Can we estimate the earnings of a movie ?
A Network Tour of Data Science - Project

Group 40 : Romain Caristan, Benoît Pasquier, Hugo Kohli, Roc Arandes

1 Introduction

1.1 Goal

In this project, we have attempted to find a way to estimate the earnings of a movie, depending on the company and the people being involved in it, including actors, directors and producers. In order to do so, we have employed data that was extracted from the TMDB Kaggle dataset. As constructing a good graph is as important as visualizing it, several construction methods will be explored and used to extract as much information as possible. The nodes of our graph are movies, and this will be a constant throughout the presented work.

1.2 Data acquisition and exploration

Description

As stated in the previous section, during this project we have chosen to continue to work with the *TMDB 5000 Movie Dataset* from Kaggle. This one, being already a subset of the more complete *TMDB Dataset*, allowed us to reduce the computational time of the different methods shown in the following chapters.

Pre-processing

In the first place, the data was not complete and sometimes inconsistent with its budget/revenue units. Thus, we chose to remove the values expressed in millions for more consistency of the future results. The movies with a 0 $ budget/revenue and the movies which budget/revenue smaller than 1000 were dropped from the dataset. By doing so, a totality of 1592 movies, representing the 34% of our dataset were removed. Moreover, we created another feature for each movie representing the return on the initial investment, henceforth called *earnings*. For the rest of the project, we have chosen to keep the first 5 actors, first 5 characters, the director, the producer and the production company name (which is considered as a person in the first model) as we consider that no more than the first 5 roles in a film play have a preponderant role in the success of the film. This is a subjective choice and could be changed if wanted.

2 Average earnings model

Our first attempt to estimate the earnings of a movie consisted in computing the weights of each feature (*actor* / *character* / *director* / *producer* / *company*) as the average of the earnings over all the movies in which he/she had contributed. During these process, some other processing steps were required such as removing the duplicates of certain members that had different roles on the same movie (i.e. actor and director at the same time).

2.1 Graph construction

First of all, the previously found weights were used to create our first adjacency matrix, which edges were computed as the sum of the weights of the common features between the two movies in question. The second step of this first process was to extract the biggest component of the adjacency matrix previously calculated. While we know that it can be obtained using NetworkX,

we chose to use our own functions which allow to extract the indexes of the nodes that belong to the biggest component more easily.

2.2 Graph visualization and analysis

After computing the normalized Laplacian and its respective eigenvectors, we created a *signal* employing a trivial threshold of earnings (i.e. x 1.5 of the budget) in order to label the movies as $+1$ (earnings bigger than the threshold) or -1 otherwise.

Furthermore, a GFT between the *labels signal* and the eigenvectors of our biggest component was computed and used to extract the eigenvectors that had the highest correlation with our signal. Finally, the two eigenvectors that corresponded the best to our signal, were used as an eigenmap, and the values of the labels were applied as colors.

As we can observe in the notebook, the proposed graph does not seem appropriate to estimate the earnings of a movie. The inconvenient of this model (which is further discussed in the notebook) is the outliers. If a person played on a single movie with very high earnings, its weight also becomes very high. Indeed, when we analyzed which features had the highest values, we found that they were all members of the cast of Paranormal Activity, a home-made amateur movie that became famous and which is the movie with the highest earnings ratio of our dataset.

Remark: The implementation, plots and further detailed explanations of the mentioned steps can be found in the notebook.

3 BRH models

After the first unsuccessful attempt, we tried to create a more developed method of attributing the weights to the features, so that the outliers (such as the cast of Paranormal Activity) would not be on the top of our list. Our objective was to create what we called an "index" that would allow us to give a weight to the features in function of their earnings but also rewarding their consistency. In order to do so, we were inspired by the mathematical model "H-index" (used in the scientific community to measure the productivity of a researcher), and we developed our own metric.

This one is explained in detail on the notebook, but it can be summarized as follows:

1. (Taking an actor as example) The list of movies in which he/she participated is retrieved and divided in *positive* and *negative* earning movies.

2. The earnings were up-scaled of 100 (representing the %) and both positive and negative indexes were computed.

3. If an actor had a positive index of 25, this meant that he/she participated in at least 25 movies that made at least 25% of earnings. Similarly, if an actor had a negative index of 5, this would mean that he had participated in 5 movies that at least lost 5% their budget.

4. Finally, these two indexes were added, and the final values were shifted so that the minimum index value was 0 (as we want to construct an undirected graph i.e all the edges will need to be positive).

The obtained results were astonishing, attributing the highest values to the companies/actors/directors that one would expect (such as Universal, Brad Pitt, Steven Spielberg, etc). We decided to name our method *BRH-index*.

3.1 BRH first graph: adding edges

3.1.1 Graph construction

While the new way of attributing a weight to each feature seemed promising, the dilemma resided then in how to treat these different weights to determine a final edge value between

two movies. The first attempt we tried consisted in **adding** the BRH-indexes of the different common features between two movies, and constructing a new adjacency matrix, and extracting its corresponding biggest component as done for the previous method presented in section 2.

3.1.2 Graph visualization and analysis

In this case, when showing the graph in the eigenmap composed by the two first eigenvectors (of non-null eigenvalue), we could clearly identify at least 5 major clusters, as shown in Fig. 1.

The clusters being so clear, we investigated them and found that each of the three clusters on the extremities was formed by one of the top three production companies: Universal, Paramount and Twentieth Century Fox. This comes from the fact that when using the BRH-index while considering the companies as "individuals", these ones received indexes that are not comparable to those of the actors/directors. Indeed, a production company can participate in 10 movies at the same time, while an actor/director cannot. From this point, we decided to recalculate our features by not taking into account the production companies, and therefore only using actors, characters, directors and producers.

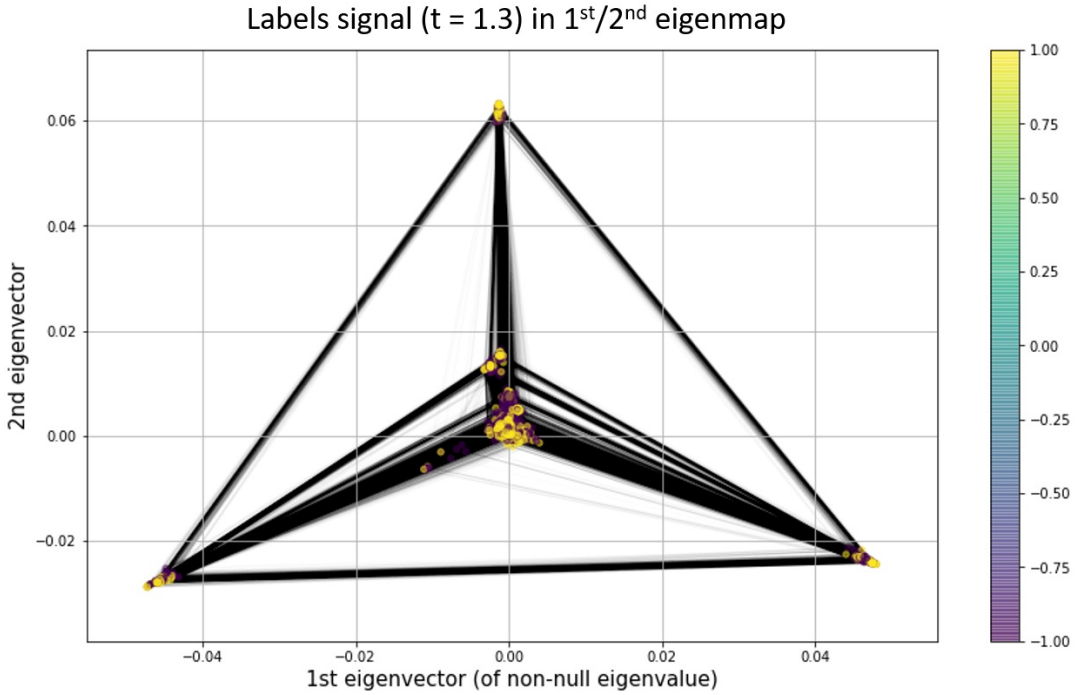

Figure 1: Labels signal represented in the two first and second eigenvectors eigenmap (adding edges model)

3.2 BRH second graph: adding edges (no companies)

After removing the production companies from our features table, we wanted to observe if the nodes of the graph would still cluster into potentially meaningful groups of movies. Therefore, the same steps previously explained for the other methods were applied. Unfortunately, when removing the production companies of our features and showing the 1st/2nd eigenmap, we could observe that almost the totality of the movies were gathered in a big cluster, with just several isolated outliers along the first and second eigenvectors.

While the BRH-index looked promising and potentially useful, we didn't achieve the results we were expecting yet. Moreover, after analyzing the previous presented graph, we believed that not only the production companies but also the fact that we were adding the indexes of each

feature when defining an edge was negatively affecting the obtained results. Similarly to the problem observed with the first method (section 2), when adding all the BRH indexes we are not taking in account the number of elements in common that the two movies have, meaning that a single very high value feature, would suppose a high-value edge.

3.3 BRH final graph: averaging edges

The final attempt of constructing a graph consisted in calculating the edges by **averaging** the values of the BRH-indexes previously found (also without the production companies).

We can observe an increase of values separated from the principal cluster but despite this new way of calculating edges, the graph is still not separable. However, this change has induced an improvement and is a step in the right direction, but the way of calculating the edges should be still improved to have better separable clusters.

Because of time limitation, we chose to keep this model and further investigate its interpolation potential.

3.4 Interpolation on final BRH graph

Our final model now being fixed, we wanted to see if we were able to provide an analysis over the earnings of the movies. To do so, we computed the *binarized_labels*, set to +1 if a movie had positive benefit or to -1 if it lost money. We saw that around 75% of our movies were earning money, letting the other 25% be defaulting movies. With these new labels computed we did a p-norm interpolation to be able to retrieve the labels of a missing part of the data as in Milestone4. We chose to use the F1-score instead of the relative error here to evaluate our results as it is a good metric to classification problems. Computing the F1-score for different values of the *mn_ratio* (proportion of the labels we know a priori) and *threshold* (value used as a boundary between the two labels) gives acceptable results, with the parameters *mn_ratio*=0.3 and *threshold*=0 we get a *F1-score*=0.89. We chose to use the 2-norm interpolation, providing more consistent results and a threshold set to 0.

Rather than a score only, which gives us good insights but cannot really be interpreted, it is more meaningful to compute the proportion of well computed labels after the interpolation. We observe that with 50% of prior knowledge about the benefits of the movies we are able to retrieve 90% of the labels as show in Fig. 2.

Another important information we can extract is the proportion of False Positives, i.e. the labels that are labeled as +1 but their true value is -1. This corresponds to say that a movie will earn money but it is losing in the end. This information is useful as it can have a high cost, in particular for all the people involved in those movies. This case is worse than saying that a movie will lose money but in reality it does not. The results, shown in Fig. 3, tell us that again with 50% of the data we were able to get back around 70% of the negative labels, and that we need at least 80% of the data to retrieve 85% of the negative labels. This result can be explained by the fact that in our data most of the movies (75%) can be considered as successful as they are earning money in the end. But the fact that the data is arranged in such a way when looking at the success of a movie is still good news for the movie industry.

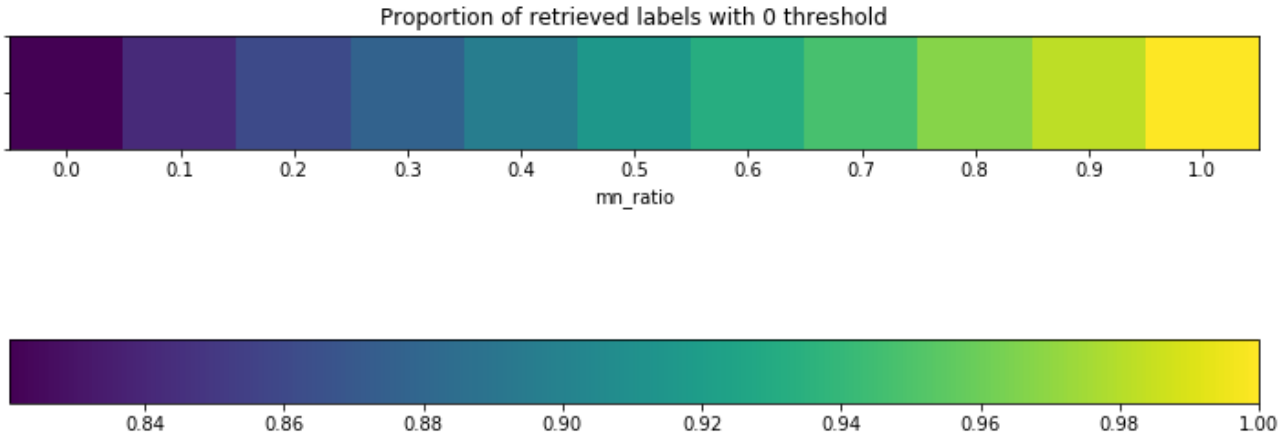

Figure 2: Proportion of all retrieved labels in terms of the percentage of prior knowledge about the labels.

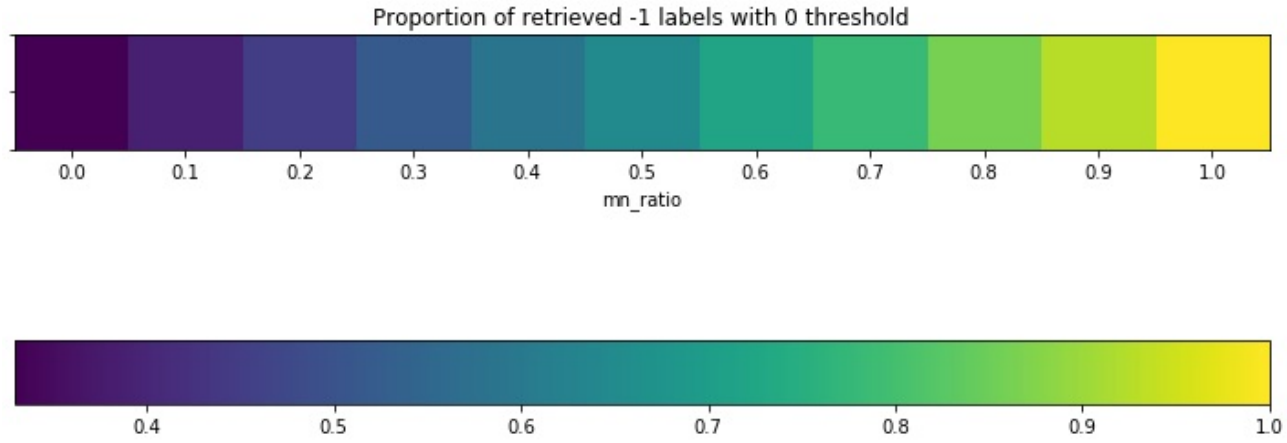

Figure 3: Proportion of retrieved negative labels in terms of the percentage of prior knowledge about the labels.

4 Conclusion

During this project, we attempted to find a way to evaluate the return on investment of a movie according to the participating members of this one. Different network constructions have been investigated in order to best fulfill this initial goal.

We came with a creative algorithm to compute the weights of the different features, giving more importance to successful and recurrent people (actors, producers and characters), which gave meaningful and realistic results. Despite this promising algorithm, we could not find a way to compute the edges so that the obtained graph offered meaningful separable clusters. The best results were obtained when computing the edges by averaging the indexes of the common features between the two movies.

Nevertheless, we have further investigated the interpolation potential of our model. It appeared that from a subset of 50% of our graph, we were able to correctly retrieve 90% of the labels indicating if a movie is profitable or defaulting.

To conclude, we did not succeed to evaluate the exact earning of a movie depending on its position on the graph, but we could quite well evaluate if a movie would make profit or not. Moreover, as already mentioned, while our BRH index seems promising, the edge computation could be further investigated in order to achieve better results.