Predicting global video game sales: Analyzing genre, publisher, platform, and region influences

Tanner Patrom

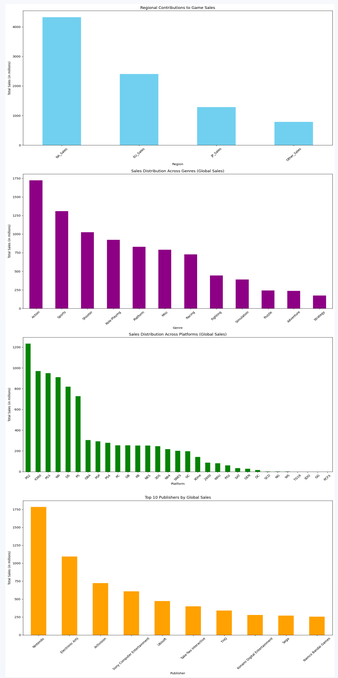
Arkansas State University

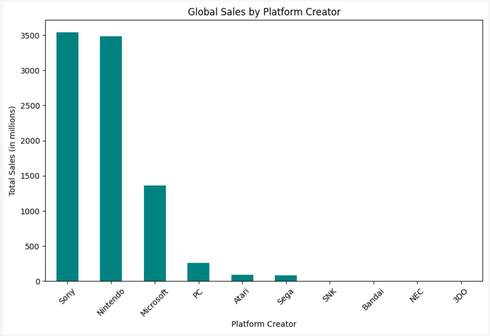
INTRODUCTION

The video game industry, from its infancy to its modern-day status as a pillar of entertainment for families and individuals, has seen a remarkable evolution. Early games adopted a simple gameplay, limited by the software and hardware of its time, that were often enjoyed by users in solidarity or in small social groups. The industry had a limited audience that mainly consisted of enthusiasts. Consoles would not be found in a large portion of households and at the earliest stage required participants to travel to arcades which housed the machines powering developed games. Publishers and development teams were small and unaware of the impact that their work would have as the industry matured.

Forwarding to the present, with the help of advancements in technology, the video game sector has evolved into an industry generating millions in revenue. Games are often characterized by cutting edge graphics, realistic physics, and complex AI, supported by powerful consoles and mobile devices. Creators offer complex storytelling and artistic expression within their developed games. Culturally, video games no longer encompass small audiences. There are organized multiplayer gaming events and streaming platforms that offer more engagement from the community. Like it or not, video games are recognized as an integral part of global culture, influencing societal trends and values.

Video game sales are heavily influenced by multiple factors, including technological advancements, changing consumer demographics, and the growing variety of consoles offered to gamers. Among these factors, three stand out as being key elements affecting a video game's success, genre, publisher, and platform. Monetarily, it is of great value to a company to be able to estimate the number of sales. Worldwide, video games generated 406.2 billion dollars in revenue with an upward trend in revenue in the coming years. It is estimated that in 2029 video games would have generated 666.69 dollars in revenue [1]. This study aims to identify patterns, correlations, and trends that could inform predictive models and strategic decision-making in game development and marketing.





OBJECTIVES

In our analysis we aimed to predict video game sales using platform creator, publisher, genre, and regional sales data as our predictors. The goal was to identify which factors significantly influence global sales. The choice of models and techniques kept this goal in mind. A linear regression and random forest model were chosen to try and accommodate the mixture of categorical and numerical data. Linear regression, with its simplicity and interpretability, was useful in understanding the direct influences of the chosen predictors on a video games success. Multiple linear regression models were created to evaluate the usefulness/influence of certain chosen predictors. The predictors in the first model were genre, publisher, and platform creator. The predictors in the second model were North American, Europe, Japan, and other sales. The predictors in the third model were a combination of the two using genre, publisher, platform creator, and one of the four regional sales attributes. Seeing as there was bound to be non-linearity among the predictors, a random forest model was chosen for its robustness and ability to capture complicated relationships between the predictors. The primary metrics for evaluation were Mean Squared Error (MSE) and an R^2 score. A lower MSE would indicate the model has a closer prediction to the actual sales figure. An R^2 score, ranging from 0 to 1, indicates how well the variance in global sales is explained by the model. In the results we will be using R^2 to evaluate the explanatory power of each model and the MSE for prediction accuracy.

MATERIALS and METHODS

The language of choice in the analysis was Python. For the exploratory data analysis *Pandas, Numpy, Matplotlib, and Seaborn* were used. Building the models involved *Pandas and SciKit-Learn*. (**There will be more to follow here after building out the Random Forest).** The code will be made available on GitHub through the following public repository [3].

The dataset [2] contains a list of video games with sales greater than 100,000 copies. Fields include rank, name, platform, year, genre, and publisher. Fields also list sales from North America, Europe, Japan, and other countries, with a global sales field summing the prior. The data was gathered by scraping the website vgchartz.com. There are 16,598 records. The rank column was nothing more than just an incremental number ordered from greatest global sales of a video game to the least global sales. There was not much use found to that column, so it was not used in the analysis. There were also 329 missing values found, 271 coming from the year column, and 58 coming from the publisher column. A year attribute will not have much of an effect on our analysis and would be hard to correct without searching for every game and correcting. A publisher is needed for the analysis but would again be hard to correct without researching for each. With a total of 16,598 entries the rows with missing values could be dropped, with the dataset now containing 16,291 entries. The percentage of missing values in year and publisher columns was relatively small (1.63% and 0.35% respectively), so this approach shouldn't significantly impact the overall analysis.

RESULTS

The first linear regression model reported.

- MSE = 1.01×10^{23}
- $R^2 = -2.36 \times 10^{22}$

These results would suggest the model was not an appropriate fit when using the predictors genre, publisher, and platform creator. This outcome is likely due to the non-linear relationships between the features and our target variable.

The second linear regression model reported.

- MSE = 2.87×10^{-5}
- $R^2 = 0.9999932860147335$

As a reminder the features used in this model were all the regional sales data and targeted to predict the global sales data. The model fits very well. This result is expected however, as global sales are directly correlated with the sum of regional sales. The model is tailored to the data and is not of much use.

The third linear regression model reported.

- NA_Sales: MSE: 0.316, R^2: 0.925
- EU_Sales: MSE: 0.343, R^2: 0.919
- JP_Sales: MSE: 2.779, R^2: 0.350
- Other_Sales: MSE: 1.280, R^2: 0.700

Needing to complete the random forest analysis and add graphs

CONCLUSIONS

Needing to still finish

REFERENCES

1. https://www.statista.com/statistics/1344668/revenue-video-game-worldwide/
2. https://www.kaggle.com/datasets/gregorut/videogamesales