BONNELL Hugo BLANCHET Lucas PAROISSIN Théo

Data and AI – Project

Machine Learning to understand the video games' industry

WHY

Why games?

The video game industry is a subject we're keen on because we spent part of our childhoods playing different games, on multiple devices. Having grown up, we'd now like to discover what variables made the success of the games we loved to play as kids.

The economy of the video game industry is now more important than ever. What if we could figure out how to make the most of it, knowing what game the public expects in advance?

So far what happened?

We found some similar projects to ours on GitHub. Usually, they use only one database at a time and don't individually dig into each game's features to find out why it was successful. These projects yielded general results, such as "What are the most profitable genres", "What are the most profitable platform" or "Does critics affect games sale".

We'll try to dig their research deeper, merging databases and asking deeper and more time-related questions. Moreover, in the projects we found, the context was often left out from the analysis. We'll integrate it, analyzing not only the games, but also the market and the publishing companies, trying to merge and link the results we get to have a deeper understanding of our subject.

What are we asking ourselves?

Is the company responsible for a game's success or the game itself?

Do successful games set trends or do trends make games successful?

Is the success of a game only decided by its genre and platform? What other features of a game can make it successful?

Is there still space for little developers on the market or is it already overcrowded?

Was it easier to make a successful game back in the days or nowadays?

Does making a successful game imply a huge investment?

Is the success of a game due to the game itself or the way it's sold? (Marketing)

What games were turning points in the industry? Can we predict in advance if an upcoming game will be one?

Did these “turning point” games impact in any kind of way the means necessary to make a successful game afterwards?

Is the gaming industry driven more by innovation or by meeting consumer expectations?

Are indie games more likely to push boundaries and introduce innovative concepts compared to big-budget AAA titles?

Are sequels more likely to succeed than original IPs in the gaming industry? What factors contribute to their success or failure?

How much does timing (release date) impact a game's success in the market? Is there an ideal time to release a game?

Do game reviews and critic scores significantly influence a game's commercial success?

In a highly competitive market, is there a particular niche or underserved gaming demographic that presents an opportunity for success?

WHAT

By answering these questions, our goal is to determine the major factors that predict the popularity, and therefore the success or otherwise, of a video game. At the end, our idea would be to build a software based on Machine Learning that from a predicted year of publishing, a geographical release place, and other relevant variables; yields us a game genre to develop alongside with its expected return on investment.

Indeed, we know that game trends evolve over time and are subject to fads. We want to study whether these waves of enthusiasm follow well-defined cycles. Obviously, we hope that it could be possible to predict them in advance, to choose the genre and features of our future game.

For example, we know that in the years 1995-2003 the ‘sports’ game genre experienced a peak in popularity, and then it was the turn of action games from 2003.

We also believe that the popularity of a game can vary greatly from one region to another. For example, Grand Theft Auto V sold very well in the USA, but the result was much more mixed in India. Our studies will enable us to determine which strategy would be the most convincing: choosing a game that we are sure will be a success in a specific region, even if it means abandoning certain other regions, or, on the other hand, betting on a game whose success may not be as flamboyant but will affect more people.

HOW

Data Collection:

To answer all our questions, check our hypothesis and build our project we need large datasets. In front of the complexity of our task, we know that we might need to scrap data from multiple datasets.

That is why, using lots of sites like Kaggle, Statista and mostly GitHub. We gathered comprehensive datasets that include historical data on video game releases, sales, user ratings, genres, release locations, critics scores, game features and other relevant variables.

Data Preprocessing:

When merging datasets, we quickly were missing some data.

There are several ways of dealing with this. Either assign an arbitrary value equal to the average of its pairs, so as not to destabilise the whole. However, this runs the risk of distorting certain studies later and seems to be unwise. Another option is simply to delete the lines of data where information is missing and use only the complete lines when merging data.

As we have a very large number of rows, we feel that the second option is the best for global merging. Deleting unfilled rows when merging our datasets won't affect the outcome of our study too much, as we would still have a few thousand data entries.

If we want more entries, we can study the datasets separately when possible. For instance, when looking at the number of sales in a raw form, we don't have to merge the 16 000 rows "sales" dataset to the 2 000 rows "game reviews" dataset, and thus, have more data for the sale analysis on its own.

Data Splitting:

After identifying the variables we think might have a thorough impact on a game's popularity, we'll split the dataset into training, and test sets. The training set will be used to train our machine learning model, and the test set to evaluate our model's performance.

Model Selection and Training:

We'll then choose and train an appropriate machine learning algorithm on the training dataset using the chosen features and target variable(s).

Deployment:

Afterwards, we'll develop a simple software application that takes user input for predicted year of publishing, geographical release place, and any other relevant variables. The application should then use the trained machine learning model to predict the ideal video game genre and expected ROI.

Impact of the first meeting:

After discussing the preparation and basis of our project with the professor, we realized that our research had several biases, in particular linked directly to video game publishers. Some, such as Nintendo and Activision, are so big and so well-known they're almost certain to meet success when publishing a game.

The video games' sales could be directly linked to the company releasing the game rather than the game itself, which would then bias our results.

Faced with this problem, we decided to focus our study on games from independent or little-known developers, to eliminate all the big publishers. These small ones are much closer to our position, and it seems to us that focusing on them makes much more sense.

Although, eliminating completely the big publishers isn't good either. We'll try and see if there is a link between the games they publish and the other successful games. We'll try and answer the question: are the big publishers deciding the trends?

Sources and links:

Our GitHub depository :
https://github.com/Hu9o73/Video-Games-Industry-Analysis-and-Forecast

Datasets about games and the gaming industry:
https://github.com/leomaurodesenv/game-datasets/blob/master/README.md

Similar projects:
https://github.com/dsintheocean/game-market-analysis
https://github.com/chantellechow/online_game_store_marketing/