A Network Tour of Data Science

EE - 558

Project : Will my new videogame be successful?

Students
Maxime Bonnesoeur
maxime.bonnesoeur@epfl.ch
Michael Hodara
michael.hodara@epfl.ch
Tarik Kamel
tarik.kamel@epfl.ch

Professor
Pierre Vandergheynst
Pascal Frossard
Andreas Loukas
Michaël Defferrard
Volodymyr Miz

January 10, 2020

1 Introduction

More and more video game platforms are making a principle out of physical dematerialization of games that are now only accessible via download on an online portal.

One of these platforms is Steam. Steam is a video game digital distribution service launched as a standalone software client which provides automatic updates for their games. It is today the largest digital distribution platform for PC gaming with over a billion registered accounts with 90 million monthly active users. But each game on this platform has its own parameters and thus its own success. *What parameters influence the most its popularity ?* This question is indeed quite interesting because today's video games' success may depend also a lot on the reviews that similar games receive from their platform. We shall then try to answer this question among others.

In this project, we will attempt to find a way to estimate the success of a released game according to different factors such as the category the game belongs to, its price, the gameplay, or time period of release. In order to do so, we have scrapped our way through the Steam dataset of Kaggle [1] to gain a bit of knowledge.

2 Story

The video game market has been growing exponentially for the last decades. But what if every video game became digitalized and that the CD form video game would no longer existed, 5 years from now? Will they join the group of "old-fashioned" objects that the VHS case, the typewriter or the flip phone already belong to? This is actually possible since the digital download of the video games through platform like Steam is gradually taking over the CD market and perhaps the video game box wrapped under the Christmas tree will be replaced by a piece of paper containing a voucher code.

We won't focus on how probable this scenario can happen but rather on how it sets up and what influence it has. Indeed, when releasing a game, the original developer (that can be a big company, an independent programmer or other) seek for a certain **success**.

2.1 Parameter of success

The success can be defined by very different ways. Usually, the better a game is, the more likely it is to be successful. What can define a good game? Today good video games can be defined by four parameters:

- Great Gameplay : the game has a specific way in which players interact with it. Gameplay is the pattern defined through the game rules and is defined by connection between player and the game
- A Great Story: A great story can keep the player immersed in the world you've created who feels attached to the characters and want to continue playing to see how the story unfolds
- Great Art Style: Very sophisticated graphics help the total immersion of the player in the video game which can benefit in addition to the playability of a pleasant feeling of reality.
- The Player Should Work for It: a too easy task can make the game boring.

But a "good game" as defined above is not necessary a successful game. Indeed: first of all, there are many different types of game: adventure, action, RPG (role playing game), simulation, sport, racing etc.. and therefore, many different players enjoy different games. Not everyone likes an action game like Call of Duty and not everyone likes a RPG game like Skyrim but they both accomplish what they're trying to provide for the player while having enough of each key element to be successful. But, the category the game belongs has an impact on the eventual success of a game. Furthermore, not every successful games is from has a specific category.

The notion of success is taken only in the point of view of the outside customer and thus we shall neglect the elements of scope (if the game achieved its objectives within the given framework), schedule (If the developers managed to hit their milestones on time) and team satisfaction. Not every developer is aiming to have its game to be a blockbuster and often do it for personal satisfaction of programming no matter what the result is.

Therefore, the principal criteria we can set our researches on are the Customer satisfaction through a certain **rating** based on a reviews left by the user and also the **economical revenue** after selling. The way the rating is calculated is explained in [2] whilst the economic success is the product of the number of owners and the price of the game.

External factor will therefore have an impact on the success of a game: the price, the date it was released (which year and which period during the year), to which category it belongs, but not only: the developer and its reputation, how many games it has already released and what success they faced can also have an influence on the future success of other games.

2.2 Availability of dataset

We might think that one of the problems we may encounter is that by choosing Steam platform, we might be taking a bias since it is a platform developed by VALVE, which is also a publisher of Video Games. Thus, their games are supposedly more valued than the others. Furthermore, the games we will study are exclusively for PC user. We shall take this into account in our study.

3 Data analysis

3.1 Data Gathering and cleaning

We first gathered the Data by using different APIs to get some data from the Steam and SteamSpy website. This data was then pre-processed to be usable. The main advantages of parsing the data ourselves was to be able to only consider the games that are available in Switzerland. Moreover, the actual span of our data is extended until January 2020 instead of March 2019. We were then able to use the data of 2019 as wlel in our analysis.
The data we gathered, being well organised as a network in terms of observations and features, still needs cleaning as it contains a lot of features we won't require in our project.

The first step of cleaning consists in getting for every game its rating score based on SteamDB method [2]. Then, we chose to remove the games that were not available in english, french or german as we took the assumption that a successful game would require international understanding, therefore requiring the English language. Then, as Steam offers a lot of game types, we kept only what we strictly considered as games and removed the additional contents from games, and custom developed software (mods). Also, we wanted to be able to take advantage of the different genres that each game belongs to. For that, we created a column for each genre and for each game, put a 1 in the column if it belongs to the mentioned genre (one hot encoding). We only considered the PC games owned by more than 20′000 people as we try to understand what a successful game is and what makes a game unsuccessful (having lower amount of owners). Finally, formatted the release date to be exploitable for our analysis. This processing led us to obtaining a dataset formed of :

- 7868 distinct PC games, playable in English
- 14 features (publisher of the game, platform, clear genres, number of owners, release year, release month)
- 2 evaluation metrics (economic success, steam rating)

3.2 Exploration

In this section, we focused on analyzing how the features we chose to work with had an impact on the two evaluation metrics we established, and most importantly, which of these features had the most impact, in order to reduce, if possible, the amount of data we will process in the next steps of the project. To do so, we chose to use Machine Learning by applying a Lasso Regression technique which will give a coefficient of weight for each feature on the chosen metric. It consists in first splitting the dataset into training and testing set, finding an optimal regularization parameter α that is associated with the lowest root mean square error in the range.

This technique has been done for both our metrics of success and we were able to see that :

- for the **SteamDB ratings metric**, the number of owners has the most impact which is coherent with our first observations, that the more people own a game, the best probability it has of being a successful game. Moreover, it helps us see that there is a correlation between the rating and the price. Some games genres, such as Casual, or Simulation seem to have a greater importance than others.
- for the **economic success**, putting aside the features that are part of its calculation, we can see that the casual and simulation game genres have a bigger impact on the economic success than the rest, which tells us that some genres are more associated with success.

3.3 Graph

We can use Gephi to visualize our graphs. Among all the layouts available, we focused on Yifang Hu layout. Just like the ForceAtlas, this layout is a Home-brew layout of Gephi. It is made to spatialize either Small-World or Scale-free networks by focusing on the real data that is useful to explore but that is better for large networks than Atlas force. It is even better than ForceAtlas2 where Barnes-Hut calculation is used for replacing the "attraction" and "repulsion" forces by a "scaling" parameter.

The first interpretation it allowed us to have is that we are more in a "small-world" spatialization rather than a scale-free one. Indeed, we notice on the figures above a large clustering rate as we have one element but not fully connected. The categories are well distinguished as shown in 1

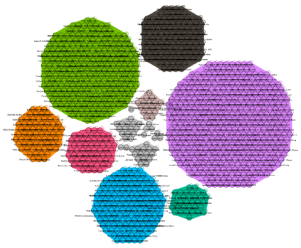

Figure 1: Cluster categories (Circular Pack Layout)

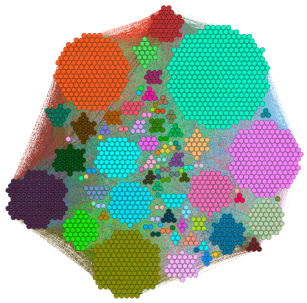

Figure 2: Modularity between clusters (Circular Pack Layout)

Figure 3: Modularity between clusters (Yifang Hu layout)

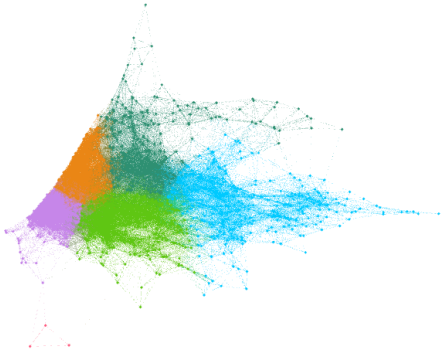

Figure 4: Modularity clusters in Yifang Hu layout

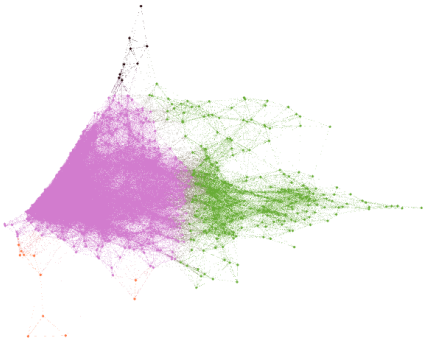

Figure 5: Leiden Algorithm clusters with Yifang Hu layout

The connections among each nodes are represented in the layout 2. Among the 2000 best rated games, we can distinguish 3 categories of games in them:

- a mass of well-rated games
- a level of game more restricted and better rated and gear
- some rare games super well rated with a lot of owner.

The 4 and 5 represent in colors the difference when using Leiden algorithm. We notice less clusters in it. It is also a good way to finds well-connected communities and split clusters generating *subset optimal clusters*. It is very fast and can optimize modularity.

The different Figures allow us to notice a bit of a trend on the platform with very "trendy" games that centralise the bulk of traffic and the more "standard-kind" games are more on the outskirts.

Unfortunately, with the data we have now, we are not able to identify what characterises each game stratum. These are probably marketing arguments that we cannot evaluate with this dataset.

3.4 Exploitation

3.4.1 Graph analysis

In this section, we went on with the cleaned the data and tried forming a network. For that we considered a subset of our data composed of the 2000 best rated games. When using the economic success metric, we only consider the premium games as we don't have any information on the money generated by free games with in-game advertisement or payment.

We then use a RBF kernel to set the edge weights $w_{ij} = \exp(-||x_i - x_j||_2^2 / \ 2\sigma^2)$ of our adjacency matrix and threshold the ones with the smallest magnitude by setting $\sigma = 0.6422$ from the variance of the pair wise distance and $\epsilon = 0.85$ to put the unwanted corresponding edges weights to 0. A sparse matrix has then been obtained when increasing epsilon. The observation of our adjacency matrix tells us that our network is connected with one gigantic component, built in a pyramid fashion with the best rated games at the tip. We can also see a pattern inside our adjacency matrix, showing interactions within the inner groups. By then analyzing the degree distribution of our graph, and the citation graph moments we were able to asses that our graph, composed of 1667 nodes and 100939 edges is connected, probably related to the **small world model**. The clustering coefficient of approximately 0.63% tells us that two thirds of the graph is connected. We can finally say even though our graph is not fully connected, all the elements are connected to each other in some way.

3.4.2 Spectral clustering

Observing our graph, based on the economic success and the ratings of the games did not let us perceive any apparent cluster as our data is quite sparse and close from each other. We chose therefore to do a spectral clustering based on the computation of the normalized Laplacian and K-Means clustering to check if our findings are coherent with the one we found with our observations with Gephi in section 3.3. This analysis helped us distinguish that our games were divided in three clusters :

- The first category, in yellow are the game that have the best grade out of all other games. We can definitely see that here, having an huge grade is synonym of being a huge economic success with a large community of gamers buying this game whatever its price.
- The second category in blue are the games with a really good grades and a pretty good economic success.
- The last category of games in purple have the largest economic success span of all of the clusters. This cluster is our largest cluster.

This plot gives us information about probably the most important criteria on Steam which is the user's rating and the way a game is promoted by the players' community in their reviews. In fact, as it can be seen here, a good rating is always correlated with a huge number of sales which will increase the economic success of the game.

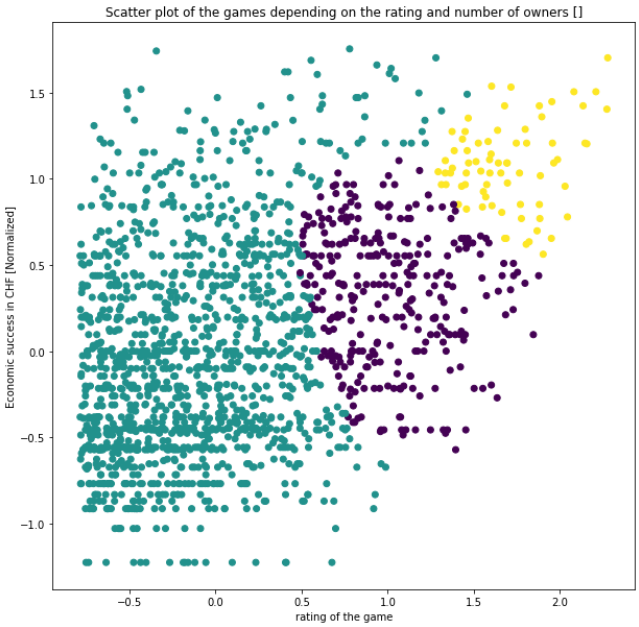

Figure 6: Spectral Clustering of the 2000 best rated games, based on their economic success and their SteamDB rating

4 Discussion and conclusion

Through our graph analysis, we were unfortunately not able to find an absolute given parameter that explained the success of a game. We were able to show the huge dependency of the number of owners and the overall rating of a game on its economic success.
It currently seems that the high success rate of a video game depends on parameters that could not be depicted with the provided data on Steam. The game's marketing campaign, the promotion plan, advertisement and review could be a critical criteria to define the success of a game. By pulling another database of external reviews of the games or even statistics about in-game purchasing of free games we probably would have been able to extend our analysis and the parameters that are more related to the success of any game.
One thing that our study showed was the fact that ultimately, the success of a game was determined by how good it will be received by the gamer's community. The rating of a game definitely defines its success but it has to be put in consideration with the fact that the game needs to be rated by a handful of players buying it before knowing the community's opinion. Therefore, for a higher success, a first conception of "success" has to be achieved.
To conclude, this project was a tremendous opportunity to experience the data scientist approach; find an interesting dataset, setting an objective, exploring it and then exploiting it to reach our goals. We will for sure come better prepared for new adventures of data science and appreciated the network tour.

References

[1] Kaggle Steam Dataset
`https://www.kaggle.com/nikdavis/steam-store-games` by Nik Davis, May 2019

[2] Introducing Steam Database's new rating algorithm.
`https://steamdb.info/blog/steamdb-rating/` By SteamDB Team September 27, 2017

[3] Knuth: Computers and Typesetting,
`http://www-cs-faculty.stanford.edu/~uno/abcde.html`

[4] Scikit Learn Documentation on Lasso Regression,
`https://scikit-learn.org/stable/modules/generated/sklearn.linear`$_m$*odel.Lasso.html*`sklearn.linear`$_m$*odel.Las*