Social Network Analysis of the Professional Community Interaction -Movie Industry Case

By:

22PD22 - Mithun Senthil V

22PD28 - Arun Prakash S

ABSTRACT:

Minimize the number of unsuccessful titles being released in OTT platforms (Netflix, Hulu, HBO Max, and Amazon Prime,etc.) by analyzing the movie industry community structure by creating a network graph with “actor”-“casting director”-“talent agent” - “director” as the nodes in the communication graph and show that usage of additional knowledge leads to better movie rating prediction.

(PDF) Social Network Analysis of the Professional Community Interaction -Movie Industry Case. Available from: https://www.researchgate.net/publication/356284173_Social_Network_Analysis_of_the_Professional_Community_Interaction_-Movie_Industry_Case

DATASETS:

The dataset used here is made from three major datasets. IMDb, Rotten Tomatoes, and IMDb Pro datasets.

Features in the dataset:

- **Poster_Link**: URL link to the movie's poster.
- **Series_Title**: Title of the movie or series.
- **Released_Year**: The year the movie was released.
- **Certificate**: Age rating of the movie (e.g., PG, PG-13).
- **Runtime**: The duration of the movie in minutes.
- **Genre**: Categories or types of the movie (e.g., horror, drama).
- **IMDB_Rating**: IMDb rating of the movie.
- **Overview**: A brief synopsis or description of the movie.
- **Meta_score**: Metascore rating from Metacritic.
- **Director**: Name of the director.
- **Star1, Star2, Star3, Star4**: Names of the main actors/actresses.
- **No_of_Votes**: Number of user votes on IMDb.
- **Gross**: Worldwide gross income of the movie.
- **Casting Director**: Name of the casting director.
- **IMDb_id**: Unique IMDb identifier for the movie.
- **TMDB_ID**: Unique identifier for The Movie Database (TMDb).
- **Budget**: Production budget of the movie.
- **Revenue**: Revenue generated by the movie.
- **Movie_Success**: Label indicating whether the movie was successful or not.

Movie industry Network Model:

Number of nodes: 3539
Number of edges: 13861
Average clustering coefficient: 0.83

Types of edges:

Director <—-> Actor link.
Because directors hire actors.
Star<—-> Star link.
Indicates both Star are participating in same movie.
Casting director <—> Star link.
Casting directors approach Stars, & them movie roles.
Director< —-> casting director link.
Directors hire casting directors.
—> (Unidirectional) <—> (Bidirectional)

The graph has a core-power law structure.
Top three person:

Francine Maisler (casting director for “The Usual Suspects” and another 76 titles)
Janet Hirshenson (casting director for “A Beautiful Mind”)
Robert De Niro (star for “Heat”).

Node2Vec Algorithm (Graph Random Walk):
Node2Vec is an algorithm which is used to create vectors of low-dimensions from the nodes in graph.
SInce almost all machine learning models need vectors as input. We have used Node2Vec algorithm for node embedding. It encodes each node into an embedding space while preserving the network structure information

We have planned to implement the Node2Vec algorithm for **networkx library(available in Py3)**

```
from node2vec import node2vec
```

Parameters of Node2Vec algorithm:

Node2Vec algorithm requires the following parameters to conduct random-walk over graph

We need:

- Number of walks
- Length of each walk
- Dimensions

In the research paper, they have fixed **number of walks=100** and its **lengths=10**, so we have decided to follow through with the same parametric values. So that we could build a **80-Dimensional vector** which fully encapsulates the attributes/feature of a single node in the graph network.

Machine learning Models:

Various Machine learning models can be used to predict whether the movies titles will be successful or not. The performance of the predictive models can be increased using the insights that we observed from the Social Network Model. We consider the director, actor, casting director and writers to see if these affect the result or not. Since the dataset here is unbalanced (the number of successful and unsuccessful titles in the dataset is not equal) it can be converted into a balanced dataset using the SMOTE algorithm.

For the prediction various types of predictive models can be used.

1. Gradient boosting
2. Random Forest Models
3. Decision Tree Model
4. Neural Network Models

These models are trained with the dataset and without SNA parameters and these models are compared with the models that are trained with the dataset and with SNA parameters to see if the addition of the SNA parameters increases the efficiency of the models.

The gradient boosting models gives a 3.07% gain due the addition of SNA parameters. Random Forest models gives a 3.13% loss due the addition of SNA parameters. The Decision Tree Model gives a 0.66% loss due the addition of SNA parameters. Neural Network Models gives a 4.49% gain due the addition of SNA

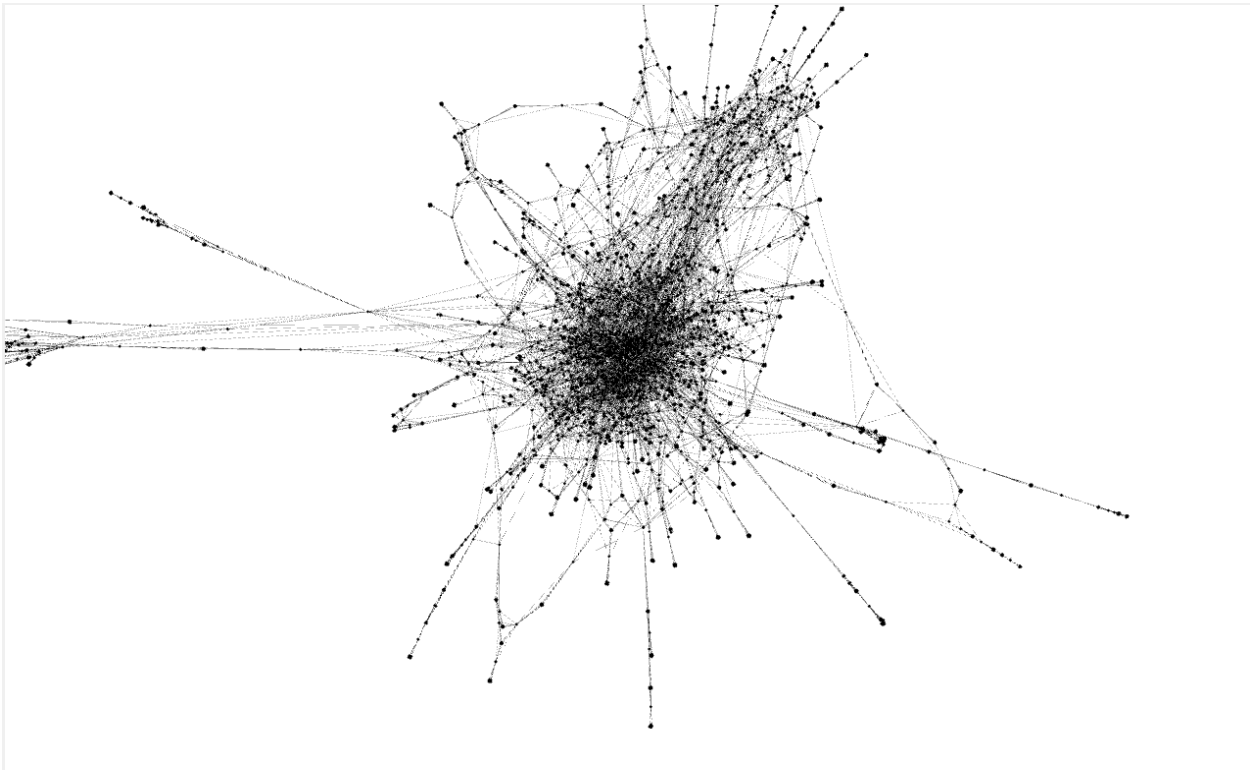
parameters. The neural network model that is trained with the dataset and SNA parameters has the highest accuracy of all the models that we saw before. The neural network model that is trained with the dataset and SNA parameter has a considerably higher accuracy than the other models. Overall the neural network model with SNA parameters give the highest accuracy of all the models that are consider with and without SNA parameters.

Results:

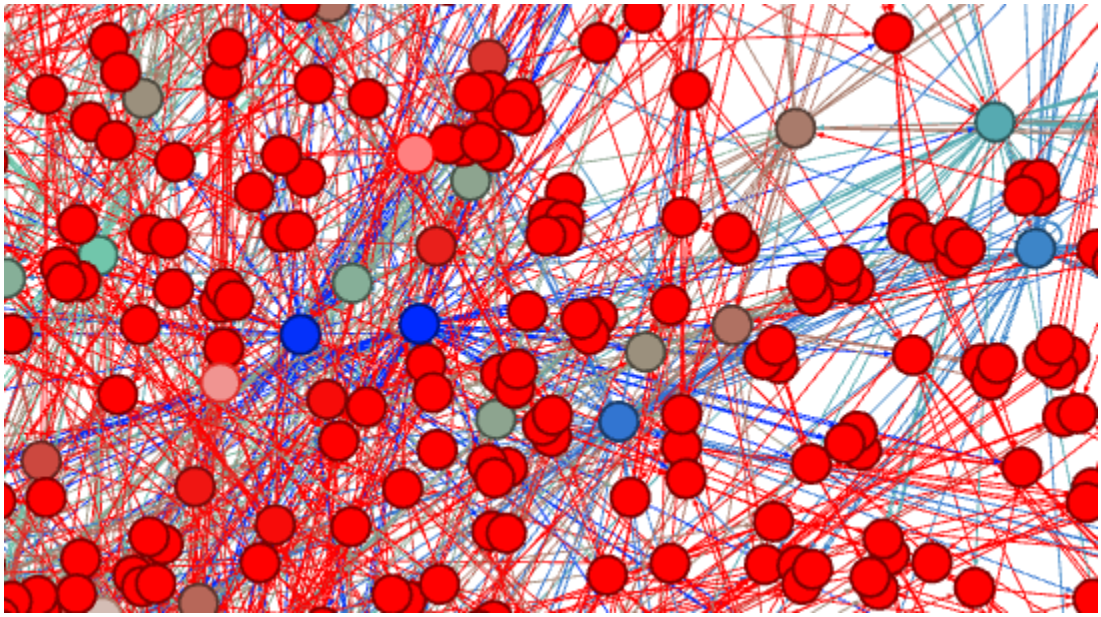
The result of the machine learning models with SNA parameters.
Random Forest: -3.13% gain in accuracy with embeddings
Gradient Boosting: 3.07% gain in accuracy with embeddings
Decision Tree: 0.66% gain in accuracy with embeddings
Neural Network: 4.49% gain in accuracy with embeddings

Visualization:

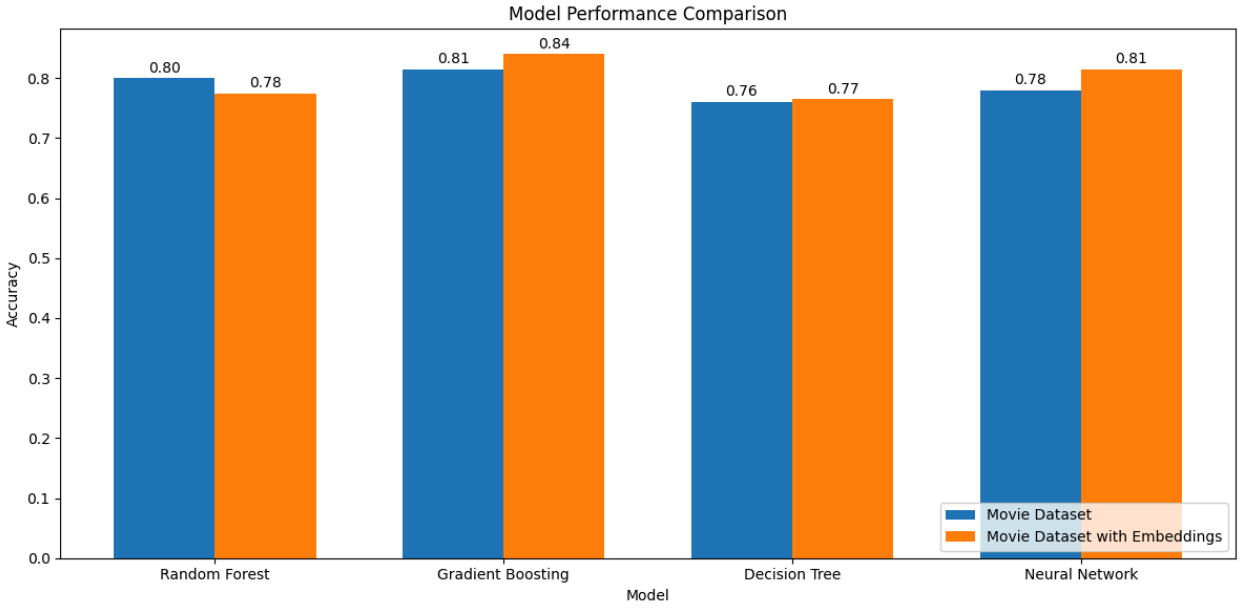
The movie industry graph in force atlas2 Layout



The graph with nodes colored, red color is used for low degree nodes, green for node with medium degrees and blue has very high degree.



The accuracy of the machine learning models visualized



Inference:

Dataset with embeddings generated by Node2Vec algorithm provides a model with more accuracy, compared to the dataset with no embedding.
We have embedded vectors of 80-dimensions for all actors in the graph, with the use of Node2Vec algorithm.

- If these vectors contain more negative values, it will affect cosine dissimilarity measure among different nodes in the network.
- Some mini-clusters are formed because, few people have only done a single movie together, and haven't participated in other movies or communities in the industry. These clusters does not affect the network majorly.
-The role of casting directors is really crucial in the movie and this has been reflected in our project.

Conclusion:

From our extensive reading through the paper, we inferred that the Movie Industry follows common social network properties that we are already familiar with, the movie industry follows a power-law distribution like most social networks. The features selected for our joint dataset also plays a real major role, they influence the IMDb rating of movies.It is important that we include crucial features of the directors and casting directors.Casting Directors play a vital role in a movies' success.

Our learnings from this were really useful, we were exposed to how social networks can be treated as a machine learning problem. New algorithms like the Node2Vec algorithm, which converts all nodes in social-network model to a n-dimensional vector, so that we could proceed with machine learning approach.
The unique data about casting directors and talent agents has been collected by scrapping web pages.The models used in this research paper were new & crucial. Random Forest Model is a really useful model used in machine learning and it will definitely be important added to it, Gradient-Boosting method, which is a great optimization technique is also used.