History of movie success through GSP

Project Report – A Network Tour of Data Science

by CIONCA Alexandre, DONZIER Paul, DE GOUMOËNS Frédéric & FLUHR Hugo

1 Introduction

The movie industry was born more than a century ago in France. It went through multiple revolutions before becoming the cinema we know today. As a group of cinephiles, we thought it would be interesting to learn more about the evolution of this industry throughout the last century. More precisely, we want to see if successful movies produced at a given time period have the same characteristics as successful movies from another time period. To investigate this matter, we worked on the TMDb dataset. We focused particularly on genres, keywords, production companies and casts provided in this database. We built metrics to measure film success and attempted to compare the influence of different features on a film's success in different time periods.

2 Methods

Our overall strategy to investigate our dataset is as follows. First, we have to define a way to assess the success of a movie. To do so, we decided to consider two features of a movie: its weighted rating and its return on investment (ROI). The weighted rating (WR) is a quantity defined to take into account the vote average as well as the number of votes for each film. Indeed, a movie rating is relevant only if it has been rated by a certain number of users. Therefore, the WR will be brought closer to the mean over all movies if a movie has a small number of votes. The weighted rating and ROI are defined as:

$$WR = \frac{n_{votes}}{n_{votes}+n_{min}} R + \frac{n_{min}}{n_{votes}+n_{min}} C \qquad ROI = \frac{Revenue - Budget}{Budget}$$

where n_{vote} = number of votes for the movie, n_{min} = min number of votes to be considered, R = average rating of the movie and C = average rating computed across all movies. Additionally, the dataset is divided into subdatasets corresponding to four time periods covering the last century.[1]

1920–1960	Classical Hollywood cinema	**1960–1980**	New Hollywood cinema
1980–2000	Boom of bluckbusters movies	**2000–today**	Modern (digital) cinema

3 Data Exploration

Now that the metrics (**WR, ROI**) and the periods where we evaluate them are set, we select four features whose influence we want to analyze : **actors, production companies, genres and keywords**.
To make the exploration easier, we removed all unreleased movies, movies in a language other than English or with a vote count lower than 5. With this cleaned dataset, we can now look at our metrics to have a better idea of what they represent (Figures 1, 2, 3 & 4).

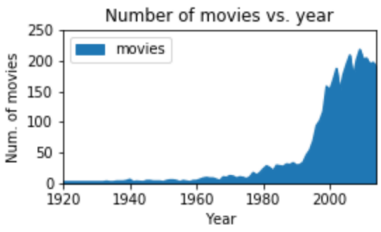


Figure 1: Movies over the years

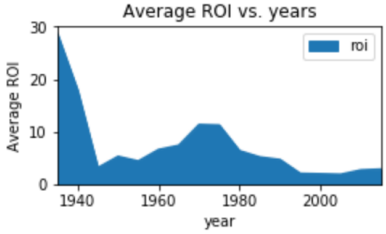


Figure 2: ROI over the years

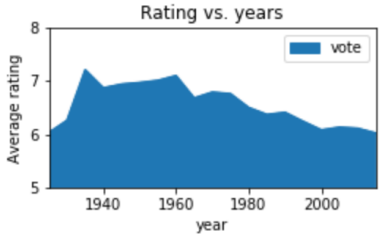


Figure 3: Ratings over the years

[1] "Cinema of the United States" Wikipedia. 2-Jan-2019.

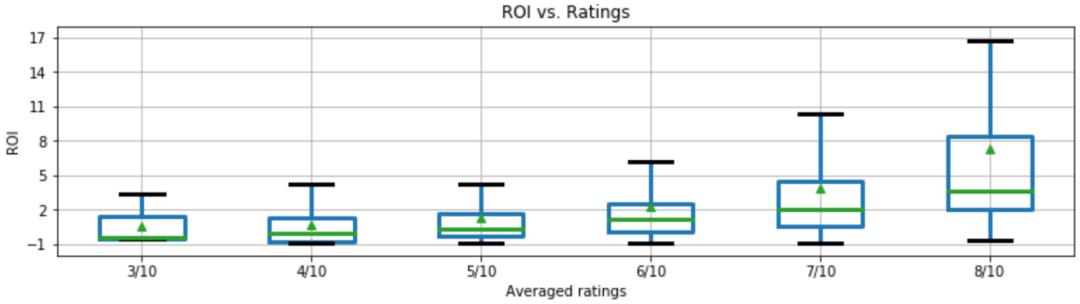


Figure 4: Boxplots of the relation between the ROI and the rating of a movie (1920 – 2015)

Comparing the ROI vs. the rating of a movie (Figure 4), we realize that a good movie is more likely to give a higher return on investment, which follows our intuition. It is also interesting to observe the correlations between the numerical features in the dataset (Figures 5 & 6). We can clearly see that correlations are not constant, this is a strong indication of how much the cinema industry has changed over the years.

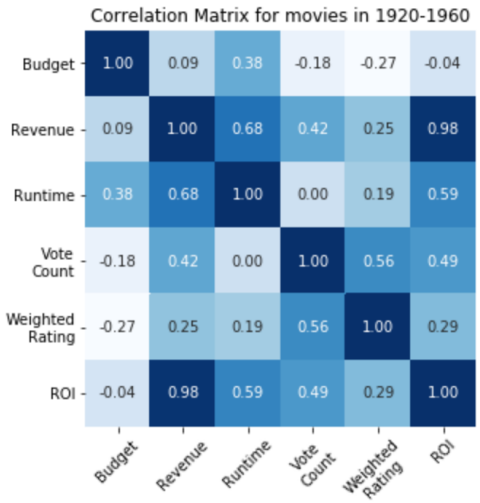


Figure 5: Correlation of features in 1920 – 1960

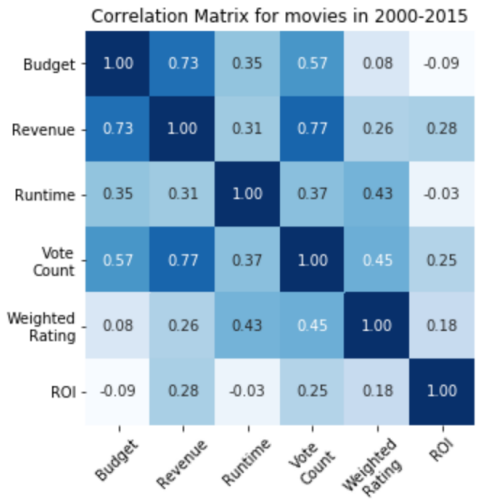


Figure 6: Correlation of features in 2000 – 2015

3.1 Actors & Production companies

Who has never seen a movie just because it was made by a company they liked or because their favorite actor was playing in it? As a first approximation, we can show that some actors or companies are more likely to give a better/more profitable movie than others (Figures 7 & 8). In the next sections of the report, we will try to see if we can observe a relation between actors and success and between production companies and success from a graph analysis perspective.

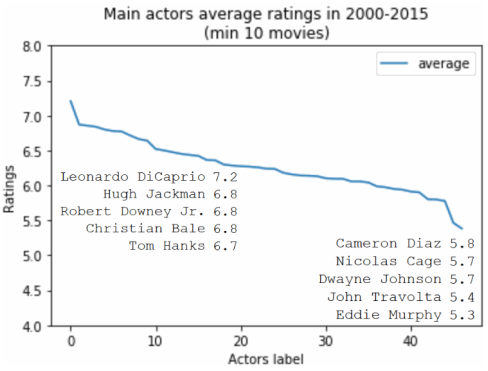


Figure 7: Average rating of actors in 2000 – 2015

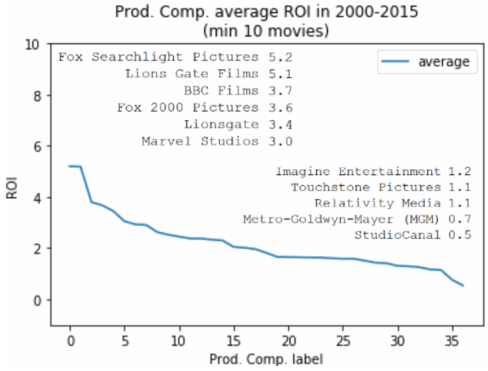


Figure 8: ROI of companies in 2000 – 2015

3.2 Genres & Keywords

Are some genres of keywords more enjoyed by the viewers or more profitable? Did this evolve over the years? Here, we try to give a first element of answer. We begin by giving the most present genres for our four time periods (Figure 9). We then get a first feeling of how the genre of a movie can influence its ROI (Figures 10 & 11). We can see that some genres in the 1920 – 1960 period could really influence the ROI but that nowadays no dependence is visible.

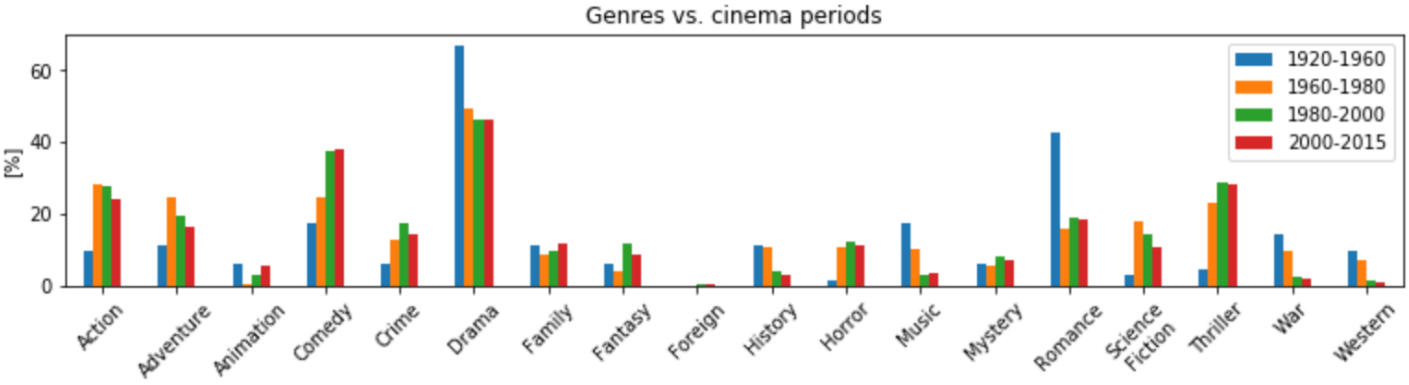


Figure 9: Different genres representation for the four predefined periods

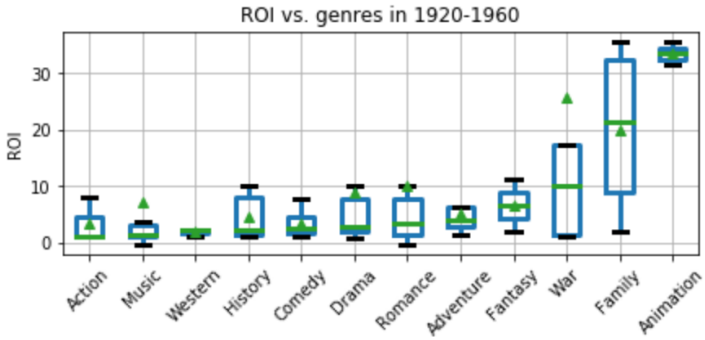


Figure 10: ROI dependence with genres in 1920 – 1960

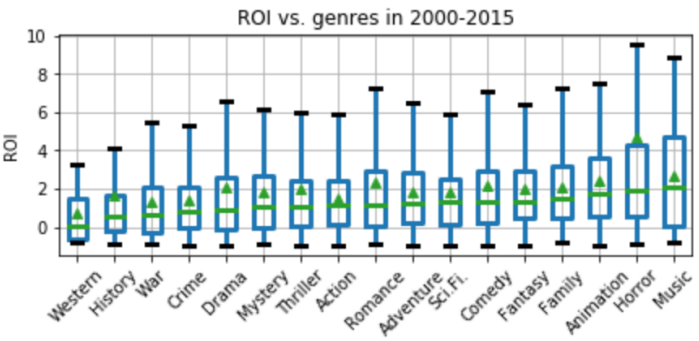


Figure 11: ROI dependence with genres in 2000 – 2015

4 Graph perspective

We now perform a similar analysis using the graph theory framework. We build a similarity graph for each period for our 4 features (16 graphs in total). For each graph, the edges are defined differently. In all of our graphs, the nodes are the movies and the links' weights are proportional to the number of shared items of the feature of interest (number of common actors, genres, etc.).
The isolated nodes of the graphs are removed. Then, each adjacency matrix is refined to keep only its giant component. After this, we can consider the two success metrics defined earlier as signals on our sixteen graphs. This means that we have 32 signals to consider. We are now interested in assessing the relation of the feature used to build the graph with each signal. To do so, we explored three different methods.
The first one is to compute the smoothness using the unnormalized Laplacian. More precisely, we have the following property [1]:

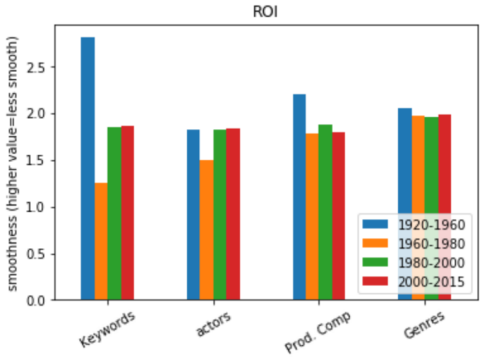
$$f^T L f = \sum_{i,j} W_{i,j}[f(j) - f(i)]^2$$

This quantity is linked to the smoothness of the graph since it gives the sum of the $[f(j) - f(i)]^2$ on the graph where $[f(j) - f(i)]$ is the variation of the signal along the edge linking nodes i and j. The output of the function will be larger for graphs with a lot of edges. Therefore, we normalize this value by the sum of all weights of the graphs to get our final value. It is important to note that the larger the smoothness value, the less smooth the signal is across the graph. A smooth signal on the graph indicates that localized relations exist between the feature used to build the graph and the considered signal.
The second method was based on an evaluation of the percentage of energy located in the first third of

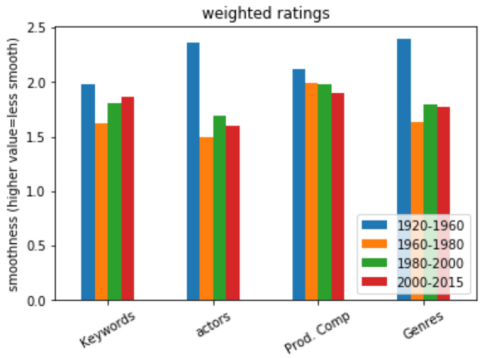
the spectrum (after removing the continuous component). Indeed, if the GFT of a signal over a graph shows large coefficients for low frequencies, this signal should have a smooth behaviour on the graph.
Our last method is less conventional; inspired by the fourth milestone, we decided to use Transductive Learning to assess the smoothness of signals over our graphs. Indeed, the chosen method attempts to recover missing values of a signal over the graph by minimizing the p-norm of its gradient on the graph. This method should thus produce best results (a lower reconstruction error) for signals that are regular on the graph. We chose the L-2 norm (interpolation by Tikhonov regularization).
Looking at the signal's smoothness over the graph, we assess the influence of the features used to build the graph and the signal. If these markers are smooth, we can for each of the two success markers, look at the region of the graph with the highest success and extract the corresponding coordinates. These coordinates give us information about the cast, genres, keywords or production company (depending on the graph) that are linked to a successful movie.
An interesting application is then to use these graphs to build a success prediction system. Knowing the cast, the genre, the keywords and the production company of an unreleased movie we could localize it on our four graphs (for each period) and look at the value of the success markers at these coordinates. Taking into account the values of these 2 success markers, we could compute an estimation of the success that the movie will achieve. Furthermore, we could imagine to extend our project to predict what features will have the most influence on movie success in coming years.

5 Results and Discussion

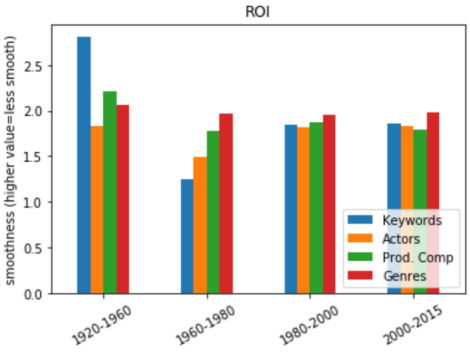
We can now investigate the results obtained for our different methods.



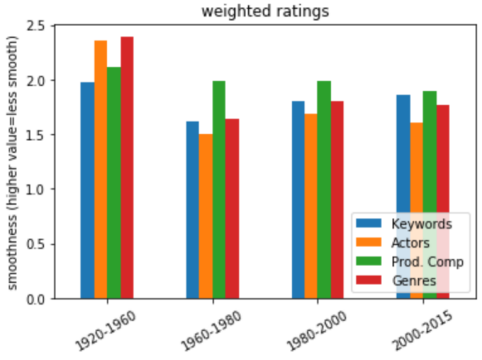
(a) Smoothness for the ROI signal



(b) Smoothness for the weighted rating signal



(c) Smoothness for the ROI signal



(d) Smoothness for the weighted rating signal

Figure 12: Smoothness results obtained with the first method using the unnormalized Laplacian. Note that the axes are different for the two rows of figures.

Concerning the first method, the results are shown in Figure 12. As we can see on the ROI histogram (Figure 12a), keywords were a very bad predictor for ROI in 1920-1960, a good one in 1960-1980 before

stabilizing for the two last periods. The ROI smoothness for the other features remained pretty stable over time. However, the results obtained for the first time period may not be very relevant since the number of nodes in the corresponding graphs are very low. In the case of the weighted ratings (Figure 12b), the signal's smoothness is not very good in 1920-1960 but gets better and stabilizes during the following time period. Overall, the best predictor (best smoothness) for the weighted ratings nowadays seems to be the actors playing in the movie even though they had little influence for old movies.
Concerning Figure 12c, for the two last periods, all features gave equally smooth ROI signal on the graphs. However, for movies from 1960-1980, ROI was a lot smoother for actors. Looking at the histogram 12d, it appears that for every time period the best weighted rating predictors were the same: actors, followed by keywords and genres and finally production company.
Here are some results for the second method :

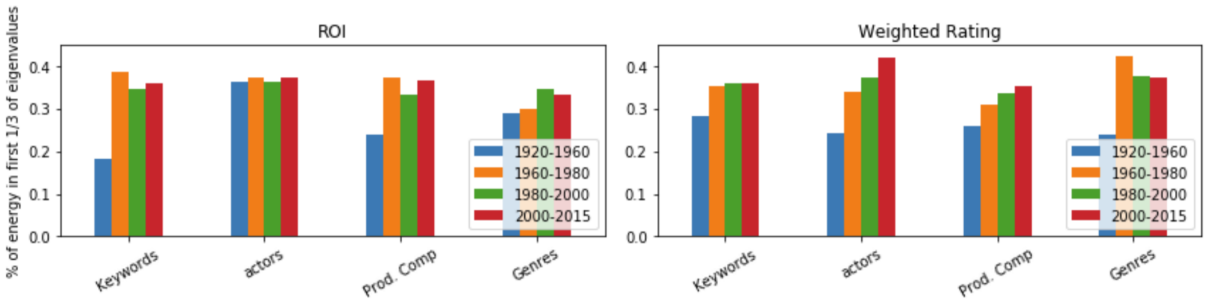


Figure 13: Percentage of energy located in first third of eigenfrequencies spectrum, comparison of different periods for a given feature.

As seen on fig.13, results are not striking. Nevertheless, it is still possible to see some tendencies. Looking at the weighted rating, we can see a net increase on the influence of the actors throughout the years. Moreover, actors seems to have the most influence on movie rating in the 21st century while 1960-1980 movies appears to be more influenced by genres.
Since most of the results for our third method to assess smoothness were not satisfactory, we do not present them here. A discussion of these results can be found in our Jupyter notebook 3.3. We also make a detailed comparison of the results produced by the first and third method for certain combinations of graphs/signals.

6 Conclusion

Using three different methods, we investigated the possibility to predict a movie's success. Computing the smoothness of the "success" signals and looking at the GFT of our graphs gave rather similar results. This gave us an idea of the importance of using different methods to consolidate our results. However, we could only observe tendencies but no concrete results. Moreover, comparing smoothness of different signals proved to be challenging. Indeed, our signals ROI and WR had different means, variance and overall distribution. Additionally, our dataset was quite heterogeneous in the sens that we only had a small number of movies from 1920-1960 and a lot more for the other time periods; we hade to figure out how to normalize our measures to get comparable results. It is also important to mention that some of the films had incomplete data. Possible improvements could come from building more complex graphs based on multiples features, therefore encapsulating more information. Finally, the dream of building an effective strategy to predict an unreleased movie's success seems to be realistic but would require more resources and a deeper analysis.

References

[1] Shuman, Narang, Frossard, Ortega, and Vandergheynst. The emerging field of signal processing on graphs. *IEEE Signal Processing Magazine*, 2013.