Predicting global video game sales: A data analysis of genre, publisher, platform, and sales region influences

Tanner Patrom

**Abstract**

This study aims to predict global video game sales using different machine learning techniques and a dataset comprised of various video game metrics. Initially, the predictive capability of linear regression models was explored using platform creator, genre, and publisher. However, these models underperformed, indicating these features did not do well in predicting global video game sales. Modified linear regression models incorporated regional sales data, significantly enhancing model accuracy as observed by improved mean squared error (MSE) and R² values.

To further explore model capabilities, random forests models were trained adapting similar approaches used when observing linear regression techniques. These models again showed limited success when only using the platform creators, genre, and publisher, but improved with the inclusion of regional sales data, confirming the importance of this metric in sales prediction using this data set. The study highlights the necessity of using a mix of sales data and other influential factors to accurately predict game sales. It also emphasizes the limitations of using only the categorical data platform creator, genre, and publisher without considering other characteristics such as platform availability, the price of a video game, the average play time of a video game, or more consumer demographic data.

**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 would 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.

**Methods**

In our analysis we aimed to predict video game sales using platform creator, publisher, and genre data as our predictors. The goal was to identify these factors as significantly influencing 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. Two random forest models were generated in order to evaluate their accuracy. The first would use the features genre, publisher, and platform creator. The second would use genre, publisher, platform creator, and North American sales. The primary metrics for evaluation were Mean Squared Error (MSE) and an R2 score. A lower MSE would indicate the model has a closer prediction to the actual sales figure. An R2 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 R2 to evaluate the explanatory power of each model and the MSE for prediction accuracy.

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***.** 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.

Of the data in the platform column, there were many legacy consoles listed. Some of these consoles were very popular in their time and had games that sold many copies. A heavy emphasis on a platform value of “PS2” would not be useful in our analysis. A new column was instead created “Platform\_Creator” of which a platform was mapped to the creator of that console i.e. “Sony” to “PS2”. Categorical variables also had to be encoded at times into numerical variables for model use.

In the exploratory data analysis, a series of graphical techniques were employed to understand the underlying structure and distribution of the data set. Figures 1, 2 and 3 show us the distribution of the number of games sold by each platform, genre, and publisher. Saturating the market were action and sports games released on platforms such as the Nintendo DS, Sony’s PlayStation, and Microsoft’s X-Box. There are also numerous publishers with EA, Activision, Namco Bandai, and Ubisoft leading the number of releases. Figures 4, 5, 6, and 7 modeled the total number of sales per each region, genre, platform, and publisher, giving us insight into top performing platforms and genres from a financial perspective. Action and sports games sold the most units and Nintendo was the top performing publisher. Of particular note is North America contributing the most to global sales. Figure 8 is made up of the new platform creator column simply modeling the total sales of each platform aggregated into a single creator.

**Results**

The results from each of the models will be listed below but can also be found aggregated in Table 1.

The first linear regression model reported.

* MSE: 6.45 × 1020
* R2: -1.51 × 1020

The second linear regression model reported.

* MSE: 2.87 × 10-5
* R2: 0.999

Figure 10 also shows almost perfect predictions being made on a scatter plot.

The third linear regression model reported.

* North American sales: MSE: 0.316, R2: 0.925
* European Sales: MSE: 0.343, R2: 0.919
* Japan sales: MSE: 2.779, R2: 0.350
* Other sales: MSE: 1.280, R2: 0.700

Figure 9 shows close predictions being made by the model as compared to the actual values. We can compare this with the near perfect values being predicted from the second linear regression model.

The first random forest model reported

* MSE: 3.0288
* R2: 0.095

Figure 11 gives us knowledge into the features that are being weighted heavily by the random forest.

The second random forest model reported.

* MSE: 0.679
* R2: 0.795

In relation the Figure 11, Figure 12 shows how much of a difference North American sales figures are weighted when added to the model.

**Discussion**

Our first analysis technique was linear regression. The results from the first linear regression model 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. As a reminder the features used in the second linear regression model were all of the regional sales data and were targeted to predict the global sales data. The model fits very well with a very low MSE and an R2 value approaching 1. 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 incorporates direct sales data from each of the four regions, along with platform creator, genre, and publisher. From the reported metrics, incorporating a sales region showed us that it is a more direct predictor of global sales compared to categories like platform creator, genre, and publisher. Regional sales figures are shown to have a more straightforward and stronger predictive relationship with the overall global sales, making the model more effective. Figure 9 models a scatter plot for the visualization of the predicted values vs. the actual values. Each point on the graph represents an observation from the dataset, where the x-coordinate is the actual value, and the y-coordinate is the predicted value. The black dashed line represents where the points would lie if the predicted values were exactly equal to the actual values. The closer the scatter points are to this line the better the model is at making predictions. Compared with Figure 10 where we can see the almost perfect predictions being made by the overfit second liner regression model.

Observing our next analysis technique, the first random forest model would suggest that the model might still not be effective in predicting the global sales of video games accurately, based solely on the predictor’s platform creator, genre, and publisher. The MSE values indicate a high amount of error relative to the sales figures, and the low R-squared value points to a model that does not capture much of the variability in the sales data. In observation of Figure 11, feature importance’s are modeled. These values represent the importance or contribution of each feature to the model in making its predictions. The model is doing its best to make accurate predictions and we can see it using popular parameters discovered in the exploratory data analysis. For the second random forest we again approached a better fitting model by adding more features that could influence game sales, such as the success in a particular region. Given that North America made up a strong portion of global video game sales it was added as a predictor. The improvements across all metrics suggest that North American sales is highly predictive of global sales. The model now not only predicts more accurately but also provides predictions that are close to actual sales figures. Seeing as North American sales(or possibly any other region) and global sales seem to be highly correlated this would indicate a limitation in the model and our prediction. There is a heavy reliance on regional sales figures to predict global sales figures. This is also reflected in another feature importance graph. In Figure 12 we can see North American sales figures rated much higher in importance than other features. The aim was to try and predict global sales independent of any particular region's performance.

For further analysis and an improvement in the predictive capabilities of the models chosen, without relying so heavily on regional sales data, we could try to enhance the dataset with additional relevant features that capture more dimensions of the factors influencing game popularity and market success. As examples, some points to capture that could help in further research are platform availability, the price of a video game, the average play time of a video game, or more consumer demographic data.

**References**

1. Clement, J. (2024, April 8). *Global video game industry revenue 2029*. Statista. https://www.statista.com/statistics/1344668/revenue-video-game-worldwide/
2. GregorySmith. (2016, October 26). *Video game sales*. Kaggle. https://www.kaggle.com/datasets/gregorut/videogamesales
3. https://github.com/PhysicalBit7/Intro-to-Data-Science

**Tables**

|  |  |  |
| --- | --- | --- |
| Table 1: Model performance given varying features | | |
| Model Name | MSE | R2 Score |
| Linear Regression 1 | 6.45 × 1020 | -1.51 × 1020 |
| Linear Regression 2 | 2.87 × 10-5 | 0.999 |
| Linear Regression 3(NA) | 0.316 | 0.925 |
| Linear Regression 3(EU) | 0.343 | 0.919 |
| Linear Regression 3(JP) | 2.779 | 0.350 |
| Linear Regression 3(Other) | 1.280 | 0.700 |
| Random Forest 1 | 3.0288 | 0.095 |
| Random Forest 2 | 0.679 | 0.795 |

**Figures**

**A graph of a number of numbers

Description automatically generated with medium confidence**

*Figure 1: Top platform by number of games*

*A graph of a number of green bars

Description automatically generated with medium confidence*

*Figure 2: Number of games by genre*

**A graph of blue and white bars

Description automatically generated**

*Figure 3: Top publishers by number of games*

**A graph of sales and sales

Description automatically generated with medium confidence**

*Figure 4: Regional contributions to game sales*

*A graph of a bar graph

Description automatically generated with medium confidence*

*Figure 5: Sales distribution across genres*

**A green and white graph

Description automatically generated**

*Figure 6: Sales distribution across platforms*

A graph of a bar graph

Description automatically generated with medium confidence

*Figure 7: Top 10 publishers by global sales*

*A graph of a bar graph

Description automatically generated with medium confidence*

*Figure 8: Global sales by platform creator*

*A dotted line with blue dots

Description automatically generated*

*Figure 9: Linear regression model 3 scatter plot including North American sales figures*

*A line graph with dotted line

Description automatically generated*

*Figure 10: Linear regression model 2 scatter plot*

*A graph of a number of people

Description automatically generated*

*Figure 11: Top 20 feature importance’s in the first random forest model*

*A graph with many names

Description automatically generated with medium confidence*

*Figure 12: Top 20 feature importance’s in the second random forest model that includes North American sales figures*