Community Detection And Link Prediction In Social Networks

Bachelor of Engineering

in

Information Technology

by

Rahul Thorat	117A3060
Krishna Tiwari	218A3076
Abhilash Kanaujia	113A3034

Under the Guidance of:

Prof. Seema Redekar

Department of Information Technology

SIES Graduate School of Technology

2020-2021

CERTIFICATE

This is to certify that the project entitled **"Community Detection And Link Prediction In Social Networks"** is a bonafide work of the following students, submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **Bachelor of Engineering** in **Information Technology**.

Rahul Thorat	**117A3060**
Krishna Tiwari	**218A3076**
Abhilash Kanaujia	**113A3034**

Prof. Seema Redekar
Internal Guide

Dr. Atul N Kemkar
Principal

PROJECT REPORT APPROVAL

This project report entitled ***"Community Detection And Link Prediction In Social Networks"*** by following students is approved for the degree of ***Bachelor of Engineering*** in ***Information Technology***.

Rahul Thorat	**117A3060**
Krishna Tiwari	**218A3076**
Abhilash Kanaujia	**113A3034**

Name of External Examiner: _____

Signature: _____

Name of Internal Examiner: _____

Signature: _____

Date:

Place:

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have appropriately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Rahul Thorat	117A3060	_____
Krishna Tiwari	218A3076	_____
Abhilash Kanaujia	113A3034	_____
		Signature

Date:

ACKNOWLEDGEMENT

We wish to express our deep sense of gratitude and thank to our Internal Guide **Prof. Seema Redekar** for her guidance, help and useful suggestions, which helped in completing our project work in time. We would like to thank our HOD **Dr. K. Lakshmisudha** for her guidance. We also thank to our **Principal Dr.Atul Khemkar**, for extending his support to carry out this project.

Also we would like to thank the entire faculty of Information Technology department for their valuable ideas and timely assistance in this project, last but not the least, we would like to thank our teaching and non teaching staff members of our college for their support, in facilitating timely completion of this project.

Project Team
Rahul Thorat
Krishna Tiwari
Abhilash Kanaujia

ABSTRACT

Social Networks have been an important aspect in our lives since the beginning of time. These social networks include physical or digital networks formed over time due to certain similarities. Due to the nature of these networks being so large there can be many communities identified within these networks. Community detection is an important part in various fields such as marketing, recommendation systems, healthcare, detecting terrorist activities, Link prediction and many more. Community detection in social networks helps to understand the network structure and analyze the network properties. Link prediction in social networks is important with respect to identifying future links and connectivity in the network to predict the future of the network. It is used in spam detection, disease prediction, advertising, recommender systems and many more. One way to extract information from communities for the mentioned uses includes techniques like data extraction and data mining. Both these techniques require huge amounts of time and is difficult to perform on large datasets. To overcome this, the use of machine learning algorithms proves essential to saving time and accurately extracting important information from such large datasets. In this paper we aim to identify and predict social communities present in large networks and accurately identify the future links in these networks through various machine learning models and algorithms.

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

In this digital age, there is a huge amount of heterogeneous data available which can be put to good use if used carefully. In the analysis of social network data, recognizing groups of similar nodes is a difficult task. Using this data leads to enhance the quality of community discovery. The analysis of social networks, mainly based on graph theories and sociological analysis, aims to study different aspects of these networks. The main factors are network detection, identity of influential actors, and the observe and prediction of the evolution of networks.

Datasets

The most simple and customary sort of datasets are spreadsheet or CSV format. Therein one file is organized as tables of rows and columns. Datasets are the fastest and most efficient way to work with logically grouped data in your application. We have taken the datasets of social network pages which shows the graph relates the knowledge objects within the save to a series of nodes and edges, the sides representing the relationships between the nodes. The relationships permit information that's saved to couple together directly and, in many cases, retrieved with one operation. The results are obtained on the dataset Fb-pages-food network from the network repository website.

Community Detection Algorithms

Girvan-Newman algorithm:
In this algorithm the edges between nodes which have the highest betweenness centrality are removed consecutively until two or more communities are establish. In this paper, this algorithm rule is employed for network community detection.

Triangle counting:
In this algorithm, a triangle refers to a set of three nodes or vertices of the triangle and all of the nodes are connected to other nodes. This technique is useful to classify a node into three general communities and is often used for fraudulent website detection.

K-1 colouring algorithm:
This algorithm assigns a colour to every node in the graph while making sure that the colours used are as few as possible. This is a NP complete problem and that's why this is a greedy algorithm.

1.2 Need of Project

Community detection is an important aspect to help detect the structure of the network and identify nodes based on their similarities. With today's digital age it becomes an important aspect to detect communities which can be applied to various fields such as healthcare, marketing et cetera. Link prediction serves an important role to help detect the future of the community and the network. Data mining becomes a tedious and almost impossible task to help detect communities and the links between their nodes. This is where machine learning models and various community detection algorithms come into picture to save time and accurately predict the future of these networks. By Finding the Influential nodes based on various centrality models for communities. Accurately predict links between various food joints based on their labels. Should be able to able to accurately classify restaurants between drive thru (fast food) or dine in. Link prediction will help us to predict whether there will be links between two nodes based on the attribute information and the observed existing link information.

1.3 Existing System

The existing system propose a framework to predict the occurrence of different events and transition for communities in dynamic social networks. The framework incorporates key features related to a community – its structure, history, and influential members.

1.4 Proposed System

In addition to the existing system our system automatically detects the communities and classify them. We propose to accurately detect communities based on dining types: Fast food or Dine in restaurants, accurately depict the influential nodes based on various centralities and form together links which are required for marketing strategies. We use data from Facebook pages, to build a recommendation system. That provides personalized restaurant recommendations to users. Since different people have different food preferences and dietary restrictions. We wanted to help a group of users find a restaurant that all will like. We could try several methods: one is finding the restaurant that maximizes average happiness.

By Finding the Influential nodes based on various centrality models for communities. Accurately predict links between various food joints based on their labels. Should be able to able to accurately classify restaurants between drive thru (fast food) or dine in. Link prediction will help us to predict whether there will be links between two nodes based on the attribute information and the observed existing link information.
Link prediction not only can be used in the field of social network but can also be applied in other fields. Real-time data from Facebook involves communities of various sizes and it is necessary to detect the overlapping nature of communities. A graph-oriented solution requires the usage of link-weights which is crucial to determine communities of high importance. The aim of this work is to successfully determine communities for user networks based on key elements of Facebook Restaurant Pages Network accurately detect communities based on dining.

1.5 Objective

- To develop a system to accurately detect communities based on dining types: Fast food or Dine in restaurants, accurately depict the influential nodes based on various centralities and form together links which are required for marketing strategies.
- Find Influential nodes based on various centrality models for communities.
- Accurately predict links between various food joints based on their labels.
- Should be able to able to accurately classify restaurants between drive thru (fast food) or dine-in.

1.6 Applications

The application of the proposed system covers a few areas such as marketing applications by identifying most influential nodes, detection of accurate communities, identifying future links in a network. Also, the link prediction to be performed with greater accuracy. Link prediction has found varied uses, however any domain throughout that entities act throughout a structures approach will relish link prediction.A common applications of link prediction is up similarity measures for cooperative filtering approaches to recommendation. Link prediction is to boot often utilised in social networks to counsel friends to users. it is also been used to predict criminal associations. In biology, link prediction has been used to predict interactions between proteins in protein-protein interaction networks. Link prediction has additionally been used to infer interactions between medication and targets victimization link prediction Another application is found along prediction in scientific co-authorship networks.

Chapter 2

Review of Literature

Qi Chen et al. [1] The primary emphasis of this paper is on the study of identifying patterns of activity and forecasting the potential configuration of advanced networks in overlapping substructures. It examines a few of the best algorithms for detecting overlapping communities in advanced networks.

X.Ma et al. [2], This paper investigates two evolutionary non-negative matrix factorization mechanisms for detecting complex populations by clustering, mapping, and other techniques. The algorithm used in this paper proved to be more accurate than many state-of-art approaches.

Hao Shao et al. [3] This paper describes a relation prediction algorithm for unsupervised networks. It focuses on unsupervised machine learning models for detecting node connections accurately.

Junming Shao et al. [4] For connection prediction and group identification, this paper employs cluster-driven low rank Matrix completion. The proposed algorithm in this paper outperformed many previously studied algorithms in terms of accuracy.

David Liben Nowell. [5] This paper focuses on Link estimation using various coefficients such as Jaccard's coefficient, based on a large number of near neighbours, and various paths such as page rank, reaching time, commuting time, and so on. This paper also discusses approaches focused on node neighbourhoods, such as common neighbours. The accuracy of these forecasts was equivalent to 16 percent, and there is still potential for significant progress in the algorithms used in the paper. Often, the time frame or optimization of these algorithms can be done to increase the time complexity so that they operate much faster on massive data sets.

Mohsen Shahriari et al. [6] This paper proposes a two-step method for locating overlapping communities in signed social networks. Furthermore, assess the significance of three node classes: additional, overlapping, and intra. The results show that overlapping nodes can predict signals more accurately than intra and extra nodes. In addition, anger was applied as a measure to test errors in fuzzy group identification in signed social networks.

Le Yu et al. [7] This paper focuses on a population identification thesis based on dynamic network analysis. It suggests a novel algorithm for detecting overlapping communities. The proposed algorithm, as opposed to traditional algorithms relying on node clustering, is based on connection clustering. The connection clustering would reflect groups of links with similar properties. The algorithm employs a genetic operation to cluster on links. An effective coding schema for number of communities can be automatically detected.

Kamal Sutaria et al. [8] The aim of this paper is to explain the interpersonal interaction between a community of active actors representing various types of structures. Many real-world structures, such as human cultures and various types of components, may be modelled as social networks. Social network research provides key words to a forum for industry to produce product surveys and promote the introduction of new technologies to the public body. This method differs from conventional clustering.

Jaewon Yang et al. [9] This paper reflects on the basic methods for uncovering operational concepts in networks. To create communities based on edge structure and node attributes, an accurate, scalable, and efficient algorithm for detecting overlapping communities in networks was created. This model integrates with the network configuration and node characteristics, resulting in more precise community identification and greater robustness.

Lei Tang et al. [10] This paper is written from the standpoint of data mining. It yields graph-based group identification strategies as well as numerous important extensions for dealing with complex, heterogeneous networks in social media.

Chapter 3

Design and Functionality

3.1 Flowchart

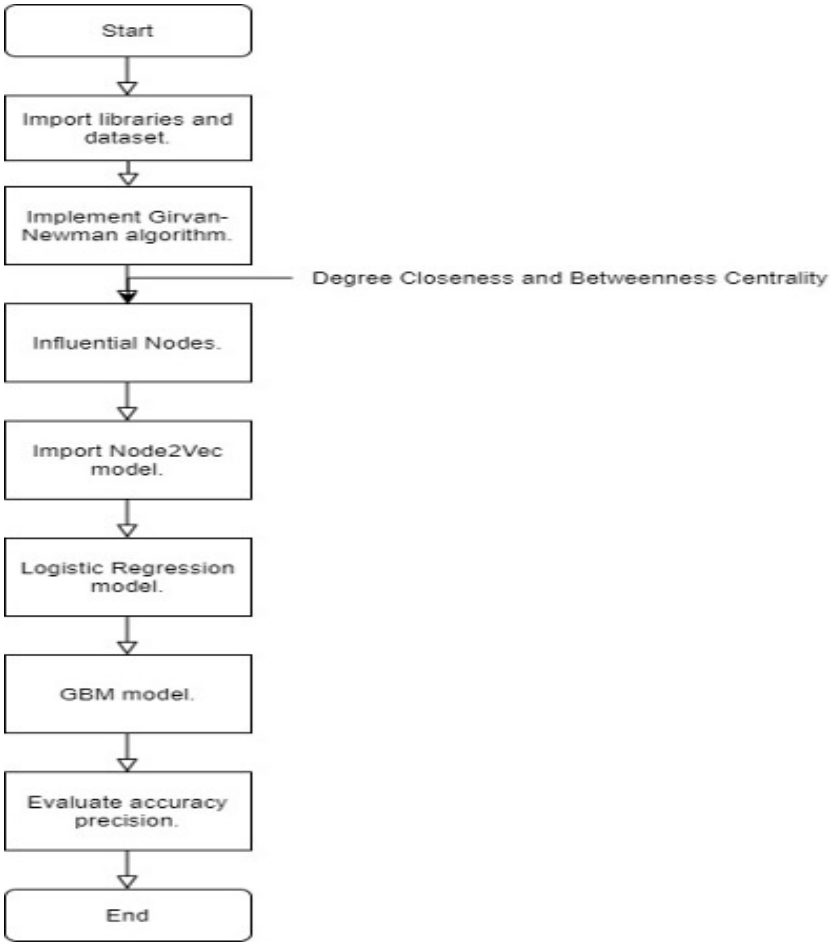


Figure 3.1 : Flowchart

3.2 Functioning of System

3.2.1 Logistic Regression -

Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits. 1 / (1 + e power -value) Where e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.

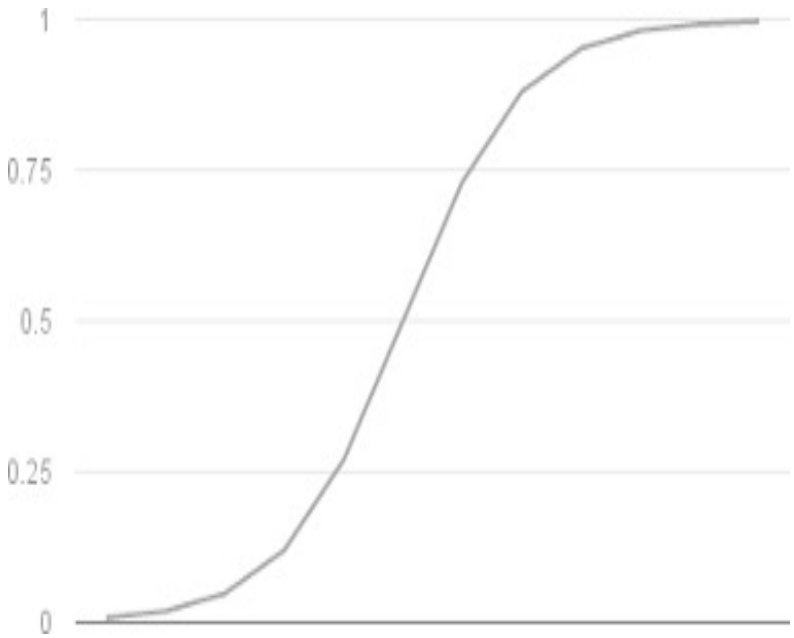


Figure 3.2 : Logistic Function

3.2.2 Node2Vec

Node2Vec is an algorithmic framework for representational learning on graphs. Given any graph, it can learn continuous feature representations for the nodes, which can then be used for various downstream machine learning tasks. The node2vec framework learns low-dimensional representations for nodes in a graph by optimizing a neighborhood preserving objective. The objective is flexible, and the algorithm accommodates for various definitions of network neighborhoods by simulating biased random walks. Specifically, it provides a way of balancing the exploration- exploitation tradeoff that in turn leads to representations obeying a spectrum of equivalences from homophily to structural equivalence. After transitioning to node v from t, the return hyperparameter, p and the in out hyperparameter, q control the probability of a walk staying inward revisiting nodes (t), staying close to the preceding nodes (x1), or moving outward farther away (x2, x3).

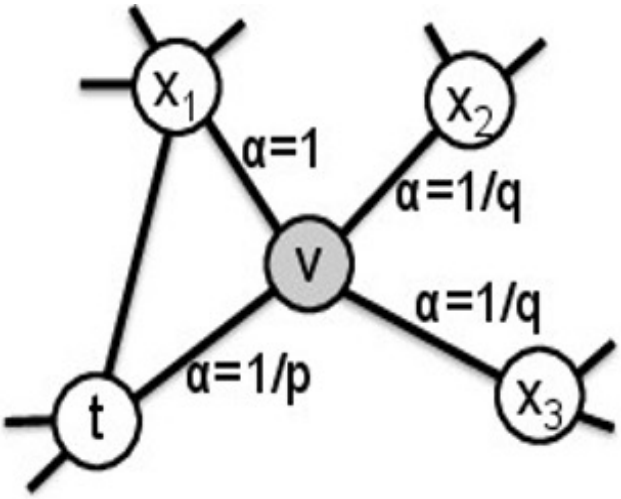


Figure 3.3 : Node2Vec Model

3.2.3 LightGBM Model :

LightGBM is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm, used for ranking, classification and many other machine learning tasks. Since it is based on decision tree algorithms, it splits the tree leaf wise with the best fit whereas other boosting algorithms split the tree depth wise or level wise rather than leaf- wise. So, when growing on the same leaf in Light GBM, the leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence results in much better accuracy which can rarely be achieved by any of the existing boosting algorithms. Also, it is surprisingly very fast, hence the word 'Light'.

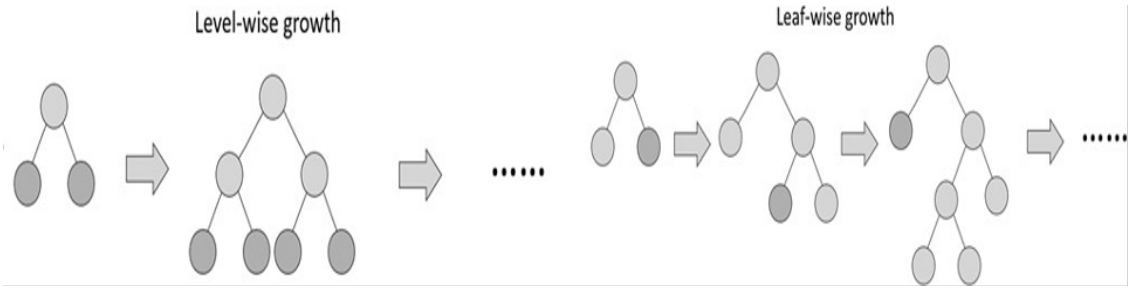


Figure 3.3 : LightGBM Model

3.2.4 Influential Nodes–

Influential nodes based on degree centrality -
Degree centrality is a simple count of the total number of connections linked to a vertex. It can be thought of as a kind of popularity measure, but a crude one that does not recognize a difference between quantity and quality.
Degree centrality does not differentiate between a link to the president of the United States and a link to a high school dropout. Degree is the measure of the total number of edges connected to a particular vertex. For directed networks, there are two measures of degree. In-degree is the number of connections that point inward at a vertex. Out-degree is the number of connections that originate at a vertex and point outward to other vertices.

Influential nodes based on closeness centrality-
Closeness centrality indicates how close a node is to all other nodes in the network. It is calculated as the average of the shortest path length from the node to every other node in the network. Closeness centrality measures each individual's position in the network via a different perspective from the other network metrics, capturing the average distance between each vertex and every other vertex in the network.
Assuming that vertices can only pass messages to or influence their existing connections, a low closeness centrality means that a person is directly connected or "just a hop away" from most others in the network.

Influential nodes based on Betweenness Centrality -
The betweenness centrality captures how much a given is in-between others. This metric is measured with the number of shortest paths that passes through the target node. This score is moderated by the total number of shortest paths existing between any couple of nodes of the graph. The target node would have a high betweenness centrality if it appears in many shortest paths.

Chapter 4

Results and Discussions

4.1 Testing

Software Testing is defined as an activity to check whether the actual results match the expected results and to ensure that the software system is defect free. It involves the execution of a software component or system component to evaluate one or more properties of interest. Software testing also helps to identify errors, gaps, or missing requirements in contrary to the actual requirements. It can be either done manually or using automated tools.
There are two major types of software testing, namely functional and non-functional testing. Functional Testing is the type of testing in which the system is tested against the functional requirements and specifications. It ensures that the requirements or specifications are properly satisfied by the application. On the other hand, non-functional testing is the testing of a software application or system for its non-functional requirements i.e. the way a system operates, rather than specific behaviors of that system. Say, for ex- ample, software performance, it is a broad term that includes many specific requirements like reliability and scalability.

4.2 Results and Discussions

Community detection:
The two communities created by the Girvan-Newman algorithm are depicted in the diagram below. Dine-in restaurants are represented by blue nodes, while fast-food restaurants are represented by red nodes.

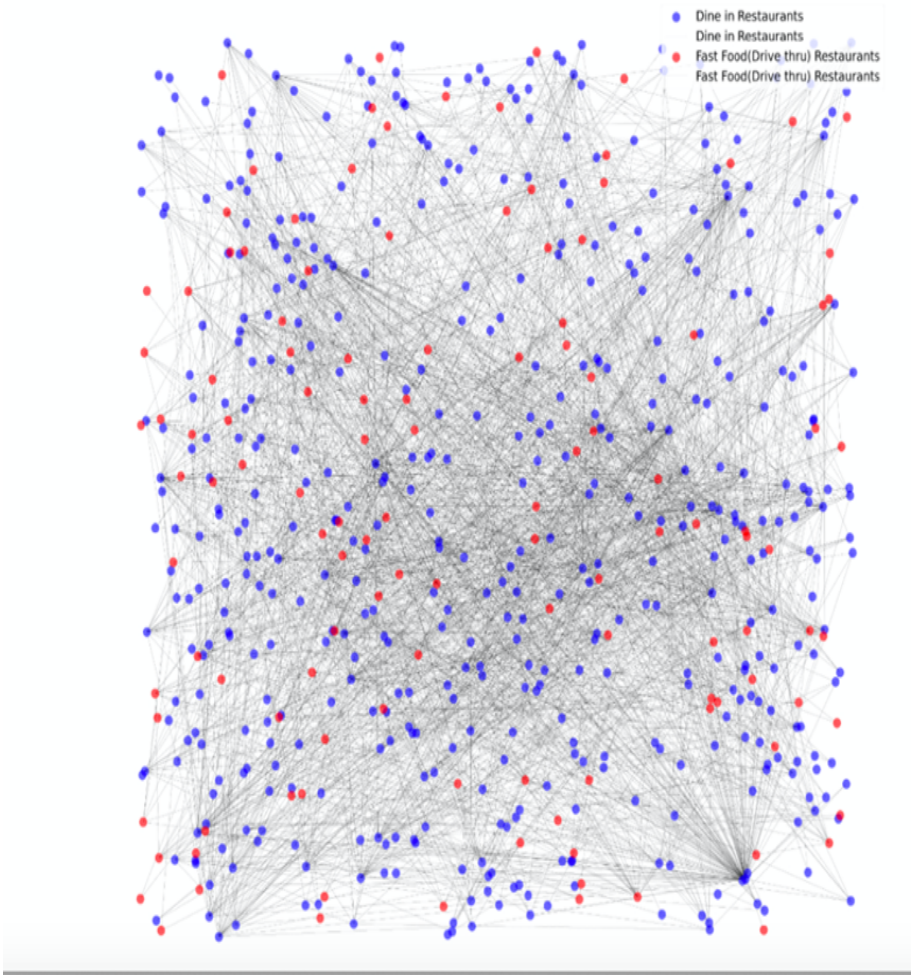


Figure 4.1: Community Detection

Influential Nodes:
Influential for both the communities are based on three different centralities: Degree, Betweenness, Closeness centrality.

```
In [15]: blue_community = []
         red_community = []
         x = centrality.keys()
         for node in x:
             if node in node_groups[0]:
                 blue_community.append(node)
             else:
                 red_community.append(node)

In [16]: print ("Most influencing node in Blue Community according to degree centrality:")
         node = centrality_dict.keys()
         for n in node:
             if n in blue_community:
                 print(n)
                 print(centrality_dict[n])
                 break;

         Most influencing node in Blue Community according to degree centrality:
         265
         0.21647819063004847

In [17]: print ("Most influencing node in Red Community according to degree centrality:")
         node = centrality_dict.keys()
         for n in node:
             if n in red_community:
                 print(n)
                 print(centrality_dict[n])
                 break;

         Most influencing node in Red Community according to degree centrality:
         545
         0.035541195476575124
```

Figure 4.2.1: Degree centrality

In [23]:
```
blue_community = []
red_community = []
x = centrality.keys()
for node in x:
    if node in node_groups[0]:
        blue_community.append(node)
    else:
        red_community.append(node)
```

In [24]:
```
print ("Most influencing node in Blue Community according to betweeness centrality:")
node = centrality_dict.keys()
for n in node:
    if n in blue_community:
        print(n)
        print(centrality_dict[n])
        break;
```

```
Most influencing node in Blue Community according to betweeness centrality:
265
0.34990766617377767
```

In [25]:
```
print ("Most influencing node in Red Community according to betweeness centrality:")
node = centrality_dict.keys()
for n in node:
    if n in red_community:
        print(n)
        print(centrality_dict[n])
        break;
```

```
Most influencing node in Red Community according to betweeness centrality:
618
0.0932726061636337
```

Figure 4.2.2: Betweeness centrality

In [19]:
```
blue_community = []
red_community = []
x = centrality.keys()
for node in x:
    if node in node_groups[0]:
        blue_community.append(node)
    else:
        red_community.append(node)
```

In [20]:
```
print ("Most influencing node in Blue Community according to closeness centrality:")
node = centrality_dict.keys()
for n in node:
    if n in blue_community:
        print(n)
        print(centrality_dict[n])
        break;
```

```
Most influencing node in Blue Community according to closeness centrality:
265
0.33137044967880086
```

In [21]:
```
print ("Most influencing node in Red Community according to closeness centrality:")
node = centrality_dict.keys()
for n in node:
    if n in red_community:
        print(n)
        print(centrality_dict[n])
        break;
```

```
Most influencing node in Red Community according to closeness centrality:
618
0.2443742597710225
```

Figure 4.2.3: Closeness centrality

Link Prediction:
The figures below represent the roc-auc score after fitting the logistic regression model for our dataset. It is roughly 81.33 percent. The curve of the auc score is visualized below.

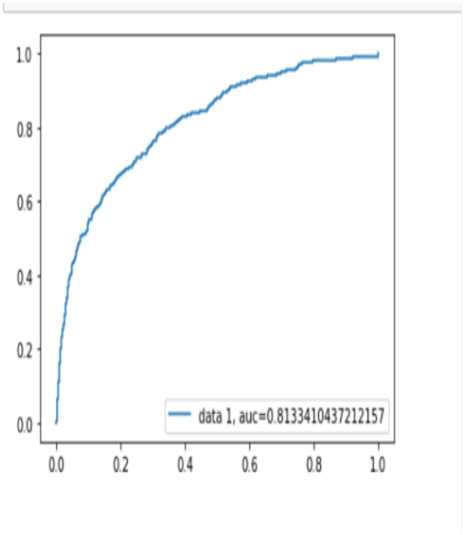


Figure 4.4: ROC AUC curve

The confusion matrix for logistic regression for actual links vs predicted links is represented below, based on which the accuracy, precision and recall is calculated. The accuracy is roughly 75.04 percent and the precision and recall stand at 17.46 percent and 71.02 percent respectively.

The confusion matrix for lightGBM model for actual links vs predicted links is represented below, based on which the accuracy, precision and recall is calculated. The accuracy is roughly 95 percent and the precision and recall stand at 69 percent and 54 percent respectively.

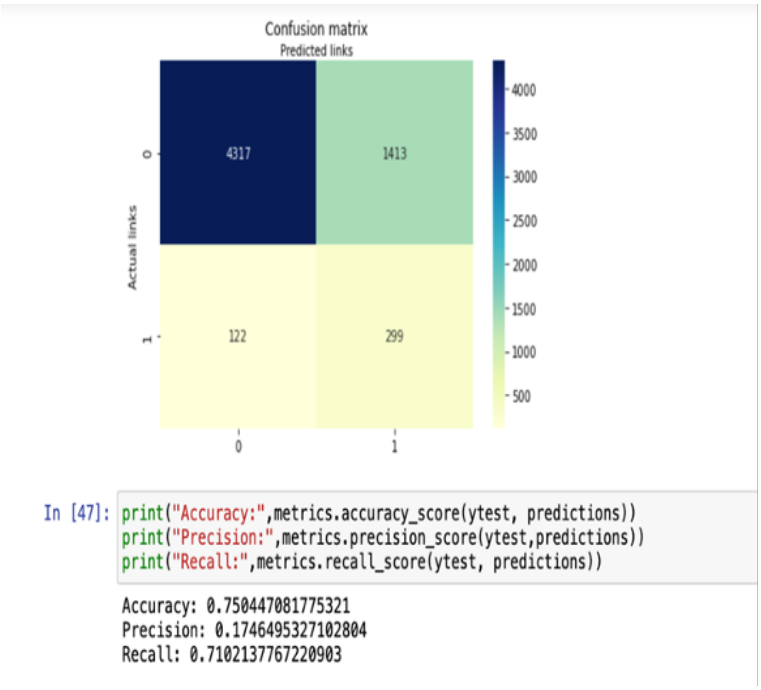
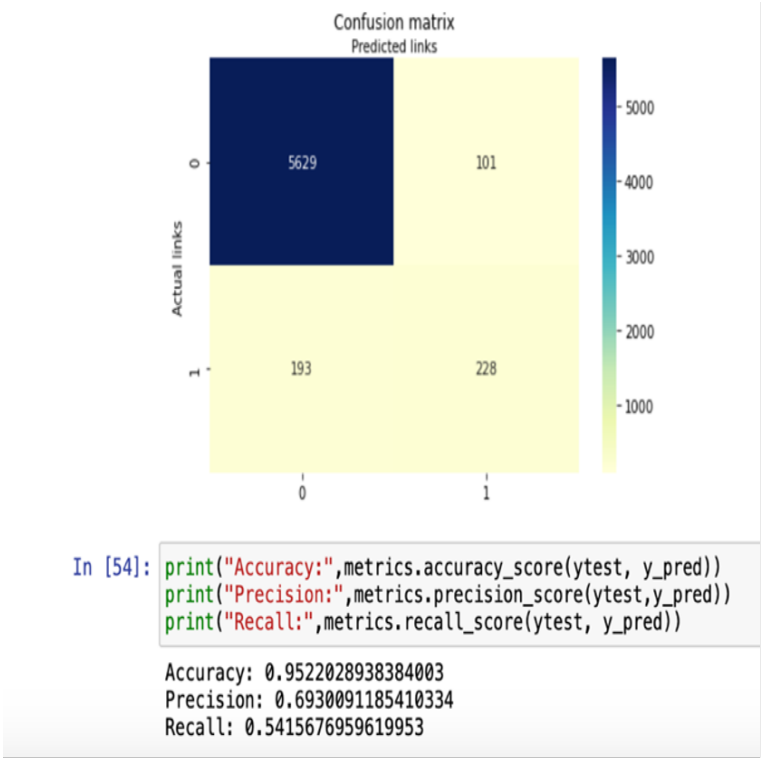


Figure 4.5: Confusion matrix for logistic regression




```
In [54]: print("Accuracy:",metrics.accuracy_score(ytest, y_pred))
         print("Precision:",metrics.precision_score(ytest,y_pred))
         print("Recall:",metrics.recall_score(ytest, y_pred))

         Accuracy: 0.9522028938384003
         Precision: 0.6930091185410334
         Recall: 0.5415676959619953
```

Figure 4.6: Accuracy and confusion matrix for lightGBM

Chapter 5

Conclusion

In the proposed project we explored how Community Detection on Facebook Restaurant Pages Network and item graphs could be used to improve link prediction between various food joints based on their labels. The Influential nodes required for marketing can be accurately found out based on various centrality measures and community detection algorithm to increase the statistics for marketing. We can conclude that the LightGBM model works excellently on this data set and as intended. Influential node detection for particular communities is determined and community detection accurately predicts the similarities between nodes for further classification.

Bibliography

References

[1] Qi Chen; Lingwei Wei, "Overlapping Community Detection of Complex Network", 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), Gold Coast, QLD, Australia, DOI: 10.1109/PDCAT46702.2019.00102.

[2] X. Ma; D. Dong, "Evolutionary Nonnegative Matrix Factorization Algorithms for Community Detection in Dynamic Networks", 2017 IEEE Transactions on Knowledge and Data Engineering, Pages: 1045 – 1058, DOI: 10.1109/TKDE.2017.2657752. of Engineering and Technology (IRJET).

[3] Hao Shao; Lunwen Wang; Jian Deng, "A Link Prediction Algorithm by Unsupervised Machine Learning", 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), DOI: 10.1109/CISCE.2019.00145. Pages 3382-3388. https://doi.org/10.24963/ijcai.2019/469.

[4] Junming Shao; Zhong Zhang; Zhongjing Yu; Jun Wang; Yi Zhao; Qinli Yang, "Community Detection and Link Prediction via Cluster-driven Low-rank Matrix Completion", Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), Pages 3382-3388, https://doi.org/10.24963/ijcai.2019/469.

[5] David Liben-Nowell; Jon Kleinberg, "The Link Prediction Problem for Social Networks", CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management, November 2003, Pages 556–559, https://doi.org/10.1145/956863.956972.

[6] Mohsen Shahriari; Ralf Klamma "Signed social networks: Link prediction and overlapping community detection", 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Pages: 1608-1609, DOI Bookmark: 10.1145/2808797.2809357.

[7] Le Yu; Bin Wu; Bai Wang "LBLP: link-clustering-based approach for overlapping community detection", Tsinghua Science and Technology, DOI:10.1109/TST.2013.6574677.

[8] Kamal Sutaria; Dipesh Joshi; C.K. Bhensdadia; Kruti Khalpada, "An Adaptive Approximation Algorithm for Community Detection in Social Network", 2015 IEEE International Conference on Computational Intelligence and Communication Technology, DOI: 10.1109/CICT.2015.103.

[9] Jaewon Yang; Julian McAuley; Jure Leskovec, "Community Detection in Networks with Node Attributes", 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, DOI: 10.1109/ICDM.2013.167.

[10] Lei Tang; Huan Liu, Community Detection and Mining in Social Media, Morgan & Claypool, 2010.

[11] Q. Liu; G. Liu and X. Chu, "Comparison of different spatial resolution bands of SPOT 5 to plant community patch detection," 2012 5th International Congress on Image and Signal Processing, 2012, pp. 1029-1033, doi: 10.1109/CISP.2012.6469748.

Plagiarism report

Figure 5.1: Plagiarism report

Title of the Paper : Community Detection And Link Prediction on Social Networks