A person walking in a city

Description automatically generatedExploring the Dynamics of Video Game Sales

*A Data Analytics Project*

*By Adeel Shah*

Table of Contents

[Introduction 3](#_Toc140693088)

[Literature Review 5](#_Toc140693089)

[Dataset Description 9](#_Toc140693090)

[Visuals 12](#_Toc140693091)

[Results 18](#_Toc140693092)

[Evaluation Metrics 21](#_Toc140693093)

[Conclusion 24](#_Toc140693094)

[References 26](#_Toc140693095)

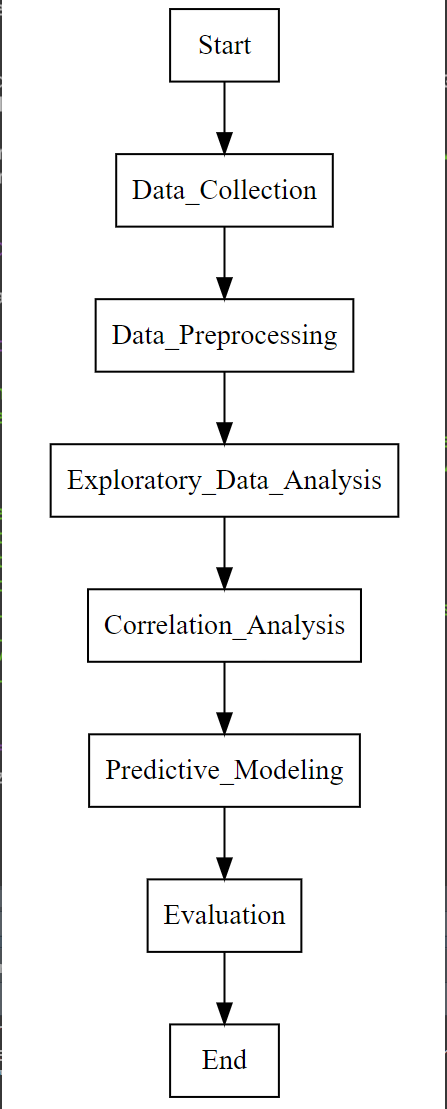
# Introduction

"What are the key factors influencing the sales performance of video games, and how can they be leveraged to optimize marketing strategies within the industry?" The purpose of this project is to answer this question and gain insights on the dataset I have chosen. Video Games can be seen as an interactive digital product that generates vast amounts of data and revenue in modern society due to how consumers interact with them. Types of video games, their genres, and their popularity in sales depending on region are all important factors in contributing to how widespread video games are in today’s age. A data-driven approach to how well video games perform as digital products can be used to identify trends, predict sales patterns, improve marketing strategies, and ultimately drive revenue growth in the industry.

The purpose of this data project will be to develop a predictive model on sales of video games using the “Video Game Sales Dataset Updated -Extra Feat” by Ibrahim Muhammed Naeem on Kaggle. (2023) The predictive model will assist with market analysis of video game sales as well as analysis of trends in video games, such as genre and popularity by platform. By analyzing a comprehensive and updated dataset of video game sales, this project will aim to identify key influencing factors that contribute to overall sales of video games.

The dataset I have chosen provides sales information on various video games worldwide across multiple platforms, developed by many different studios. The dataset features 16 columns and over 11000 values, with the dependent variables being the “NA\_Sales,” “EU\_Sales,” JP\_Sales,” “Other\_Sales,” & “Global\_Sales” columns for the predictive analysis portion of this project. The Theme of my data analytics project will be predictive analysis and correlation Analysis on influential variables within the dataset. By analyzing important attributes such as genre, platform, year of release, critic scores, and user scores, this project will aim to uncover patterns and relationships that contribute to sales performance.

This project begins with a thorough data preprocessing phase, where the dataset is carefully cleaned and prepared for analysis. Missing values are handled, outliers are identified, and necessary transformations are applied to ensure the quality and integrity of the data. Following data preprocessing, exploratory data analysis techniques are employed to gain a comprehensive understanding of the dataset's characteristics and uncover initial insights into the sales patterns of video games.

The findings of my predictive analysis will provide valuable insights for video game developers and publishers in understanding dynamic market insights on the types of video games they should release. This Data Analytics Project will delve into factors influencing video game sales by analyzing the dataset through preprocessing, exploratory data analysis, correlation analysis, and predictive modeling. The Following graph shows the methodology process of the data Project:

In this data project, various machine learning algorithms were employed for the purpose of the predictive analysis portion. The algorithms utilized included: Random Forest, Regression, Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost). These algorithms were chosen based on their suitability for the dataset and their ability to handle both numerical and categorical variables. The Random Forest algorithm was used to capture complex relationships and interactions among the selected features, while Regression model was employed to estimate the impact of individual predictors on the target variable. Support Vector Regression was utilized for nonlinear regression tasks, and Extreme Gradient Boosting was employed to build a powerful ensemble model. By exercising multiple algorithms, a thorough analysis of the dataset was conducted, enabling the identification of significant predictors and the development of accurate predictive models.

# Literature Review

The article “What Makes a Blockbuster Video Game? An Empirical Analysis of US Sales Data” by Joe Cox (2013) attempts to provide information on video games that have sold tremendously well, gaining the “blockbuster” title commonly associated with movies. The article states that a comparatively recent addition to these more familiar and mainstream forms of entertainment is video gaming. The article goes on to suggest that 14% of titles generate over 70% of industry sales due to the platform and the publisher of the games. The article also states that in the sample of the video game titles analysed, it was acknowledged that the top 5% of titles had been responsible for more than 50% of software revenue, whereas only 10% of games earn a profit and nearly half of all titles sell fewer than 10,000 copies. When we look at Figure 1 of “NA\_Sales” from the Video Game Sales Dataset, we can see that the conclusions in this article hold truth. Notable investigations in Joe Cox’s own studies are that major publishers, particularly Nintendo, are found to be associated with significant differences in the unit sales of video game titles. Cox found in each of his logistic regression models that platform of release is found to have a significant influence on the likelihood of a video game title becoming a “blockbuster”. Cox infers that a game that has the best chance of becoming a blockbuster would need to be released on popular platforms by a major publisher, while also offering a high-quality gameplay experience. This study investigates the factors that affect the likelihood of a video game becoming a blockbuster.

A picture containing text, screenshot, diagram, font

Description automatically generated

Figure

The purpose of the article “The Impact Of Platform On Global Video Game Sales,” by Babb et al. (2013), is to analyze the video game industry with a focus on comparing global sales across different gaming platforms. This article carries useful information that will help in better understanding specifically the “Platform” and “Global\_Sales” columns in the Video Game Dataset, and their correlations. The purpose of this research is to compare global video game sales by gaming platform for the years 2006 through 2011. In this study, a Kruskal-Wallis test is used to examine the sales performance of eight different gaming platforms. The findings reveal that Nintendo's Wii emerged as the top-selling global platform. Following closely behind is the Nintendo DS, which falls into the second tier. The third tier includes the Xbox 360, Sony PlayStation 3, and personal computer (PC). The fourth tier comprises the Sony PlayStation 2 and Sony PSP. Lastly, the lowest sales tier is occupied by the retired sixth-generation Nintendo GameCube. When we look at the most common platforms in the Video Game Dataset in Figure 2, we can see that the findings in this article are relevant, though not entirely accurate, to the data that will be analyzed further for this project; the Dataset covers a broader range of entries from 1980 to 2016.A picture containing text, screenshot, diagram, font

Description automatically generated

Figure

While previous studies have explored the relationship between product sales and Internet search data, the application of time series analysis in empirical research related to the video game industry has been limited. Specifically, an important research question is whether the current weekly sales volumes of video games are influenced by past volumes of Internet searches, or if it works the other way around. By addressing this question, Ruohonen & Hyrynsalmi (2017) in “Evaluating the use of internet search volumes for time series modeling of sales in the video game industry.” make a valuable contribution to the field of video game sales modeling in general. The article utilizes two time series models using volume of video game games and volume of google searches to analyze the correlation and importance of the two. In one specific study, the article claims that both sales of the video game: “God of War II” and its search frequency were relatively similar in their respective time series patterns. This article presents another dimension of influencing factors, notable popularity in search terms for video game releases and their correlation with sales.

Similar to Cox’s (2013) use of “blockbuster” in his article “What Makes a Blockbuster Video Game? An Empirical Analysis of US Sales Data,” Binken et al. (2009) use the term “superstar” with a parallel meaning in their study: “The Effect of Superstar Software on Hardware Sales in System Markets.” This article holds importance to the overall analysis of the video game dataset due to findings by Binken et al. (2009) which illustrates the correlation between user score, or “quality” as they described it, and the overall sales of a video game. Observations we can make from this study is that the market exhibits a non-proportional response of video game sales to video game quality. Additionally, there is a consistent and increasing relationship between video game quality and the number of video game units sold. Binken et al. (2009) suggest a scarcity of high-quality products, which contributes to the increasing returns associated with quality. These conclusions support the notion that the video game industry operates as a “superstar” industry, where a small number of highly rated games enjoy significant popularity and generate disproportionate returns.

Haviv et al.’s (2020) “Intertemporal Demand Spillover Effects on Video Game Platforms” provides information on factors that affect sales indirectly in the market. The article goes over platform sales and the platform economy which are aware of the hardware/software spillover effect, namely, that an increase in sales on one side of the platform will boost demand on the other. When we look at video games sold on a certain platform in and around the time of its year of release, we should naturally see an increase in sales of video games as the frequency of platform increases. Haviv et al. suggest that the combined variables of platform and release year play an important factor in the spikes of sales data we see in video game sales in and around a multiyear gap. Understanding and leveraging the spillover demand effect is important for game developers and publishers when planning their marketing and release strategies. By strategically timing the release of their games to coincide with highly anticipated titles or by creating synergies between their games and popular franchises, they can capitalize on the spillover effect and potentially boost their sales.

This literature review highlights articles that study the multiple factors of video game sales and their importance. Understanding and using those factors are essential in the analysis of the Video Game Dataset and the commercial sales of video games. Through an examination of these studies and research articles, it becomes apparent that video game sales are influenced by a combination of factors including online popularity through search terms (Ruohonen & Hyrynsalmi 2017), hardware platform sales, both indirect (Haviv et al. 2020) and direct (Binken et al. 2009), year of release, and critical reviews. The findings imply that game quality also plays a crucial role in attracting consumers and driving sales. My personal findings with the Video Game Dataset will prove to look for likeness and patterns in the data studied in the literature reviewed, but more importantly look towards the data that does not provide likeness in the results found by the authors of the literature.

# Dataset Description

When we look at the profile report done in Python through the pandas\_profiling package, we can discover many interesting insights from the Video Game dataset. When we look at the common values of various attributes like: Year of Release, Platform, Publisher, we get a better understanding on the time frame or when most of these products were released, as well as which platforms and publishers were dominating the market in that similar timeframe.



Figure

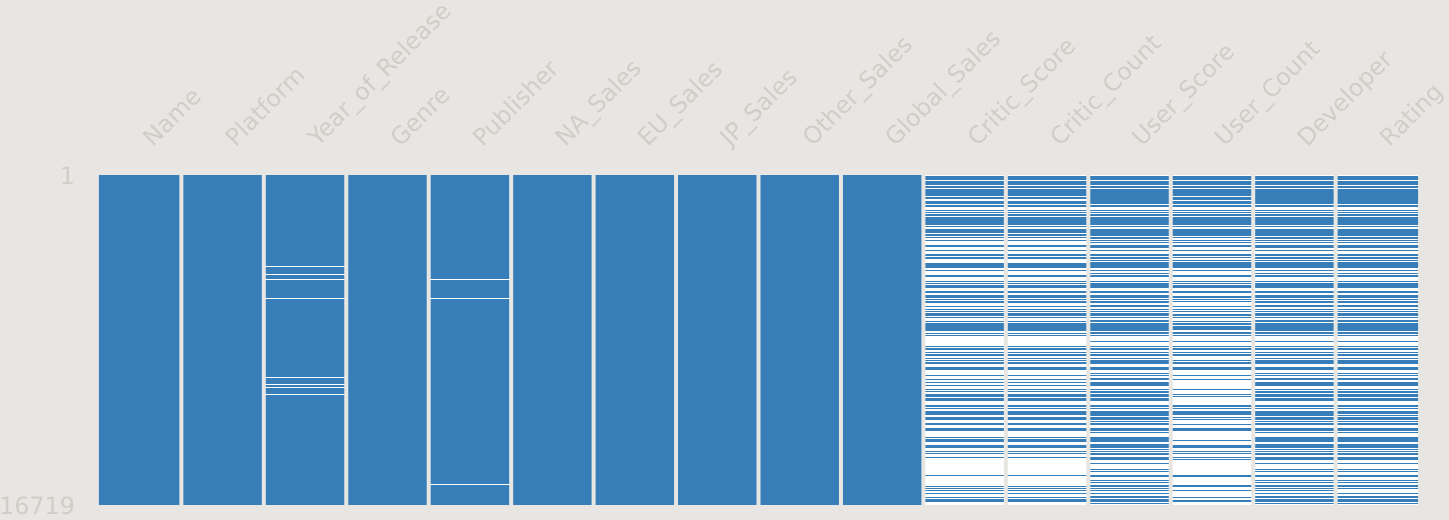
As shown in Figure 3, which shows common values for Year of Release within the dataset, we can see that most titles in the dataset were released in the year 2008 and the top five release years are from the late 2000s. In Figure 4 we can see that the publisher with the most sales is very clearly Nintendo. While both Nintendo and Electronic Arts dominate the global sales of video games, Nintendo sells a whopping 2.5x more than the third most profitable publisher: Activision. Insights such as these will provide the opportunity for predictive analysis on the publishers with the most data associated with them during the next step of this Data Analysis project.

A screen shot of a computer

Description automatically generated with low confidence

Figure

In Figure 5 we can see that the completeness of the dataset, represented by the missing values matrix, is focused on the sales categories which will provide sufficient material to draw predictive and correlation analysis. The less complete data fields of critic score, user score, and critic count and user count will not be the most significant variables to draw conclusions from due to the lack of entries, but these fields will still be used to draw correlation analysis from other variables, such as ratings and sales.



Figure

In fact, upon plotting the graphs for sales by user score and critic score, a noticeable trend emerged. Both distributions exhibited a right-skewed pattern, indicating that as the score decreased, the sales also declined. This observation aligns with the general expectation that higher-rated games tend to garner more sales. However, an intriguing finding emerged when examining the peak of sales in relation to the scores. Surprisingly, the peak sales were concentrated around the 80s in both user and critic scores, suggesting that the highest-rated titles may cater to a more niche audience rather than appealing to the broader market. As a consequence, these highly rated titles may experience lower overall sales despite their exceptional scores. It is noteworthy that when examining the data closely, we observed that the highest-rated titles from the users' perspective tended to have modest sales figures. On the other hand, titles with slightly lower ratings received recognition from critics but still exhibited relatively lower sales figures compared to titles with slightly lower ratings. This finding suggests that highly acclaimed titles may possess a level of prestige and critical acclaim but may not resonate as strongly with the wider audience in terms of sales. The following figure visualizes this point: A graph of a performance

Description automatically generated

Figure

# Visuals

When we explore the dataset, we can find correlations that may prove useful to the feature selection process once we move into the machine learning algorithm phase of analysis for this dataset. Exploring the dataset started with finding correlations and insights that directly visualized sales through the other categorical variables in the dataset. Looking at year of release and global sales, we can see that the sales of video games peaked in 2008, and the A graph of sales

Description automatically generatedyears where sales were at the highest, were in an around the late 2000s as we can see in Figure 7. However, upon further consideration, the use of release year in my machine learning algorithm did not seem appealing, as the factors of year don’t logically hold weight in predictive analysis, rather they provide an idea of the history behind sales patterns in and of itself.

Figure 7

When looking at the best features to use for machine learning algorithms, I decided to go with platform, genre, publisher and developer, with the target variable being global sales. Although publisher and developer overlap in terms of sales patterns, as many companies do both publishing of games and developing them in-house, I found some variance between the two variables in terms of sales, notably that the top selling developers and top selling publishers were not entirely identical. The developer and publisher of a video game are key players in the industry, responsible for creating and distributing games. By incorporating these variables, we can examine the impact of different developers and publishers on sales performance. The bar plots of sales per developers and sales per publishers highlight the variation and potential overlap among top performers, indicating the importance of considering both aspects in our A graph of different colored bars

Description automatically generatedA graph of sales

Description automatically generatedanalysis.

Figure

Figure

The genre of a video game provides insights into its content and target audience, allowing us to explore the relationship between different genres and their respective sales. By including genre as a feature, we can assess how different genres contribute to the overall sales patterns and identify genres that are particularly successful or have untapped potential. Action, the genre which most games fall under, is the dominant and highest selling genre globally, except for in Japan. Figure 10 shows the popularity in sales among all the genres globally. The analysis of regional sales in relation to genre reveals intriguing patterns and disparities among different markets. One noteworthy observation is the disproportionate popularity of the role-A pie chart with text and numbers

Description automatically generatedplaying genre in Japan compared to Europe and the US. Role-playing games tend to enjoy significantly higher sales in Japan, reflecting the cultural preferences and strong demand for immersive storytelling and character-driven experiences in the Japanese gaming market. This finding, illustrated in Figure 11, highlights the importance of considering cultural factors and regional preferences when analyzing sales trends. On the other hand, the shooter genre exhibits interesting variations across regions. While it ranks among the highest-selling genres in Europe and the US, after the action genre, it experiences relatively lower sales in Japan specifically. This disparity may be attributed to differences in gaming preferences and cultural influences. Shooters often emphasize fast-paced action and competitive gameplay, which may resonate more with the Western gaming audience compared to Japanese gamers who may lean towards other genres, evidently role-playing. Additionally, genre analysis reveals that the "Other\_Sales" category, representing sales in regions other than Europe, Japan, and the US, closely mimics the sales in the US. This suggests that the gaming markets in other regions may exhibit similar preferences and consumption patterns to the US market. These findings underscore the importance of understanding the interplay between genre and regional sales dynamics in the video game industry. By recognizing these distinct patterns and regional variations we can conclude that genre plays a large factor in the sales of video games, and that publishers and developers can maximize the product appeal depending on the genre and region.

Figure 10

A graph of different sizes and shapes

Description automatically generated Finally, I looked at platform as the last feature to use for machine learning algorithms I implemented onto the dataset. Analysis on the platform variable within the dataset produced interesting results, especially around how a lot of titles in the dataset were released on more than one platform. The choice of including "Platform" as a feature for analysis was motivated by its potential impact on the sales performance of video games. I wanted to explore how the availability of a game on multiple platforms or its exclusivity to a particular platform could influence its sales. To conduct my own analysis on the "Platform" variable, I first explored the distribution of video games across different platforms. I found that many titles were released on multiple platforms, indicating a wide range of options for consumers. I created a new variable called "Multi-platform" that indicated whether a game was released on multiple platforms or not. This additional variable allowed me to capture the potential influence of multi-platform availability on the sales figures. One notable finding from the analysis was that multi-platform games generated approximately 30% more sales, as seen in Figure 12, compared to games that were not available on multiple platforms. This observation highlighted the importance of catering to a wider audience by releasing games on multiple platforms, as it significantly contributed to higher sales figures. This finding highlighted the value of platform availability as a critical consideration in driving sales success in the video game industry. The analysis of the "Platform" A red and blue graph

Description automatically generatedvariable, along with the creation of the "Multi-platform" variable, and the subsequent decision tree modeling and confusion matrix evaluation, provided insights into the role of platform availability in video game sales.

Figure 11

Figure 12

Based on the confusion matrix and the evaluation metrics obtained, we can interpret the performance of the classification model for predicting multi-platform availability in video games. The model achieved an accuracy of 0.602, revealing that it correctly classified approximately 60% of the instances. The precision of 0.617 suggests that when the model predicted a game to be multi-platform, it was correct around 62% of the time. The recall, or sensitivity, of 0.863 indicates that the model identified a high proportion of actual multi-platform games correctly. The specificity of 0.222 suggests that the model had difficulty correctly identifying non-multi-platform games. Finally, the F1 score of 0.720 provides a balanced measure of precision and recall, indicating the overall effectiveness of the model. These results highlight the model's ability to identify multi-platform games with a relatively high recall rate, but also the challenges it faced in correctly identifying non-multi-platform games. These challenges are justified considering many developers and publishers release titles as both exclusives and multiplatform games.

# Results

The initial analysis of the video game dataset involved conducting an Analysis of Variance (ANOVA) to explore the relationship between the selected variables (Genre, Platform, Publisher, Developer) and the target variable: Global Sales. The ANOVA results revealed significant differences across the multiple selected features. These findings indicated that these variables played a crucial role in explaining the variation in video game sales performance. In fact, the results of the ANOVA indicated a high F-value between all four of the selected variables, as well as a very small p-value, well below 0.001, signifying a great difference between the variables in relation to sales. Building upon this initial exploration, further investigation was conducted using various machine learning algorithms, including Regression, Random Forest, SVR, and XGBoost. These algorithms were employed to develop predictive models and gain deeper insights into the factors driving sales in the video game industry. To assess the performance and generalizability of these models, rigorous cross-validation techniques have been applied, yielding valuable evaluation metrics and stability analysis. The following section will provide detailed discussions on the performance, effectiveness, and stability of each machine learning algorithm, shedding light on their respective contributions in predicting video game sales and informing strategic decision-making in the industry.

The linear regression model was employed using the features of developer, genre, and platform to predict the global sales. To prepare the data, the character columns were transformed into factors, and a 70/30 train-test split was performed. The linear regression model yielded the following results: root mean squared error (RMSE) of 2.620288, coefficient of determination (R-squared) of 0.191112 and mean absolute error (MAE) of 1.673714. These metrics provided an initial assessment of the model's performance in predicting global sales, which will be further evaluated and compared with other machine learning models in the subsequent evaluation metrics section.

A graph with numbers and dots

Description automatically generatedFor the random forest algorithm, I applied one-hot encoding to the categorical variables of genre, platform, and developer. This encoding transformed the data into a matrix format suitable for the random forest algorithm, enabling improved prediction accuracy in theory. The encoded data was then split into separate train and test sets. Subsequently, I trained the random forest model and obtained the following results: The model consisted of 100 trees, and during each split, 111 variables were considered. The mean of squared residuals (MSR), a measure of the average squared difference between predicted and actual values, was 0.6867701. Additionally, the model explained approximately 93.13% of the variability in the data, as indicated by the % Var explained metric. These results provided insights into the performance and explanatory power of the random forest model for predicting global sales. Initially, with an MSE of 1.938, the model had larger prediction errors, indicating less accuracy in capturing the underlying patterns in the data. As the MSE decreased to 0.686, it suggests that the model's predictions are closer to the actual values, indicating improved performance, as seen by Figure 13. A lower MSE indicates a better fit of the model to the training data, as it indicates smaller average squared differences between predicted and actual values. However, it's important to consider the other evaluation metrics, such as R-squared, to assess the overall performance and predictive power of the model which I will describe in the next section.

Figure 13

For the SVR (Support Vector Regression) algorithm, the data preparation steps were similar to the random forest algorithm, involving data cleaning, one-hot encoding, and data splitting. Once the SVR model was trained, the initial evaluation metrics were calculated. The RMSE (Root Mean Squared Error) of 0.9786961 indicates the average magnitude of the model's prediction errors. A lower RMSE value suggests that the model's predictions are closer to the actual values. The R-squared value of 0.1596329 measures the proportion of variance in the target variable (global sales) that can be explained by the model. This value indicated the goodness of fit of the model to the data, which was not very high, suggesting this model by itself was not a good fit for the data. The SVR model parameters, such as epsilon (0.1), gamma (0.003012048), and the number of support vectors (568), provided insights into the specific configuration of the model. These parameters control the trade-off between model complexity and accuracy, as well as the flexibility in capturing the underlying patterns in the data.

Lastly, I used the XGBoost algorithm for the analysis of the video game dataset. Like the previous algorithms, I performed the necessary data preparation steps, including data cleaning, one-hot encoding, and data splitting. For the XGBoost model, I defined the key parameters: eta (learning rate) as 1, max depth as 3, and rounds as 100. During model training, I set the number of rounds to 10. The evaluation of the XGBoost model revealed an improvement in the RMSE (Root Mean Squared Error) from 4.037 to 2.677 as the model iterated through the 10 rounds. The decreasing RMSE values indicate that the XGBoost model's predictions became more accurate and closer to the actual values with each iteration. Although the accuracy nearly doubled after 10 iterations, the results of this model produced diminishing returns with more iterations set, as using 100 iterations would bring the RMSE from its original 4.037 to just 1.863.

# Evaluation Metrics

A graph of a graph

Description automatically generated Regarding the regression model evaluation metrics, the initial RMSE of 2.893 indicates the average difference between the predicted and actual values of the target variable (global sales) is approximately 2.893 units. After predicting on the test data, the RMSE decreases to 2.367, suggesting an improvement in the model's predictive accuracy. The R-squared value, a measure of the proportion of variance explained by the model, decreases from 0.222 to 0.111, indicating that the model explains about 11.1% of the variability in global sales. The MAE (mean absolute error) decreases from 1.844 to 1.087, implying that, on average, the model's predictions are off by approximately 1.087 units. Applying cross-validation to the regression model provides an additional set of evaluation metrics, which are even higher than the initial model. This suggests that the model's performance may vary when tested on different subsets of the data. Furthermore, when examining Cook's distance, three observations stand out as influential data points, as shown in Figure 14. These observations have a notable impact on the regression coefficients and may have a disproportionate influence on the model's predictions. In terms of the three areas of evaluation, the regression model demonstrates effectiveness by providing reasonable predictions of global sales, although the explained variance is relatively low. However, it is important to note that the model's performance may vary when tested on different subsets of data, indicating some instability. From an efficiency perspective, the training and testing time for the regression model is relatively quick. However, the presence of influential data points, as indicated by Cook's distance, suggests the need for further investigation into potential outliers or influential observations.

Figure 14

Evaluating the random forest model, the MSE (mean squared error) increases from 0.531 to 0.653, indicating a larger average difference between the predicted and actual values of the target variable (global sales). The SSR (sum of squared residuals) is 3257.889, representing the sum of the squared differences between the predicted and actual values. The SST (total sum of squares) is 4277.262, which quantifies the total variability in the target variable. The R-squared value of 0.238 suggests that the random forest model explains approximately 23.8% of the variability in global sales. Performing cross-validation on the random forest model with 5 folds allows for a more comprehensive evaluation of its performance. The optimal value of “mtry”, which determines the number of variables considered at each split, is determined to be 165. The cross-validated model yields an RMSE of 2.732, indicating the average prediction error of approximately 2.732 units. The R-squared value improves to 0.373, suggesting a higher proportion of variance explained by the model. The MAE (mean absolute error) is 1.489, indicating the average absolute difference between the predicted and actual values. From an effectiveness standpoint, the random forest model shows some improvement in terms of predictive performance, with a higher R-squared value and reduced RMSE compared to the initial model. However, the results may not be satisfactory, especially considering the long compute time required to train the random forest model on a dataset of this scale. In terms of efficiency, the extended compute time for cross-validation raises concerns about the viability of using the random forest algorithm for this particular dataset.

The SVR model attempts to capture the complex relationships between the selected features (developer, genre, and platform) and the target variable (global sales) by mapping the data points to a higher-dimensional space. Evaluating the SVR model, the RMSE increases slightly from 0.997 to 1.102 on the test set, indicating a slightly higher average prediction error. The R-squared value decreases from 0.173 to 0.128, suggesting that the model explains a smaller proportion of the variance in global sales. However, it is important to note that these metrics should be interpreted in the context of the dataset and the complexity of the relationships being modeled. Performing cross-validation on the SVR model with sigma = 0.003335581 and C = 1 provides further insights into its performance. The RMSE value of 3.505 indicates the average prediction error, which is relatively higher compared to the initial model. The R-squared value of 0.080 suggests that the model accounts for approximately 8% of the variance in global sales. The MAE value of 1.624 represents the average absolute difference between the predicted and actual values. In terms of effectiveness, the SVR model shows moderate performance in predicting global sales based on the selected features. The relatively lower R-squared value indicates that the model may not capture the full complexity of the relationships, possibly due to the inherent noise or other factors in the dataset. As for efficiency, the training and testing times of the SVR algorithm are typically lower compared to more computationally intensive models like random forest. For this analysis, the SVR model exhibits relatively low stability. These metrics suggest that the model's performance varies significantly across different subsets of the data, indicating instability in its predictions.

In the context of this dataset, XGBoost was employed to predict global sales based on various features. The evaluation metrics obtained from the initial evaluation of the XGBoost model, including an RMSE of 2.773, MSE of 10.203, MAE of 1.716, and R-squared of 0.084, provided insights into the performance of the model, namely that it doesn’t perform well with this dataset. By employing cross-validation with tuned hyperparameters such as nrounds, max\_depth, eta, gamma, colsample\_bytree, min\_child\_weight, and subsample, we further evaluated the XGBoost model's performance. The cross-validated metrics of RMSE: 2.886894, R-squared: 0.3376801, and MAE: 1.604984 provided a more robust assessment of the model's effectiveness. From an effectiveness perspective, the XGBoost model demonstrates reasonable performance, as indicated by the relatively low RMSE and MAE values compared to other models. The R-squared value of 0.3376801 suggests that the model explains approximately 33.77% of the variance in global sales. In terms of efficiency, the training and testing times for the XGBoost model were relatively efficient, making it a practical choice for this dataset. Lastly, the stability of the XGBoost model is evident from consistent results across different subsets of the data, as demonstrated by the cross-validated metrics. The XGBoost algorithm has shown promising results in predicting global sales in this dataset. Its effectiveness, efficiency, and stability make it a valuable tool for analysis and prediction tasks.

# Conclusion

This data project aimed to investigate the key factors influencing the sales performance of video games and explore how they can be leveraged to optimize marketing strategies within the industry. Through extensive data cleaning and preprocessing, as well as in-depth exploratory data analysis, valuable insights were gained from the dataset. Four machine learning algorithms, namely Regression, Random Forest, SVR, and XGBoost, were employed to predict global sales based on various features. Each algorithm provided its own unique perspective and performance metrics, contributing to a comprehensive understanding of the dataset. Amongst the algorithms utilized, the XGBoost model emerged as the most favorable performing algorithm. It showcased relatively lower values of root mean squared error (RMSE) and mean absolute error (MAE), indicating more accurate predictions. Additionally, its higher R-squared value suggested a better fit of the model to the data, as opposed to the other algorithms.

Throughout this project, working with this dataset has been a rewarding experience, given my personal interest and passion for video games. Despite its limitations, such as the data being dated and containing missing information, that did not hinder the overall data analysis process. In conclusion, this data project sheds light on the factors influencing video game sales and provides valuable insights for industry professionals. The XGBoost algorithm, with its superior performance, demonstrates its potential in optimizing marketing strategies within the video game industry. This project sets a foundation for future research and analysis, paving the way for further exploration of this fascinating domain.

# References

GitHub: <https://github.com/Adeelshah001/Data-Project.git>

Ibrahim Muhammad Naeem. (2023). <i>Video Game Sales Dataset Updated -Extra Feat</i> [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/4984906.

Cox, Joe. (2014), What Makes a Blockbuster Video Game? An Empirical Analysis of US Sales Data. *Manage*. Decis. Econ., 35: 189-198. <https://doi-org.ezproxy.lib.torontomu.ca/10.1002/mde.2608>

Babb, J., Terry, N., & Dana, K. (2013). The Impact Of Platform On Global Video Game Sales. The International Business & Economics Research Journal (Online), 12(10), 1273. <https://doi.org/10.19030/iber.v12i10.8136>

Ruohonen, J., Hyrynsalmi, S. Evaluating the use of internet search volumes for time series modeling of sales in the video game industry. *Electron Markets* **27**, 351–370 (2017). <https://doi-org.ezproxy.lib.torontomu.ca/10.1007/s12525-016-0244-z>

Binken, J. L. G., & Stremersch, S. (2009). The Effect of Superstar Software on Hardware Sales in System Markets. Journal of Marketing, 73(2), 88–104. <https://doi-org.ezproxy.lib.torontomu.ca/10.1509/jmkg.73.2.88>

Avery Haviv, Yufeng Huang, Nan Li (2020) Intertemporal Demand Spillover Effects on Video Game Platforms. Management Science 66(10):4788-4807. <https://doi-org.ezproxy.lib.torontomu.ca/10.1287/mnsc.2019.3414>