Social Network Analysis on the Relationship Between Actors for Predicting a Movie's Box Office Success

Jessica Nguyen | Jnguye84@gmu.edu
CDS Department | George Mason University

ABSTRACT

In this research project, I dove into the factors influencing financial outcomes in the movie industry by specifically focusing on who are the most significant actors that best predict a movie's box office success?

Many studies have found that neural networks increase the accuracy behind predicting variables to a box office success. Through social network analysis on the relationships between actors in movies and random foresting to validate my results, I found that it is beneficial to analyze the importance of an actor through their scores on betweenness centrality which is "is a way of detecting the amount of influence a node has over the flow of information in a graph"(neo4j). I also found that it's much harder to predict a flop movie than a decently earning movie or a blockbuster, with the prediction error rate of blockbusters being half of the flops.

My research findings would fit the lens of The Transforming Hollywood Conference hosted by Denise Mann and Henry Jenkins, professors from UCLA and USC, where their conference overview states that they focus on "the massive cultural-industrial shift underway as digital distribution platforms harness algorithmic technologies to green-light projects" (Transforming Hollywood). These research findings can provide valuable insights to movie producers and studios, guiding them in making informed casting decisions to optimize their film's box office success.

title	Summarized Stars	revenue	budget	profit	class
Toy Story	[Tom Hanks, Tim Allen, Don Rickles, Jim Varney...	373554033.0	30000000.0	343554033.0	2
Jumanji	[Robin Williams, Jonathan Hyde, Kirsten Dunst,...	262797249.0	65000000.0	197797249.0	2
Waiting to Exhale	[Whitney Houston, Angela Bassett, Loretta Devi...	81452156.0	16000000.0	65452156.0	1
Heat	[Al Pacino, Robert De Niro, Val Kilmer, Jon Vo...	187436818.0	60000000.0	127436818.0	2
Sudden Death	[Jean-Claude Van Damme, Powers Boothe, Dorian ...	64350171.0	35000000.0	29350171.0	1
...	...	...	...	...	...
Sivaji: The Boss	[Rajinikanth, Suman, Shriya Saran, Vivek, Ragh...	19000000.0	12000000.0	7000000.0	2
All at Once	[Andrey Muravyov, Yuliya Khlynina, Anton Shurt...	3.0	750000.0	-749997.0	0

LITERATURE REVIEW

Dezhou found through their linear analysis that actors were the best influence in social media towards movies (tweets about a movie, Instagram posts about a movie). They then used social network analysis to group actors based on their social media activity (10-dimensional measurement features include fans number, post number, followers, average post interval, etc. (Dezhou, 2006).

As mentioned before, Quader found that cast/star power amounted to 11.63 percent of the forecasting accuracy level of 48.41 percent overall. They separated box office successes into 5 classes where when seeing the accuracy of one class away, their neural network came out to 84.1 percent accuracy- a large improvement from their binary classes of not vs. is box office success. In my research, I separated it only into three classes (flops, decently earning movies, and blockbusters) which may be why Quader found a greater improvement than I did.

RESEARCH QUESTION/HYPOTHESIS

How Can I Use the Relationships of Actors within Hollywood To Predict a Movie's Box Office Success?

METHEDOLOGY

There was a retrieved Kaggle dataset of 10000 IMDB movies and their characteristics, including Genre, Production, Crew, Cast, Ratings, Revenue, Gross Income, and etc. First, it was string values that looked like dictionaries within the Cast column of the dataset. I then needed to change them into lists.

After that process, the 'Summarized Stars' column was pivoted into 36 rows where each row included a cast member with either a 1 or 0 for if they were present in the movie. These were binary classification variables. It amounted to over 50,000 columns of actor names as the columns, however, due to the CPU expense, I truncated my methods to only use the first 500 actor columns.

I also added a 'class' column within the dataset. There will be three classification factors. I determined this by first finding the profit of a movie, which was the revenue column minus the budget column. Afterward, I used the profit column and classified in this way: anything negative was 0 for flop, anything within the IQR was 1 for decently performing, and anything else was 2 for blockbuster hit.

0 (Flops):1609
1(Decently Preforming):1876
2(Blockbusters): 1908

Table 1: Summary Statistics on the Profits

Min.	1st Qu.	Median
Mean	3rd Qu.	Max.
-87982678	-5084484	4041495
44995927	44030899	857100000

I wanted to also focus on the top actors based on the movies that they've been in and what those movies have had in gross profits. It's interesting that 3 of the celebrities here were in the Harry Potter Franchise and probably made the list when they were only early adults.

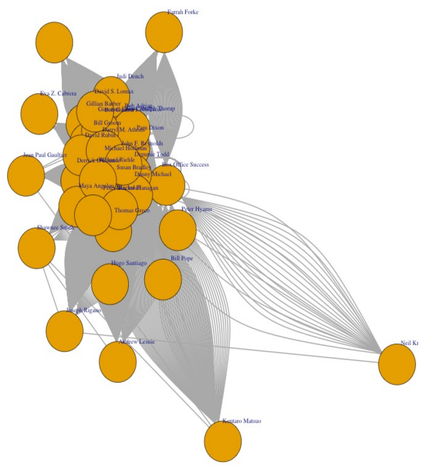
Table 2: After Performing Data Wrangling on Original Dataset

Rank	Total Gross Profit	Total IMDB$_{Rating}$
1	Joe Russo	Humphrey Bogart
2	Mark Ruffalo	Toshirō Mifune
3	Daniel Radcliffe	Emma Watson
4	Emma Watson	Charles Chaplin
5	Rupert Grint	Ethan Coen

In the past, I used just these variables to perform PCA, however, it was unsuccessful with no change in the prediction model. Therefore, I wanted to use social network analysis and find the metrics of between-centrality scores for each actor. Between-Centrality is, "is a way of detecting the amount of influence a node has over the flow of information in a graph"(neo4j). To do this, the nodes would represent actors and the links would be movies where those actors had worked together. I would get this information through the rows.

For instance, Margot Robbie was in Babylon with Leonardo DiCaprio. Therefore, two nodes would represent those actors and a link would connect them. Margot Robbie was also in Barbie with Rylan Gosling. If this was the entire social network, Margot Robbie would have the greatest between-centrality.

Then, I created a dictionary where the keys were the actors' names and the values were their between centrality scores. I would then go through every movie take note of the cast and sum up all their centrality scores. This was put in a new column named 'Total Celeb Betweenness'. Afterward, I would perform Random Forest to validate if it helped my model.



Before: OOB estimate of error rate: 60.89 percent

class	flop	decent	blockbuster	class.error
flop	16	40	68	0.8709677
decent	22	69	88	0.6145251
blockbuster	31	64	116	0.4502370

After: OOB estimate of error rate: 60.31 percent

class	flop	decent	blockbuster	class.error
flop	14	35	75	0.8870968
decent	14	80	85	0.5530726
blockbuster	35	66	110	0.4786730

DISCUSSION/CONCLUSIONS

I first performed Random Foresting on the model without the 'Total Celeb Betweenness'. This means that the features I did have were: belongs.to.collection, budget, popularity, revenue, vote.average, and vote.count.

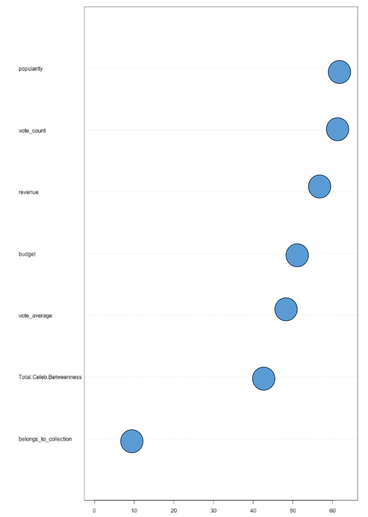
Compared to other research papers, a ~40 percent accuracy rating is not bad to start with. Within one study, the neural networks correctly predicted "58.41 percent exact match" while the other techniques such as the linear model predicted, "56.16 percent exact"\cite{Quader et al.}. Another study exclusively used neural networks to predict box office success and got an accuracy of 63.15 percent (Zhou et al. 2019).

RESULTS

OOB Estimate of Error Rate is 1-accuracy. This means it changed from **39.11 percent to 39.69 percent accurate**. **The 'Total Celeb Betweenness' column also scored a 60 on the MeanGini response, the same as the column 'vote_average'.** This is a very high score with the other ones being around 70-80. It makes sense that it improved the accuracy. If the collective sum of all the actors involved in the movie are well connected (have high betweenness centrality scores), they are most likely going to be well respected by audiences and more likely to have a positive influence on the box office revenue.

As we can see the accuracy scores for blockbusters are the highest in both scenarios as well as the decently made movies. This means that it's pretty easy to know when something will be a blockbuster, but not so much if it will be a flop. With the improved 'Total Celeb Betweenness' variable, the blockbuster prediction improved by 2 percent.

While the results were not as dramatic as I'd like them to be, I'm happy to have contributed to the expansion of using social networks to perform machine learning and data validation. In previous research, I hadn't seen any do it based on a movie's total betweenness centrality measures. The summing up into the 'Total Celeb Betweenness' method was a new one, and it showed that there is a positive correlation between high centrality measures and a movie being a box office success.



REFERENCES