Carnegie Mellon University

Global Sales Prediction for Video Games

Final Report

Data Science

May 2023

1. Background and Motivation, introduction

The video game industry is a vast and multifaceted sector, with its market size in the United States currently reaching $47 billion and expected to maintain rapid growth in the foreseeable future. Driven by rapid technological advancements and ever-changing consumer preferences, the industry is in a state of continuous transformation. To better understand this field, it is crucial to explore the key aspects that shape it.

To study this field, the following key aspects are worth focusing on: development platforms, game publishers, sales, developers, and game ratings. Development platforms include various consoles, such as Sony's PlayStation and Microsoft's Xbox, each with unique features, user experiences, and audiences. Game publishers are responsible for financing, marketing, and distributing games across different platforms, and they may sign exclusive agreements with specific platforms or create games for different platforms. Game developers are the creative forces behind video game design, programming, and production. In the video game industry, developers can work independently, as part of an in-house team for publishers, or in collaboration with publishers. Game ratings come from various groups, representing their evaluations of a particular game. The aspects we talked about before significantly impacted sales. Additionally, there may be significant differences in the performance of the same game between different countries and regions due to certain factors, such as economic levels and cultural differences.

2. Research Questions

The purpose of this study is to identify key predictive factors that influence game sales by analyzing data sets and developing several models to predict global game sales. We also want to conduct exploratory data analysis to find potential correlations between features and explore the real-life meanings behind the relationships. Additionally, by fitting the dataset with different models and fine-tuning hyperparameters, we would like to evaluate the model performance of each model. Ultimately, we will combine the findings from the above research to provide recommendations to game developers or publishers.

3. Literature review

In recent years, the video game industry has experienced exponential growth, and various studies have explored different aspects of the industry. The literature review will provide an overview of the current state of knowledge in this field by examining research on development platforms, game publishers, sales, developers,

and game ratings. Kretschmer et al. (2012) analyzed the competition between Sony PlayStation and Microsoft Xbox, emphasizing the role of network effects and exclusive content in driving platform adoption. Teixeira and Karahanna (2018) investigated factors influencing consumer choice of gaming platforms, such as brand loyalty and the size of the game library. Nieborg (2015) discussed the importance of publishers for game sales, highlighting strong correlations between publishers, developers, and platforms with game sales. Mäntymäki and Salo (2013) focused on the impact of exclusive games and cross-platform releases on game sales. Whitson (2013) discussed the role of independent game developers in subverting traditional power structures in the industry, with these small developers being more innovative compared to well-established, richly resourced larger developers. Wertz (2015) examined the relationships between critical reviews, user ratings, and sales, suggesting that both types of evaluations influence consumer behavior.

These literature sources provide valuable insights into various aspects of the video game industry, such as development platforms, game publishers, sales, developers, and game ratings. Building on this content, we will further explore these key predictive factors by analyzing datasets and developing predictive models to estimate the impact of these factors on game sales.

4. Data Sources and Pre-processing

In this project, we aimed to predict video game sales using various features such as genre, platform, publisher, and critic scores. We utilized the comprehensive Video Game Sales and Ratings Dataset, which includes 16,598 data entries on video games before cleaning the data, covering 15 aspects such as title, platform, launch year, genre, publisher, sales figures, and scores from Metacritic and ESRB. This data resource allows researchers and analysts to explore the connections between game ratings, player involvement, and financial success. Detailed regional sales data facilitates the analysis of regional performance and preferences, and the integration of critic and user ratings offers a comprehensive perspective on a game's overall reception. Additionally, the dataset contains ESRB age ratings, providing crucial information on the intended audience for each game. In summary, this dataset offers a rich source of data for examining trends and patterns in the gaming industry, making it an indispensable tool for both professionals and enthusiasts. We developed machine learning models to provide insights into the factors that contribute to the success of video games in the market. By understanding these factors, we can answer our research question: What are the key drivers of video game sales, and how can we leverage this information to develop successful games?

In order to improve the quality of our dataset and enhance the predictive power of our machine learning models, we focused on improving the dataset by addressing missing values and transforming features to enhance the predictive power of our machine learning models. We filled in missing values by analyzing the distribution of features, using median values for user count, JP sales, critic count, and user score, and mean values for the remaining columns (this is based on the distribution of the variables, if it follows a more like a normal distribution it will be filled in with the mean value, otherwise, median, see Appendix 1). We excluded regional sales features (NA sales, EU sales, Other sales, JP sales) in our final regression model due to their strong linear relationship with global sales as in **Figure 1** shows strong correlations between global sales and sales from different regions of the world. Additionally, transformed the year of release into game age to be a more interpretable variable to include in the model.

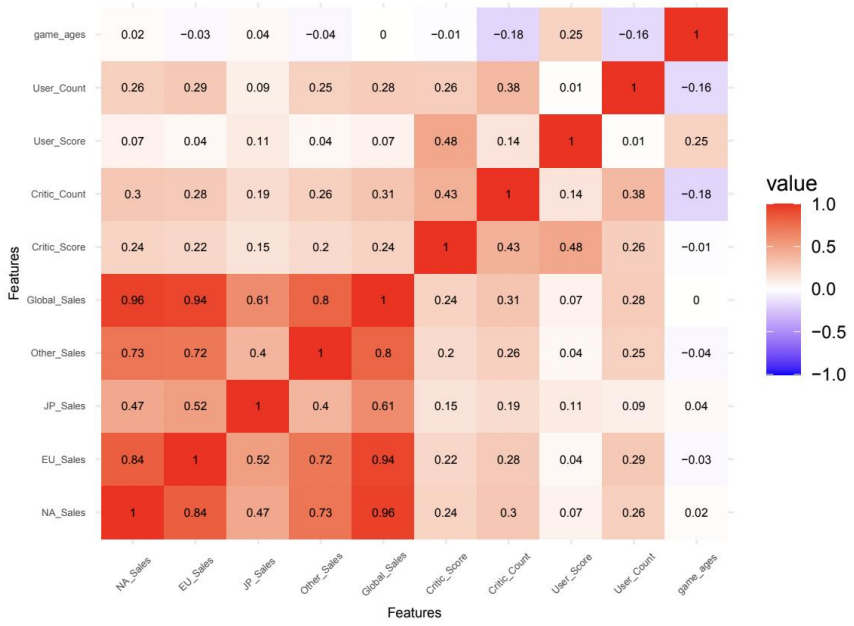


Figure 1: Video games heatmap

During the Data Pre-processing phase, we performed several feature engineering tasks to improve the predictive power of our models. First, we encoded categorical features such as genre, platform, and publisher using one-hot encoding (dummy coding) to convert them into numerical values that can be used by machine learning algorithms. Additionally, we grouped publishers with fewer than 10 entries into a new "Other Publisher" category. Next, we examined the distribution of our target variable, video game sales, and discovered that it was heavily skewed. To address this issue, we applied a log transformation to the sales data, creating a more normally distributed target variable that is more suitable for linear models. The log transformation was necessary as the Global Sales feature did not have a normal distribution (see **Appendix 2**).

We began our analysis with an Exploratory Data Analysis (EDA) to understand the data and identify trends. Plots of the dataset can be found in the code. For example, figure 2 shows a boxplot of Global Sales by Platform. We noticed that certain genres, platforms, and publishers tended to have higher sales than others, suggesting that these factors may play a significant role in determining a game's success. Additionally, we observed a positive correlation between critic scores and sales, implying that well-reviewed games are more likely to sell well. Through the use of plots, it helped us gain a better understanding of the relationships between various features and the target variable, video game sales.

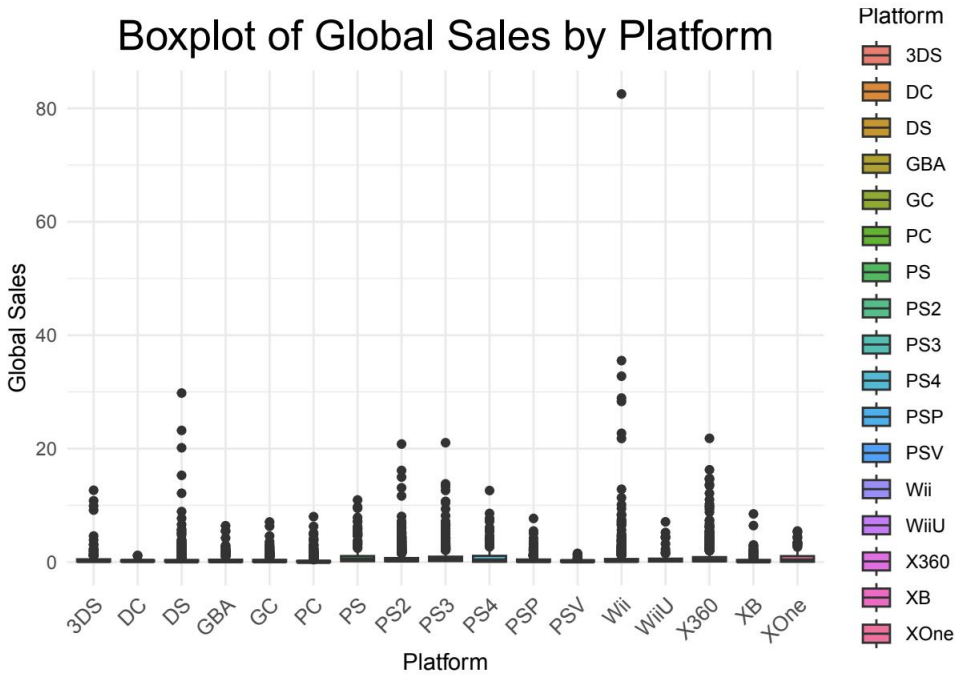


Figure 2: Boxplot of Global Sales by Platforms

Moreover, we joined other data sources obtained from a Wikipedia page titled "List of best-selling game consoles". This dataset contains information about various game consoles, including the platform name, the number of units sold, the company that produced the console, and the console's release year. The features in the second dataset include Platform, Units sold in millions, Firm that manufactured the game console, and Released which is the year the game console was released.

The two datasets were joined on the "Platform" feature using a left join, which combined the data based on matching platform names from both datasets. This process enriched the original dataset with additional information about the game consoles, such as the manufacturer and the total number of units sold. A bar plot and a pie chart were created to visualize the global sales data. The bar plot in **Figure 3** displays the global sales of video games by platform manufacturer (Firm). This provides insights into which manufacturer's consoles

have the highest total sales. The pie chart in **Figure 4** shows the proportion of global sales by platform manufacturer, giving a clear representation of the market share of each manufacturer.

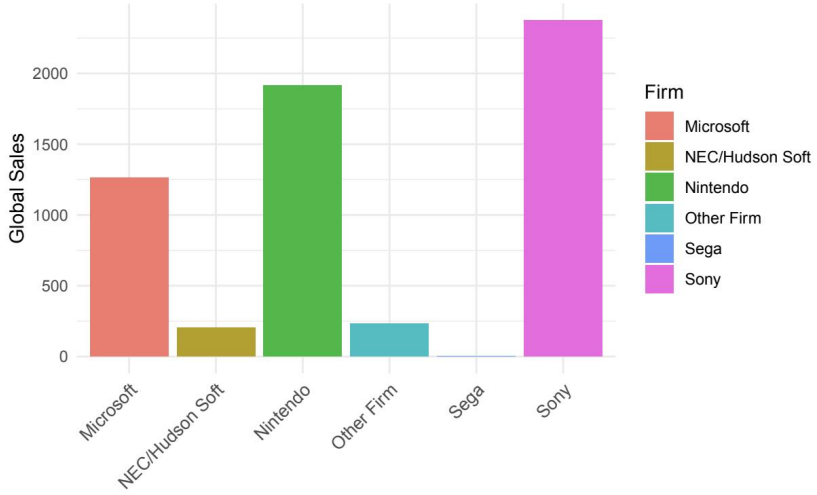


Figure 3: The global sales of video games by platform manufacturer

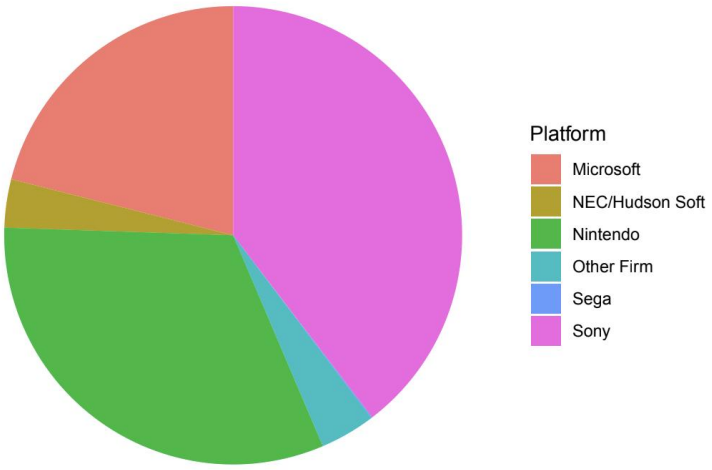


Figure 4: Proportion of global sales by platform manufacturer

By incorporating this additional data source and visualizing the data, the analysis can provide more context and understanding of the video game market, allowing for better decision-making and predictions.

By carefully examining our data, including handling missing values, performing feature transformation, and implementing feature engineering techniques such as log transformation and dummy coding, we tried to make our models make accurate predictions. Our analysis of the data allowed us to understand the key factors influencing video game sales and informed our selection of appropriate machine learning models to answer our research question. Through this process, we built a solid foundation for developing models that can provide valuable insights into the video game market, guiding the development of successful games in the future.

5. Analysis

5.1 Model Selection

The model we chose includes linear regression, LASSO (Least Absolute Shrinkage and Selection Operator) regression, and Gradient Boosting regression.

5.1.1 Linear and LASSO Regression

We initiated our analysis using linear regression as a foundation to obtain a fundamental grasp of the data and establish a benchmark model. Subsequently, we incorporated a Lasso regression model to refine the feature selection process, yielding a more streamlined and comprehensible model. Given that dummy variables within the dataset can contribute to an expanded set of predictors and a more intricate model, Lasso regression effectively filters out superfluous noise among the predictors and isolates the mnjhn pertinent features for model inclusion. This approach mitigates overfitting and bolsters the model's generalization capabilities.

5.1.2 Gradient Boosting Regression

Linear and LASSO regression are effective in scenarios with linear or near-linear predictor-target relationships but may falter when dealing with intricate, nonlinear relationships. Our exploratory analysis showed weak correlations between individual predictors and global sales (highest correlation coefficient = 0.31, **Figure 1**), suggesting the presence of more sophisticated relationships not captured by simple regression. Therefore, we chose Gradient Boosting Regression, an ensemble method that combines multiple weak learners to model complex correlations and nonlinear patterns, to provide a more comprehensive representation of the dataset's relationships.

Gradient Boosting regression is a tree-based approach, which works by building an ensemble of decision trees, where each tree is trained on the residuals of the previous tree. This allows the model to learn over time and improve the predictions based on the errors of the previous tree. It has the advantage of automatically learning and incorporating interactions between features. This also makes it a good approach for generating feature importance. By analyzing the ensemble of decision trees, it becomes possible to quantify the frequency and impact of each feature across all trees. This information facilitates the identification and ranking of predictors based on their relative significance within the model, offering valuable insights into the key drivers of the target variable.

5.1.3 SHAP

In addition to the three algorithms we implemented, we also tried to explain the output of our gradient boosting model using SHAP (SHapley Additive exPlanations). It can serve as a powerful tool to help us in the decisions making process of complex models(such as gradient boosting regression). It is based on the concept of Shapley values from cooperative game theory, which assigns a value to each feature in a prediction based on the contribution of that feature to the difference between the predicted output and the expected output. Using that method, we can have a better understanding of how much each feature contributed to the final prediction of global sales and determine whether the contribution is negative or positive. This can provide us with a better understanding of the model and allow us to analyze the features more thoroughly.

5.2 Evaluation Metrics Selection

We decided to use MSE(Mean Squared Error) and R^2(R-squared) as our decision metrics.

MSE quantifies the average squared difference between the actuarial and predicted values of the target variables. A lower MSE suggests a better-fitting model because the predictions will be closer to the true values. However, this metric can be sensitive to outliners because it penalizes large errors more severely than small ones.

R squared is the coefficient of determination; it measures the proportion of the total variance in the target variable that can be explained by the predictor variables in the model. Which ranges from 0 to 1, where larger values indicate that the model explains a larger fraction of the variation in the target variable and means the model fits the data better. However, it can increase with the addition of more predictor variables in the model, so it is important to consider other evaluation metrics, such as MSE alongside R squared to have a better understanding of the model performance.

5.3 Model Fitting and Fine Tunning

5.3.1 Model Fitting for the First Dataset

We fitted the Video Game Sales with Ratings dataset (downloaded from Kaggle) using linear regression, with log-transformed global sales as the target variable and critic score, critic count, user score, user count, age of the game, and dummy-coded variables for platform, genre, game rating, and publisher as predictors. We grouped publishers with less than 10 occurrences into a new category, Other Publishers, to reduce the number of features. We splited the dataset into training (80%) and testing (20%) before fitting the model. We plotted a scatter plot to visualize the performance and calculated the MSE and R-squared scores for test set.

The next thing we did was fit the same dataset with a LASSO regression model, plot the scatter plot, and calculate the MSE and R squared scores for further evaluation.

After that, we decided to fit the dataset with a Gradient Boosting regression model and print the MSE and R square score. We also tried to fine-tune the Gradient Boosting regression model using a 5-fold cross-validation method. This method allowed the model to validate the performance by assessing multiple validation sets and utilizing the entire training dataset for both training and validation, resulting in a more reliable model and reduce the risk of overfitting.

We tuned two key hyperparameters for our Gradient Boosting regression model: learning rate (0.01, 0.1, 0.2) and max_depth (4, 6, 8). The learning rate is the step size in gradient-based optimization, influencing how quickly the model adapts to the data. A lower learning rate makes the model more conservative, potentially getting trapped in local minima rather than finding the global minimum, and it can be computationally expensive. In contrast, a higher learning rate leads to faster convergence but may cause overshooting. The max_depth parameter determines the maximum depth allowed for each decision tree in the ensemble. A larger tree depth can capture more complex relationships in the data, but it may lead to overfitting.

We identified these two crucial hyperparameters and optimized the model using a grid search method, which evaluates all possible combinations of the parameters and selects the best combination (smallest MSE). Then, we applied the best model to the dataset and reported the MSE, R squared score, and a scatter plot.

Finally, we examined whether grouping the Publisher feature by creating a new category would influence the model performance. We fitted the original dataset without grouping the Publisher to the Gradient Boosting regression model and returned the MSE and R square score. We then compared these results with the MSE and R square of the Gradient Boosting regression applied to the dataset with feature grouping. This comparison allowed us to determine the impact of the feature grouping on the model's performance.

We also do a exploratory analysis and try to predict the critic score by fighting a linear regression model to the dataset, results are shown in the **Appendix 3.**

5.3.2 Model Fitting for the Combination of the First and Second Dataset

In Part 4, as previously mentioned, we aimed to enhance our model by incorporating additional features from an external dataset. We acquired the "List of best-selling game consoles" dataset from Wikipedia to include more information about consoles and platforms, enriching our original dataset. We then performed a left join on the datasets using the common "Platform" column.

We fitted a Gradient Boosting regression model to the updated dataset using a similar method as with the first dataset and obtained the MSE, R square, and scatter plots. By comparing these results with our previous findings, we can determine if incorporating the new dataset has improved our model's performance.
Lastly, we reported the feature importance results of our model using two methods. The first method utilizes the built-in function of the xgboost library to display a plot of ranked feature importance. To gain deeper insights into each feature's significance and the role they play in the model, we applied the SHAP algorithm to the Gradient Boosting model using the joined dataset and plotted the results. We implemented the SHAP method in Python for convenience.

6. Interpretation and Discussion of Findings and Models

6.1 Findings and Models

6.1.1 Linear Regression Model

By running the code, we first obtained the P values of each variable in the linear regression model. The size of the P value represents the probability of error in our data sampling and hypothesis. Generally, if the P value is less than 0.05 (or another predetermined significance level), it is considered significant which means that the correlation is not caused by chance. Among all variables, the following variables have stronger correlations compared to other variables: Critic_Score, Critic_count, User_score, User_count, game_ages, Publisher_groupNitendo.

Next, we calculated the Mean Squared Error (MSE) and R-squared value of the linear regression model for the test set and plotted the scatterplot. In R, the MSE represents the average squared error between the predicted and actual values. The standard deviation of the dataset is 0.3045, while the MSE is 0.11, lower than the dataset's standard deviation. This means that the model's prediction error is smaller than the dispersion of the dataset, which implies that the model has good predictive ability. However, the R-squared value of this model is only 0.4307019, meaning that the model can only explain 43% of the target variable's variance. This value indicates that the performance of this model needs further evaluation.

Figure 5: scatter plot of the predict values and actual values for the Linear regression model

6.1.2 LASSO Regression Models

Considering that the dummy variables in the dataset might increase the number of predictor variables, leading to a more complex model, we decided to use the LASSO regression model to screen features, select the most important ones, reduce the number of dummy variables, decrease the model's complexity, and improve the model's predictive accuracy. By using LASSO regression models, we obtained new MSE and R-squared values for the test set. In the new model, the MSE value remains at 0.11, with no significant improvement, and the R-squared value only shows a slight increase, reaching 0.4308226, which is almost the same as in the linear regression model. The scatter plot also confirms this. Therefore, we can only conclude that there is still room for improvement in the regression model.

Figure 6: scatter plot of the predict values and actual values for the LASSO regression model

6.1.3 Gradient Boosting Regression

We used a gradient boosting regression model to fit the dataset, printed the MSE and R-squared values for the test set, with MSE being 0.06780442 and R-squared being 0.6502253. Compared to linear regression and LASSO regression, these values have changed significantly, and the model's accuracy has improved markedly. The reason for this phenomenon is that gradient boosting regression has significant advantages in handling non-linear relationships compared to linear regression and LASSO regression, and it is more robust to outliers and noise. However, it also has some drawbacks, as it requires tuning multiple hyperparameters, such as the learning rate, tree depth, and the number of weak learners. Improper tuning may lead to overfitting risk.

To avoid overfitting, we tried to fine-tune the hyperparameters using cross-validation and kept the parameters with the lowest MSE for XGBoost. In the model, we adjusted the learning rate and maximum depth using cross-validation, and obtained the best learning rate: 0.2, best max tree depth: 8. The tunning is based on selecting the lowest MSE score. We selected the best XGBoost model, printed the MSE and R-squared values again, and the R-squared value changed to 0.6800238, while the MSE changed to 0.06202793. Compared to before the adjustment, the changes were more apparent, and the model's accuracy improved again. By plotting a scatter plot and comparing it to the scatter plots of the linear regression and Lasso regression models, the improvement in accuracy can be clearly seen, indicating that the gradient boosting regression model is more suitable for predicting global sales.

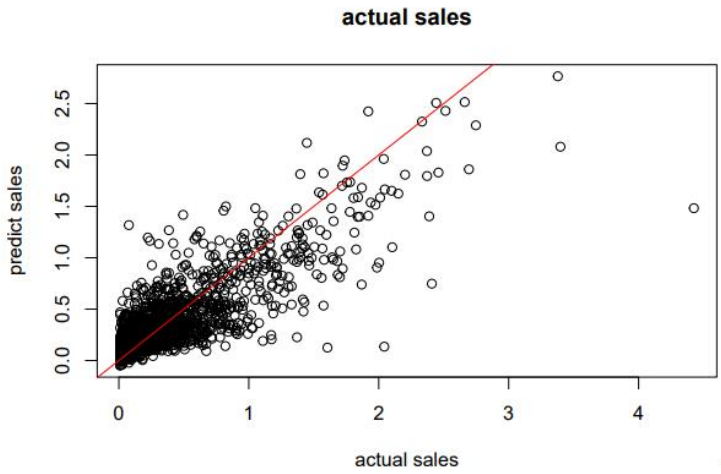


Figure 7: scatter plot of the predict values and actual values for Gradient Boosting Regression model (Group Publisher)

Before further discussion, we also considered whether the step of grouping publishers would affect the final result (during data processing, we merged publishers with a frequency of fewer than 10 appearances) and attempted to use the dataframe without the grouping of publishers to see if that changed the result. We printed the MSE and R-squared values, and the R-squared value was 0.679009, while the MSE value changed to 0.06222466. Compared to the values after grouping publishers, there was almost no change, meaning that the changes in the publisher group do not have a strong impact on the model's performance, the scatter plot also confirms this, so the grouping method works.

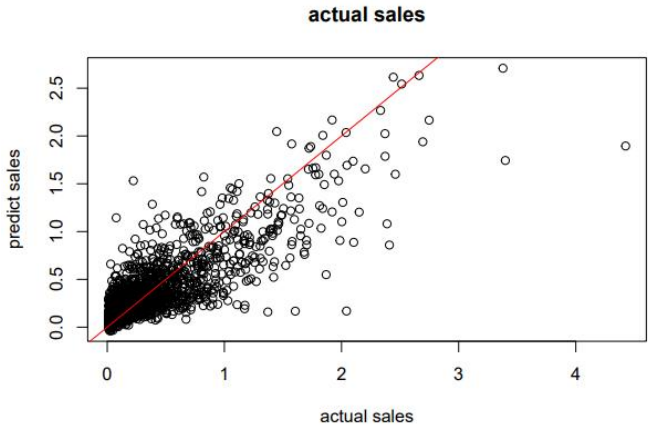


Figure 7: scatter plot of the predicted values and actual values for the Gradient Boosting Regression model (Not Group)

After determining that the gradient boosting regression model is the best model, in order to enhance the model and enrich our dataset, we obtained the "Best Game Console List" dataset from Wikipedia and added it to our original dataset.

We performed the same fitting process of Gradient Boosting regression and obtained new MSE and R-squared values, as well as plotted a new scatter plot. The new R-squared value for the test set is 0.6591916, and the new MSE value is 0.06606629. Compared to the previous values, there is no significant decrease, so we believe that the combined dataset is still valid, and due to the increased data volume, our model's performance has also improved.

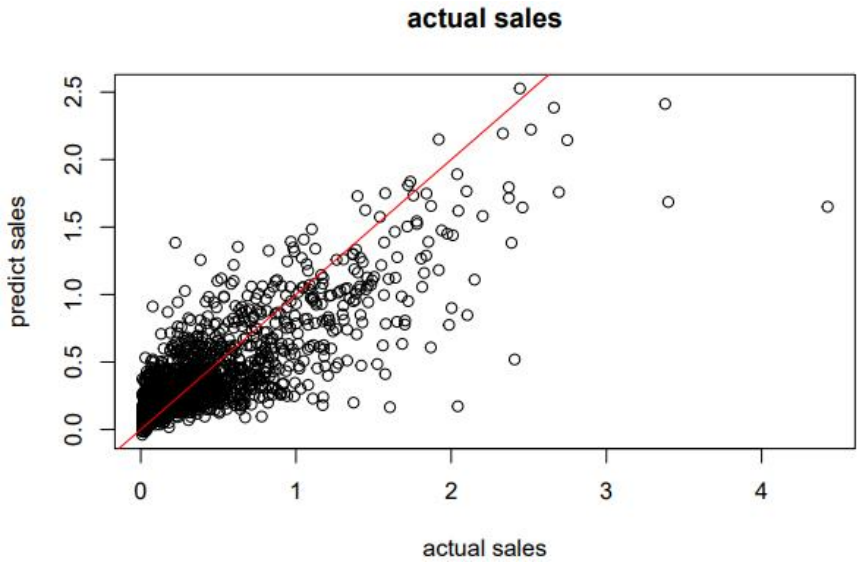


Figure 8: scatter plot of the predict values and actual values for Gradient Boosting Regression model (Add data set)

6.2 Feature Importance Analysis

Firstly, we used the built-in function in the xgboost library to display a chart ranking feature importance.

Generally speaking, if a model has many features and we want to find the most important ones, we can choose to filter them based on the cumulative percentage, which is usually 80% to 90%. Here, we select the top 5 features with a cumulative percentage of 80% as the most important features. They are, in descending order: Units_sold, game_ages, Publisher_groupedNintendo, GenreMisc, and RatingM.

It should be noted that feature importance doesn't necessarily indicate whether these features have a positive correlation with sales or not. They could have a positive correlation or a negative correlation, and we need to combine other research tools to make a judgment.

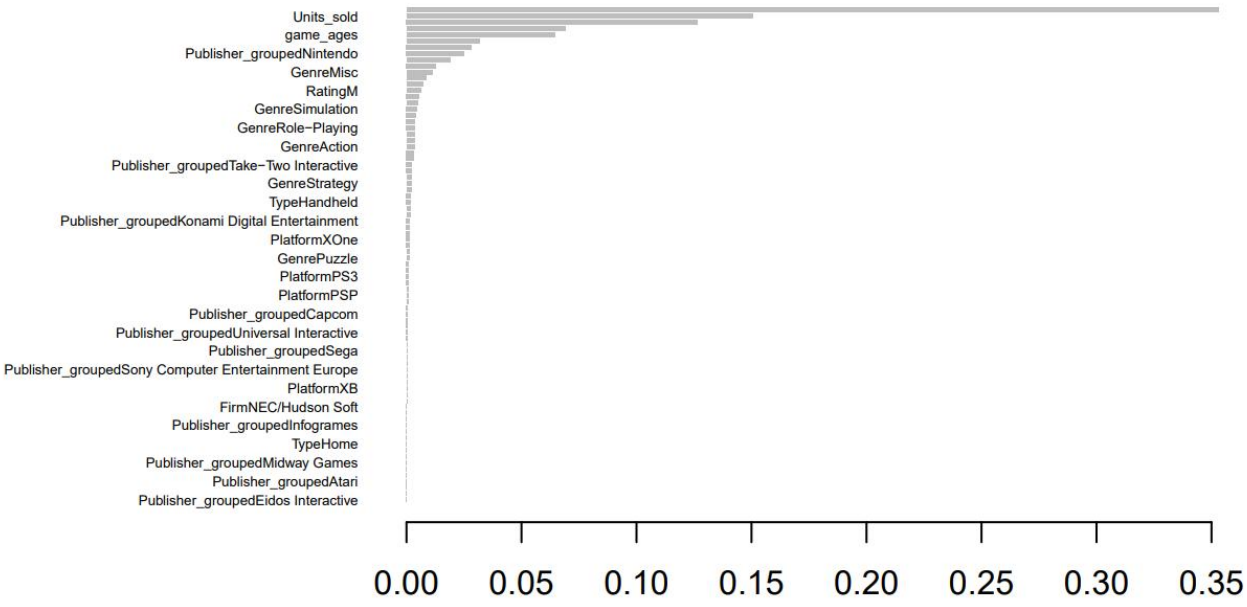


Figure 9: feature importance from the gradient boosting method

To further understand the relationship between these features and global sales, we used the SHAP algorithm, which can help us determine the positive or negative contributions of feature values in predictions. As shown in the figure below, from blue to red, the feature importance gradually increases, where a positive value represents a positive correlation between feature importance and global sales, and a negative value represents a negative correlation.

As shown in the figure below, from blue to red, the feature importance gradually increases, where a positive value represents a positive correlation between feature importance and global sales, and a negative value represents a negative correlation.

Combining the conclusions from **Figure 9**, we find that Units_sold, game_ages, Publisher_groupedNintendo, and GenreMisc have a positive correlation with global sales, while RatingM has an opposite relationship with global sales.

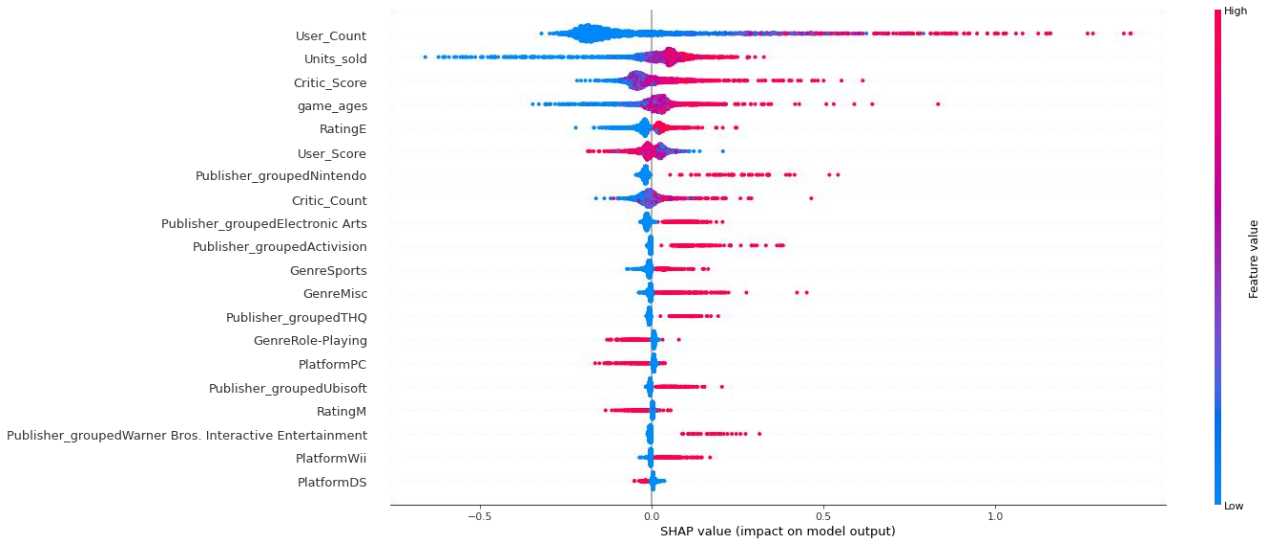


Figure 10: feature importance with SHAP model-> gradient boosting

6.3 Recommendations

Based on the aforementioned research, we can comprehensively study and discuss the factors that influence video game global sales and provide suggestions to game developers or publishers. Units_sold represents the sales of gaming consoles, which means that the choice of a game's launch platform needs to be carefully considered. Choosing a gaming console with the highest possible sales can boost game sales. Game_ages represents the sales duration of a game; the longer a game is on the market, the higher its sales will be. However, the total sales of most products tend to increase over time, so this feature does not have practical significance for discussion. Publisher_groupedNintendo represents one of the game publishers, Nintendo, which implies that games published by Nintendo often have higher global sales. Game developers can increase cooperation with Nintendo. GenreMisc represents the Misc game genre, which means that video games with multiple elements are more popular than those with single-themed content. Game developers should include more diverse gameplay when developing new games, as this can greatly help boost sales. RatingM represents the M-rated game category, which is only available to players aged 17 and older. This feature has a negative correlation with global sales, which means that game publishers need to actively lobby the Entertainment Software Rating Board (ESRB). Developers should reduce elements related to violence, blood, explicit content, and/or crude language in their games to avoid their games being classified as M-rated during game rating processes.

7. Limitations

While our study provides valuable insights into the factors that influence global video game sales, there are several limitations that should be acknowledged. First, the data used in the study has a limited time frame, which may not capture the full dynamics of the industry. As the video game industry evolves rapidly, incorporating more recent data would provide a better understanding of the current state of the market. Secondly, we believe that more features such as demographic features like the play's gender, age, and region, and qualitative factors such as graphics or storytelling can add to the dataset and improve the model performance. Possible solutions include test data processing and sentiment analysis to analyze user reviews and add that information to our predictive model.

In addition, the models we chose and the way we did feature construction can influence the overall performance. For instance, our Gradient Boosting model may require more hyperparameter tuning to avoid overfitting. And by incorporating one-hot encoding, we added a lot of features to the dataset, which will also cause overfitting and increase the model complexity. In sight of that, we can explore more models in the future and manually select the more relevant features from our feature importance analysis.

8. Conclusion

In conclusion, our study has explored the key predictive factors that influence global game sales by analyzing datasets and developing various models. Our findings suggest that units_sold, game_ages, Publisher_groupedNintendo, GenreMisc, and RatingM are the most important features that influence game sales. By employing the gradient-boosting regression model, we were able to explain approximately 68% of the target variable's variance, and achieve a MSE score of 0.066.

Based on these findings, we recommend that game developers and publishers consider launching their games on platforms with high console sales, cooperating with Nintendo, creating games with diverse gameplay elements, and avoiding M-rated content. By considering these factors, developers and publishers can potentially increase their game sales and achieve greater success in the competitive video game market.

Future research could expand the scope of the study by incorporating more recent data, considering qualitative factors, and exploring more sophisticated models to improve predictive accuracy. Additionally, researchers could investigate the impact of emerging technologies, such as virtual reality and cloud gaming, on the video game industry and its sales dynamics. In summary, while our study sheds light on some factors affecting global

video game sales, it is essential to approach these findings with caution due to the limitations mentioned above. Game developers and publishers should consider these insights as a starting point for understanding the market dynamics and further investigate specific aspects of their games and target markets to devise effective strategies for success.

Reference:

Kretschmer, T., Claussen, J., & Peukert, C. (2012). Piracy and box office movie revenues: Evidence from Megaupload. International Journal of Industrial Organization, 30(2), 188-199.

Teixeira, R., & Karahanna, E. (2018). Explaining the adoption of social network games: A multi-theoretical perspective. Journal of the Association for Information Systems, 19(9), 861-891.

Nieborg, D. B. (2015). Prolonging the magic: The political economy of the 7th generation console game. Media, Culture & Society, 37(7), 966-987.

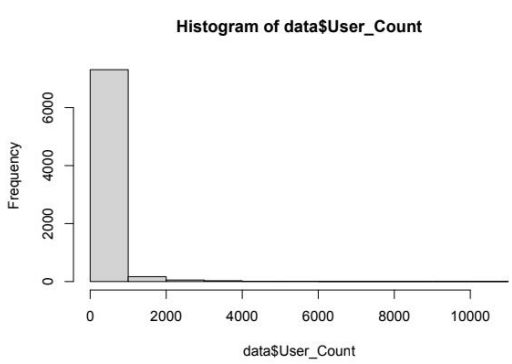
Mäntymäki, M., & Salo, J. (2013). Purchasing behavior in social virtual worlds: An examination of Habbo Hotel. International Journal of Information Management, 33(2), 282-290.

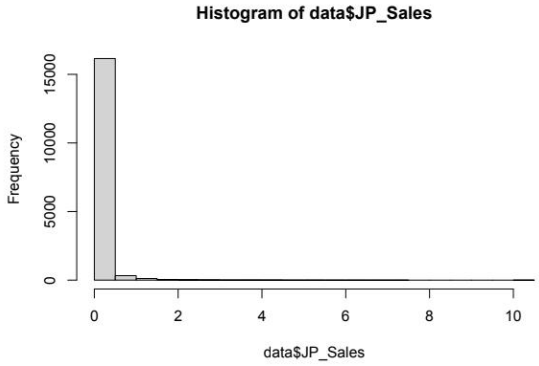
Whitson, J. R. (2013). The cultural economy of Indie game development. Media, Culture & Society, 35(3), 333-345.

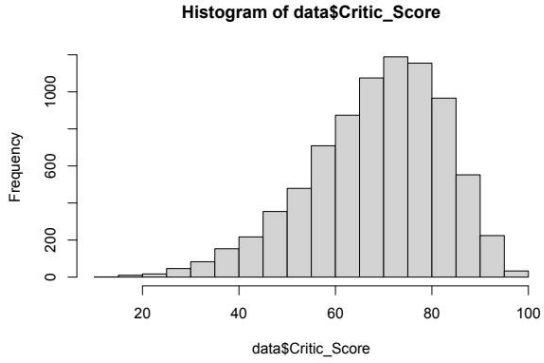
Wertz, J. (2015). How the relationship between video game sales and user ratings has evolved over time. Journal of Information Systems Applied Research, 8(3), 4-15.

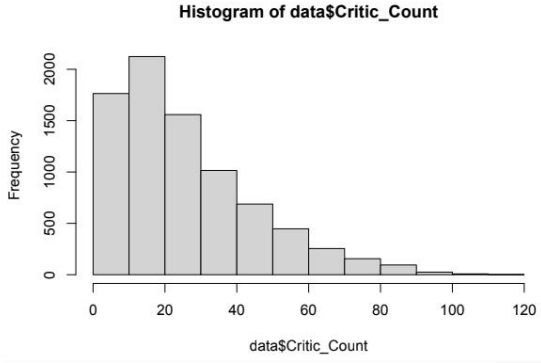
Appendix:

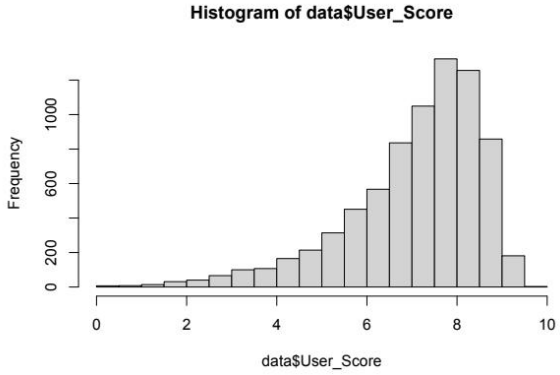
1. Distribution of numerical variables that have missing values in dataset

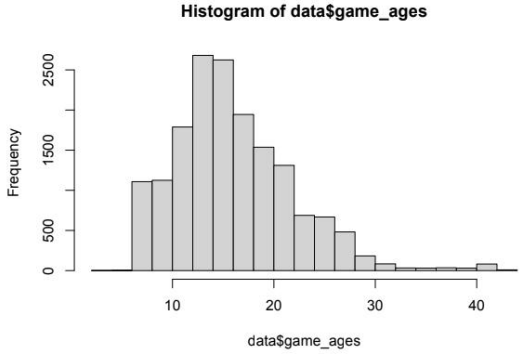




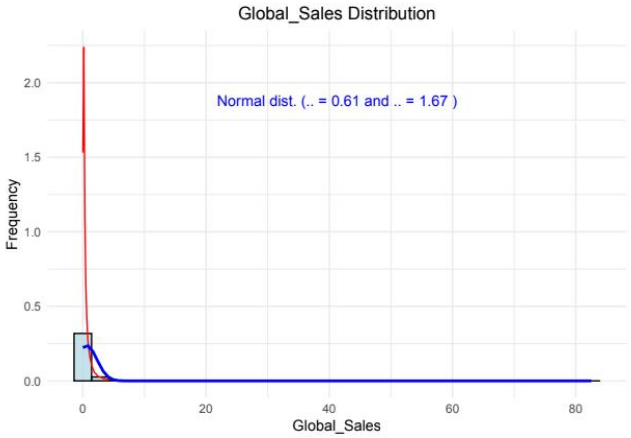


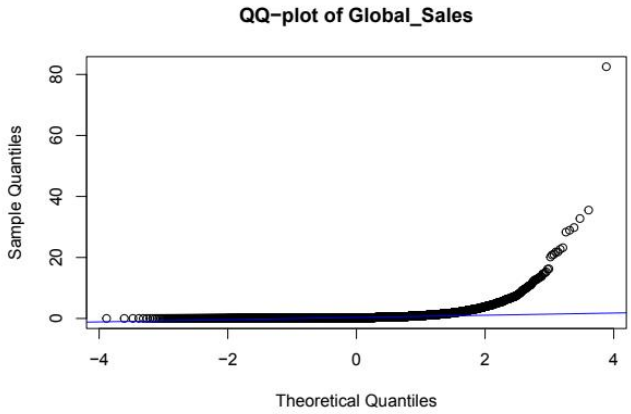




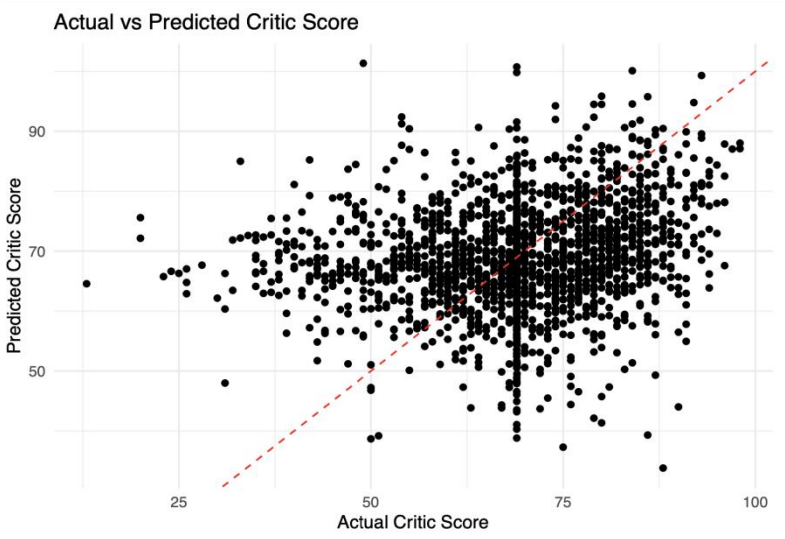


2. Distribution of Global Sales





3. Scatter plot to see the relationship between actual and predicted critic scores



---

4. Model Comparison Table

Matrics	Linear regression	LASSO regression	Gradient Boosting regression (one dataset)	Gradient Boosting regression (joined dataset)
R square	0.43	0.43	0.68	0.66
MSE	0.11	0.11	0.062	0.066