GitHub Link :
https://github.com/dheerajreddykomandla/Feature-Engineering

CSCE 5222
Feature Engineering
Project Proposal

1. Project Title and Members:

- **Project Title: Feature Engineering on Facebook Dataset**
- **Team Number :** Team 15
- **Team Members:**
  Dheeraj Reddy Komandla (11526265)
  Prem Sai Vuppula (11528103)
  Sravan Boinapalli (11553755)
  Viveksen Naroju (11611662)

2. Idea Description:

The idea in this project is about feature engineering, it involves adding new features by extracting hidden information from the existing data. Some of the popular feature engineering techniques are One Hot Encoding, TFIDF, and Word2Vec Etc., Here in this project we would like to select a unique dataset and extract the important insights from the data.

3. Goals and Objectives:

We as a team did lot of Research on this project, when we are going through the project we found couple of challenging tasks. In this project we have choose to perform feature engineering on graph dataset.

The Graph dataset contains only 2 columns source and destination. Here source and destination represents the id's of each social media user. Each row represents the source id following the destination id. It is really where interesting and challenging part to do feature engineering with 2 columns dataset.

4. Motivation

Generally feature engineering is mostly performed on the structured data or unstructured data like image and audio files. We can find lot of popular techniques available in the internet. Our main motivation in the project is taking a challenging dataset and performing feature engineering on the dataset, because it has only two columns of data. We as a team felt that it will be very interesting to extract hidden features from the dataset having only two columns.

5. Significance:

Here, in this project the concept is straightforward, we are using feature engineering drastically where we can justify the importance of that. It is a Process of making decisions from the data and to ensure the model is working correctly. In this project data source and the ways the data has been processed and managed in Unique, because the dataset is having only two columns. Extracting few feature from the dataset is very important for building predictive models in machine learning, it is proven that machine learning models do perform well when we can able extract the hidden information from the dataset. It is not just only for predictive analysis it is also a key factor for descriptive analysis. Based on the Extracted features we can perform exploratory data analysis (EDA) and provide significant insights. Such analysis will be helpful to take Business Decisions.

6. Literature Survey:

In this project we are using kaggle dataset. The dataset was organically provided by the Facebook. This kaggle repository is actually a challenge given by the Facebook to recommend the friends based on the existing friend connections. Each row in the dataset is nothing but the friend connection from source to destinations. Since the connection is like followers and followees, we assume the dataset is related to Instagram.

As we did research, our team found some of the existing solutions for the above dataset the links for those are given below.

https://www.kaggle.com/code/genialgokul1099/social-network-graph-link-prediction

https://www.kaggle.com/code/curioso/link-prediction-facebook

https://www.kaggle.com/code/ajaysh/stackoverflow-tag-prediction

https://www.kaggle.com/code/vohoangbaoduy/lab3-link-prediction

7. Objectives:

The main objective in this project is to built the feature engineering to the dataset. In general based on the dataset structure and type we have different kinds of feature engineering techniques,

Example 1: If the given dataset is of type text then we will perform the feature engineering techniques like TFIDF, Word2Vec.

Example 2: If the dataset is of type image we will convert the image into numpy array.

Example 3: If the dataset is of type audio file then we have special feature engineering technique called NFCC Features.

In the same way for the graph dataset, we have some graph data specific features engineering techniques. Our main objective is to implement those feature engineering on our Facebook graph dataset. In the part of our research we found some of the graph feature engineering techniques.

- jaccard followers
- jaccard followees
- cosine followers
- cosine followees
- number of followers source
- number of followees source
- number of followers destination
- number of followees destination
- is following back

- shortest path between source and destination

These are the initial feature Engineering techniques we would like to implement on top the given Facebook dataset. If you find future more techniques we will implement those techniques as well.

8. Features:

As we already discussed above, we are working on the graph dataset. General graph dataset Will have nodes and edges. Hence in our dataset we have only two features. They are:

- Source
- Destination

The important point to be noted in the given graph dataset is that the edges are not bidirectional. This represents there is no rule that if a follows b and b follows a. These are some of the important characteristics.

9. Expected outcome:

As we already mentioned that we did some research and found some important graph feature techniques which are listed in the objective sections. Here we are expecting to implement all the above listed objectives, and extract the important hidden features by using above listed Feature Engineering techniques. We would also like to plot some features that are extracted which will help us to understand how the extracted features are behaving and its importance.

10. References:

https://towardsdatascience.com/feature-extraction-for-graphs-625f4c5fb8cd

https://www.learndatasci.com/glossary/jaccard-similarity/

https://www.geeksforgeeks.org/find-the-jaccard-index-and-jaccard-distance-between-the-two-given-sets/

https://www.geeksforgeeks.org/cosine-similarity/

https://datascience.stackexchange.com/questions/5121/applications-and-differences-for-jaccard-similarity-and-cosine-similarity

https://betterprogramming.pub/5-ways-to-find-the-shortest-path-in-a-graph-88cfefd0030f