.Line2D at 0x7fe9d56ddc90>,  
<matplotlib.lines.Line2D at 0x7fe9d56e6210>,  
<matplotlib.lines.Line2D at 0x7fe9d5678750>,  
<matplotlib.lines.Line2D at 0x7fe9d5678c90>,  
<matplotlib.lines.Line2D at 0x7fe9d5692210>,  
<matplotlib.lines.Line2D at 0x7fe9d5692750>,  
<matplotlib.lines.Line2D at 0x7fe9d569fc90>,  
<matplotlib.lines.Line2D at 0x7fe9d56aa210>,  
<matplotlib.lines.Line2D at 0x7fe9d563a750>,  
<matplotlib.lines.Line2D at 0x7fe9d563ac90>,  
<matplotlib.lines.Line2D at 0x7fe9d5653210>,  
<matplotlib.lines.Line2D at 0x7fe9d5653750>]
```



```

<matplotlib.lines.Line2D at 0x7fe9d5666c90>,
<matplotlib.lines.Line2D at 0x7fe9d566e210>,
<matplotlib.lines.Line2D at 0x7fe9d55ff750>,
<matplotlib.lines.Line2D at 0x7fe9d55ffc90>,
<matplotlib.lines.Line2D at 0x7fe9d5618210>,
<matplotlib.lines.Line2D at 0x7fe9d5618750>,
<matplotlib.lines.Line2D at 0x7fe9d5629c90>,
<matplotlib.lines.Line2D at 0x7fe9d55b2210>,
<matplotlib.lines.Line2D at 0x7fe9d55c3750>,
<matplotlib.lines.Line2D at 0x7fe9d55c3c90>,
<matplotlib.lines.Line2D at 0x7fe9d55de210>,
<matplotlib.lines.Line2D at 0x7fe9d55de750>,
<matplotlib.lines.Line2D at 0x7fe9d55efc90>,
<matplotlib.lines.Line2D at 0x7fe9d5577210>,
<matplotlib.lines.Line2D at 0x7fe9d5588750>,
<matplotlib.lines.Line2D at 0x7fe9d5588c90>,
<matplotlib.lines.Line2D at 0x7fe9d55a1210>,
<matplotlib.lines.Line2D at 0x7fe9d55a1750>,
<matplotlib.lines.Line2D at 0x7fe9d5532c90>,
<matplotlib.lines.Line2D at 0x7fe9d553b210>,
<matplotlib.lines.Line2D at 0x7fe9d554a750>,
<matplotlib.lines.Line2D at 0x7fe9d554ac90>,
<matplotlib.lines.Line2D at 0x7fe9d5567210>,
<matplotlib.lines.Line2D at 0x7fe9d5567750>]]}

```



In [40]:

```
#PCA and Logistic Regression
```

```

steps = [('pca', PCA(n_components=19)), ('m', LogisticRegression())]
model = Pipeline(steps=steps)
model.fit(X_scaled, y_scaled)
y_pred = model.predict(X_test)
accuracy_score(y_test, y_pred, normalize=True, sample_weight=None)

```

/usr/local/lib/python3.7/dist-packages/sklearn/utils/validation.py:993: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().

y = column_or_1d(y, warn=True)

/usr/local/lib/python3.7/dist-packages/sklearn/linear_model/_logistic.py:818: ConvergenceWarning: lbfgs failed to converge (status=1): STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:

<https://scikit-learn.org/stable/modules/preprocessing.html>

Please also refer to the documentation for alternative solver options:

https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

```
extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG,
```

Out[40]:

0.84108671380722

In [41]:

```
confusion_matrix(y_test, y_pred)
```

Out[41]:

```
array([[2062, 150],
       [ 277, 198]])
```

In [42]:

```
precision_score(y_test, y_pred)
```

Out[42]:

0.5689655172413793

In [43]:

```
recall_score(y_test, y_pred)
```

Out[43]:

0.4168421052631579

In [44]:

```
f1_score(y_test, y_pred)
```

Out[44]:

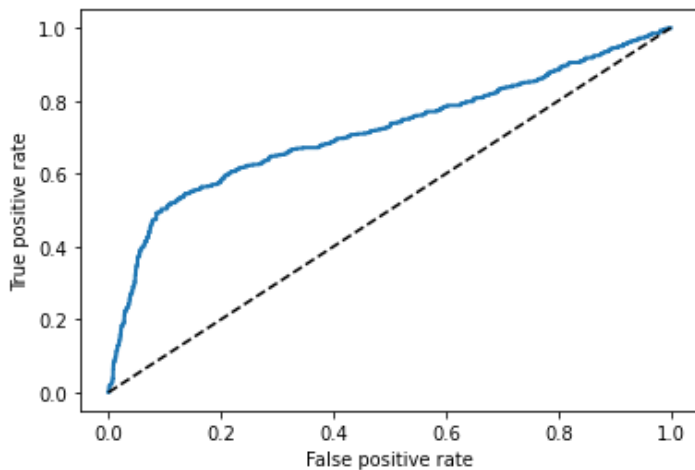
0.48116646415552855

In [45]:

```
y_pred = model.decision_function(X_test)
fpr, tpr, thresholds = roc_curve(y_test, y_pred)
```

```
def plot_roc_curve(fpr, tpr, label=None):
    plt.plot(fpr, tpr, linewidth=2, label=label)
    plt.plot([0, 1], [0, 1], 'k--')
```

```
plot_roc_curve(fpr, tpr)
plt.xlabel("False positive rate")
plt.ylabel("True positive rate")
plt.show()
```



In [46]:

```
#Show ROC Score
roc_auc_score(y_test, y_pred)
```

Out[46]:

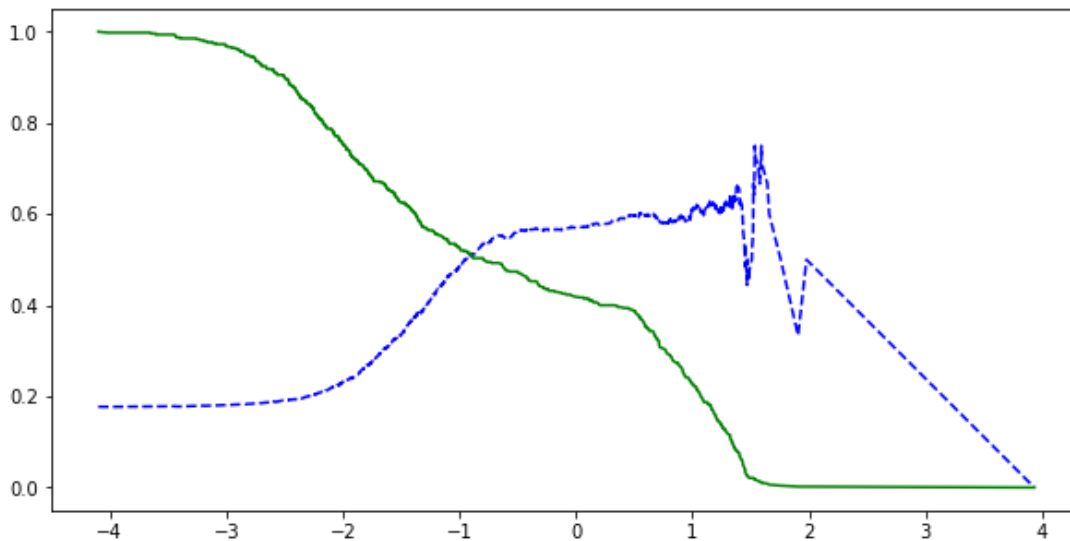
0.7164047777672028

In [47]:

```
#Precision Recall Curve Plot
precisions, recalls, thresholds = precision_recall_curve(y_test, y_pred)
```

```
def plot_precision_recall_vs_threshold(precisions, recalls, thresholds):
    plt.plot(thresholds, precisions[:-1], "b--", label="Precision")
    plt.plot(thresholds, recalls[:-1], "g-", label="Recall")

plt.figure(figsize=(10,5))
plot_precision_recall_vs_threshold(precisions, recalls, thresholds)
plt.show()
```



5. Softmaxt regression:

1. How softmax regression is related to logistic regression? What library can you use for softmax regression?

Softmax Regression is a generalization of Logistic Regression and uses k dimensionality to summarize a vector of arbitrary values bounded in a binary range.

We can use TensorFlow for softmax regression

6. KNN (Always use rebalanced training set for training, if it is not specified which training set to use)

1. Use sklearn's KNN classifier to train (with k= 10) and predict the model based on the unbalanced training set (the training set before rebalancing) and test it and show the confusion matrix and classification report
2. Use sklearn's KNN classifier to train (with k= 10) and predict the model based on the rebalanced training set and test it and show the confusion matrix and classification report
3. Use grid search to tune the following hyperparameters of KNN: number of neighbors (between 1 and 20), weights (uniform or distance), and metrics (Euclidean, Manhattan, or Minkowski)istance) to use for KNN. While creating an instance of GridSearchCV, use multiple evaluation metrics such as AUC and accuracy based on the example available at [Link to Sklearn](#) Link to sklearn (Links to an external site.). Also some helpful links and codes: [GitHub Example](#) and [Youtube Video](#)
4. The above grid search process can take a couple of minutes. After completing the process, print the best_params_
5. Based on the result from grid search, use the parameters to train a model, test it with test set, and then print the confusion matrix and classification report. Also, show the AUC of ROC.
6. Use PCA and based on that train model, test it and then print the confusion matrix and classification report. Also, show the AUC of ROC.
7. A short discussion on the 4 models and their differences.

In [48]:

```
#KNN based on unbalanced set
```

```

classifier = KNeighborsClassifier(n_neighbors=10)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
print(y_pred)

```

```

/usr/local/lib/python3.7/dist-packages/sklearn/neighbors/_classification.py:198: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().
    return self._fit(X, y)

```

```
[0. 0. 0. ... 0. 0. 0.]
```

In [49]:

```

print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))

```

```

[[2197   15]
 [ 466    9]]

```

	precision	recall	f1-score	support
0.0	0.83	0.99	0.90	2212
1.0	0.38	0.02	0.04	475
accuracy			0.82	2687
macro avg	0.60	0.51	0.47	2687
weighted avg	0.75	0.82	0.75	2687

In [50]:

```
# KNN based on balanced set
```

```

classifier = KNeighborsClassifier(n_neighbors=10)
classifier.fit(X_scaled, y_scaled)
y_pred = classifier.predict(X_test)
print(y_pred)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))

```

```

/usr/local/lib/python3.7/dist-packages/sklearn/neighbors/_classification.py:198: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().
    return self._fit(X, y)

```

```

[0. 0. 0. ... 1. 0. 0.]
[[1455  757]
 [ 253  222]]

```

	precision	recall	f1-score	support
0.0	0.85	0.66	0.74	2212
1.0	0.23	0.47	0.31	475
accuracy			0.62	2687
macro avg	0.54	0.56	0.52	2687
weighted avg	0.74	0.62	0.67	2687

In [51]:

```
#Using Grid Search
```

```

knn_params = {
    "n_neighbors": range(1, 10),
    "weights": ["uniform", "distance"],
    "metric": ["euclidean", "manhattan", "minkowski"]
}
scoring = {"AUC": "roc_auc", "Accuracy": make_scorer(accuracy_score)}
knn = KNeighborsClassifier()
grid_search = GridSearchCV(estimator=knn, param_grid=knn_params, n_jobs=-1, cv=3, scoring=scoring, refit="AUC", return_train_score=True)

```

```

grid_results = grid_search.fit(X_scaled, y_scaled)
final_model = knn.set_params(**grid_results.best_params_)
final_model.fit(X_scaled, y_scaled)
y_pred = final_model.predict(X_test)
print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(grid_results.best_params_)

```

```

/usr/local/lib/python3.7/dist-packages/joblib/externals/loky/process_executor.py:705: UserWarning: A worker stopped while some jobs were given to the executor. This can be caused by a too short worker timeout or by a memory leak.

```

```

"timeout or by a memory leak.", UserWarning
/usr/local/lib/python3.7/dist-packages/sklearn/neighbors/_classification.py:198: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().
return self._fit(X, y)
/usr/local/lib/python3.7/dist-packages/sklearn/neighbors/_classification.py:198: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().
return self._fit(X, y)

```

	precision	recall	f1-score	support
0.0	0.86	0.70	0.77	2212
1.0	0.26	0.49	0.34	475
accuracy			0.66	2687
macro avg	0.56	0.59	0.55	2687
weighted avg	0.76	0.66	0.69	2687

```

[[1541  671]
 [ 242  233]]
{'metric': 'manhattan', 'n_neighbors': 7, 'weights': 'distance'}

```

In [52]:

```

#After Grid Search
print(grid_results.best_params_)

```

```

{'metric': 'manhattan', 'n_neighbors': 7, 'weights': 'distance'}

```

In [53]:

```

#Train the model after grid search

classifier = KNeighborsClassifier(n_neighbors=9, weights='uniform', metric='manhattan')
classifier.fit(X_scaled, y_scaled)
y_pred = classifier.predict(X_test)
print(y_pred)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))

```

```

/usr/local/lib/python3.7/dist-packages/sklearn/neighbors/_classification.py:198: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().
return self._fit(X, y)

```

```

[1. 0. 0. ... 1. 1. 0.]
[[1481  731]
 [ 220  255]]

```

	precision	recall	f1-score	support
0.0	0.87	0.67	0.76	2212
1.0	0.26	0.54	0.35	475
accuracy			0.65	2687
macro avg	0.56	0.60	0.55	2687
weighted avg	0.76	0.65	0.68	2687

In [54]:

```

#Validation ROC curve

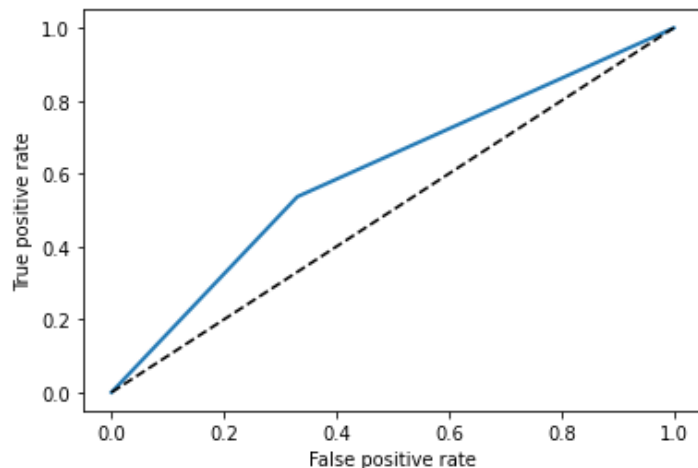
```

```
#plotting ROC curve
```

```
fpr, tpr, threshold = roc_curve(y_test, y_pred)
```

```
def plot_roc_curve(fpr, tpr, label=None):  
    plt.plot(fpr, tpr, linewidth=2, label=label)  
    plt.plot([0, 1], [0, 1], 'k--')
```

```
plot_roc_curve(fpr, tpr)  
plt.xlabel("False positive rate")  
plt.ylabel("True positive rate")  
plt.show()
```



```
In [55]:
```

```
roc_auc_score(y_test, y_pred)
```

```
Out[55]:
```

```
0.603185971257257
```

```
In [56]:
```

```
#using PCA to train model
```

```
pca = KNeighborsClassifier()  
pca.fit(X_scaled, y_scaled)  
y_pred = pca.predict(X_test)  
print(y_pred)  
print(confusion_matrix(y_test, y_pred))  
print(classification_report(y_test, y_pred))
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/neighbors/_classification.py:198: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().
```

```
    return self._fit(X, y)
```

```
[0. 0. 0. ... 1. 1. 0.]
```

```
[[1452  760]
```

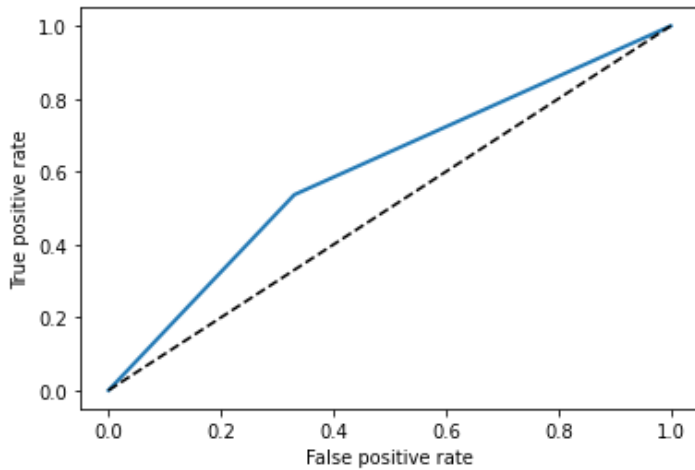
```
 [ 258  217]]
```

	precision	recall	f1-score	support
0.0	0.85	0.66	0.74	2212
1.0	0.22	0.46	0.30	475
accuracy			0.62	2687
macro avg	0.54	0.56	0.52	2687
weighted avg	0.74	0.62	0.66	2687

```
In [57]:
```

```
def plot_roc_curve(fpr, tpr, label=None):  
    plt.plot(fpr, tpr, linewidth=2, label=label)  
    plt.plot([0, 1], [0, 1], 'k--')
```

```
plot_roc_curve(fpr, tpr)
plt.xlabel("False positive rate")
plt.ylabel("True positive rate")
plt.show()
```



In [58]:

```
roc_auc_score(y_test, y_pred)
```

Out[58]:

0.5566308175502047

7. Naive Bayes

In [59]:

```
#Training a model with GaussianNB
```

```
gnb = GaussianNB()
gnb.fit(X_scaled, y_scaled)
y_pred= gnb.predict(X_test)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
[[1151 1061]
 [ 155  320]]
```

	precision	recall	f1-score	support
0.0	0.88	0.52	0.65	2212
1.0	0.23	0.67	0.34	475
accuracy			0.55	2687
macro avg	0.56	0.60	0.50	2687
weighted avg	0.77	0.55	0.60	2687

```
/usr/local/lib/python3.7/dist-packages/sklearn/utils/validation.py:993: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
```

In [60]:

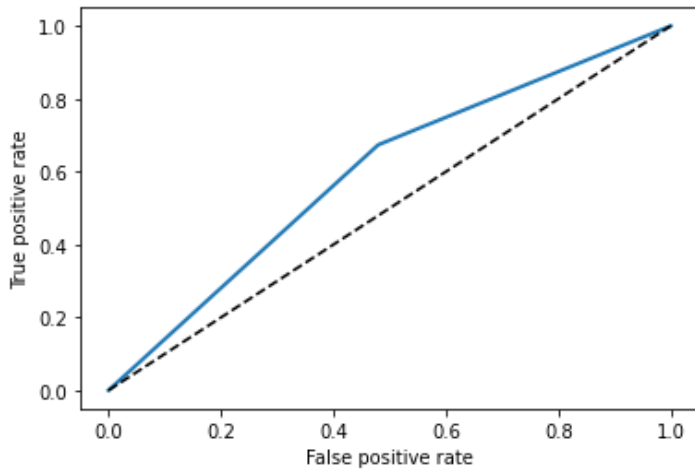
```
#Plotting and printing ROC Curve
```

```
y_pred = gnb.predict(X_test)
fpr, tpr, thresholds = roc_curve(y_test, y_pred)
```

```
def plot_roc_curve(fpr, tpr, label=None):
    plt.plot(fpr, tpr, linewidth=2, label=label)
    plt.plot([0, 1], [0, 1], 'k--')
```

```
plot_roc_curve(fpr, tpr)
```

```
plt.xlabel("False positive rate")
plt.ylabel("True positive rate")
plt.show()
```



In [61]:

```
roc_auc_score(y_test, y_pred)
```

Out[61]:

0.5970138954982392

In [62]:

```
# counting misclassification
y_pred = gnb.predict(X_test)
y_new=np.ravel(y_test)
p = (y_new)!= (y_pred)
print(p.sum())
```

1216

In [63]:

```
#Train with CategoricalNB

cnb = CategoricalNB()
cnb.fit(X_scaled, y_scaled)
y_pred= cnb.predict(X_test)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
[[1968  244]
 [ 254  221]]
```

	precision	recall	f1-score	support
0.0	0.89	0.89	0.89	2212
1.0	0.48	0.47	0.47	475
accuracy			0.81	2687
macro avg	0.68	0.68	0.68	2687
weighted avg	0.81	0.81	0.81	2687

/usr/local/lib/python3.7/dist-packages/sklearn/utils/validation.py:993: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().

```
y = column_or_1d(y, warn=True)
```

In [64]:

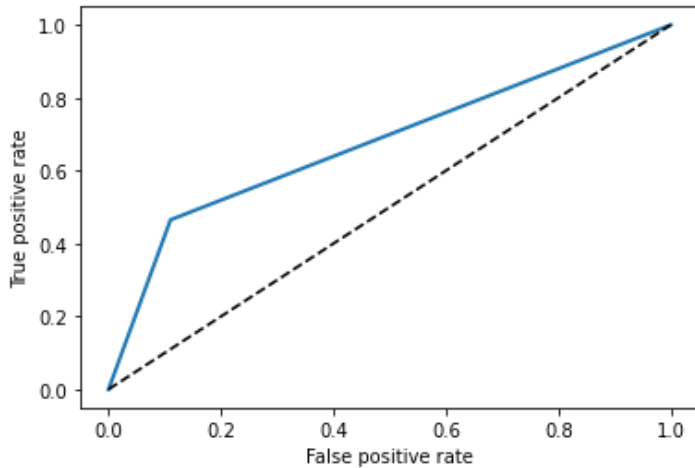
```
y_pred = cnb.predict(X_test)
fpr, tpr, thresholds = roc_curve(y_test, y_pred)

def plot_roccurve(fpr, tpr, label=None):
    plt.plot(fpr, tpr, linewidth=2, label=label)
```



```
plt.plot([0, 1], [0, 1], 'k--')
```

```
plot_roc_curve(fpr, tpr)
plt.xlabel("False positive rate")
plt.ylabel("True positive rate")
plt.show()
```



In [65]:

```
roc_auc_score(y_test, y_pred)
```

Out[65]:

0.6774778718949271

In [66]:

```
# counting misclassification
y_pred = gnb.predict(X_test)
y_new=np.ravel(y_test)
p = (y_new)!= (y_pred)
print(p.sum())
```

1216

8. Support Vector Machine

In [69]:

```
#Building support vector machine using SVC and Grid Search
```

```
model = SVC(kernel = 'sigmoid')
param_grid = {'C': [0.1,1, 10, 100]}
grid = GridSearchCV(SVC(probability=True, random_state=1), param_grid, n_jobs=-1, refit=
True, verbose=1)
grid.fit(X_scaled,y_scaled)
```

Fitting 5 folds for each of 4 candidates, totalling 20 fits

```
/usr/local/lib/python3.7/dist-packages/sklearn/utils/validation.py:993: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
y = column_or_1d(y, warn=True)
```

Out[69]:

```
GridSearchCV(estimator=SVC(probability=True, random_state=1), n_jobs=-1,
              param_grid={'C': [0.1, 1, 10, 100]}, verbose=1)
```

In [70]:

```
,grid.best_params_
```

Out[70]:

```
{'C': 100}
```

```
{'C': 100}
```

```
In [71]:
```

```
#Testing the model
```

```
model = SVC(kernel = 'rbf', C=100)
model.fit(X_scaled, y_scaled)
y_pred= model.predict(X_test)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/utils/validation.py:993: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
```

```
[[2080  132]
 [ 277  198]]
```

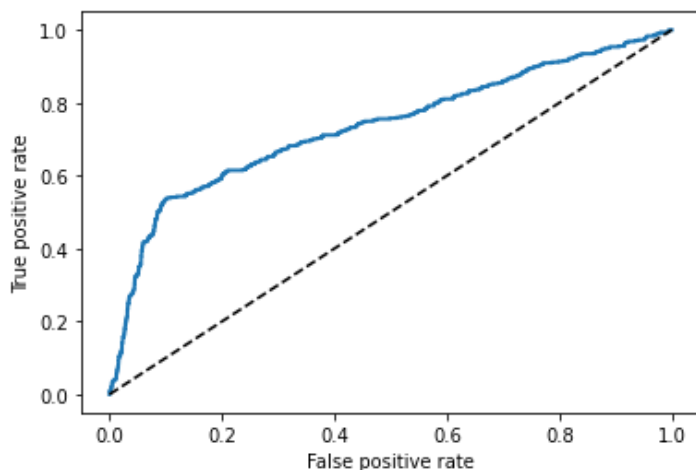
	precision	recall	f1-score	support
0.0	0.88	0.94	0.91	2212
1.0	0.60	0.42	0.49	475
accuracy			0.85	2687
macro avg	0.74	0.68	0.70	2687
weighted avg	0.83	0.85	0.84	2687

```
In [72]:
```

```
y_pred = model.decision_function(X_test)
fpr, tpr, thresholds = roc_curve(y_test, y_pred)
```

```
def plot_roc_curve(fpr, tpr, label=None):
    plt.plot(fpr, tpr, linewidth=2, label=label)
    plt.plot([0, 1], [0, 1], 'k--')
```

```
plot_roc_curve(fpr, tpr)
plt.xlabel("False positive rate")
plt.ylabel("True positive rate")
plt.show()
```



```
In [73]:
```

```
roc_auc_score(y_test, y_pred)
```

```
Out[73]:
```

```
0.7344670219853431
```

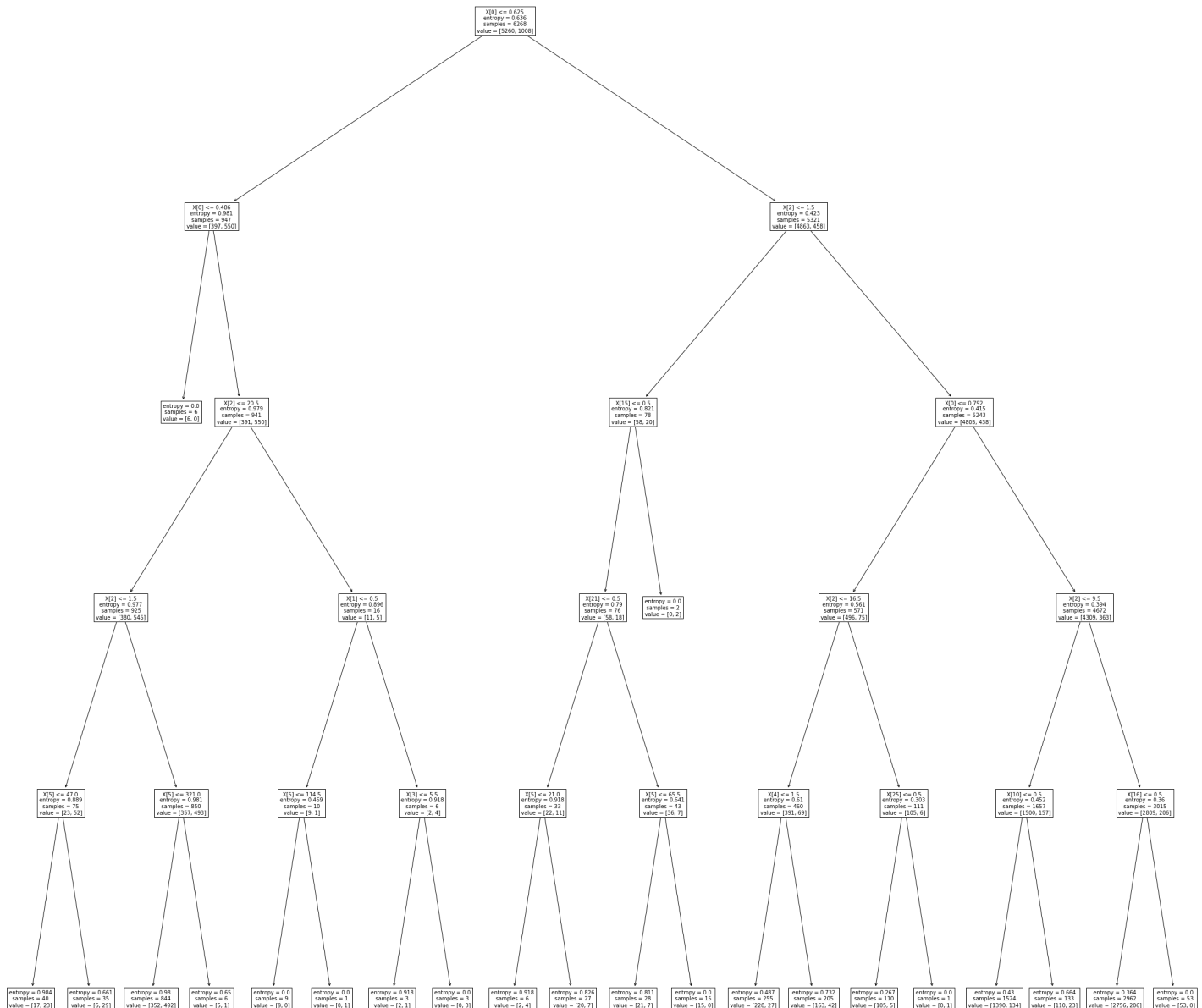
9. Decision Tree

1. Build a decision tree model using sklearn's `DecisionTreeClassifier`. Use the unbalanced training set, entropy as the criterion. Try with different `max_depth` (or use grid search). After building model, test it and print the confusion matrix and classification report. Also, plot ROC curve and show the AUC of ROC, and the count of the number of misclassification. Show the decision tree. (you can simply import tree from sklearn and call `tree.plot_tree` with your model and the call `plt.show`. At the beginning of this process, use `plt.figure` to change the figsize)
2. Perform the same tasks as 9.1 with the balanced training set
3. Discuss any difference and also discuss part of the tree of 9.2

In [74]:

```
#Decision tree model with DecisionTreeClassifier
from sklearn.tree import DecisionTreeClassifier

tree_clf = DecisionTreeClassifier(max_depth=5, min_samples_leaf=1, criterion="entropy")
tree_clf.fit(X_train, y_train)
plt.figure(figsize = (40,40))
tree.plot_tree(tree_clf)
plt.show()
```



In [75]:

```
y_pred= tree_clf.predict(X_test)
```

```
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
[[2045  167]
 [ 228  247]]
```

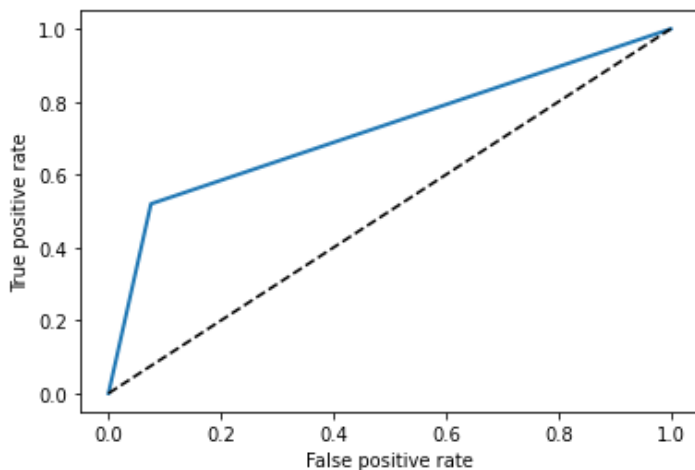
	precision	recall	f1-score	support
0.0	0.90	0.92	0.91	2212
1.0	0.60	0.52	0.56	475
accuracy			0.85	2687
macro avg	0.75	0.72	0.73	2687
weighted avg	0.85	0.85	0.85	2687

In [76]:

```
y_pred = tree_clf.predict(X_test)
fpr, tpr, thresholds = roc_curve(y_test, y_pred)
```

```
def plot_roc_curve(fpr, tpr, label=None):
    plt.plot(fpr, tpr, linewidth=2, label=label)
    plt.plot([0, 1], [0, 1], 'k--')
```

```
plot_roc_curve(fpr, tpr)
plt.xlabel("False positive rate")
plt.ylabel("True positive rate")
plt.show()
```



In [77]:

```
roc_auc_score(y_test, y_pred)
y_new=np.ravel(y_test)
p = (y_new) != (y_pred)
print(p.sum())
```

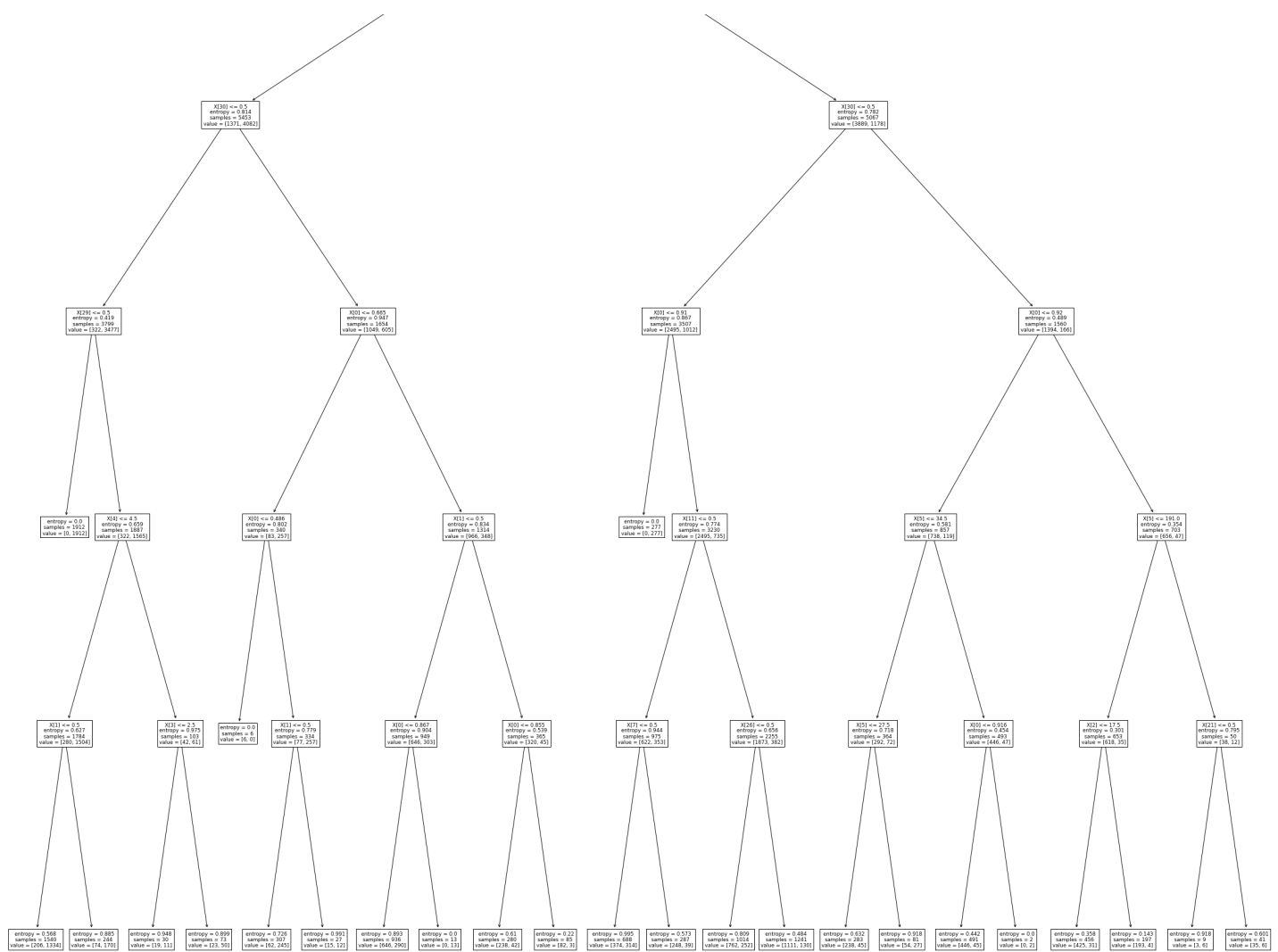
395

In [78]:

#Now with balanced dataset

```
tree_clf = DecisionTreeClassifier(max_depth=5, min_samples_leaf=1, criterion='entropy')
tree_clf.fit(X_scaled, y_scaled)
plt.figure(figsize = (40,40))
tree.plot_tree(tree_clf)
plt.show()
```

100% == 0.978
entropy = 1.0
samples = 1000
value = [1280, 520]



In [79]:

```
y_pred= tree_clf.predict(X_test)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
[[2054  158]
 [ 236  239]]

              precision    recall  f1-score   support

    0.0               0.90       0.93       0.91       2212
    1.0               0.60       0.50       0.55        475

 accuracy               0.85       2687
 macro avg              0.75       0.72       0.73       2687
weighted avg              0.84       0.85       0.85       2687
```

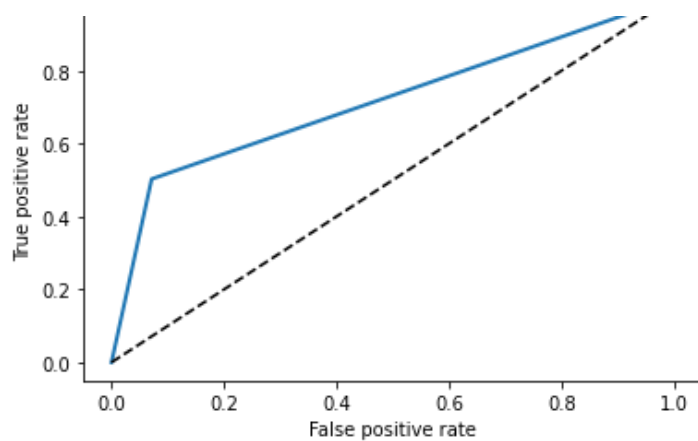
In [80]:

```
y_pred = tree_clf.predict(X_test)
fpr, tpr, thresholds = roc_curve(y_test, y_pred)

def plot_roc_curve(fpr, tpr, label=None):
    plt.plot(fpr, tpr, linewidth=2, label=label)
    plt.plot([0, 1], [0, 1], 'k--')

plot_roc_curve(fpr, tpr)
plt.xlabel("False positive rate")
plt.ylabel("True positive rate")
plt.show()
```





In [81]:

```
roc_auc_score(y_test, y_pred)
```

Out[81]:

0.7158646616541354

In [82]:

```
y_new=np.ravel(y_test)
p = (y_new) != (y_pred)
print(p.sum())
```

394

Discuss any difference and also discuss part of the tree of 9.2

AUC is different between trees used. Also the balanced dataset seems to make the tree look more even.

10. Random Forest

In [92]:

```
# Grid Search with RandomForestClassifier
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(random_state=42, n_jobs=-1)
param = {
    'max_depth': [2, 3, 5, 10, 20],
    'min_samples_leaf': [5, 10, 20, 50, 100, 200],
    'n_estimators': [10, 25, 30, 50, 100, 200]
}

gridSearch = GridSearchCV(estimator=rf, param_grid=param)
gridSearch.fit(X_scaled, y_scaled)
gridSearch.best_estimator_
```

/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().

```
estimator.fit(X_train, y_train, **fit_params)
```

/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().

```
estimator.fit(X_train, y_train, **fit_params)
```

/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().

```
estimator.fit(X_train, y_train, **fit_params)
```

/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
the shape of y to (n_samples,), for example using ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using ravel().  
    estimator.fit(X_train, y_train, **fit_params)
```


[illegible]

[illegible]

[illegible]


```
the shape of y to (n_samples,), for example using.ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using.ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using.ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using.ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using.ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using.ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using.ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using.ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using.ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using.ravel().  
    estimator.fit(X_train, y_train, **fit_params)  
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_validation.py:680: DataCo  
nversionWarning: A column-vector y was passed when a 1d array was expected. Please change  
the shape of y to (n_samples,), for example using.ravel().  
    estimator.fit(X_train, y_train, **fit_params)
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
rf.fit(X_scaled, y_scaled)
y_pred= rf.predict(X_test)
```

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().

In [97]:

```
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
[[2055  157]
 [ 235  240]]
```

	precision	recall	f1-score	support
0.0	0.90	0.93	0.91	2212
1.0	0.60	0.51	0.55	475
accuracy			0.85	2687
macro avg	0.75	0.72	0.73	2687
weighted avg	0.85	0.85	0.85	2687

In [100]:

```
roc_auc_score(y_test, y_pred)
```

Out[100]:

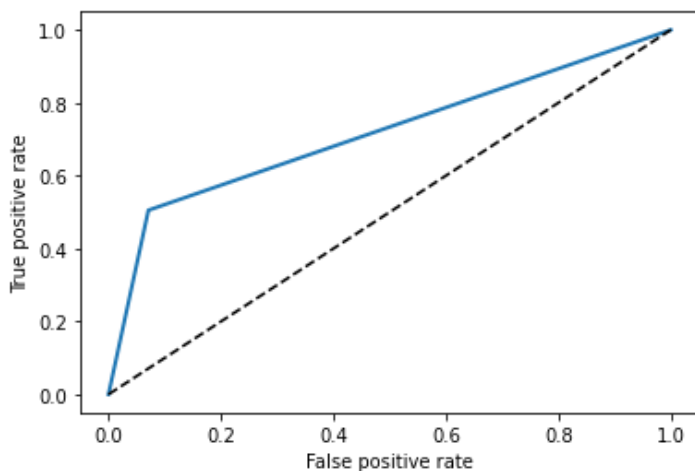
0.7171433330160845

In [102]:

```
y_pred = rf.predict(X_test)
fpr, tpr, thresholds = roc_curve(y_test, y_pred)
```

```
def plot_roc_curve(fpr, tpr, label=None):
    plt.plot(fpr, tpr, linewidth=2, label=label)
    plt.plot([0, 1], [0, 1], 'k--')
```

```
plot_roc_curve(fpr, tpr)
plt.xlabel("False positive rate")
plt.ylabel("True positive rate")
plt.show()
```



11. Boosting Algorithms

In [104]:

```
#Train an AdaBoostClassifier model with some manual/grid search-based parameters and then
test it and then print the confusion matrix and classification report.
#Also, plot ROC curve and show the AUC of ROC, and the count of the number of misclassifi
```

```

cation.
from sklearn.ensemble import AdaBoostClassifier

ada = AdaBoostClassifier(random_state=42)
param = {
    'learning_rate': [.01, 0.2, 0.5, 0.7, 0.9, 1.0, 1.5],
    'n_estimators': [10, 25, 30, 50, 100, 200]
}

gSearch = GridSearchCV(estimator=ada, param_grid=param, cv=3, n_jobs=-1, verbose=1, scoring='accuracy')
gSearch.fit(X_scaled, y_scaled)
grid_search.best_estimator_

```

Fitting 3 folds for each of 42 candidates, totalling 126 fits

```

/usr/local/lib/python3.7/dist-packages/sklearn/utils/validation.py:993: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)

```

Out[104]:

```
KNeighborsClassifier(metric='manhattan', n_neighbors=7, weights='distance')
```

In [108]:

```

ada = AdaBoostClassifier(random_state=42, learning_rate=0.5, n_estimators=200)
ada.fit(X_scaled, y_scaled)
y_pred= ada.predict(X_test)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))

```

```

/usr/local/lib/python3.7/dist-packages/sklearn/utils/validation.py:993: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)

```

```

[[2053  159]
 [ 228  247]]

```

	precision	recall	f1-score	support
0.0	0.90	0.93	0.91	2212
1.0	0.61	0.52	0.56	475
accuracy			0.86	2687
macro avg	0.75	0.72	0.74	2687
weighted avg	0.85	0.86	0.85	2687

In [110]:

```
roc_auc_score(y_test, y_pred)
```

Out[110]:

```
0.7240596745027125
```

In [111]:

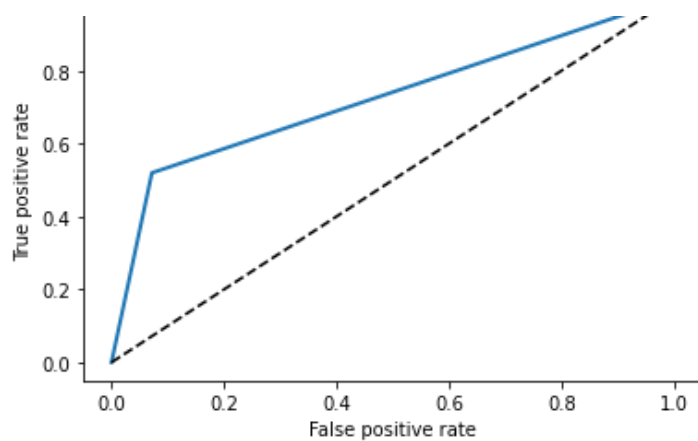
```

y_pred = ada.predict(X_test)
fpr, tpr, thresholds = roc_curve(y_test, y_pred)

def plot_roc_curve(fpr, tpr, label=None):
    plt.plot(fpr, tpr, linewidth=2, label=label)
    plt.plot([0, 1], [0, 1], 'k--')

plot_roc_curve(fpr, tpr)
plt.xlabel("False positive rate")
plt.ylabel("True positive rate")
plt.show()

```



In [113]:

```
y_new = np.ravel(y_test)
p = (y_new) != (y_pred)
print(p.sum())
```

387

In [116]:

```
#Gradient BoostingClassifier
from sklearn.ensemble import GradientBoostingClassifier

grad = GradientBoostingClassifier(random_state=42)
param = {
    'learning_rate': [0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2],
    'n_estimators' : [10, 25, 30, 50, 100, 200],
    'max_depth'    : [3, 5, 7]
}

gSearch = GridSearchCV(estimator = grad, cv = 3, param_grid = param, n_jobs=-1, scoring=
'accuracy')
gSearch.fit(X_scaled, y_scaled)
gSearch.best_estimator_
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/ensemble/_gb.py:494: DataConversionWarning
: A column-vector y was passed when a 1d array was expected. Please change the shape of y
to (n_samples, ), for example using ravel().
y = column_or_1d(y, warn=True)
```

Out[116]:

GradientBoostingClassifier(learning_rate=0.075, max_depth=7, random_state=42)

In [117]:

```
roc_auc_score(y_test, y_pred)
```

Out[117]:

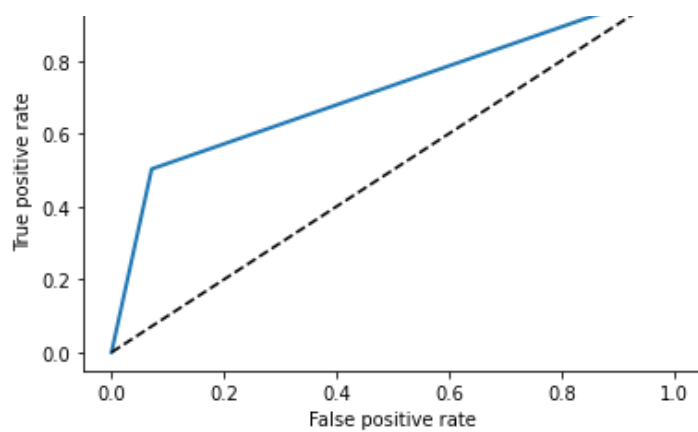
0.7240596745027125

In [118]:

```
y_pred = tree_clf.predict(X_test)
fpr, tpr, thresholds = roc_curve(y_test, y_pred)

def plot_roc_curve(fpr, tpr, label=None):
    plt.plot(fpr, tpr, linewidth=2, label=label)
    plt.plot([0, 1], [0, 1], 'k--')

plot_roc_curve(fpr, tpr)
plt.xlabel("False positive rate")
plt.ylabel("True positive rate")
plt.show()
```



In [119]:

```
y_new = np.ravel(y_test)
p = (y_new) != (y_pred)
print(p.sum())
```

394

12. Final Discussion

Finally, briefly discuss your finding such as which model could be most suitable for this given scenario and what could be your future work based on this experiment.

A Decision Tree could be pretty useful in both visualizing and determining a set of variables which could lead to scenarios where an employee might look for a new job.