A
Summer Internship Report
On
"Link Prediction for Social Networks"

(CE446 – Summer Internship - II)

Prepared by

Jay Asodariya(17CE004), Prit Gopani(17CE034), Yash Makadia(17CE053)

Under the Supervision of

Prof. Minal Shah &
Prof. Aayushi Chaudhari

Submitted to

Charotar University of Science & Technology (CHARUSAT)
for the Partial Fulfillment of the Requirements for the
Degree of Bachelor of Technology (B.Tech.)
for Semester 7

Submitted at

Accredited with Grade A by NAAC
Accredited with Grade A by KCG

U & P U. PATEL DEPARTMENT OF COMPUTER ENGINEERING
(NBA Accredited)
Chandubhai S. Patel Institute of Technology (CSPIT)
Faculty of Technology & Engineering (FTE), CHARUSAT
At: Changa, Dist: Anand, Pin: 388421.
2020

अमृतं तु विद्या

CHARUSAT

CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

Accredited with Grade A by NAAC
Accredited with Grade A by KCG

CERTIFICATE

This is to certify that the report entitled “**Link Prediction for Social Network**” is a bonafied work carried out by **Jay Asodariya (17ce004) , Prit Gopani (17ce034) , Yash Makadia (17ce053)** under the guidance and supervision of **Prof. Minal Shah** & **Mr. Bhavin Satashiya** for the subject **Summer Internship – II (CE446)** of 7th Semester of Bachelor of Technology in **Computer Engineering** at Chandubhai S. Patel Institute of Technology (CSPIT), Faculty of Technology & Engineering (FTE) – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred by the examiner(s).

Under the supervision of,

Minal..

Prof. Minal Shah
Asst. Prof.
U & P U. Patel Dept. of Computer Engineering
CSPIT, FTE, CHARUSAT, Changa, Gujarat

FOR, CODE INFOSYS
Satas
PROPRIEROR

Mr. Bhavin Satashiya
Project Manager
Development Department
Code Infosys

Dr. Ritesh Patel
Head - U & P U. Patel Department of Computer Engineering,
CSPIT, FTE, CHARUSAT, Changa, Gujarat.

Chandubhai S. Patel Institute of Technology (CSPIT)

Faculty of Technology & Engineering (FTE), CHARUSAT

At: Changa, Ta. Petlad, Dist. Anand, Pin: 388421. Gujarat.

CODE INFOSYS

455- 456, Royal Arcade, Sarthana, Varachha Road, Surat
🌐 www.codeinfosys.com
📱 +91 960 1600 568

Date :21/06/2020

SUMMER INTERNSHIP COMPLETION CERTIFICATE

This is to certify that Mr. Jay Asodariya B.Tech (CE) student of Chandubhai S. Patel Institute of Technology, CHARUSAT, Changa has done a summer internship on “Link Prediction for Social Network” at CODE INFOSYS from 17th May to 17th June.

During this period of his internship program with us, he was found punctual, hardworking, in-squitive.

We wish him all the best for his bright career.

Thanking you.

FOR, CODE INFOSYS

Satas

PROPRIEROR

Bhavin Satashiya
CODE INFOSYS

CODE INFOSYS

455- 456, Royal Arcade, Sarthana, Varachha Road, Surat
www.codeinfosys.com
+91 960 1600 568

Date :21/06/2020

SUMMER INTERNSHIP COMPLETION CERTIFICATE

This is to certify that Mr. Prit Gopani B.Tech (CE) student of Chandubhai S. Patel Institute of Technology, CHARUSAT, Changa has done a summer internship on “Link Prediction for Social Network” at CODE INFOSYS from 17th May to 17th June.

During this period of his internship program with us, he was found punctual, hardworking, insquitive.

We wish him all the best for his bright career.

Thanking you.

FOR, CODE INFOSYS

PROPRIEROR

Bhavin Satashiya
CODE INFOSYS

CODE INFOSYS

455- 456, Royal Arcade, Sarthana, Varachha Road, Surat
🌐 www.codeinfosys.com
📱 +91 960 1600 568

Date :21/06/2020

SUMMER INTERNSHIP COMPLETION CERTIFICATE

This is to certify that Mr. Yash Makadia B.Tech (CE) student of Chandubhai S. Patel Institute of Technology, CHARUSAT, Changa has done a summer internship on “Link Prediction for Social Network” at CODE INFOSYS from 17th May to 17th June.

During this period of his internship program with us, he was found punctual, hardworking, in-squitive.

We wish him all the best for his bright career.

Thanking you.

FOR, CODE INFOSYS

PROPRIEROR

Bhavin Satashiya
CODE INFOSYS

ACKNOWLEDGMENTS

With immense pleasure, We would like to present this project report on “**Link Prediction for Social Networks**” for Code Infosys at Surat. Simply put, We could not have done this work without the help we have received cheerfully from guide of Code Infosys.

As a student of Chandubhai S Patel Institute of Technology (CSPIT), We are highly thankful to **Dr. Ritesh Patel**(Head of Department CSPIT, Charusat University, Changa.who allowed me to work at Code Infosys.

We would specially like to thank Mr. Bhavin Satashiya for proving the nice ideas to work upon. His guidance showed us the way not only to understand the how we can work from home

Lastly, I convey my regards to the whole staff, which made my stay at Extended IT Arms Solutions., a memorandum part of life.

ABSTRACT

Work From Home Summer Internship is an integral part of the CE (B.Tech) program. The objective of such an exercise is to get a first-hand exposure to the realities of the IT segment and gain an insight into the working of the corporate world and develop technical and communicational skills. The summer internship gives an opportunity to become aware of the student’s strengths and weakness as required for potential IT experts.

It has been observed that practical knowledge plays a vital role than the theoretical knowledge. This is so because practical knowledge gives chance to express our own ideas, way of thinking and different thoughts.

The project allotted to us by our guide at CODE INSOYS was “Link Prediction” and training about other technologies and implementations of it. The first part of the report comprises of Organization details which includes all the information of company profile other part consists of the projects studies.

The project study was carried on topic of “Link Prediction” for CODE INSOYS (Gujarat) located in Surat. The Link Prediction for social networks is an implementation of facebook’s Kaggle competition. It predicts the future possible link between two nodes. We have used ensemble learning approach and used Random Forest classifier to predict the missing edges between two nodes. The Confusion matrix and ROC curve at the end show the accuracy of the implementation.

Table of Contents

Acknowledgement i

Abstract ii

Description of company / organization……………………………………………v

Chapter 1 Introduction to Link Prediction 1

1.1 An Overview of Social Network Analytics...1

1.2 What is Link Prediction?...1

1.3 A Primer on Link Predictions...2

1.4 Link Prediction in Science……………………………………………………3

Chapter 2 Insights on Social Networks...5

2.1 Data Collections 5

2.2 Some Insights on Social Networks...5

Chapter 3 Link Prediction Part 1 10

3.1 Understand the Data……………………………………………………………10

3.2 Data Visualization………………………………………………………………11

3.3 Data Preparation…...……………………………………………………………13

Chapter 4 Link Prediction Part 2 14

4.1 Feature Engineering……………………………………………………………14

4.2 Model Building…………………………………………………………………17

4.3 Evaluation……………………………………………………………………18

Conclusions 21

References 22

List of Figures

Fig. 1.1 Connected Graph 2

Fig. 1.2 Connected Edges 2

Fig. 1.3 Human Metabolism 3

Fig. 1.4 Link connections in Bio Networks 4

Fig. 2.1 Sample Network(i) 5

Fig. 2.2 Sample Network(ii) 6

Fig. 2.3 Sample Network(iii) 7

Fig. 2.4 Sample Network(iv) 8

Fig. 2.5 The Lone Wolf 8

Fig. 2.6 The Social Butterfly 9

Fig. 3.1 Network Visualization 12

Fig. 3.2 Missing Edges 13

Fig. 4.1 Weakly Connected Components 15

Fig. 4.2 Random Forest 18

Fig. 4.3 Confusion Matrix 19

Fig. 4.4 ROC Curve 20

Fig. 4.5 AOC Curve 20

DESCRIPTION OF COMPANY

NAME OF THE COMPANY	**“CODE INFOSYS”**
ADDRESS	455, Royal arcade, opposite of Deepkamal mall, Sarthana Jakat Naka, Surat, Gujarat 395006
TELEPHONE	+91-9601600568
E-MAIL	satashia_bhavin@yahoo.com
LOCATION	Sarthana Jakat Naka, Surat
NATURE OF THE COMPANY	The nature of company is Software Development.
TYPE OF THE COMPANY	Private Limited Company
MAIN BUSINESS	Main business of the company is developing software for clients.
DEVELOPMENT AREAS	Android Development, IOS Development Website Development, Photoshop Design, Data Science

We code infosys formulate the business ideas with effective approach to app development and uninterrupted customer engagement. Also Provide End to end service for web development, android, iPhone, iPad, windows, PhoneGap, Native & Hybrid.

- Code Infosys building apps since 2012, we work towards bridging the latest technology with your business
- Ensures long-term partnerships with our clients with face to face communication, daily updates and faster results
- Enhance Your Efficiency and by Hiring the Expert dedicated developer
- Competitive budget and on time delivery is our plus point
- Solutions with current trends and practices of the IT industry

and commitment of the team enables us to offer comprehensive services and solutions that satisfy and delight our customers.

CHAPTER 1: INTRODUCTION TO LINK PREDICTION

1.1 An Overview of Social Network Analytics

What is a Social Network?

"A social network is essentially a representation of the relationships between social entities, such as people, organizations, governments, political parties, etc."

The interactions among these entities generate unimaginable amounts of data in the form of posts, chat messages, tweets, likes, comments, shares, etc. This opens up a window of opportunities and use cases we can work on.

That brings us to **Social Network Analytics (SNA).**

Few key benefits:

- Helps you understand your audience better
- Used for customer segmentation
- Used to design Recommendation Systems
- Detect fake news, among other things

1.2 What is Link Prediction?

Link prediction is used to predict future possible links in the network (E.g., Facebook). Or, it can be used to predict missing links due to incomplete data (E.g., Food-webs – this is related to sampling that Olivia spoke of earlier).

Social networks are popular way to interpret the interaction among the people. They can be visualized as graphs, where a vertex corresponds to a person and edge represent the connection between them.

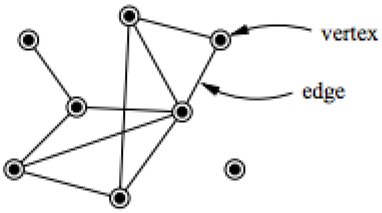


Fig 1.1: Connected Graph

Vertex = Person and edge = connection

These connections are usually made based on their(peoples) mutual interest. However, social networks are very dynamic, since new edges and vertices are added to the graph over time. Understanding the dynamics that drive the evolution of social network is a complex problem due to a large number of variable parameters. But, a comparatively easier problem is to understand the association between two specific nodes.

1.3 A Primer on Link Predictions

The objective of link prediction is to identify pairs of nodes that will either form a link or not in the future.

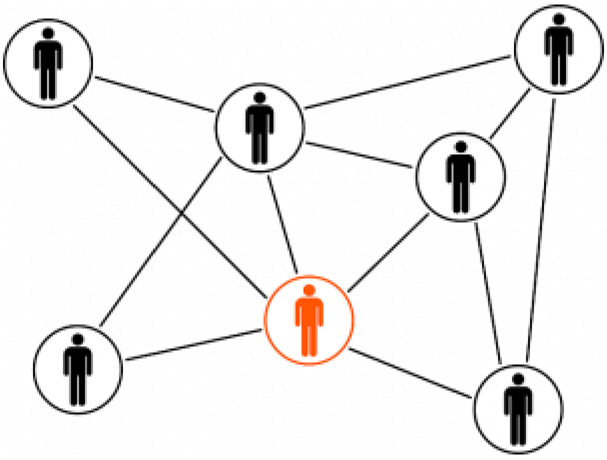


Fig 1.2: Connected edges

Link prediction has a ton of use in real-world applications. Here are some of the important use cases of link prediction:

- Predict which customers are likely to buy what products on online marketplaces like Amazon. It can help in making better product recommendations
- Suggest interactions or collaborations between employees in an organization
- Extract vital insights from terrorist networks

1.4 Link Prediction in Science

Artist's rendition of Human Metabolic network

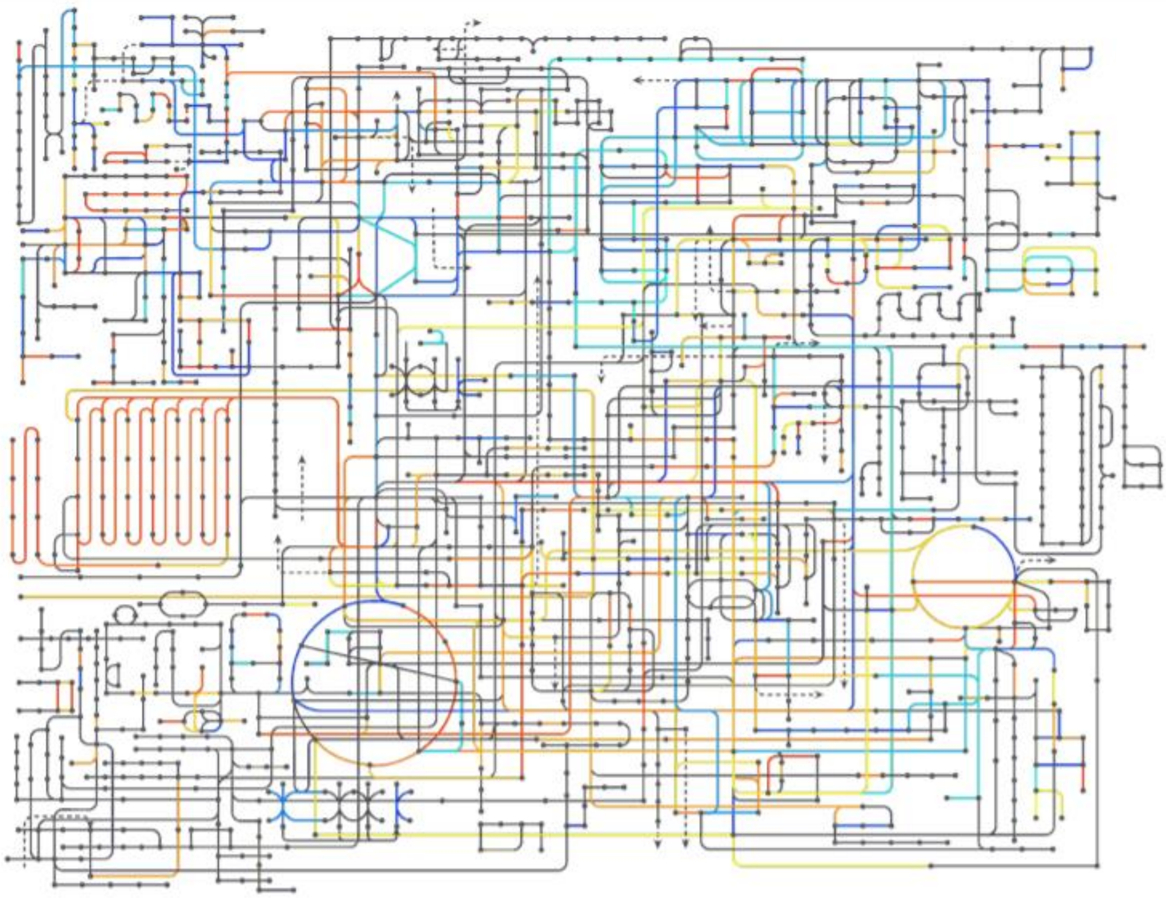


Fig 1.3: Human Metabolism

Investigating link connections in bio. networks is costly and time consuming.

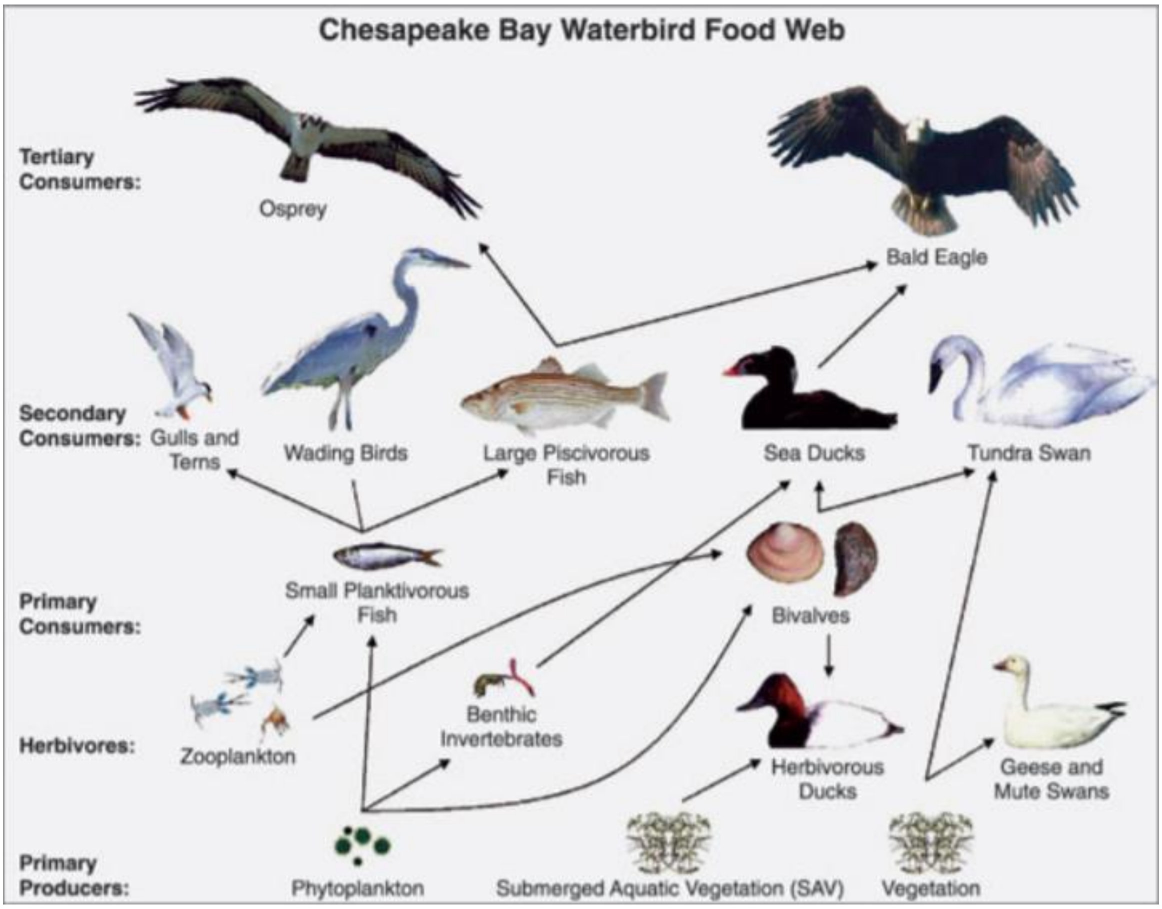


Fig 1.4: Link Connections in Bio Networks

CHAPTER 2: INSIGHTS ON SOCIAL NETWORKS

2.1 Data Collection

A couple years ago, Facebook launched a link prediction contest on Kaggle, with the goal of recommending missing edges in a social network graph. Social Networks mainly focus on building social relations among users who share common interests, background, real-life connections etc.

We have our 'train.csv' file initially which have source and destination node.

2.2 Some Insights on Social Networks

A sample of a small network is given below:

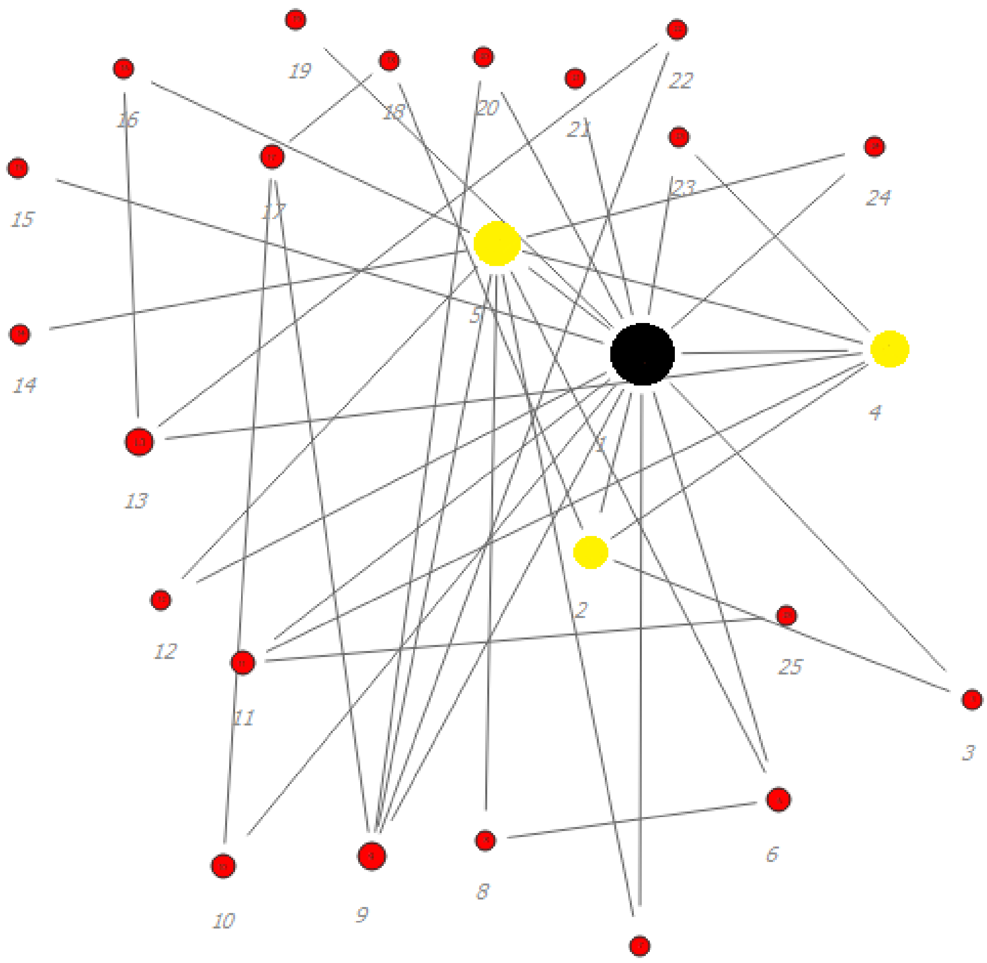


Fig 2.1: Sample Network (i)

The node in Black is the selected node from the training set connected to 3 other nodes in the network to unfold their local networks. The nodes are sized according to their number of connections.

We can see that our selected node is friends with 3 other nodes(in yellow), all having fairly large networks. Note that the edges between these nodes in this network are undirected i.e., edge between two nodes indicate both follow each other.

Since the above image does not depict the distinction between a follower and a followee, lets take a directed graph. Here, in this example, we can see the followers and following of a particular user.

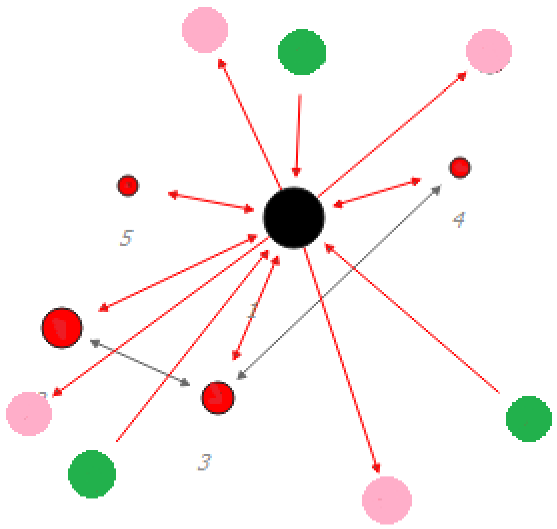


Fig 2.2: Sample Network (ii)

The selected user is the black node,it’s friends are the red nodes(users that follow and are followed back by the selected user), it’s followees are pink colored and it’s followers are green colored.

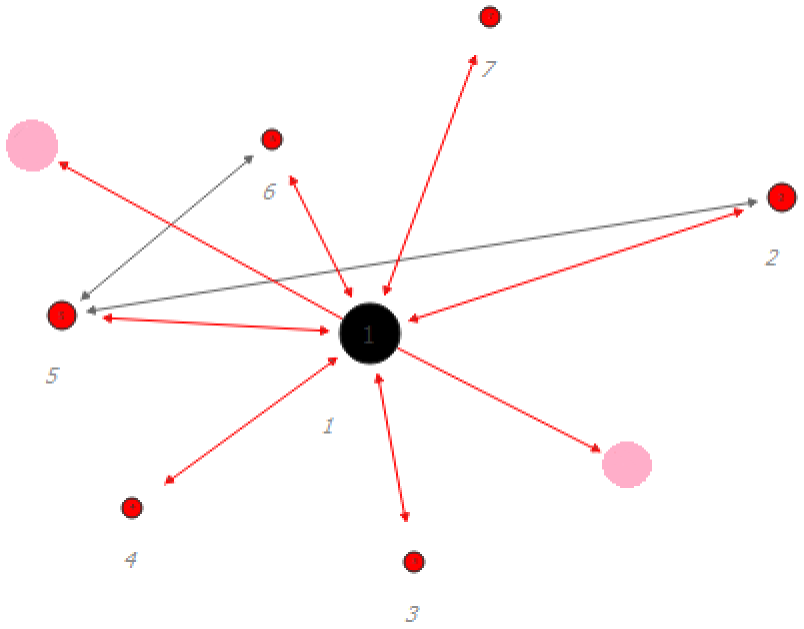


Fig 2.3: Sample Network (iii)

Here’s another network.

While the previous user had many only-followers(green nodes), this one has none. This can be considered as a node-level feature that indicates that the user is follower-hungry. And here’s another user who has nothing but followers(maybe a celebrity).

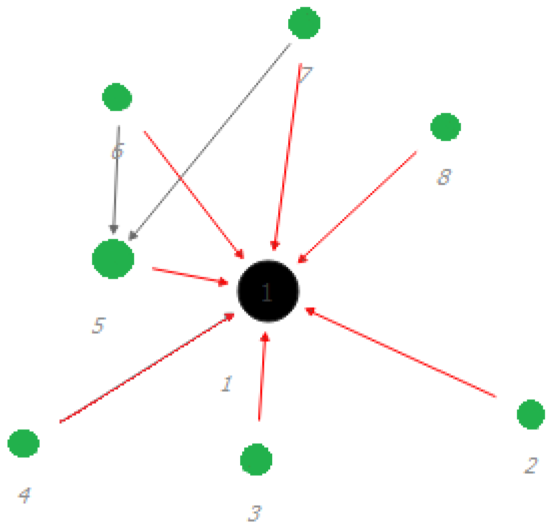


Fig 2.4: Sample Network (iv)

The Lone Wolf

Lets take a look at another graph whose network is small.

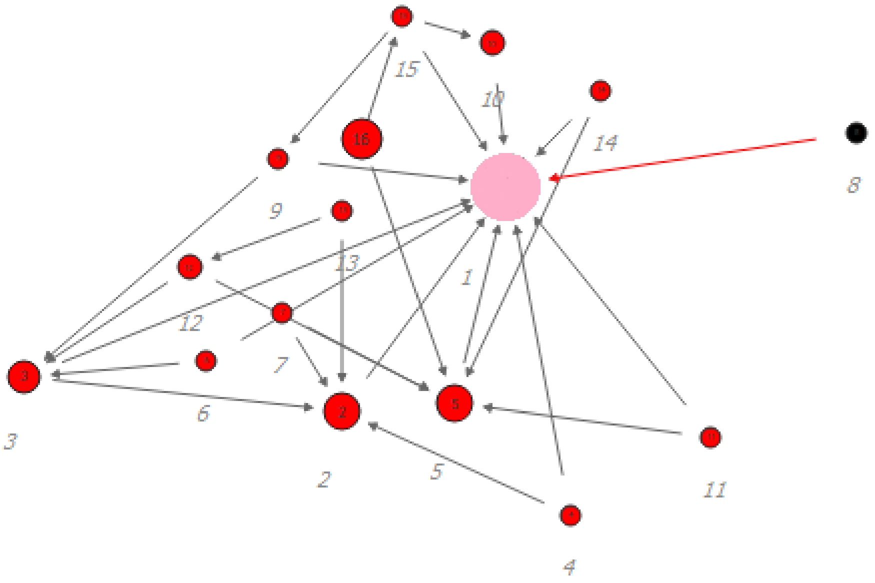


Fig 2.5: The Lone Wolf

The Social Butterfly

Whose network is a bit larger.

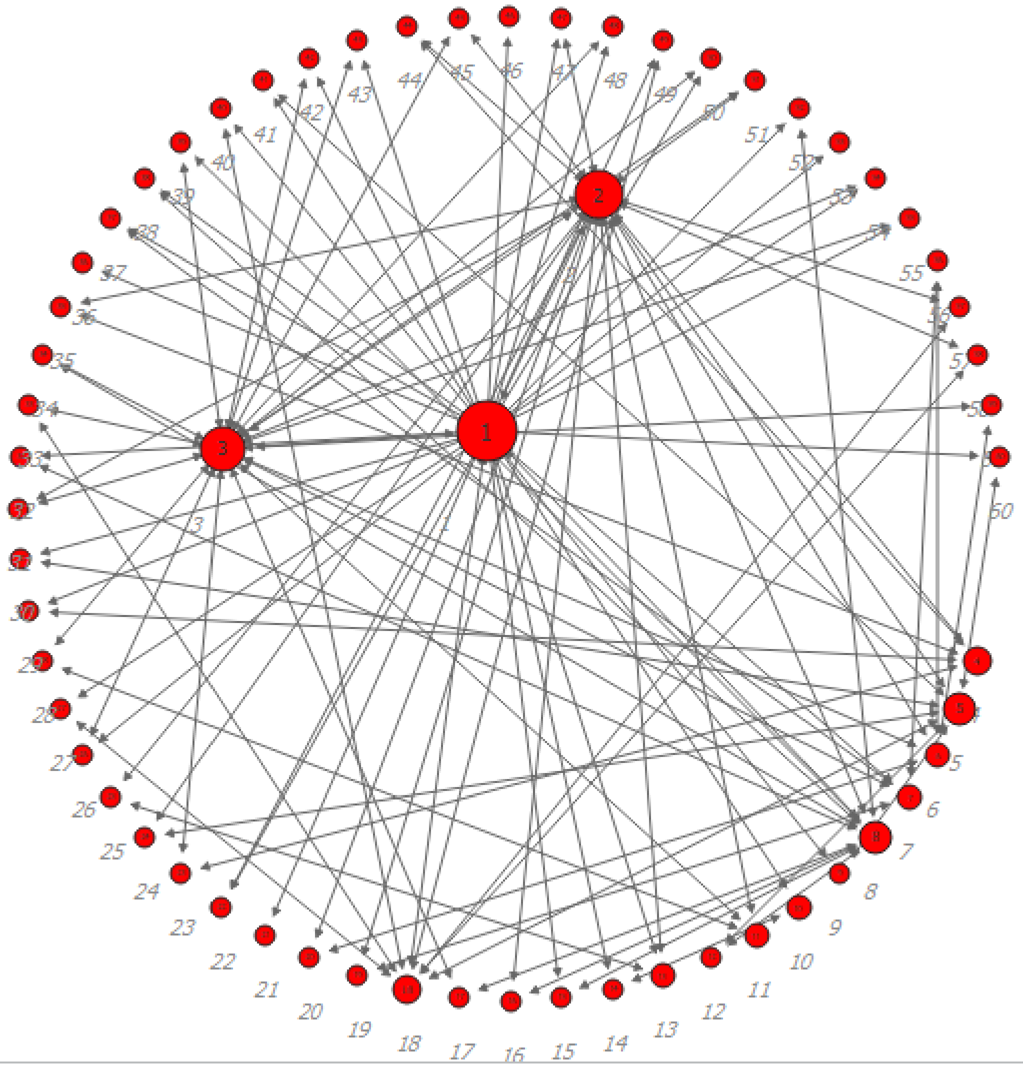


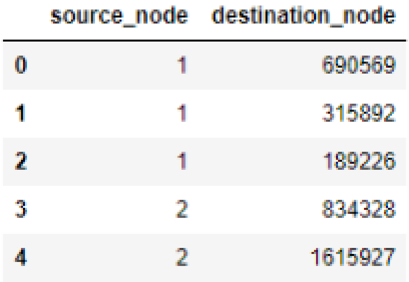
Fig 2.6: The Social Butterfly

CHAPTER 3: LINK PREDICTION PART 1

3.1 Understanding the Data

We are using train.csv file in which data of source node and destination node is given.

```
df=pd.read_csv('train.csv')
df.head()
```



	source_node	destination_node
0	1	690569
1	1	315892
2	1	189226
3	2	834328
4	2	1615927

Checking the data for any missing rows/ duplicates, if any.

```
sum(df.isna().any(1))
0
sum(df.duplicated())
0
```

There is no missing data and duplicate data. So we can say that it is a directed graph in which we are provided with two nodes; a source and a destination node. We have tried to visualize the given network using the NetworkX python library. NetworkX is a tool for creating, manipulating and perform study of structure of complex networks.

We have saved the train data by removing header and indexes. And Then by using NetworkX we have made a Digraph and then printed the information of that graph.

So, after that the total number of unique nodes in the network are 1862220 and the total number of edges in the network are 9437519.

Therefore, the total number of possible edges/links/connections in this network could be 1862220 C 2. Out of these, we are provided with only 9437519. The remaining 1862220 C 2– 9437519 edges do not exist in the network.

What do we have to predict?

It is like a binary classification problem. It takes a set of features from the users and maps them to ‘1’ if there exists a link between a pair and ‘0’ otherwise.

So, we have created an indicator variable for link which will be ‘1’ if there is a link between two nodes and ‘0’ if there is no link between the two nodes.

We will be randomly sampling 9437519 from it to get a balanced data. Therefore, the final size of the data will be 9437519 x 2 = 18875038.
So, our final data will have 9437519 rows with indicator variable ‘1’ and 9437519 rows with indicator variable ‘0’.

3.2 Data Visualization

We have taken the sample of 50 data to view the Network using NetworkX library.

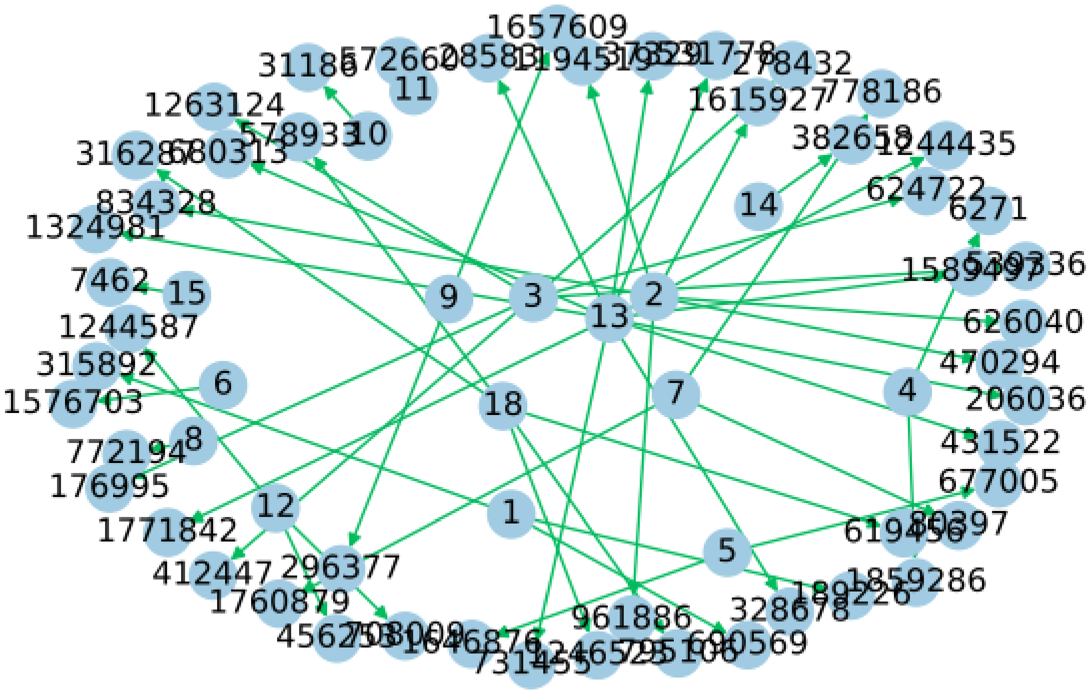


Fig 3.1: Network Visualization

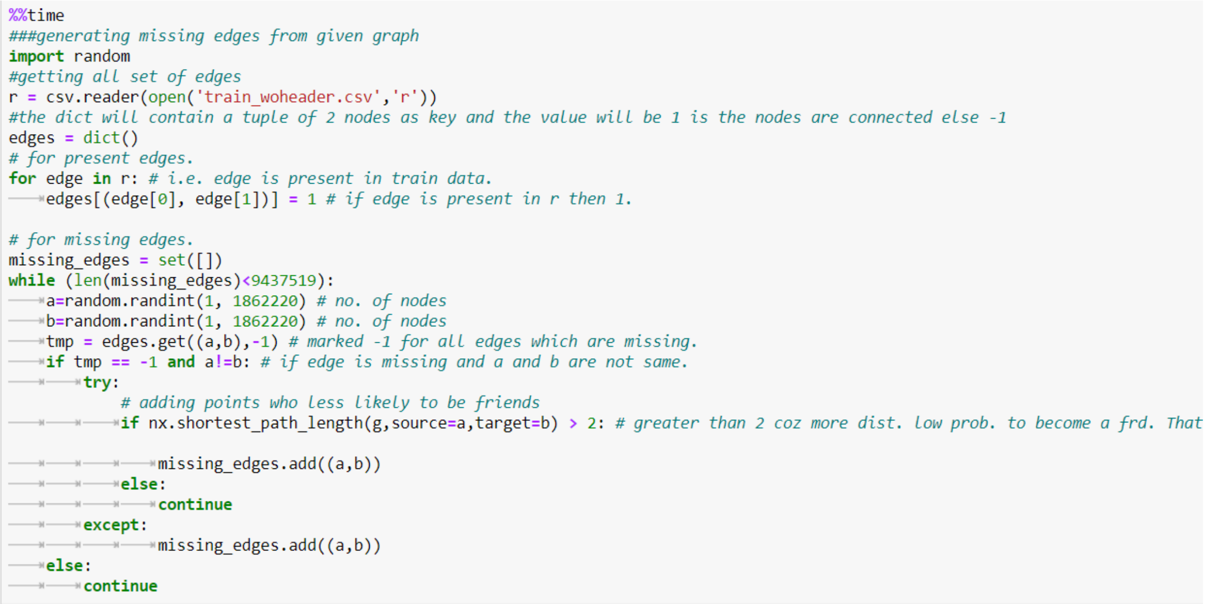
Then we had done the distribution of number of followers of each node in the training set as well as the number of followees of each nodes. We got the following observations:

1. Most of the users in this network have followers in the range of 40 to 50.
2. The maximum number of followers by a user is 552.
3. The number of followees for most users fall in the range of 40–50.
4. Maximum is 1566.

14% of the users in our data have 0 followees and 10% of users have 0 followers.

3.3 Data Preparation

Now we have to generate those missing edges.



```
%%time
###generating missing edges from given graph
import random
#getting all set of edges
r = csv.reader(open('train_woheader.csv','r'))
#the dict will contain a tuple of 2 nodes as key and the value will be 1 is the nodes are connected else -1
edges = dict()
# for present edges.
for edge in r: # i.e. edge is present in train data.
    edges[(edge[0], edge[1])] = 1 # if edge is present in r then 1.

# for missing edges.
missing_edges = set([])
while (len(missing_edges)<9437519):
    a=random.randint(1, 1862220) # no. of nodes
    b=random.randint(1, 1862220) # no. of nodes
    tmp = edges.get((a,b),-1) # marked -1 for all edges which are missing.
    if tmp == -1 and a!=b: # if edge is missing and a and b are not same.
        try:
            # adding points who less likely to be friends
            if nx.shortest_path_length(g,source=a,target=b) > 2: # greater than 2 coz more dist. low prob. to become a frd. That

                missing_edges.add((a,b))
            else:
                continue
        except:
                missing_edges.add((a,b))
    else:
        continue
```

Fig 3.2: Missing edges

We are creating a dictionary in which a pair of nodes will be the key and the value will be ‘1’ if there exists a link between them and ‘-1’ otherwise.

So, we are running a while loop to select exactly 9437519 number of missing edges and adding them into a set named ‘missing_edges’.

We first randomly select two nodes from the data and check if the edge between them is present in the dictionary that we created and we check if the two randomly chosen nodes are not the same node(a node cannot be linked with itself).

Now, we check if the shortest path between two nodes is greater than two.

CHAPTER 4: LINK PREDICTION PART 2

4.1 Feature Engineering

1. Jaccard Distance

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

The jaccard Index measures the similarity between finite sets, and is defined as the size of intersection divided by the union of the sample sets.The Jaccard distance which measures the similarity between sample sets is complimentary to jaccard index and is obtained by subtracting the jaccard index by 1.

$$d_J\ (A,B) = 1 - J(A,B)$$

2. Cosine Distance

$$CosineDistance = \frac{|X \cap Y|}{SQRT(|X| \cdot |Y|)}$$

Cosine distance is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space.

3. Page rank

PageRank (PR) is an algorithm used by Google Search to **rank** web **pages** in their search engine results. **PageRank** was named after Larry **Page**, one of the founders of Google.

PageRank works by counting the number and quality of links to a **page**(nodes in this case) to determine a rough estimate of how important the website(node) is.

4.Shortest path

Shortest path is the path between two nodes such that the sum of their weights is minimum.

5. Weakly connected components

A particular component is said to be strongly connected if there is at least one path from any given node to any other node.

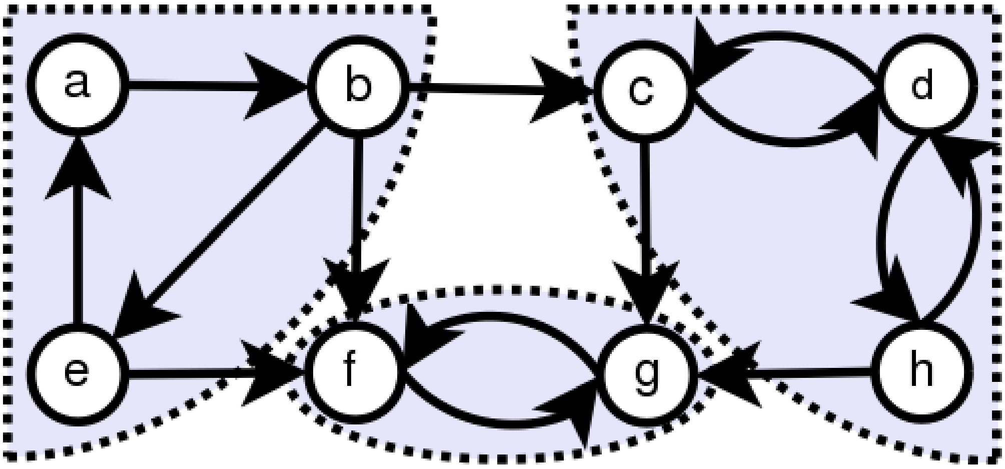


Fig 4.1: Weakly Connected Components

In the first component (consisting of a,b,c) every node is reachable from every other node in that component. We can go from a →b, e →a,b →e and b →a via e.

A directed graph is weakly connected if replacing all its directed edges with undirected edges. Therefore, every strongly connected component is a weakly connected component.However, if it is not a strongly connected component, then to check whether it is a weakly connected component remove the directions of the edges and see if still there is atleast one path from any given node to any other node.

6. Adar Index

Adamic/Adar measures is defined as inverted sum of degrees of common neighbors for given two vertices.

$$A(x,y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log |N(u)|}$$

This metric measures the closeness of two nodes based on their shared neighbors.A value 0 indicates that two nodes are not close, while higher values indicate nodes are close.

7. Follow back

This is a simple feature to find out if a node follows back after being followed by a node.

8. Katz Centrality

Katz centrality of a node is a measure of centrality in a network. Unlike typical centrality measures which consider only the shortest path between a pair of actors, Katz centrality measures influence by taking into account the total number of walks between a pair of actors.

It is similar to Google’s PageRank.

9.HITS

Hyper-link induced topic search (HITS) identifies good authorities and hubs for a topic by assigning two numbers to a node : an authority and a hub weight. Authorities estimate the node value based on the incoming links. Hubs estimates the node value based on outgoing links.

4.2 Model Building

Now we will train our machine learning model using Random Forest Classifier.

Random Forest Classifier

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working of Random Forest Algorithm

We can understand the working of Random Forest algorithm with the help of following steps −

- **Step 1** − First, start with the selection of random samples from a given dataset.
- **Step 2** − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- **Step 3** − In this step, voting will be performed for every predicted result.
- **Step 4** − At last, select the most voted prediction result as the final prediction result.

The following diagram will illustrate its working −

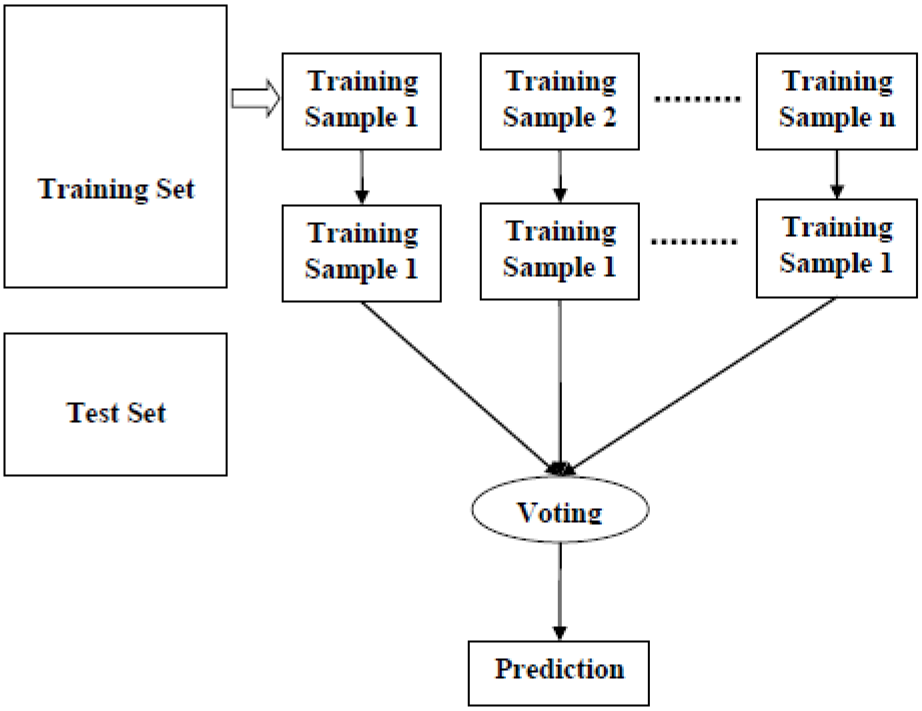


Fig 4.2: Random Forest

4.3 Evaluation

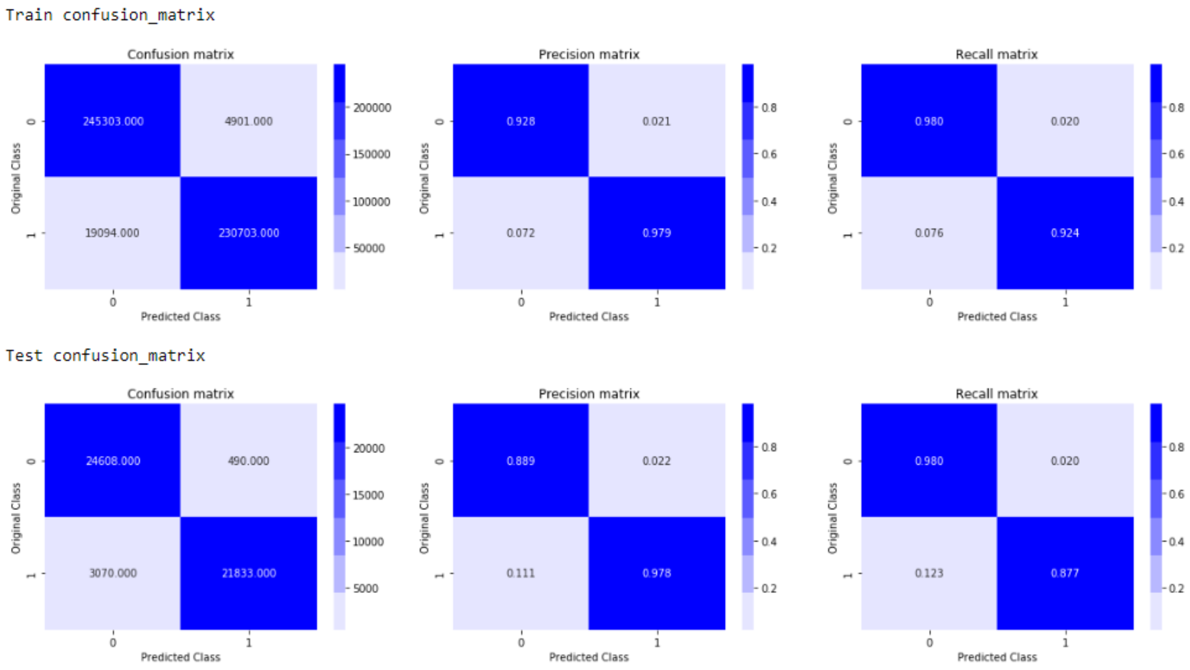
F1 score

The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

Confusion Matrix

A confusion matrix is often used to describe the performance of a classification model. is a summary of prediction results on a classification problem.

The number of correct and incorrect predictions are summarized with count values and broken down by each class.



Train confusion_matrix

Test confusion_matrix

Fig 4.3: Confusion Matrix

Observations:

1. From the test results, 490 of the points we falsely classified as ‘negative’ where as they were actually ‘positive’ points.
2. Similarly, 3070 were wrongly classified as ‘positive’

ROC curve

ROC curve is another way for measuring the performance of a classification model. Higher the area under the curve(AUC), better the model is at predicting 0s as 0s and 1s as 1s.

The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. The objective is to maximize area under the curve(AUC).

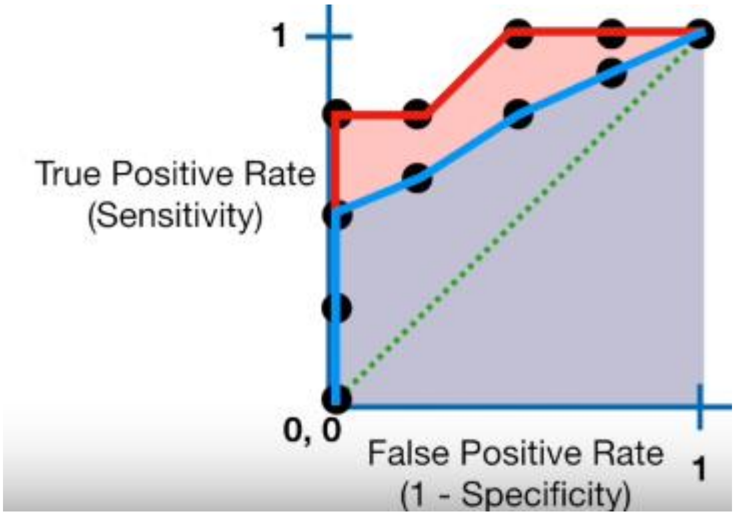


Fig 4.4: ROC Curve

This figure consists of two ROC curves(red one and a blue one). The AUC for the red curve is greater than the AUC for the blue curve. Suppose the red curve was for Random Forest and the blue one was for logistic regression, we will choose the Random Forest as our winning model as it gives a better AUC.

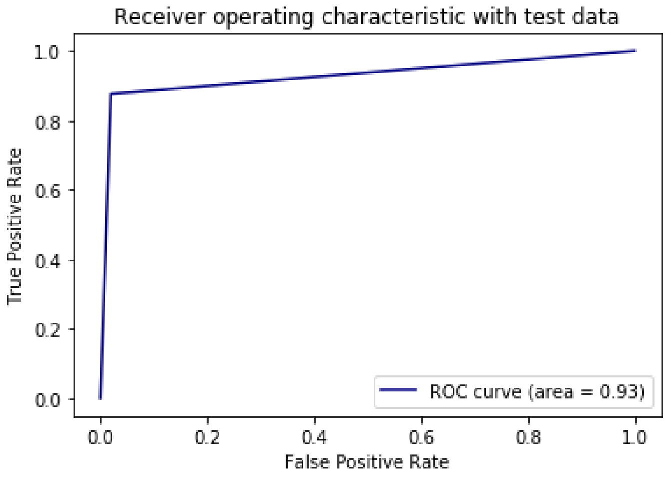


Fig 4.5: AUC Curve

CONCLUSIONS

On the whole, this internship was a useful experience. We have gained new knowledge, skills and met many new people. We have achieved several of our learning goals, however for some the conditions did not permit. We got insight into professional practice. We learned the different facets of working within company. We experienced that financing, as in many organizations, is an important factor for the progress of projects. This was a challenging work from home internship.

Related to our study we have learned more about working in team. There is still a lot to learn and to improve. The internship was also good to find out what our strengths and weaknesses are. This helped us to define what skills and knowledge we have to improve in the coming time. It would be better that the knowledge level of the language is sufficient to contribute fully to projects. I think that I could start my working career after completing my engineering. It would also be better if we can present and express myself more confidently. At last this internship has given us new insights and motivation to pursue a career in data science.

REFERENCES

i) https://networkx.github.io/documentation/stable/

ii) https://www.kaggle.com/c/FacebookRecruiting/data

iii) https://medium.com/@vgnshiyer/link-prediction-in-a-social-network-df230c3d85e6

iv) https://hackernoon.com/link-prediction-in-large-scale-networks-f836fcb05c88

v) https://ethz.ch/content/dam/ethz/special-interest/gess/computational-social-science-dam/documents/education/Spring2017/ML/LinkPrediction.pdf

vi) https://www.analyticsvidhya.com/blog/2020/01/link-prediction-how-to-predict-your-future-connections-on-facebook/

vii) https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm