Link Prediction in a Social Networks

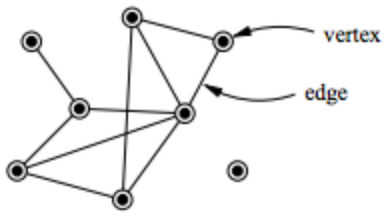
CE:350 Data Warehousing and Data Mining

Yash Makadia(17CE053), Ayushi Patel(17CE069)

Abstract

1. Introduction

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.



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.

Types of graph

There are two types of graph in graph theory.

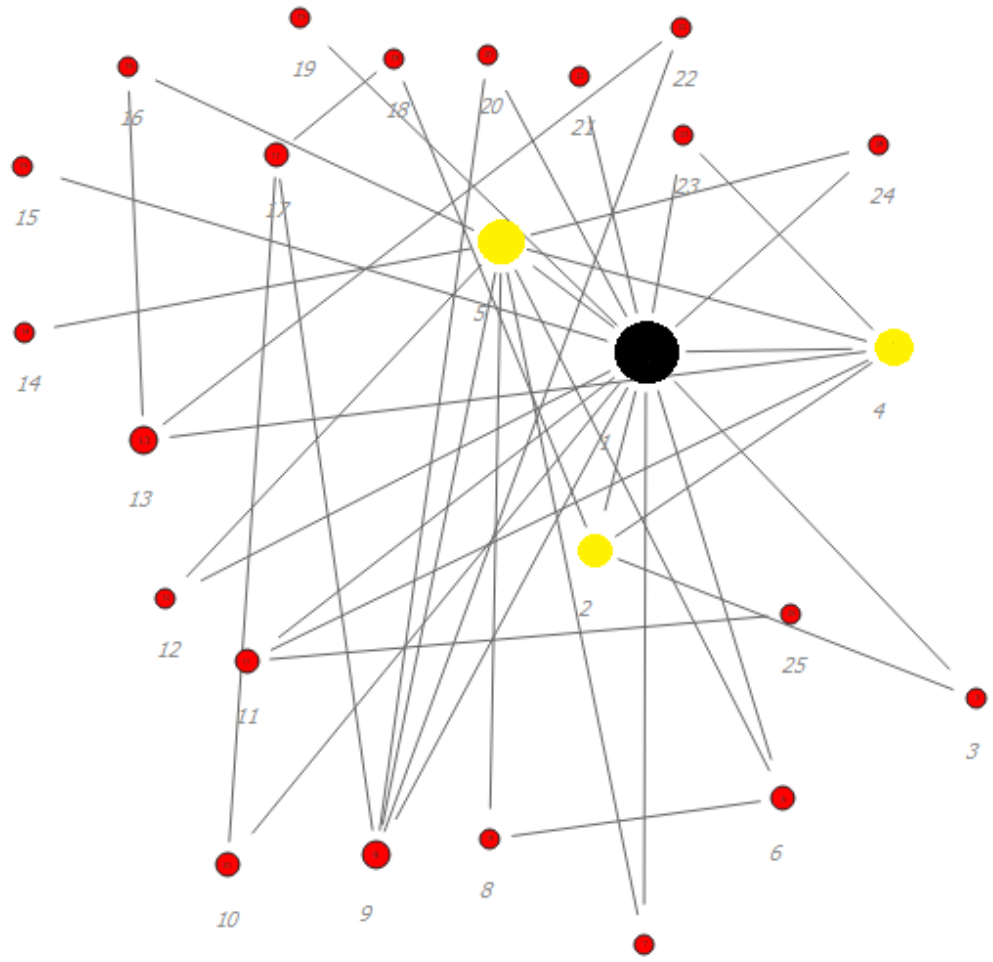
1. Undirected
2. Directed.

1. Data Collection

We have downloaded datasets from Kaggle. We only need tra

2. Some Insights on Social Networks

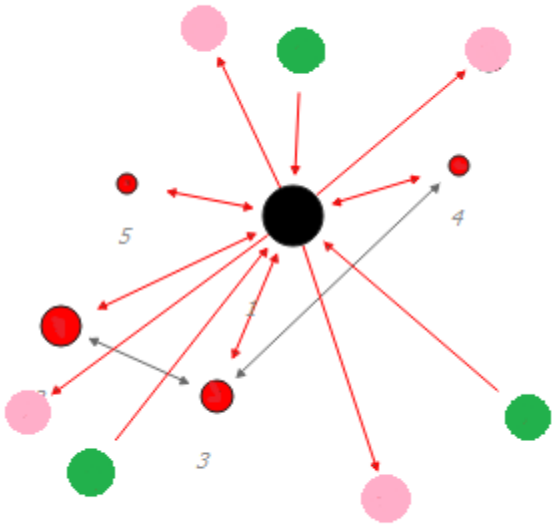
A sample of a small network is given below:



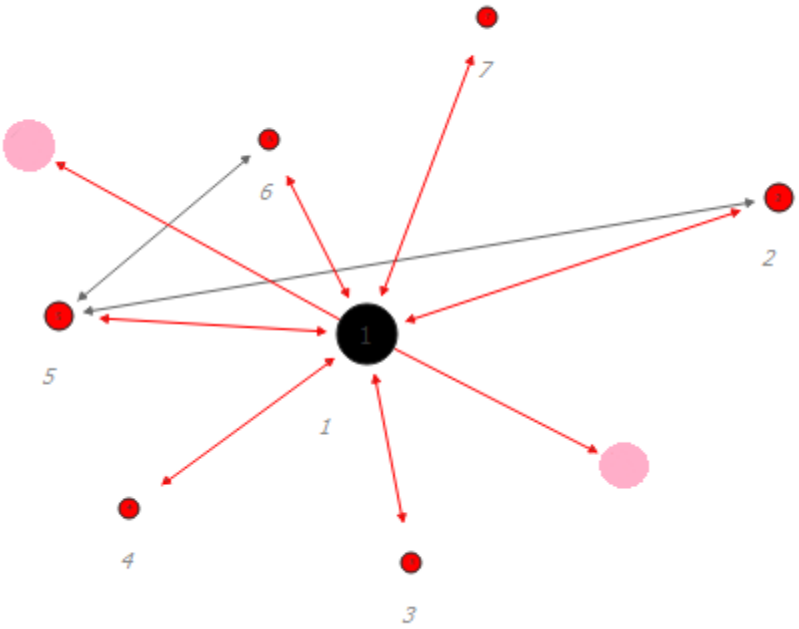
The node in Black is the selected node from the training set connected to 3 other nodes in the network to unfold their local networks. The nodes are sized according to their number of connections.

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.



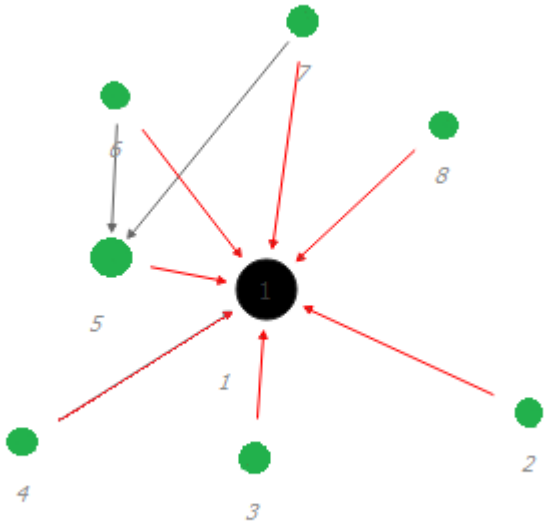
The selected user is the black node,it's friends are the red nodes(users that follow and are followed back by the selected user), it's followees are pink colored and it's followers are green colored.



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).



3. 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. I 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.

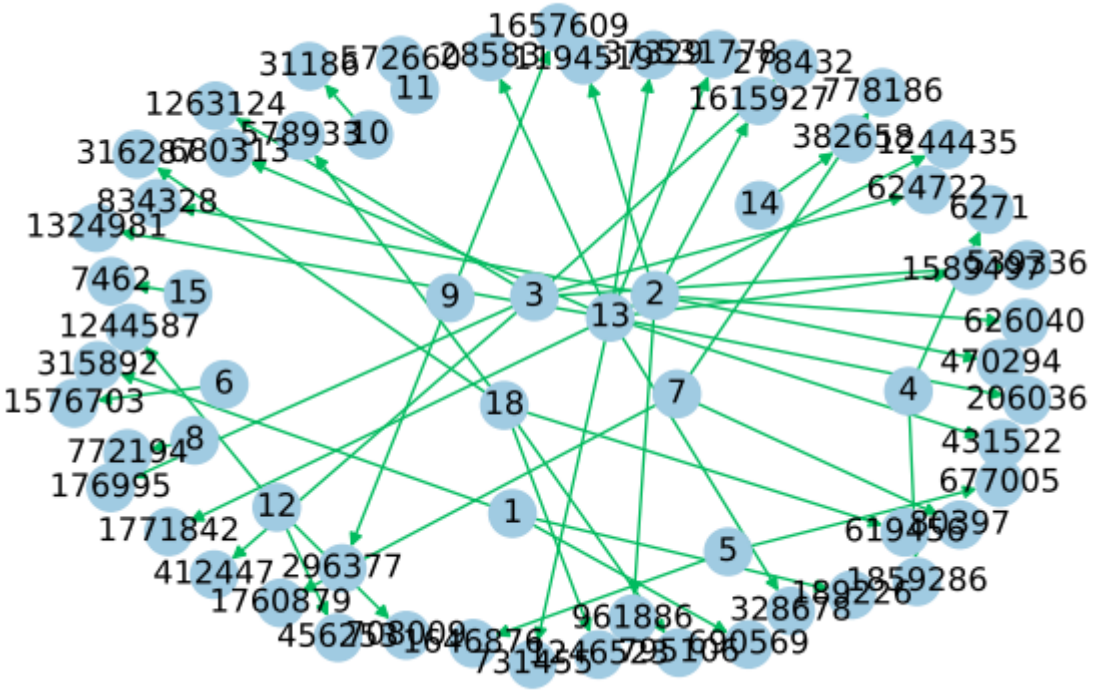
So, we have created an indicator variable for link which will be ‘1’ if there is a link between two nodes and ‘0’ if there is no link between the two nodes.

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’.

4. Data Visualization

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



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.

5. Data Preparation

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.

So look at the code for better understanding.

6. 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.

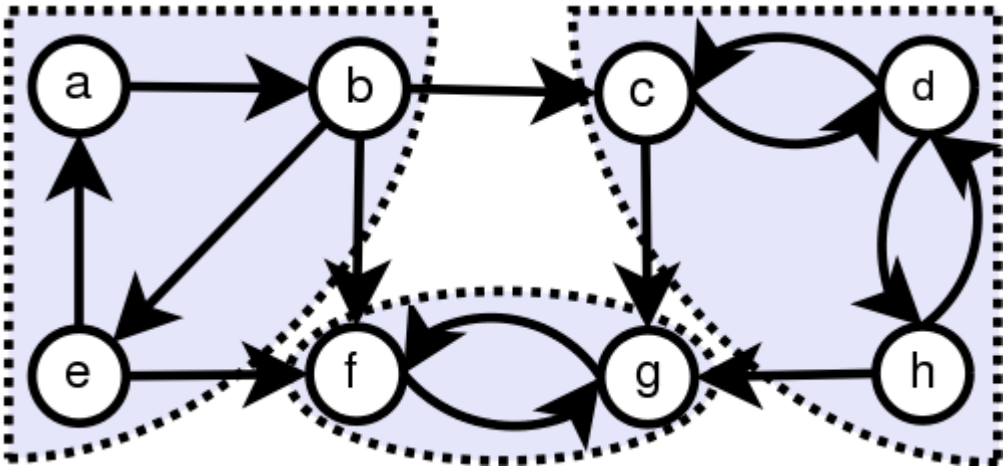
PageRank works by counting the number and quality of links to a page(nodes in this case) to determine a rough estimate of how important the website(node) is.

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.



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.

7. Model Building

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

Random Forest Classifier

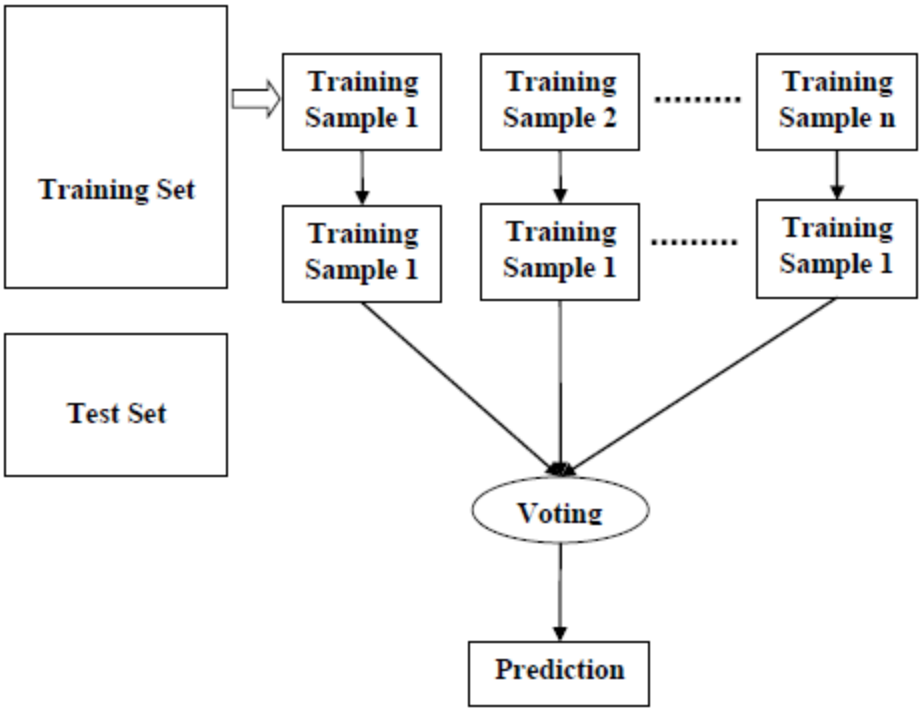
Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision 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 −



8. Evaluation

F1 score

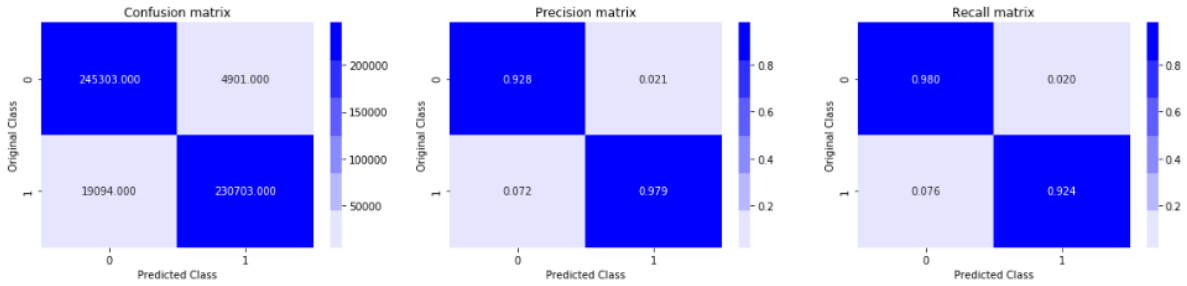
The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

Confusion Matrix

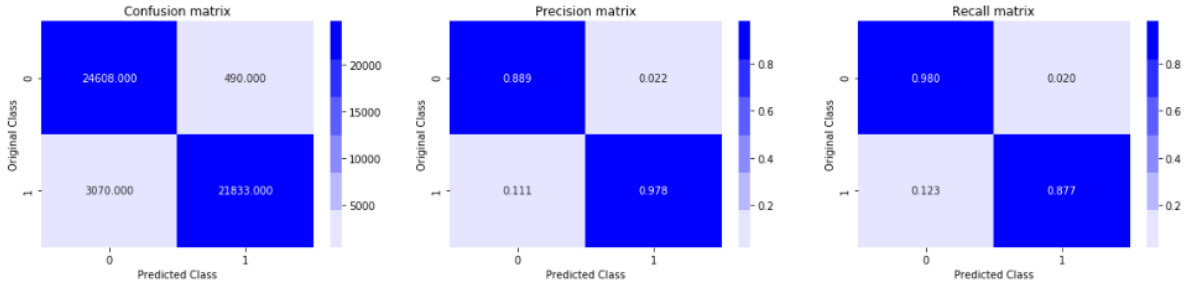
A confusion matrix is often used to describe the performance of a classification model. is a summary of prediction results on a classification problem.

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

Train confusion_matrix



Test confusion_matrix



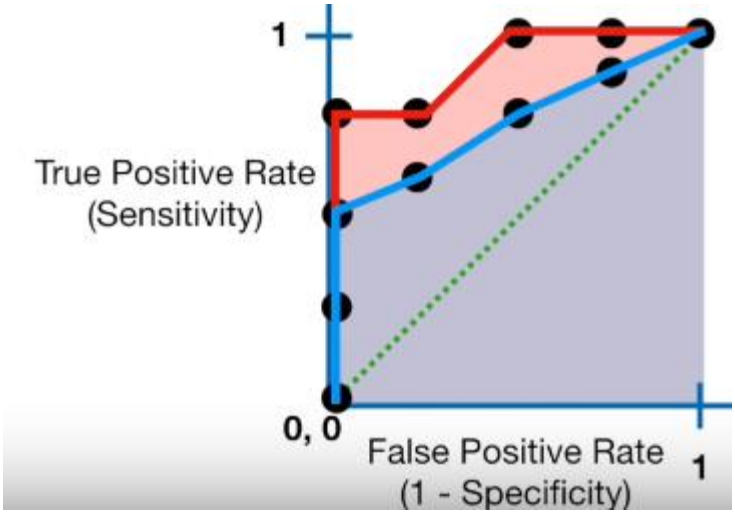
Observations:

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’

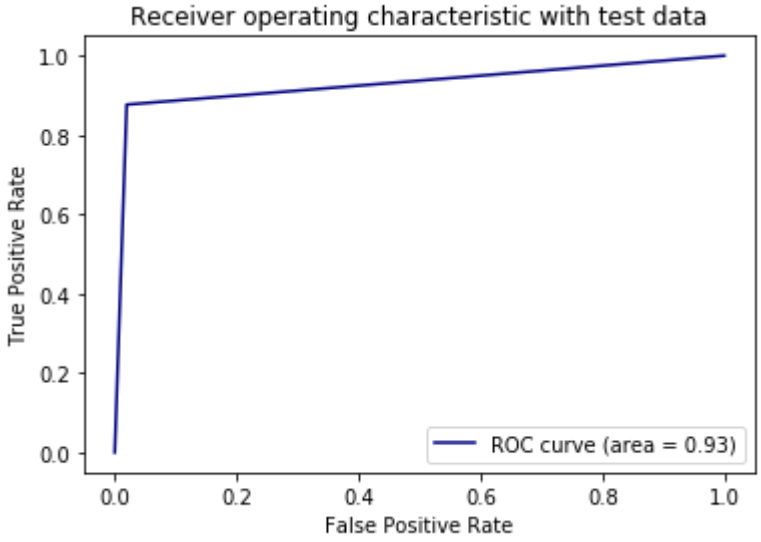
ROC curve

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).



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.



9. 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