NAME: ARUL KUMAR ARK

ROLL NO : 225229103

Lab 7: Link Prediction of future connections in Facebook

In [1]:
```
import pandas as pd
import numpy as np
import random
import networkx as nx
from tqdm import tqdm
import re
import matplotlib.pyplot as plt

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, roc_auc_score
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
```

In [2]:
```
with open("fb-pages-food.nodes", "r+", encoding="utf-8") as f:
    fb_nodes = f.read().splitlines()


with open("fb-pages-food.edges", "r+", encoding="utf-8") as f:
    fb_links = f.read().splitlines()

len(fb_nodes), len(fb_links)
```

Out[2]: (621, 2102)

In [3]:
```
node_list_1 = []
node_list_2 = []

for i in tqdm(fb_links):
    node_list_1.append(i.split(',')[0])
    node_list_2.append(i.split(',')[1])

fb_df = pd.DataFrame({'node_1': node_list_1, 'node_2': node_list_2})
```

```
100%|██████████████████████████████████████████████████████████████████████████████|
2102/2102 [00:00<00:00, 2102653.71it/s]
```

In [4]:
```
fb_df.head()
```

Out[4]:

	node_1	node_2
0	0	276
1	0	58
2	0	132
3	0	603
4	0	398

In [5]:
```
G = nx.from_pandas_edgelist(fb_df, "node_1", "node_2", create_using=nx.Graph())


plt.figure(figsize=(10,10))

pos = nx.random_layout(G, seed=23)
nx.draw(G, with_labels=False,  pos = pos, node_size = 40, alpha = 0.6, width = 0.

plt.show()
```

In [6]:
```
node_list = node_list_1 + node_list_2


node_list = list(dict.fromkeys(node_list))


adj_G = nx.to_numpy_matrix(G, nodelist = node_list)
```

In [7]:
```
adj_G.shape
```

Out[7]:
```
(620, 620)
```

In [8]:
```
all_unconnected_pairs = []


offset = 0
for i in tqdm(range(adj_G.shape[0])):
    for j in range(offset,adj_G.shape[1]):
        if i != j:
            if nx.shortest_path_length(G, str(i), str(j)) <=2:
                if adj_G[i,j] == 0:
                    all_unconnected_pairs.append([node_list[i],node_list[j]])
    offset = offset + 1
```

```
100%|████████████████████████████████████████████████████████████████| 620/620 [00:04<00:00, 136.84it/s]
```

In [9]:
```
len(all_unconnected_pairs)
```

Out[9]:
```
19018
```

In [10]:
```
node_1_unlinked = [i[0] for i in all_unconnected_pairs]
node_2_unlinked = [i[1] for i in all_unconnected_pairs]

data = pd.DataFrame({'node_1':node_1_unlinked,
                     'node_2':node_2_unlinked})


data['link'] = 0
```

In [11]:
```
initial_node_count = len(G.nodes)

fb_df_temp = fb_df.copy()


omissible_links_index = []

for i in tqdm(fb_df.index.values):


    G_temp = nx.from_pandas_edgelist(fb_df_temp.drop(index = i), "node_1", "node_


    if (nx.number_connected_components(G_temp) == 1) and (len(G_temp.nodes) == in
        omissible_links_index.append(i)
        fb_df_temp = fb_df_temp.drop(index = i)
```

```
100%|████████████████████████████████████████████████████████████| 2102/2102 [00:03<00:00, 559.95it/s]
```

In [12]:
```
len(omissible_links_index)
```

Out[12]:
```
1483
```

In [13]:
```
fb_df_ghost = fb_df.loc[omissible_links_index]


fb_df_ghost['link'] = 1

data = data.append(fb_df_ghost[['node_1', 'node_2', 'link']], ignore_index=True)
```

```
C:\Users\online.CSCENTER\AppData\Local\Temp\ipykernel_12024\3242015871.py:7: Fu
tureWarning: The frame.append method is deprecated and will be removed from pan
das in a future version. Use pandas.concat instead.
  data = data.append(fb_df_ghost[['node_1', 'node_2', 'link']], ignore_index=Tr
ue)
```

In [14]:
```
data['link'].value_counts()
```

Out[14]:
```
0    19018
1     1483
Name: link, dtype: int64
```

In [15]:
```
fb_df_partial = fb_df.drop(index=fb_df_ghost.index.values)


G_data = nx.from_pandas_edgelist(fb_df_partial, "node_1", "node_2", create_using=
```

In [19]:
```
from node2vec import Node2Vec


node2vec = Node2Vec(G_data, dimensions=100, walk_length=16, num_walks=50)


n2w_model = node2vec.fit(window=7, min_count=1)
```

```
Computing transition probabilities:   0%|          | 0/620 [00:00<?, ?it/s]

Generating walks (CPU: 1): 100%|██████████| 50/50 [00:00<00:00, 50.75it/s]
```

In [21]:
```
x = [(n2w_model.wv[str(i)]+n2w_model.wv[str(j)]) for i,j in zip(data['node_1'], d
```

In [22]:
```
xtrain, xtest, ytrain, ytest = train_test_split(np.array(x), data['link'],
                                                test_size = 0.3,
                                                random_state = 35)
```

In [23]:
```
lr = LogisticRegression(class_weight="balanced")

lr.fit(xtrain, ytrain)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:81
4: 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 (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-regressi
on (https://scikit-learn.org/stable/modules/linear_model.html#logistic-regressi
on)
  n_iter_i = _check_optimize_result(
```

Out[23]:
```
LogisticRegression(class_weight='balanced')
```

In [24]:
```
predictions = lr.predict_proba(xtest)
```

In [25]:
```
roc_auc_score(ytest, predictions[:,1])
```

Out[25]:
```
0.7900162913034287
```

In [27]:
```
!pip install lightgbm
```

```
Defaulting to user installation because normal site-packages is not writeable
Collecting lightgbm
  Downloading lightgbm-3.3.2-py3-none-win_amd64.whl (1.0 MB)
Requirement already satisfied: scipy in c:\users\online.cscenter\appdata\roamin
g\python\python39\site-packages (from lightgbm) (1.9.0)
Requirement already satisfied: numpy in c:\programdata\anaconda3\lib\site-packa
ges (from lightgbm) (1.21.5)
Requirement already satisfied: wheel in c:\programdata\anaconda3\lib\site-packa
ges (from lightgbm) (0.37.1)
Requirement already satisfied: scikit-learn!=0.22.0 in c:\programdata\anaconda3
\lib\site-packages (from lightgbm) (1.0.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\programdata\anaconda3
\lib\site-packages (from scikit-learn!=0.22.0->lightgbm) (2.2.0)
Requirement already satisfied: joblib>=0.11 in c:\programdata\anaconda3\lib\sit
e-packages (from scikit-learn!=0.22.0->lightgbm) (1.1.0)
Installing collected packages: lightgbm
Successfully installed lightgbm-3.3.2
```

In [28]:
```
import lightgbm as lgbm

train_data = lgbm.Dataset(xtrain, ytrain)
test_data = lgbm.Dataset(xtest, ytest)


parameters = {
    'objective': 'binary',
    'metric': 'auc',
    'is_unbalance': 'true',
    'feature_fraction': 0.5,
    'bagging_fraction': 0.5,
    'bagging_freq': 20,
    'num_threads' : 2,
    'seed' : 76
}


model = lgbm.train(parameters,
                   train_data,
                   valid_sets=test_data,
                   num_boost_round=1000,
                   early_stopping_rounds=20)
```

```
C:\Users\online.CSCENTER\AppData\Roaming\Python\Python39\site-packages\lightg
bm\engine.py:181: UserWarning: 'early_stopping_rounds' argument is deprecated
and will be removed in a future release of LightGBM. Pass 'early_stopping()'
callback via 'callbacks' argument instead.
  _log_warning("'early_stopping_rounds' argument is deprecated and will be re
moved in a future release of LightGBM. "
```