Machine Learning Algorithms Analyze Video Game Market Trends to predict the Top Performers

Mauro Perez

April 2025

Abstract

Video games have a huge impact on people, companies, and the general market. The aim of this project is to measure the accuracy of different algorithms and its ability to rank and predict how good a video game can be using two Kaggle data bases, named Video Games and Video Game Sales. The first data base is made up of 14,802 games with a rating and their console. The second is a database made up of video game sales with about 11,493 games. The following data bases were then merged using pandas; Since they had a different number of games, the merger was made to with specific sets of values involved and encoded, such as Score, United States, Japan, Europe, and global sales; other values that had to be encoded were the publisher, the platform, and the genre. The algorithms used and compared were logistic regression, random forest, neural networks, and gradient booster. Each model was categorized and ranked based on accuracy, time complexity, and execution with an extra of having to create a top 10 list, and then we compare that top ten list with various top ten lists across various media outlets and compare them to see which are the most accurate.

Contents

1 Preliminaries

1.1 Motivations

The motivation of this project stemmed from my infatuation for video games and to see the progression of Machine Learning and how it can help the video game world by being able to take large data sets and categorize each video game in accordance with not only just rating or sales but a combination.

To eliminate bias towards one console or the other and be able to objectively tell if one video game is better than the other. The video game world is one of opinions, as it should, however a certain bias towards one console or publisher fails to justify just how good the actual game is, maybe Machine Learning can help decide which video game to buy in the near future, however this is in no means is to discredit game review that are posted in online social media. The project is meant as a means for a quick and easy search instead of going through various websites and videos.

1.2 Data

As mentioned above two data sets were used and merged in order to create one large data set, this data set contains the following features: 'Score', 'na_sales', "jp_sales", "global_sales", 'other_sales', 'eu_sales', 'Publisher', 'Platform', and 'genre'. Japan is the hot spot for video games since most publishers and production originate from there, however in the encoding process we gave European sales and US sales more importance than other sales along with global sales since in the video game world this is a common practice, since they are part of the larger pool where video games are played.

1.3 Preprocessing

1.3.1 Encoding

For the sales part not much encoding was needed however in order to visualize the data, each genre had to be encoded, so it was encoded from numbers 1-7, in order for the algorithm to categorize each as a top game (assigned the number '1') or not (assigned the number '0'). the way each algorithm is set up is to have each feature have an importance value instilled in each feature, as such Score, North America and European sales had the most importance. The encoding process took a lot of time and research due to the importance feature that was added, in order to properly measure how each game was to be considered a top game, maybe a game that did not sell well in the US sold well in other countries but maybe that was due to external factor nor regarding the actual games themselves.

1.3.2 Data Visualization

The graphs helped tell an important story in the world of video making and how some genres in video games are being underused and underappreciated while other are being overused, however in the end when compiling a top 10 list a lot of the games are not actually from the biggest genre which is action. The need to eliminate bias is key for the algorithms to work, the meta data and facts help support the unbiased claims that most game reviews subconsciously have. **Figure 4** helps us understand the distribution of games made by genre, and we can see how Action is top in sales and in games made, another reason for a lot of game review to have biased views on other games, that is why it was identified as a bias feature, we can't have the algorithms knowing which genres are made the most, and especially not the ones which are sold the most.

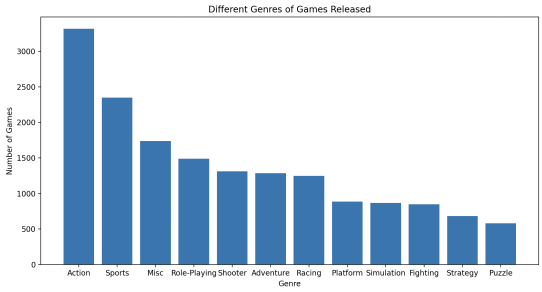

Figure 1: Includes the amount of games per genre

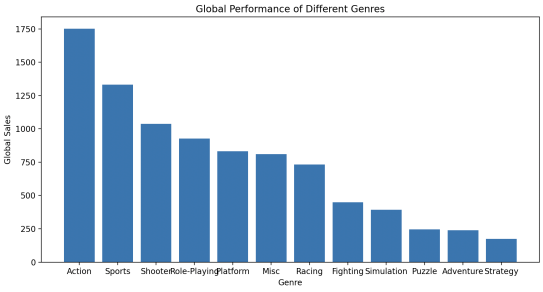

Figure 2: Includes the sales per genre

Figure 2 helps us see that Sports and Action are right in the top for sales , so it stands to reason that a lot of top games predicted would include these two genres. The goal would be to isolate these sales and not include them as deciding factors; the reason for this is the bias it can generate in the algorithms, as it sees more sales generated by these two genres, it might confuse them as top games even though the actual scores might not reflect that. This is worrying because many people end up buying games from those specific genres without even looking at reviews; it would create feature bias that would end up affecting the algorithm.

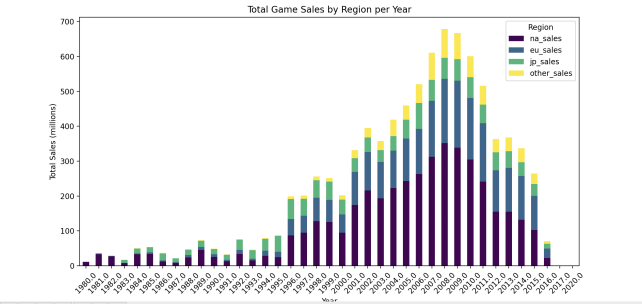

Figure 3: Amount of Regional Sales

Figure 3 tells us the amount of regional sales each year, an in these sales US and European sales by far have the most traction, therefore they warrant more importance than the other sales. It does not affect the bias trends of genres, however it creates an importance feature, since we stated that European and US sales have a much more significant impact since those are where the majority of the gamers are, the more gamers, the more accurate the the threshold can be when deciding wether the game it top or not. By giving more importance to these regional sales it can help the model make a more accurate prediction.

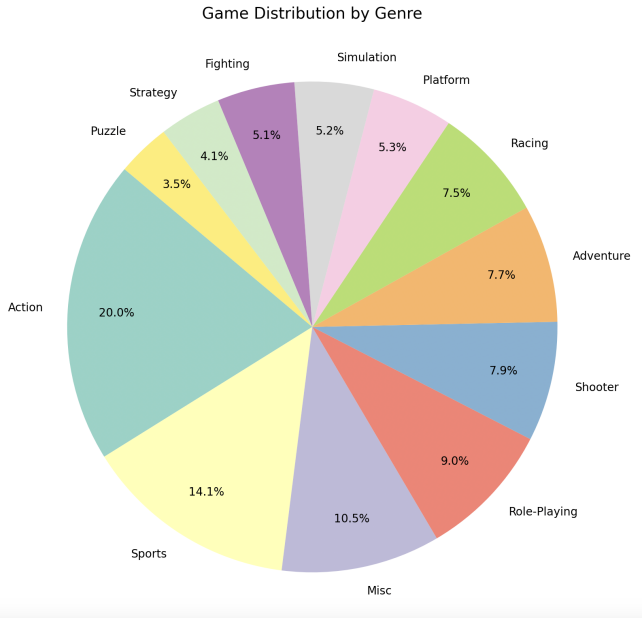

Figure 4: Genre distribution

2 Methodology

2.1 Feature Selection

In this project it was important to define the how valuable each feature was; each core feature had different values thanks to extensive research. Plain and simple United States and European sales cannot be compared to other sales, since that is where the majority of gamers and by consequence game critics reside in. Sales is a massive part of how video games are rated, the higher the sales the better the games. Above we can see a graph on regional sales; notice the sheer difference between European and US sales compared to the rest of the world, its not even a competition. Global sales will obviously have more importance than the regional ones however, for the algorithms it was important to distinguish between the two. Score was obvious to have the highest value out of all feature naturally because the higher the score the better the game. See how the pattern here is that in video game industry the numbers are black and white and paint most of the picture, obviously there will be som underlying issues like console, genre, etc. but the fact of the matter is that usually, the higher the score and sales, the better the game, as it is the same with mos products out in the open market.

2.2 Outlier Handling

The problem with video games is the bias towards a set genre, a publisher, console, or bias, etc. From the algorithms point of view it will mainly focus on Score and Sales. **Figure 5** tells us that statistically in most top 10s they will have a majority of Sports, games like Fifa, Platform games like Mario, and Shooter games like Call of duty. The predicted top 10 games, almost all of them have these three genres. As we can see they directly correlate with

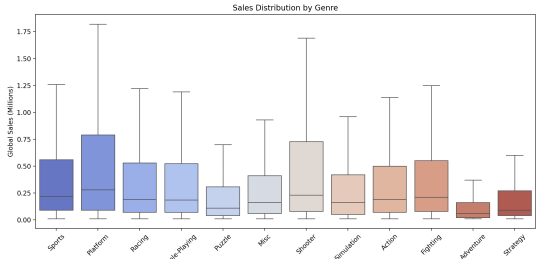

Figure 5: Sales distribution

the amount of games each one sells per year, this indicates an over saturated market filled with these genres. To have a complete and honest rating system the algorithm cannot take this into consideration just because a game is part of a genre that sells more does not mean it is a better game overall. **Figure 1** tells us the staggering difference in amount of games between these three genre and the rest, meaning that it is imperative that the algorithm finds another way to rate them. For example the sales made in countries like the US and Europe have much more of an impact and it is more indicative of the market trends the games themselves take, not the genres. **Figure 4** goes even further, with a pie chart indicating that around 20 percent of games are classified as Action, with Sports and Shooters closely behind.

3 Modeling

3.1 Models

3.1.1 Logistic Regression

Logistic Regression was used as a baseline classification model for the project. It is not as complex as Random Forest or Neural Network, and therefore it preformed considerably worse. The overall performance with the importance features and optimizations was about 0.87, which is considerably better then

0.71 without any optimization features. It helped that it was a binary case, to predict a 1 or 0 meant to predict the game as 'top' or not. The top game parameters were set at roughly more than 7 million in global sales distributed, and it must have a rating of 8 or higher, or a strong weighted sales value, meaning US or European, along with a rating of more than 8. The weighted sales part is part of the added features and shows the reason that it preformed considerably worst without it. The F1 score is the value we want to pay attention, the .87 score means that it had an 87 percent accuracy of actually sorting out the games correctly between them being and not being a top game.

3.1.2 Neural Network

Neural Network uses interconnected neurons, for this specific project only 2 layers were used, a third layer caused complications in the accuracy and F1 scores. Similar in class a Multi-layered perceptron was used. The input features for the Neural Network are the same as the ones used in the other two algorithms along with the importance features added. Removing the importance feature caused the Neural Network's accuracy score to lower, getting worse than logistic regression with importance features. The way that Neural Networks work is that they are data oriented, the less data the less they have to work with and the lower the accuracy score is, this is completely different than random forest. Neural Network needs a lot of data, but it also needs accurate data, and as we said Score and Sales are the most important features, if they are on par with the genre, platform, or console the Neural Networks take that into consideration causing their overall score to fall.

3.1.3 Random Forest Classification

With Random forest many individual trees are built, with random data samples created, this helps reduce bias, which is the overall goal for this experiment.

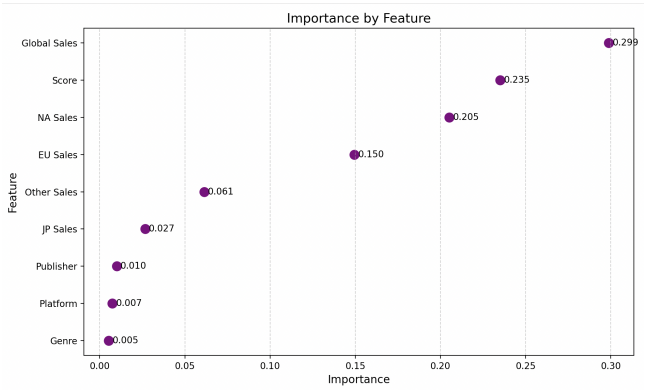

Figure 6: Importance Feature

It would be almost simulating loads of opinions from a small sample size in order for the predictions to become more accurate, this eliminates the risk of overfitting, and helps with edge cases like really good games that have low sales, or poor games that have high sales. As stated above we also gave each feature an importance coefficient, and as we can see in **table 2** ; our model was able to accurately have global , European and US sales along with score as its most important features, eliminating bias towards the publisher, the genre and the platform. Random Forest worked at about a 0.97 F1 score, a mean betrween precision and recall factor. The F1 indicates the overall ability of a model to classify games as top games or not.

Feature	**Importance Score**
Global Sales	0.298996
Score	0.235210
NA Sales	0.205338
EU Sales	0.149567
Other Sales	0.061348
JP Sales	0.026713
Publisher	0.010056
Platform	0.007435
Genre	0.005337

Table 1: Feature importance scores from the Random Forest model

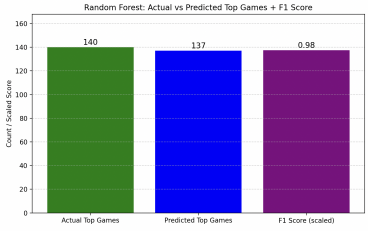

Figure 7: Random Forest

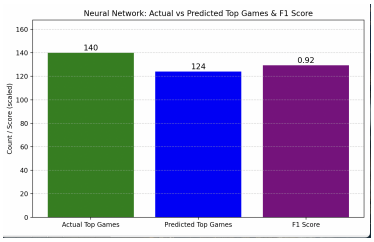

Figure 8: Neural Network

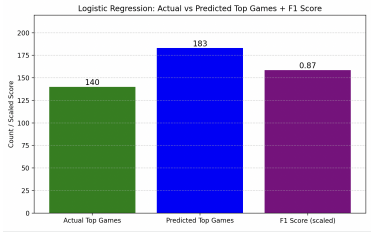

Figure 9: Logistic Regression

3.2 Modeling Result

For these we focused on F1 score, it helped us depict how each algorithm could accurately depict true positive and false negative. Essentially F1 tells us which games are top and which games are not top games. With the following figures we can clearly depict the accuracies; each model had the same features tuned since we wanted to test and verify the accuracy of each model with the same parameters. It was interesting to see way each algorithm handled each data set without the importance feature; this demonstrated how effective each algorithm was. For Random Forest it had about a 0.98 success rate on rating wether it was a top game or not based on score and sales; for logistic regression we had roughly 0.87 F1 score. For Neural Network we had an F1 score of 0.92, with 2 layers it really helped nail down this score, as more layers just complicated the reading of the data.

4 Conclusion

The three algorithms chosen were ones we discussed in class, and with this project we clearly saw the strengths and weaknesses of each one. The algorithms satisfied the need to rate each individual games and rate them as top games or not, and the data supported that they worked. This eliminates the need for long reviews, conflicting opinions and bias scores. This project feels like a genuine success since we were able to roughly predict the top games, eliminating the bias with consoles, platforms and when comparing them to the actual results the training and testing data was really conclusive and told us how each alogoirthm worked with the merged data set.

5 References

Hany, Mohamed. Video Games. Kaggle, 2021, https://www.kaggle.com/datasets/mohamedhanyyy/video-games. Accessed 4 May 2025. Pedersen, Ulrik Thyge. Video Games Sales. Kaggle, 2020, https://www.kaggle.com/datasets/ulrikthygepedersen/video-games-sales. Accessed 4 May 2025. Chinchilla, Diego. "Video Game Sales and Rating Scores: SHAP Values of Publishers, Genres and More." Medium, 22 Oct. 2021, https://medium.com/data-and-beyond/video-game-sales-and-rating-scores-shap-values-of-publishers-genres-and-more-7a3062f10046. Accessed 4 May 2025.