

# Social Media Link Prediction

<sup>1</sup>Abhisha Solanki

<sup>1</sup>Department of Computer Application, Jain University, Bangalore, India

## ABSTRACT

**Objective of Model:** To build a social media link prediction model for future friendship links predictions between any unconnected friendship nodes in a network. Nowadays, the quantity of social media network users is going sky-high that makes the network more complex through the addition of new edges. During this resulting data related to social networking, users have created a substantial amount of attention and opportunities for data mining problems that have been dedicated to computational data analysis evolution of social networking. The popular data processing problem for social media is that the new friendship links prediction between unconnected nodes and prospective connections. Instinctually, a friendship between any two social media users can be predicted premised on their similar friendship links, likes, and shared interests. However, shared interest's data generation of users can be very challenging, by the big number of practicable outcome interests. Within the previous, approaches that utilized an interest ontology of selected users are proposed to tackle problem of link prediction, nevertheless the development of the ontology is proved exorbitant computationally, also the results were not very useful of ontology for the model. Alternative, we introduce a topic modelling approach and look over some typical link prediction techniques by classifying them to predicting future friendships links premised on common interests and also from existing connections. Specifically, Latent Dirichlet Allocation (LDA) accustomed model on basis of user interests and that we create an implicit ontology on users' interest. Here, we constructing a new model for the problem of link prediction supported resulting as topic distributions and terminate the survey with a discussion on the latest advances and future research focus.

## 1. INTRODUCTION

Do you strange to think who will be your next friend on Facebook? Searching from where the following request may come? Which unconnected node pairs are credible a link in the future? Can one predict the next possible links to form within the network? [1]

Almost social media platforms and applications like Facebook, LinkedIn, LiveJournal are often framed as graphs structure. During a universe of social network, the listed users are linked together. And to figure on these graphs of networks, must beseta unique approaches and algorithms substituting the methods of traditional machine learning. Facebook, Tinder, Amazon are proposing new friendships, matches, and items to get through a model called Link Predication. It is used to predict future possible links or predict missing links because of incomplete data. Link Prediction Algorithm makes an implicit assumption of graph mechanism for growth of graph. Local methods predict links purely support the local contact structure of the triangle closing. Adjacency matrix, one can take the global structure of the graph that often based on the number of shortest paths between nodes. The proposed method will improve the accuracy of Friendship Prediction.

Advanced technology has become the integral part of our life [1]. To satisfy the need of the society, almost in each work, we use the technology [2] [3]. In current era computer science is major subject [4]. It has many real life applications such as cloud computing [5], artificial intelligence [6], remote monitoring [7], Wireless sensor network [8, 9, 10], internet of things [11, 12, 13], Neural network [14, 15], FSPP [16, 17, 18], NSPP [19, 20, 21, 22, 23], TP [24, 25, 26], internet Security [27], uncertainty [28, 29, 30, 31, 32] and so on. Technology is the mode by which user can store, fetch, communicate and utilize the information [33]. So, all the organizations, industries and also every individual are using computer systems to preserve and share the information [34]. The internet security plays a major role in all computer related applications. The internet security appears in many real-life applications, e.g., home security, banking system, education sector, defence system, Railway, and so on. In this manuscript we discuss about the protection of authentication which is a part of internet security.

### 1.1 An Overview of Social Network:

Social network like Instagram, Facebook, Tinder, LiveJournal have attracted innumerable users, according recent statistics the social networking overtaking search engines by its usage. Many social media including Facebook online services are focused on user interactions and connectivity using Link Predictions as its feature. Users in Facebook can tag friends, and also specify their target-group and shared interests in this social network. [2] We are able to see Facebook structure as graph with users relevant to point nodes in the network that represent people or other entities embedded in an exceedingly social context and edges (lines

connecting any node pairs) appreciate to friendship links between the peoples and those people edges represent interaction, collaboration or influence between users. In general, the social network is equivalent to an undirected graph. However, the edges in Facebook network are directed edges i.e., if some users are characterised as user 'A' define as a friend of user 'B', but it is not necessary that user 'B' be the friend of user 'A'[1]. This model introduced as the LINK PREDICTION, where prediction of a future friendship link between two user's 'A' and 'B' is the main objective. The massive amounts of social media data accumulated in the past few years have made the link prediction problem possible, although very picky.

In this work, we aspire the utilizing strength of approaches, tools, and algorithms of machine learning to require to take advantage of the user's profile information and graph structure of online social media service, like Facebook, which predicting new friendship links. In such social networks the users profile contains processed data into informative. As an example, specified users shared interest in Facebook act as a virtuous gauge to whether two users can be friends or not. Consequently, if two user's 'A' and 'B' have similar interests, then there is an honest chance that they'll be friends. However, the users with a common interest can be a wide range, and semantically similar interests we need to be grouped these similar interests. We exploit a topic modelling approach to accomplish this model which figures out the outline 'topics' that occur in the assembly of documents. This provide a straightforward and efficient way of capturing the semantics of user interests by grouping them into categories, and thus reducing the dimensional of the problem by using a text-mining tool. LDA is an example of topic modelling that used to classify text in a document to a specific topic. Additionally, to take advantage of the graph structure of the Facebook network by utilizing user interests and information extracted from the graph structure (conjunct friends of two unconnected user nodes) that's helpful for link prediction problem. The following terms contributions of this paper as: an approach for implement techniques of topic modelling, specifically Light-GBM, social networking user's profile data. On several social networking user's data sets of varying sizes shows the productivity of the LDA features in predicting future friendships links by observation. Experimental results on Facebook datasets obtained the most effective results by fetching information from shared interest and topic model approach by extracting data from the graph structure of the social networking. And incremental in performance of the proposed model to improves the user's number on social networking sites.

## 2. SYSTEM ARCHITECTURE

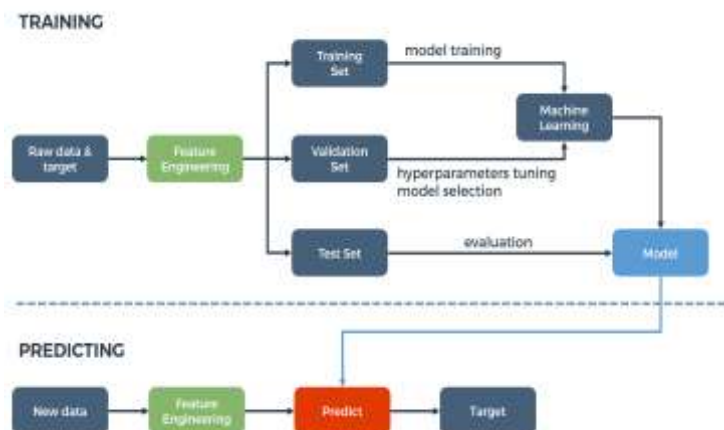


Figure 1: System Architecture

Figure 1: Raw data and Target- This gives input data to get the output. Feature engineering- Here, the job is to choose the best data in the sense the data which used more and has more importunity after choosing the best data. This data split into three sets Training set, Validation set, and Test set. Training set- we give 80% of the data to find the output. Validation set- some like training set and optional step. Test set- is executed after the convenient model, check the model functionality. We use the remaining 20% of data to test. Then the data train the model through the help of a machine learning algorithm. Prediction- it selects the best data among given data, it predicts the output and pan to model.

## 3. MODULE DIAGRAM

Figure 2: System Architecture of any network (of nodes and links) analysed and divide output into Interest Based Features (through implementing Latent Dirchlet Allocation) and Graph Based features (through degree order and distribution). Interest based features expressed by identifying and extracting features from shared interest of each user. Graph based features formed as a result of users tagging others. These are combinedly taken as input to Learning Algorithm and process to predict friendship links between users. [37]

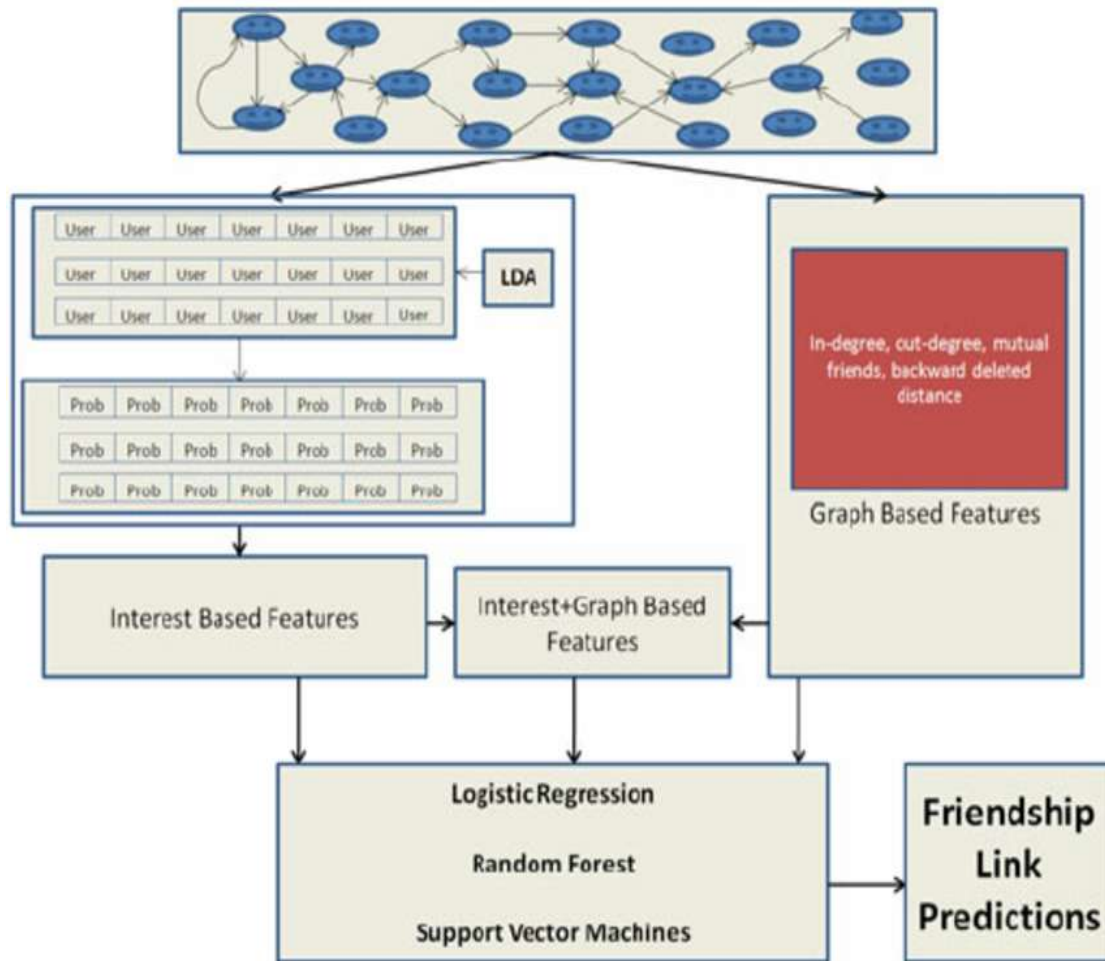


Figure 2: Modular Structure

## 4. MODULE DESCRIPTION

### 4.1 Link Prediction Primer

Link prediction among of the foremost research topics is the structure of graphs and networks. A Primer predicting targets for any social networking are presenting link prediction similarity measures, that exploit the degree distribution in the graph structure. The link prediction model is spotting the node pairs which will constitute the future friendship link or not. Degree tells how many nodes are in a network that have a particular degree (number of edges connected), it is the possibility of degree that has in graph.

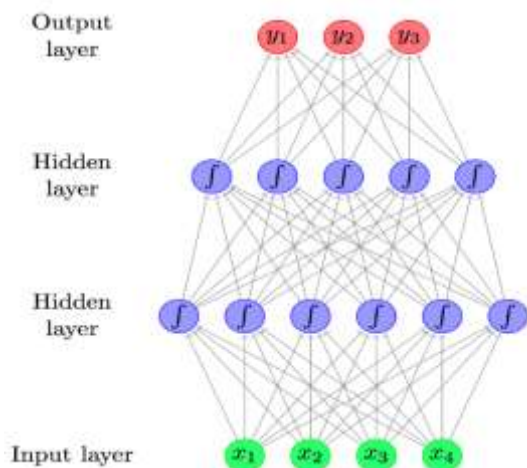


Figure 3: Levels for Prediction.



Figure 4: High influencing \ Central node.

#### 4.2 Strategy to Solve a Link Prediction Problem

Prediction by using machine learning algorithm for formation of future friendship links in pairs of the graph nodes which are unconnected in network, by representing a graph of any structured dataset that have a set of features. [38]

Let's take a graph for a better understanding. Given below is a 7-node graph and have edge connectivity between nodes are A-D, A-B, D-F, B-C, C-E, C-G. Therefore, the unconnected node-pairs are following:

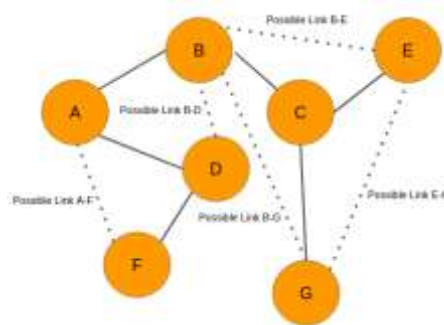


Figure 5: Graph before prediction.

By fig 5, We are plotting dotted path between unconnected nodes: F to A, D to B, B to E, G to B, G to E and label them as possible links. Missing links identification is link completion that find unobserved edges (links) that are consistent with the structure. Then estimate the reliability of given links in existing graph by following:

- Compute similarity for each node pair.
- Sort all pairs by the decreasing score (which consider by the expected number of random walk steps between particular unpaired nodes.)
- Select top n pairs with lesser walk steps as newly predicted links.

Let's say analysing data from above graph we legislated graph given below. And few new connection links are formed by red lines: A-F, B-D, B-E.

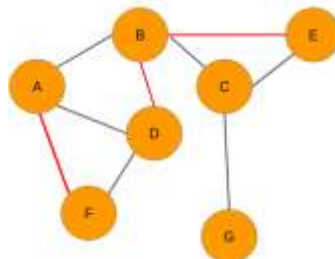


Figure 6: Graph with Predicated Links

We need a collection of regressors and variable targets to propose this machine learning approach. Therefore, we will compass from the graph. Now, we see how it works. Our objective is to search out which class output target data belongs to or predict values by mapping its optimal relationship with input predictor's data. From the existing graphon particular instantperiodp, we can extract the subsequent pair of nodes which haven't connected: F-A, D-B, B-E, G-E, B-G

Here, only considering couple of edges from the network nodes we subsequent step for all node pairs creation. There are several techniques which exploit by extracting features from the network nodes. Let's use techniques among all to build features for every pair of nodes. Still, we can't able to know target variable. At looking graph on period durations of  $p+r$ , we figure out three new links (in network) formation of the pairs F-A, D-B, and B-E respectively. Therefore, the worth of 1 we assign to them. And assign worth of 0 to the B-G and E-G node pairs because of no interconnectivity between these pair nodes.

After, it looks something like this datasheet:

Table 1. Outcomes

Features	Link (Target Variable)
Features of A-F pair	1
Features of B-D pair	1
Features of B-E pair	1
Features of B-G pair	0
Features of E-G pair	0

Here, by above table we get the target variable of all links, by this data we have a tendency to perform socialnetwork link prediction by proposed machine learning model. Thus, we like to use graphs of social networksbycompletely different two instances to extract the target variable at time presence of a link between a node pair. [39]

#### 4.3 Building a Model by Extracting data from a Graph

From above conclusion, the target variablegets labelled by accessing graph at time  $p+r$ . But,eventually for practice, we have only one dataset of graph in hand.

Let's assume we'vethe graph given below of a social network wherever the social media users are corresponding to nodes in a graph and the edges between themcorresponding some quite of relationship:

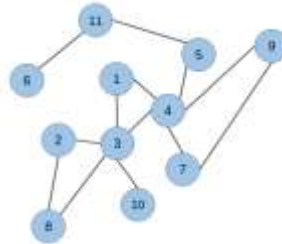


Figure 7. Network of nodes

The candidate node pairs, which can form a link at a future time are some like this (1-2), (4-10), (5 -6), (5-7), and more. We have to build a model which will predict if there would be a link between these node pairs or not. However, to build any link prediction model, it wants to create a training dataset with the help of existing graph. It can do by using a simple step. Initiallyassume how would this graph have looked seemed like at some point within the past? There would be some of those edges between the nodes because connections in a social network are built gradually over time.

Hence, some of the edges are hidden randomly from existing graph and then follow the above explained techniques for training dataset creation within the previous structure.

#### 4.4 Eliminate Links from the Graph:

We should avoid removing any edge that may produce an isolated node (node without any connectivity with any other node) while removing any link or an isolated network. Now, bring out some of the edges from graph:

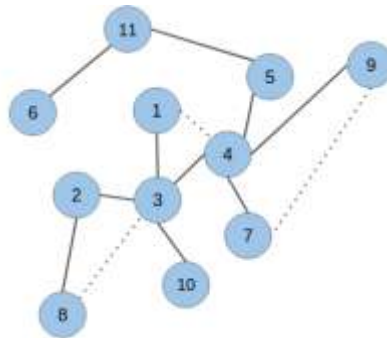


Figure 8. Network nodes second stage

Fig 8 As we noticed that the links connecting the pair of nodes (1-4), (3-8), (7-9) are detached.

#### 4.5 Add labels to extracted data

Now, in next walk we create features for all node pairs (which are without interactions between them as we done before) including those edges that have been omitted. After, edges will be labelled as value '1' (all removed edges) and the unconnected node pairs as '0' value:

Table 2: Pairing Results

Features	Link (Target Variable)
Features of pair 1-2	0
Features of pair 1-5	0
Features of pair 1-7	0
Features of pair 1-8	0



Features of pair 1-9	0
Features of pair 1-10	0
Features of pair 2-4	0
Features of pair 2-10	0
Features of pair 3-5	0
Features of pair 3-7	0
Features of pair 3-9	0
Features of pair 4-8	0
Features of pair 4-10	0
Features of pair 4-11	0
Features of pair 5-6	0
Features of pair 5-7	0
Features of pair 5-9	0
Features of pair 8-10	0
Features of pair 1-4	1
Features of pair 3-8	1
Features of pair 7-9	1

Now, it seems extremely balanced target variables by above technique. This way real-world graphs are encountered, but number of node pairs (unconnected pair of nodes) would be huge. [39] [40]

#### 4.6 Code Snapshots

```

In [1]: # Import pandas as pd
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.linear_model import LinearRegression
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]: # Load nodes details
with open("Desktop/Fb-pagets-food.nodes", 'rb') as f:
    fb_nodes = f.read().splitlines()

# Load edges (or links)
with open("Desktop/Fb-pagets-food.edges", 'rb') as f:
    fb_links = f.read().splitlines()

len(fb_nodes), len(fb_links)

Out[2]: (821, 2102)

In [3]: # Split the nodes in 2 separate lists

```

Figure 9.a. Code Snapshots

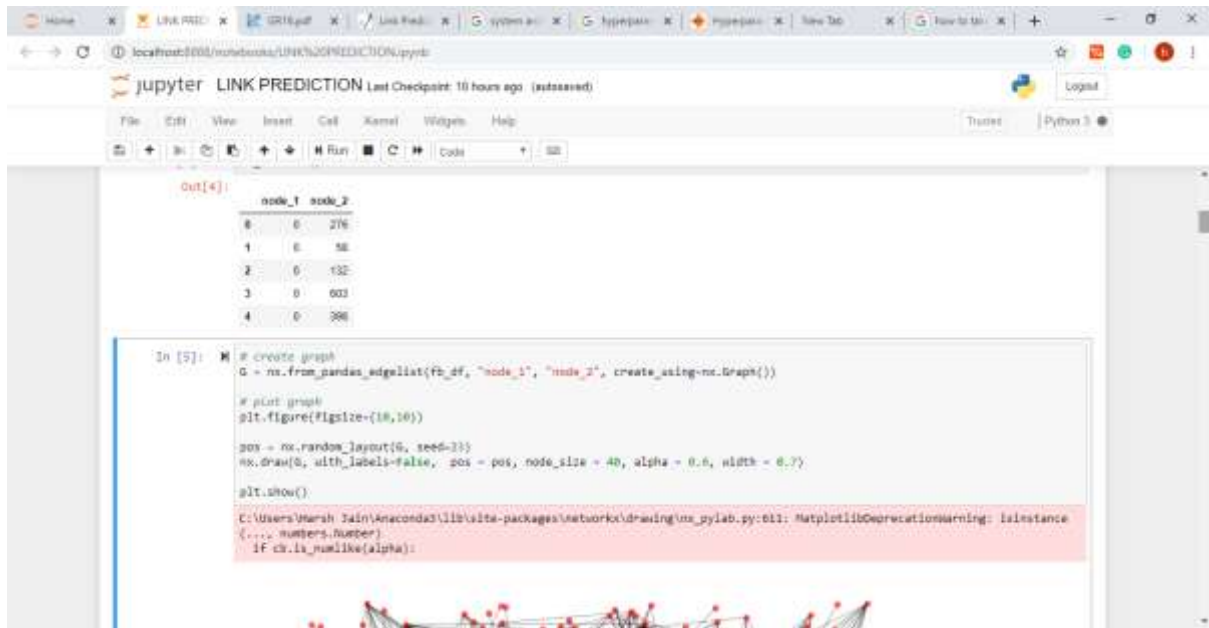


Figure 9.b. Code Snapshots

#### 4.7 GRAPH

Given undirected graph  $G = \{V, E\}$  is an unweighted graph that representing the topological structure of a social networking in which each edge represents a relationship or interconnectivity between nodes at a particular time. And all nodes denote the number of social network (Facebook) users, these users have a relationship of friendship suggestions in between these node pairs. Here, nodes are denoted by red points and edges (links) by lines. The size of fig = (10, 10), Size of node= 40, labels= false, alpha= 0.6, width= 0.7, and Position is pos = pos.

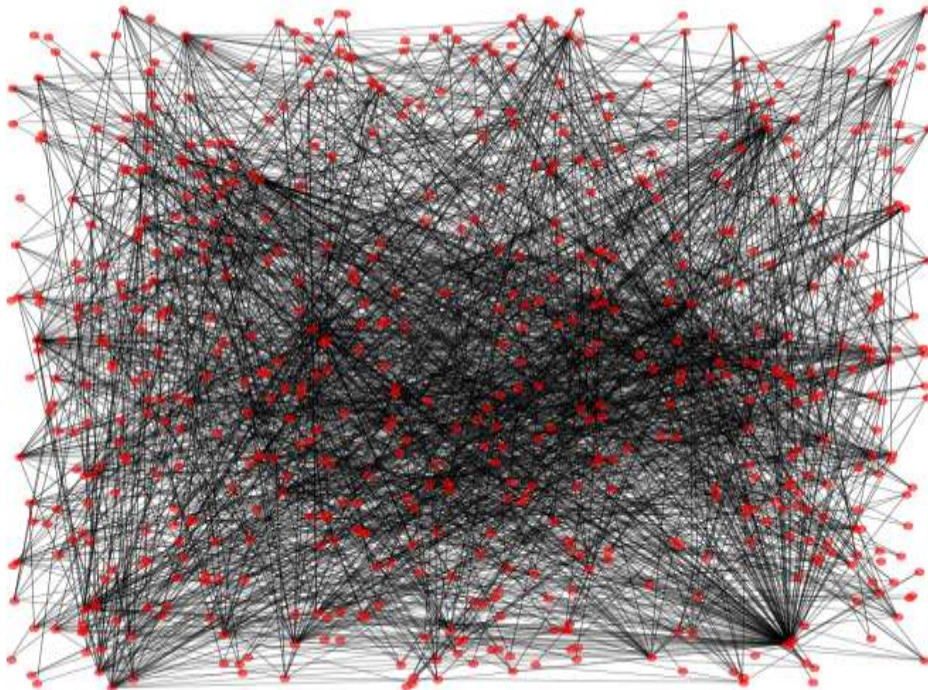
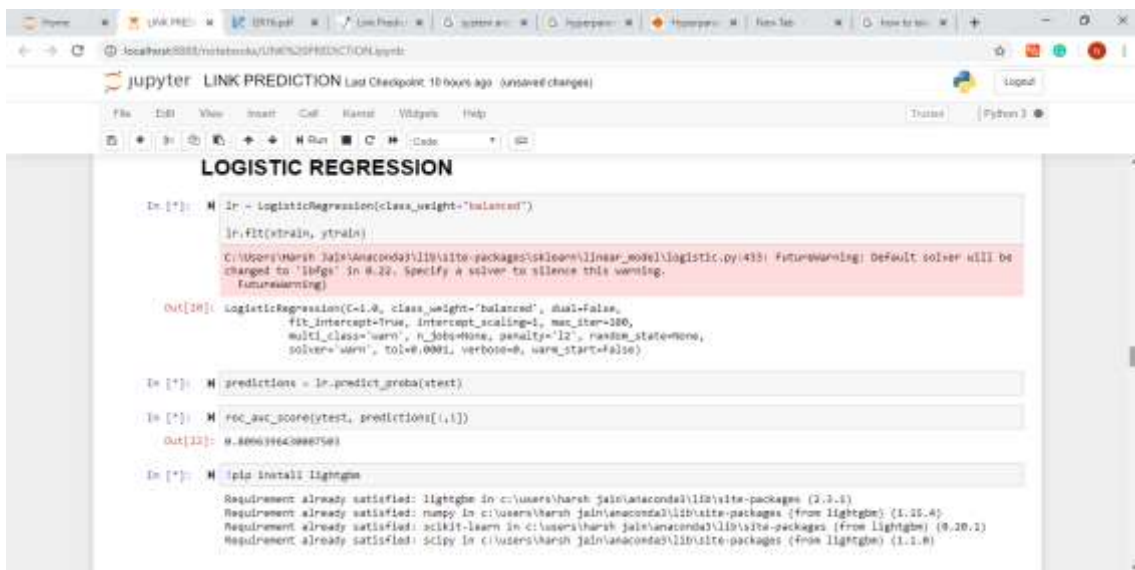


Figure 10. Network



The screenshot shows a Jupyter Notebook titled 'LINK PREDICTION'. The code in the cell is as follows:

```

In [*]: lr = LogisticRegression(class_weight="balanced")
        lr.fit(xtrain, ytrain)

Out[10]: LogisticRegression(C=1.0, class_weight='balanced', dual=False,
                             fit_intercept=True, intercept_scaling=1, max_iter=100,
                             multi_class='warn', n_jobs=None, penalty='l2', random_state=None,
                             solver='warn', tol=0.0001, verbose=0, warm_start=False)

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

In [*]: roc_auc_score(ytest, predictions[:,1])

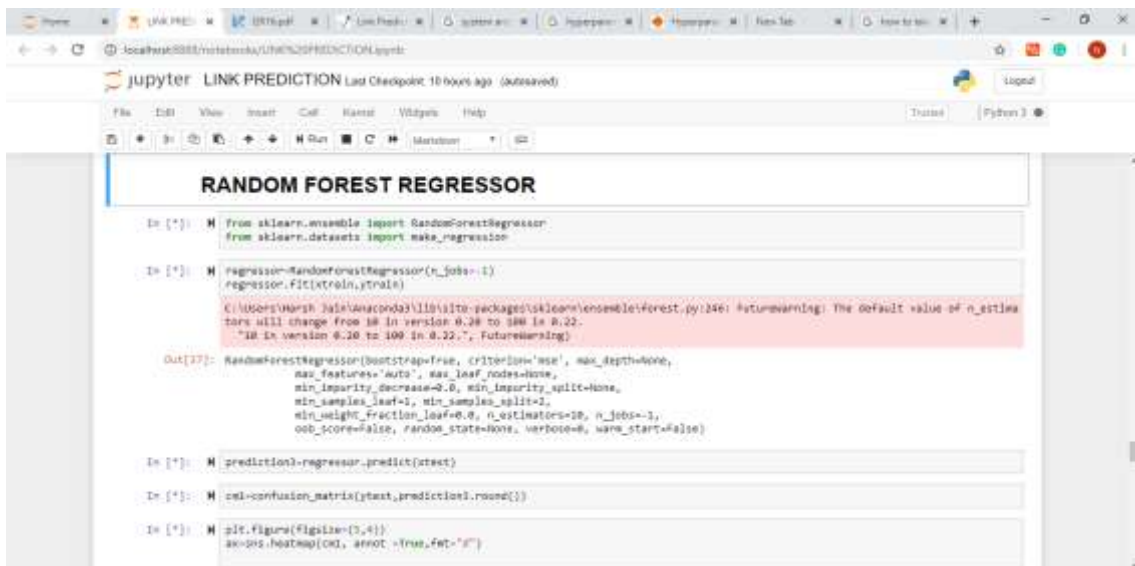
Out[12]: 0.86639640887561

In [*]: !pip install lightgbm

```

The output shows a warning from sklearn.linear\_model.logistic.py: 'FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.'

Figure 11.a. Code Snapshots



The screenshot shows a Jupyter Notebook titled 'LINK PREDICTION'. The code in the cell is as follows:

```

In [*]: from sklearn.ensemble import RandomForestRegressor
        from sklearn.datasets import make_regression

In [*]: regressor=RandomForestRegressor(n_jobs=-1)
        regressor.fit(xtrain,ytrain)

Out[17]: RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
                                max_features='auto', max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=1, min_samples_split=2,
                                min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=-1,
                                oob_score=False, random_state=None, verbose=0, warm_start=False)

In [*]: prediction3=regressor.predict(xtest)

In [*]: cm=confusion_matrix(ytest,prediction3.round())

In [*]: plt.figure(figsize=(5,4))
        ax=cm.heatmap(cm, annot=True,fmt="d")

```

The output shows a warning from sklearn.ensemble.forest.py: 'FutureWarning: The default value of n\_estimators will change from 10 in version 0.20 to 100 in 0.22.'

Figure 11.b. Code Snapshots

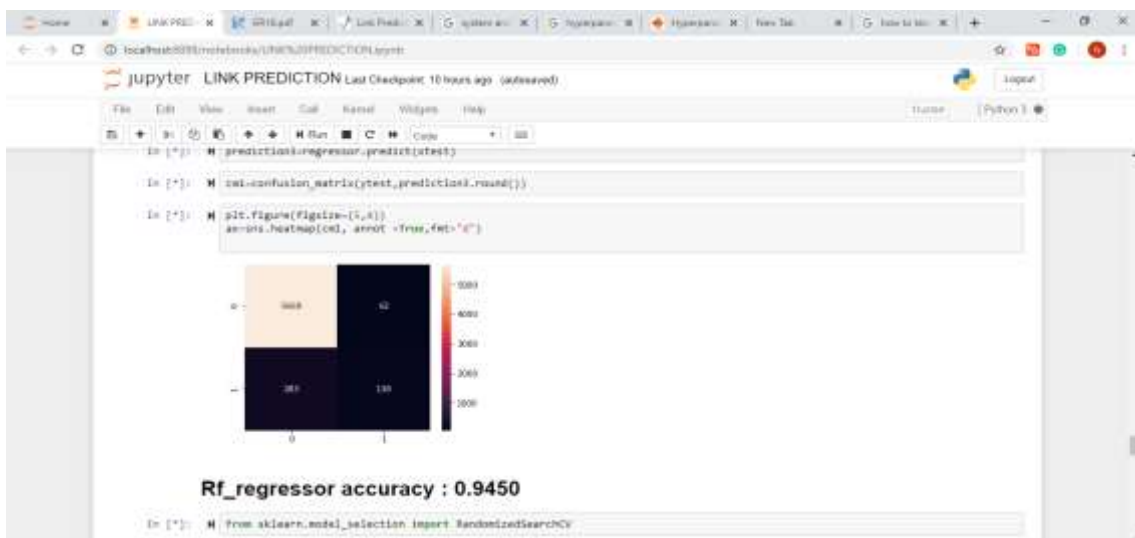


Figure 11.c. Code Snapshots



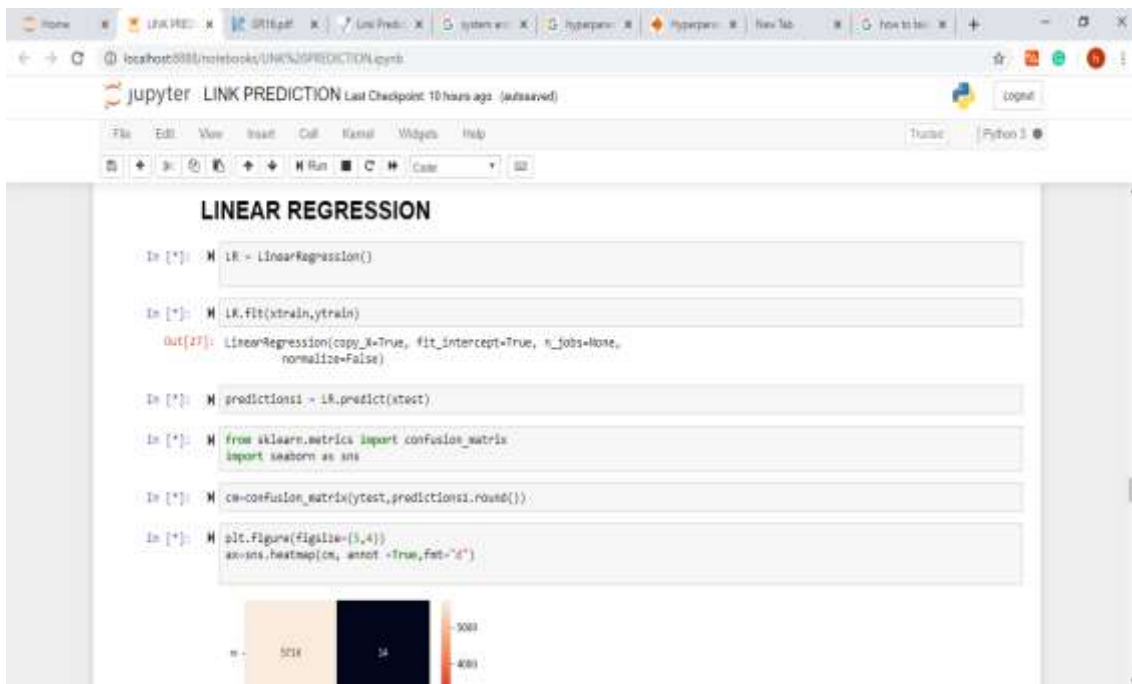


Figure 11.d. Code Snapshots

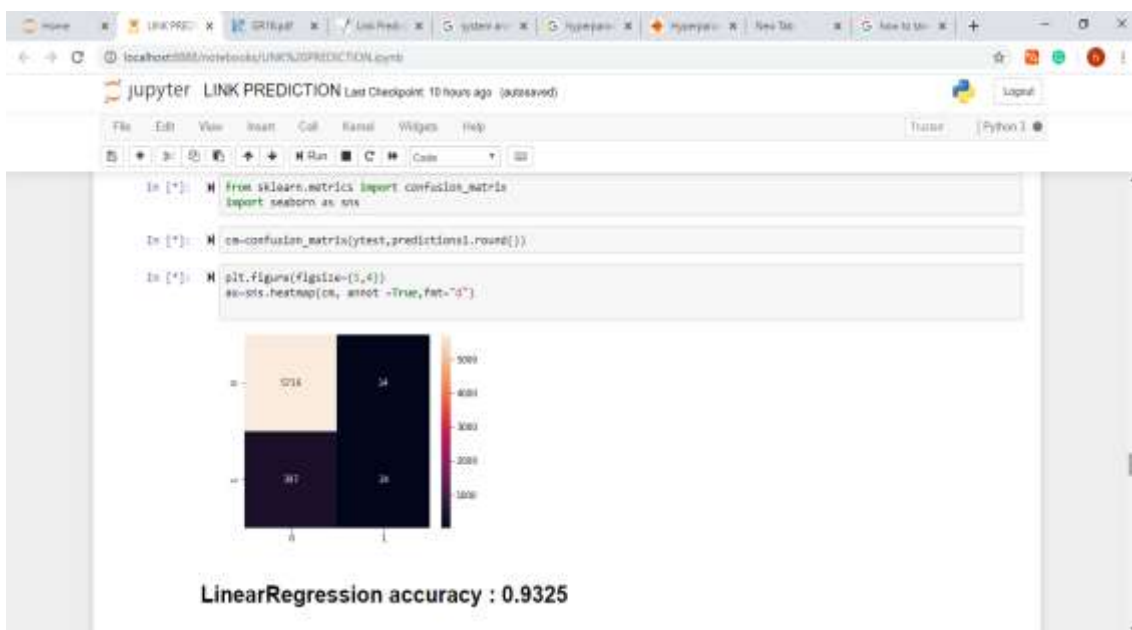


Figure 11.e. Code Snapshots

## 5. CONCLUSION

Social Media Link Prediction is that methodology which predict the potential future friendship links among the unconnected nodes within future. This algorithmic program takes both of global and native characteristics for predicting links of the networks. The time constraint in global approach is to traverse all paths of the graph structure for the link predictions. Local approaches are less efficient as they take an account solely local features of the node. This approach as a compression of all approaches provides efficient and accuracy in new friend suggestions in an exceedingly less interval of time.

## 5. REFERENCES

- [1] M. BM and H. Mohapatra, "Human centric software engineering," *International Journal of Innovations & Advancement in Computer Science (IJIACS)*, vol. 4, no. 7, pp. 86-95, 2015.
- [2] H. Mohapatra, *C Programming: Practice*, Vols. ISBN: 1726820874, 9781726820875, Kindle, 2018.
- [3] H. Mohapatra and A. Rath, *Advancing generation Z employability through new forms of learning: quality assurance and recognition of alternative credentials*, ResearchGate, 2020.
- [4] H. Mohapatra and A. Rath, *Fundamentals of software engineering: Designed to provide an insight into the software engineering concepts*, BPB, 2020.
- [5] V. Ande and H. Mohapatra, "SSO mechanism in distributed environment," *International Journal of Innovations & Advancement in Computer Science*, vol. 4, no. 6, pp. 133-136, 2015.
- [6] H. Mohapatra, "Ground level survey on sambalpur in the perspective of smart water," *EasyChair*, vol. 1918, p. 6, 2019.
- [7] H. Mohapatra, S. Panda, A. Rath, S. Edalatpanah and R. Kumar, "A tutorial on powershell pipeline and its loopholes," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 4, 2020.
- [8] H. Mohapatra and A. Rath, "Fault tolerance in WSN through PE-LEACH protocol," *IET Wireless Sensor Systems*, vol. 9, no. 6, pp. 358-365, 2019.
- [9] H. Mohapatra, S. Debnath and A. Rath, "Energy management in wireless sensor network through EB-LEACH," *International Journal of Research and Analytical Reviews (IJRAR)*, pp. 56-61, 2019.
- [10] H. Mohapatra and A. Rath, "Fault-tolerant mechanism for wireless sensor network," *IET Wireless Sensor Systems*, vol. 10, no. 1, pp. 23-30, 2020.
- [11] H. Mohapatra and A. Rath, "Detection and avoidance of water loss through municipality taps in india by using smart tap and ict," *IET Wireless Sensor Systems*, vol. 9, no. 6, pp. 447-457, 2019.
- [12] M. Panda, P. Pradhan, H. Mohapatra and N. Barpanda, "Fault tolerant routing in heterogeneous environment," *International Journal of Scientific & Technology Research*, vol. 8, pp. 1009-1013, 2019.
- [13] D. Swain, G. Ramkrishna, H. Mahapatra, P. Patra and P. Dhandrao, "A novel sorting technique to sort elements in ascending order," *International Journal of Engineering and Advanced Technology*, vol. 3, pp. 212-126, 2013.
- [14] H. Mohapatra, "HCR using neural network," 2009.
- [15] V. Nirgude, H. Mahapatra and S. Shivarkar, "Face recognition system using principal component analysis & linear discriminant analysis method simultaneously with 3d morphable model and neural network BPNN method," *Global Journal of Advanced Engineering Technologies and Sciences*, vol. 4, p. 1, 2017.
- [16] R. Kumar, S. Edalatpanah, S. Jha, S. Gayen and R. Singh, "Shortest path problems using fuzzy weighted arc length," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, pp. 724-731, 2019.
- [17] R. Kumar, S. Jha and R. Singh, "A different approach for solving the shortest path problem under mixed fuzzy environment," *International Journal of fuzzy system Applications*, vol. 9, no. 2, pp. 132-161, 2020.
- [18] R. Kumar, S. Jha and R. Singh, "Shortest path problem in network with type-2 triangular fuzzy arc length," *Journal of Applied Research on Industrial Engineering*, vol. 4, pp. 1-7, 2017.
- [19] S. Broumi, A. Dey, M. Talea, A. Bakali, F. Smarandache, D. Nagarajan, M. Lathamaheswari and R. Kumar, "Shortest path problem using Bellman algorithm under neutrosophic environment," *Complex & Intelligent Systems*, vol. 5, pp. 409--416, 2019.
- [20] R. Kumar, S. Edalatpanah, S. Jha, S. Broumi, R. Singh and A. Dey, "A multi objective programming approach to solve integer valued neutrosophic shortest path problems," *Neutrosophic Sets and Systems*, vol. 24, pp. 134-149, 2019.
- [21] R. Kumar, A. Dey, F. Smarandache and S. Broumi, "A study of neutrosophic shortest path problem," in *Neutrosophic Graph Theory and Algorithms*, F. Smarandache and S. Broumi, Eds., IGI-Global, 2019, pp. 144-175.
- [22] R. Kumar, S. Edalatpanah, S. Jha and R. Singh, "A novel approach to solve gaussian valued neutrosophic shortest path problems," *International Journal of Engineering and Advanced Technology*, vol. 8, pp. 347-353, 2019.
- [23] R. Kumar, S. Edaltpanah, S. Jha, S. Broumi and A. Dey, "Neutrosophic shortest path problem," *Neutrosophic Sets and Systems*, vol. 23, pp. 5-15, 2018.
- [24] R. Kumar, S. Edalatpanah, S. Jha and R. Singh, "A Pythagorean fuzzy approach to the transportation problem," *Complex and Intelligent System*, vol. 5, pp. 255-263, 2019.
- [25] J. Pratihari, R. Kumar, A. Dey and S. Broumi, "Transportation problem in neutrosophic environment," in *Neutrosophic Graph Theory and Algorithms*, F. Smarandache and S. Broumi, Eds., IGI-Global, 2019, pp. 176-208.