SNA Stage 2: Link Prediction

Name: Udayraj Deshmukh
Roll: 150101021
Dataset File: "higgs-mention_network"

Results below are based on a subgraph of original dataset having ~10000 nodes and ~3000 edges (taken as undirected graph).

Part 1: Topological neighborhood methods:

- The individual AUC values along with overall(average) AUC are stored in the "topological_metrics.csv" file.

- The precision/recall values are also stored in the same file in the last 3 rows.

	Common Neighbour	Jaccard Coefficient	Adamic adar	Resource allocation	Preferential Attachment
AUC	0.348230715	0.3491474621	0.3476460323	0.347632517	0.06742654652
Precision, Recall at 25%	[0.500, 0.591]	[0.500, 0.591]	[0.500, 0.591]	[0.500, 0.591]	[0.973, 0.486]
Precision, Recall at 50%	[0.500, 0.591]	[0.500, 0.591]	[0.500, 0.591]	[0.500, 0.591]	[0.812, 0.844]
Precision, Recall at 75%	[0.500, 0.591]	[0.500, 0.591]	[0.500, 0.591]	[0.500, 0.591]	[0.500, 1.000]

Part 2: Machine learning based methods

- The scores in above five topological neighborhood metrics as features along with the labels are stored in the "dataset.csv" file.
- 5-fold-cross validation was used for the below classifiers.
- For each of the fold considered as "test" dataset, the evaluation using different methods is done and reported below.
- The fold for which best accuracy was noted is highlighted for each of the classifer.

a) SVM
Fold 1 of 5:
Accuracy: 0.8276157804459692

ROC AUC score: 0.83003108003108

Cross Validation Scores
[0.81934932 0.85702055 0.82418525 0.81303602 0.83190395]

Classification Report	precision	recall	f1-score	support
0	0.76	0.96	0.85	572
1	0.95	0.7	0.81	594
avg / total	0.85	0.83	0.83	1166

Confusion Matrix	
548	24
177	417

Fold 2 of 5:
Accuracy: 0.829477292202228

ROC AUC score: 0.8216085733991192

Cross Validation Scores
[0.81934932 0.85702055 0.82418525 0.81303602 0.83190395]

Classification Report	precision	recall	f1-score	support
0	0.78	0.95	0.86	618
1	0.93	0.69	0.79	549
avg / total	0.85	0.83	0.83	1167

Confusion Matrix	
590	28
171	378

Fold 3 of 5:
Accuracy: 0.8423307626392459

ROC AUC score: 0.845842151675485

Cross Validation Scores
[0.81934932 0.85702055 0.82418525 0.81303602 0.83190395]

Classification Report	precision	recall	f1-score	support
0	0.77	0.97	0.86	567
1	0.96	0.72	0.82	600
avg / total	0.87	0.84	0.84	1167

Confusion Matrix	
550	17
167	433

Fold 4 of 5

Accuracy: 0.8251928020565553

ROC AUC score: 0.8265989507757561

Cross Validation Scores
[0.81934932 0.85702055 0.82418525 0.81303602 0.83190395]

Classification Report	precision	recall	f1-score	support
0	0.75	0.98	0.85	578
1	0.97	0.68	0.8	589
avg / total	0.86	0.83	0.82	1167

Confusion Matrix	
564	14
190	399

Fold 5 of 5:
Accuracy: 0.8226221079691517

ROC AUC score: 0.8229932152612565

Cross Validation Scores
[0.81934932 0.85702055 0.82418525 0.81303602 0.83190395]

Classification Report	precision	recall	f1-score	support
0	0.75	0.97	0.84	582
1	0.95	0.68	0.79	585
avg / total	0.85	0.82	0.82	1167

Confusion Matrix	
563	19
188	397

b) Decision Tree

Fold 1 of 5:
Accuracy: 0.8464837049742711

ROC AUC score: 0.8472869722869724

Cross Validation Scores
[0.84931507 0.85873288 0.86277873 0.83190395 0.84305317]

Classification Report	precision	recall	f1-score	support
0	0.81	0.89	0.85	572
1	0.88	0.8	0.84	594
avg / total	0.85	0.85	0.85	1166

Confusion Matrix	
509	63
116	478

Fold 2 of 5:
Accuracy: 0.8449014567266495

ROC AUC score: 0.8415610023520258

Cross Validation Scores
[0.84931507 0.85873288 0.86277873 0.83190395 0.84305317]

Classification Report	precision	recall	f1-score	support
0	0.82	0.9	0.86	618
1	0.87	0.79	0.83	549
avg / total	0.85	0.84	0.84	1167

Confusion Matrix	
555	63
118	431

Fold 3 of 5:
Accuracy: 0.8594687232219366
ROC AUC score: 0.8610537918871253
Cross Validation Scores
[0.84931507 0.85873288 0.86277873 0.83190395 0.84391081]

Classification Report	precision	recall	f1-score	support
0	0.82	0.92	0.86	567
1	0.91	0.81	0.85	600
avg / total	0.87	0.86	0.86	1167

Confusion Matrix	
520	47
117	483

Fold 4 of 5:
Accuracy: 0.8551842330762639
ROC AUC score: 0.8555994853748949
Cross Validation Scores
[0.84931507 0.85958904 0.86277873 0.83190395 0.84391081]

Classification Report	precision	recall	f1-score	support
0	0.82	0.9	0.86	578
1	0.89	0.81	0.85	589
avg / total	0.86	0.86	0.85	1167

Confusion Matrix	
520	58
111	478

Fold 5 of 5:
Accuracy: 0.8560411311053985
ROC AUC score: 0.8561943783593268
Cross Validation Scores

[0.84931507 0.85873288 0.86277873 0.83190395 0.84305317]

Classification Report	precision	recall	f1-score	support
0	0.82	0.92	0.86	582
1	0.9	0.8	0.85	585
avg / total	0.86	0.86	0.86	1167

Confusion Matrix	
533	49
119	466

c) Naive Bayes (Gaussian) method

Fold 1 of 5:
Accuracy: 0.7264150943396226

ROC AUC score: 0.7312224812224812

Classification Report	precision	recall	f1-score	support
0	0.64	0.99	0.78	572
1	0.97	0.48	0.64	594
avg / total	0.81	0.73	0.71	1166

Confusion Matrix	
564	8
311	283

Fold 2 of 5
Accuracy: 0.7446443873179092

ROC AUC score: 0.7588996763754046

Classification Report	precision	recall	f1-score	support
0	1	0.52	0.68	618
1	0.65	1	0.79	549
avg / total	0.83	0.74	0.73	1167

Confusion Matrix	
320	298
0	549

Fold 3 of 5:
Accuracy: 0.844044558697515

ROC AUC score: 0.8474118165784833

Classification Report	precision	recall	f1-score	support
0	0.77	0.97	0.86	567
1	0.96	0.73	0.83	600
avg / total	0.87	0.84	0.84	1167

Confusion Matrix	
548	19
163	437

Fold 4 of 5
Accuracy: 0.7386461011139674

ROC AUC score: 0.7410542764993743

Classification Report	precision	recall	f1-score	support
0	0.66	1	0.79	578
1	0.99	0.49	0.65	589
avg / total	0.83	0.74	0.72	1167

Confusion Matrix	
576	2
303	286

Fold 5 of 5:
Accuracy: 0.8251928020565553

ROC AUC score: 0.8249757687902017

Classification Report	precision	recall	f1-score	support
0	0.89	0.74	0.81	582
1	0.78	0.91	0.84	585
avg / total	0.83	0.83	0.82	1167

Confusion Matrix	
431	151
53	532