Final Paper

STOR 520 Yifei Zhang
November 18, 2023

INTRODUCTION

Contemporary gaming has become a powerful industry, often referred to as the modern "ninth art," and creates enormous economic value. Game sales are a direct indicator of a game's quality, popularity, and economic success. So, we selected a dataset of game sales for our analysis, and through the depth of the analysis, we raised two compelling questions.

The first question is whether it is possible to predict the global sales of this game using a range of information such as whether the game is a sequel, the genre of the game, the publisher's game platform, and so on. Furthermore, if sales in a particular region were to be used as a reference point, which region would be the most appropriate benchmark? This is a major concern for most game studios. When starting to develop a game, the game studio must consider the projected total sales to determine if the game will be profitable enough to justify the development costs. If it is possible to make an accurate prediction of global game sales without relying on regional sales data, that would be very beneficial. However, it might not easy to predict the Global Sales without the Region Sales. We know that advertising and promotional expenses play a significant role in the launch of a game. Conducting large-scale marketing campaigns simultaneously across all global regions is often impractical. Therefore, game companies first focus on key promotional activities in one region to observe the effects before further expanding to other areas Given that some games will be released in first or demo versions, it is important to consider which territories have the most strategic advantage in influencing global sales when planning the initial release of these games.

The second question concerns the categorization of game ratings: can we predict whether a game will be highly rated? Sales and ratings fluctuate over time, and high ratings can attract new players and thus potentially increase sales further. However, this also raises the question: does a high rating always mean a good game, and does a good game always receive a high rating? We cite the examples of Cyberpunk 2077 and PUBG and Call of Duty: 20, where Cyberpunk 2077 has high sales and low ratings, while PUBG and Call of Duty:20 have high sales and low ratings, and we attempt to decipher the relationship between ratings and sales, including the relationship between high ratings and low ratings. relationship between ratings and sales, including the common phenomenon of high ratings and low sales and vice versa. This insight is critical because it allows players to make informed decisions about buying games based on the interplay between their sales numbers and ratings.

Data

We have obtained two datasets from Kaggle:
Dataset1: https://www.kaggle.com/code/coffeepot/videogame-sales/input

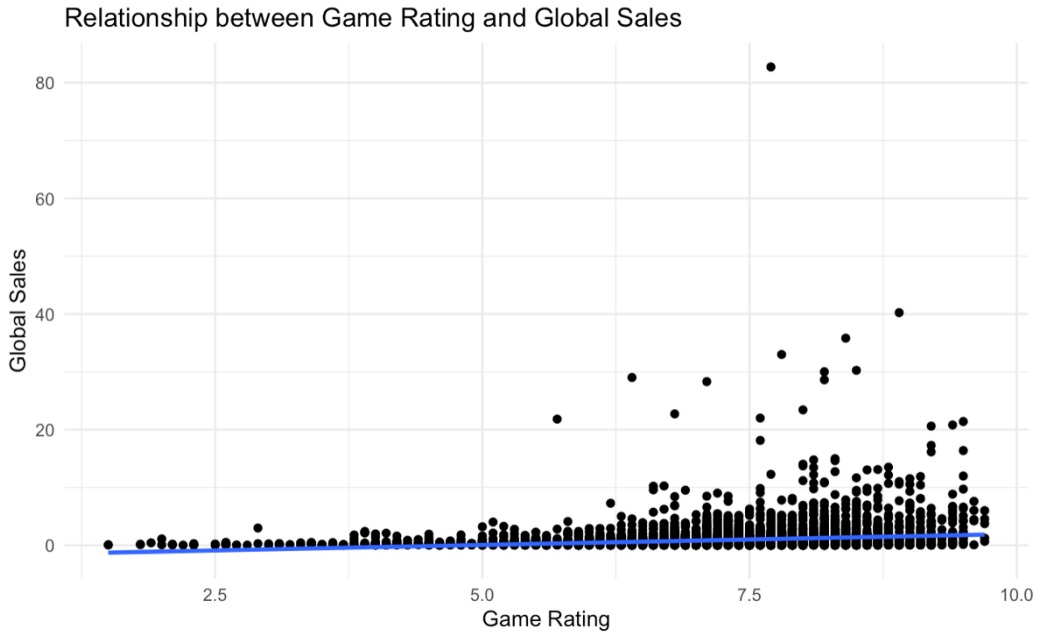
which was generated by a scrape of vgchartz.com. VGChartz is a video game sales tracking website that was launched in June 2005. It provides weekly sales figures of console software and hardware by region, as well as tools for data analysis and reviews of the data it provides.
Dataset2: https://www.kaggle.com/datasets/nyagami/video-game-ratings-from-imdb
IMDb, being a comprehensive and widely recognized database for movie and television content, is a valid source for academic research but should be used cautiously due to its user-generated and non-disclosed weighting methods for ratings. Dataset 1 contains 16,598 entries with variables 1-10 primarily including sales data, while Dataset 2 has 12,635 entries with variables 1, 2, 3, 11, and 12 coming from it. We merged the two datasets using the 'Name' variable and, after data cleaning and removing NA values, were left with 6,420 entries.

1.Name: The name of the game, a categorical variable.
2.Year: The year the game was released, a numerical variable.
3.Genre: The genre of the game, a categorical variable.
4.Publisher: The publisher of the game, a categorical variable.
5.Platform: The platform on which the game was released, a categorical variable.
6.NA_Sales: Sales in North America, in millions of units, a numerical variable.
7.EU_Sales: Sales in Europe, in millions of units, a numerical variable.
8.JP_Sales: Sales in Japan, in millions of units, a numerical variable.
9.Other_Sales: Sales in other regions, in millions of units, a numerical variable.
10.Global_Sales: Total worldwide sales, in millions of units, a numerical variable.
11.Rating: The rating of the game, a numerical variable.
12.NoR: Number of ratings, a numerical variable.
13.Platform_Category: The category of the platform based on the company, a categorical variable.
14.Sequel: Indicates whether the game is a sequel, a categorical variable.
15.High rating: Indicates whether the game has a rating above 8.0, a categorical variable.

In our academic research, we have innovatively created three additional variables derived from the existing dataset to facilitate regression analysis: 13. Platform_Category: This variable classifies gaming platforms into four categories based on the information gathered from online research about the companies that own these platforms: 1) Microsoft, 2) Sony, 3) Nintendo, and 4) Others. This categorization aims to better investigate the relationship between game sales and the platforms they are released on.14. Sequel: We determine whether a game is a sequel by examining the 'Name' variable for the presence of numbers, Roman numerals, or titles from classic gaming IPs such as Mario, Grand Theft Auto, Pokemon, etc. This allows us to identify continuations within game franchises. 15.High rating: We evaluate the game ratings, coding them as '1' (high rating) if the rating is above 4.0, and '0' (low rating) if the rating is below 4.0. This binary variable is used to analyze the impact of higher game ratings on sales and other dependent variables in our study. These tailored variables are expected to provide deeper insights and contribute significantly to the robustness of our regression models.

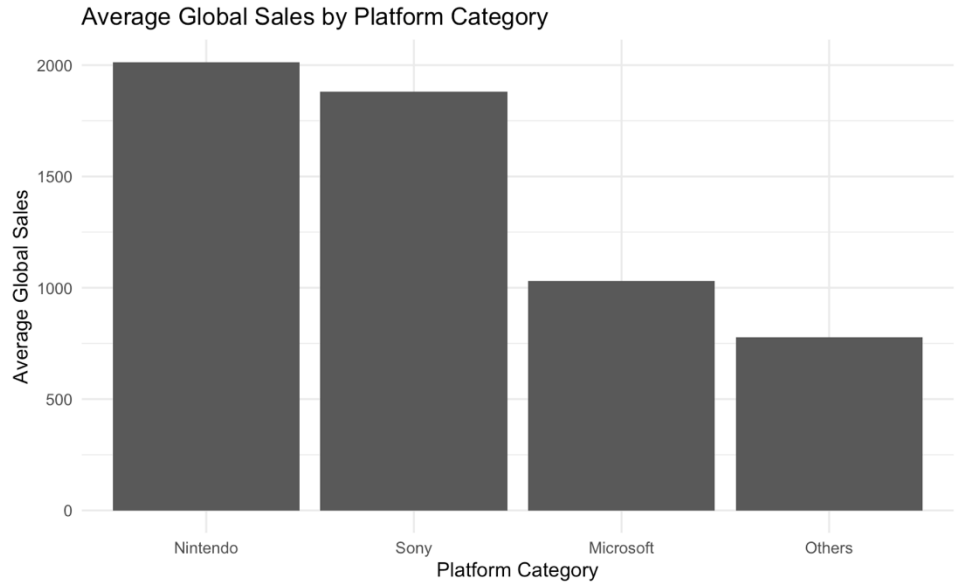
	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Rating	NoR	Platform_Category	Sequel
	<chr>	<chr>	<dbl>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>
1	Wii Sports	Wii	2006	Sports	Nintendo	41.5	29.0	3.77	8.46	82.7	7.7	3905	Nintendo	0
2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.1	3.58	6.81	0.77	40.2	8.9	6231	Nintendo	1
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.8	12.9	3.79	3.31	35.8	8.4	3944	Nintendo	1
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.8	11.0	3.28	2.96	33	7.8	1444	Nintendo	0
5	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.3	8.5	150	Nintendo	0
6	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.4	9.23	6.5	2.9	30.0	8.2	2634	Nintendo	1
7	Wii Play	Wii	2006	Misc	Nintendo	14.0	9.2	2.93	2.85	29.0	6.4	669	Nintendo	0
8	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.6	7.06	4.7	2.26	28.6	8.2	3053	Nintendo	1
9	Duck Hunt	NES	1984	Shooter	Nintendo	26.9	0.63	0.28	0.47	28.3	7.1	1710	Nintendo	0
10	Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.4	8	2045	Nintendo	1

Tabel1: The first ten variables of the organized dataset



Graph 1: Relationship between Game Rating and Global Sales

This scatterplot shows the relationship between game ratings (Rating) and Global Sales. As We can observe from the graph, most of the data points are concentrated in the lower Global Sales range, with a wide distribution of ratings. There are a few games that have achieved significant sales in the higher rating ranges, but these are not very common. In addition, the highest rated games did not always correspond to the highest sales. This suggests that there may not be a strong direct correlation between ratings and sales.



Graph 2: Average Global Sales by Platform Category

This bar chart shows the average global sales for the different gaming platform categories (Nintendo, Sony, Microsoft, Others). As you can see from the graph, Nintendo and Sony have higher average global sales than Microsoft and Other categories of platforms.

Platform_Category <chr>	Mean_Rating <dbl>	Median_Rating <dbl>
Microsoft	7.136032	7.3
Nintendo	7.014033	7.2
Others	7.381229	7.5
Sony	7.231977	7.4

Table2: Mean and Median Rating of the 4 Platform_Category

This table presents the average and median ratings of games in the four main gaming platform categories. The "Others" category has the highest average ratings, indicating that this category is likely to have higher quality games, but it might be caused by Others has the least number of Games on this Platform_Category. In contrast, games on Nintendo have the lowest ratings, but the difference is not significant. Overall, game ratings for all platforms are clustered around 7, indicating that user satisfaction with games on these platforms is relatively consistent.

Result

Firstly, we commence with the inquiry of predicting a video game's global sales based on prior variables. Further, we seek to identify which regional sales metric would best supplement the prediction of global sales. To begin, we will apply a Multiple Linear Regression (MLR) approach for forecasting global sales, initially removing the sales data from the four regions (NA, EU, JP, Other) from the variable set under consideration, which has the formula.

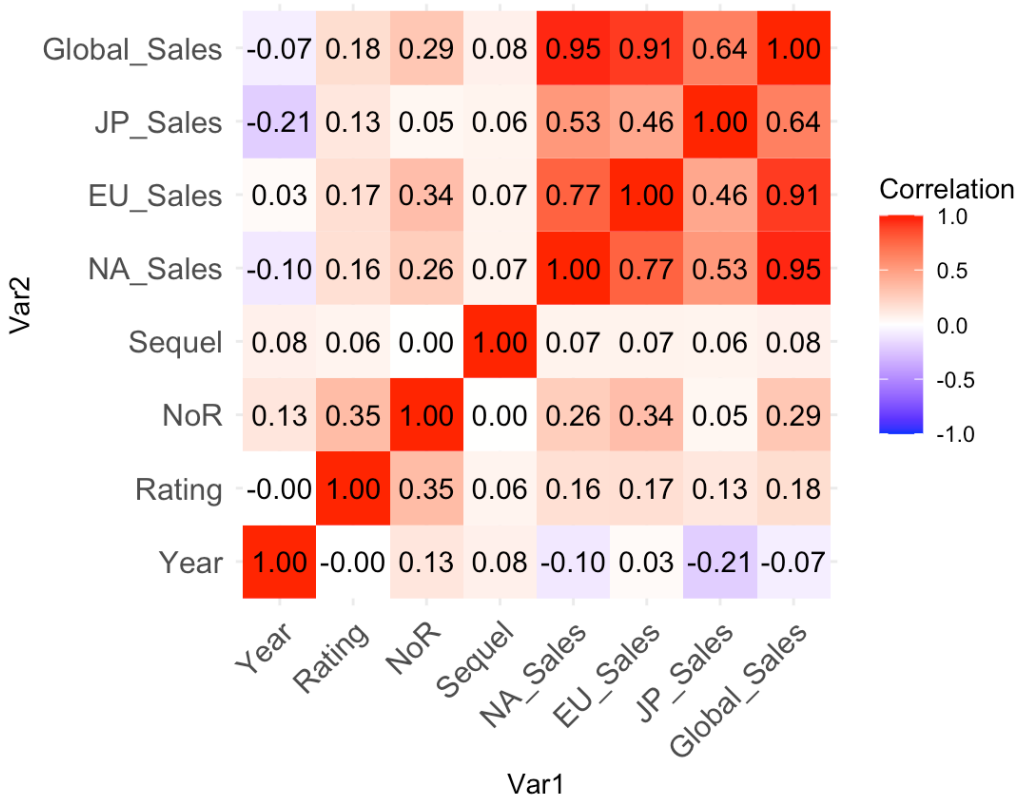
$$Global\_Sales \sim Year + Genre + Rating + NoR + Platfom\_Category + Sequel$$

For the dataset, we partitioned it into an 80% training set and a 20% test set for model training. However, post-training, we discerned that relying solely on the above data does not adequately predict global sales.

	intercept <lgl>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	RsquaredSD <dbl>	MAESD <dbl>
1	TRUE	2.016911	0.1772641	0.8409125	0.8788657	0.06428236	0.09480815

Table 3: Result of the Prediction without Region Sales

An R-squared value of only 0.1772641 implies that the model does not have a strong predictive capability. It suggests that only about 17.72% of the variability in the global sales can be explained by the model's inputs. This indicates that there are other factors not included in the model that significantly influence global sales outcomes. Thus, we plotted a correlation heatmap for all numerical variables to comprehend the relationships between them.



Graph 4: Correlation Matrix Heatmap

Based on this heatmap, we found that Global Sales is highly correlated with the 4 region sales numbers, which is obviously. Except these four variables, Global Sales is highly related to Rating and number of sales, and also the above-mentioned variable platform.

The results were unexpected as we had assumed that being a sequel might significantly impact a game's sales. However, upon examining two iconic series, Call of Duty and Mario, we discovered insights. The Call of Duty series maintained a consistently good sales volume, with an average of 13 million units sold per title, suggesting that sequels can assure a stable sales volume without necessarily standing out. Conversely, the Mario series displayed a different trend; from the highest selling 'Super Mario Bros.' at 40.24 million units to 'Super Mario Bros. 3' at 5.2 million units, indicating that sequels might not always live up to the impact of their predecessors, potentially leading to a decline in sales. This led to the conclusion that a sequel might not have a significant impact on global sales predictions.
Therefore, we shifted focus to another aspect of forecasting global sales: deciding which regional sales data to collect. Disregarding the 'Other' region for practical reasons, we considered the three major markets: NA, EU, and JP. Utilizing the MLR approach for model fitting, we observed a marked improvement in results. The best fit was for NA with an R-squared of 0.86, followed by EU with an R-squared of 0.80, and finally JP with an R-squared of only 0.43. These findings align with our earlier correlation heatmap. However, we are still not completely satisfied with this fitting process.

To enhance the fitting capability of the Multiple Linear Regression (MLR) model, we employed k-fold cross-validation, specifically using 10-fold cross-validation, to ensure that the model

generalizes effectively to unseen data. This method optimizes the MLR model by providing a robust estimate of its predictive performance. Listed from top to bottom are MLR model including NA Sales, JP Sales, EU Sales.

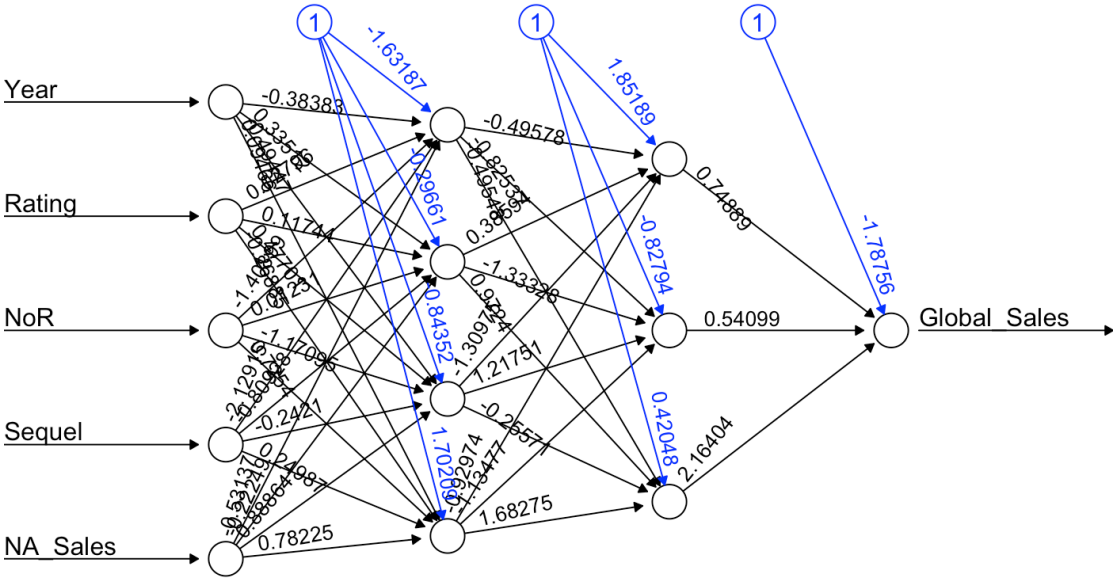
```
     RMSE  Rsquared       MAE
0.5393922 0.9068438 0.2806822
     RMSE  Rsquared       MAE
1.3104842 0.4543244 0.6806196
     RMSE  Rsquared       MAE
0.8319634 0.7835699 0.3898942
```

Table 4: Results for MLR model including Region Sales (NA, JP, EU)

We noted that only the third model, which includes the European sales (EU), exhibited a slight decline in fit, while the models including North American (NA) and Japanese (JP) sales data displayed varying degrees of improvement in fit.

In addition to Multiple Linear Regression (MLR), we also attempted to address the problem using neural network approaches. Initially, we constructed a neural network incorporating Na Sales. We use the neural net function to predict the Global Sales.

$$\text{Global\_Sales} \sim Rating + Year + NoR + Sequel + NA\_Sales$$



Graph 5: Neural Network Connection Diagram

However, after fitting with neural network models, we evaluated the outcomes using metrics such as RMSE, MAE, and MAE, RMSE=1.76 which showed significant higher compared to our previous MLR models RMSE=0.54. This indicates that neural networks may not be the optimal predictive model for this dataset. The reasons for the poor fit might be: 1. Our dataset is not sufficiently large for the neural network model, which requires a larger dataset for

implementation. 2. There is an inherent correlation between Global Sales and NA_Sales, which results in better performance for MLR.

To consider the second question of predicting whether a game will achieve a high rating, we first define a 'high rating' as a score above 9. We then recreate a column in the dataset named 'High_Rating', where ratings greater than 9 are marked as 1, and ratings less than or equal to 9 are marked as 0. For this classification task, we split the dataset into 70% for training and 30% for testing. We will use three methods to address this classification problem: logistic regression, decision trees, and Support Vector Machines (SVM). We begin with a brief description of the three models.

Logistic Regression: Logistic regression is a statistical method for binary classification problems that centers on the use of a logistic function (usually a sigmoid function) to estimate the probability of an event occurring. The model predicts the probability corresponding to the dependent variable (usually 0 and 1), and this probability can be converted into a binary output for classification purposes. Logistic regression estimates the model parameters by maximizing the likelihood function of the observations to find the value of the best-fitting coefficient.

Decision Tree: At each node, the tree selects an optimal feature to split the dataset into subsets, aiming to increase homogeneity within each subset. This recursive partitioning continues until certain criteria are met, such as a minimum number of observations in a node, a maximum tree depth, or achieving sufficiently pure nodes. The resulting tree structure visually represents the decision-making process, with each leaf node indicating a final decision or classification.Super Vector Machines (SVM): SVM are a set of supervised learning methods used for classification, regression, and outliers detection. The principle of SVM is to find a hyperplane in an N-dimensional space (N the number of features) that distinctly classifies the data points. To separate two classes of data points, SVM finds the hyperplane that maximizes the margin between the classes. The data points that are closest to the hyperplane and which influence its position and orientation are known as support vectors.
In Logistic Regression, we use glm() (Generalized Linear Models) function. The model predicts the binary outcome “High_Rating” in such a formula.

$$\text{High\_Rating} \sim Year + Genre + Rating + NoR + Platfom\_Category + Sequel$$

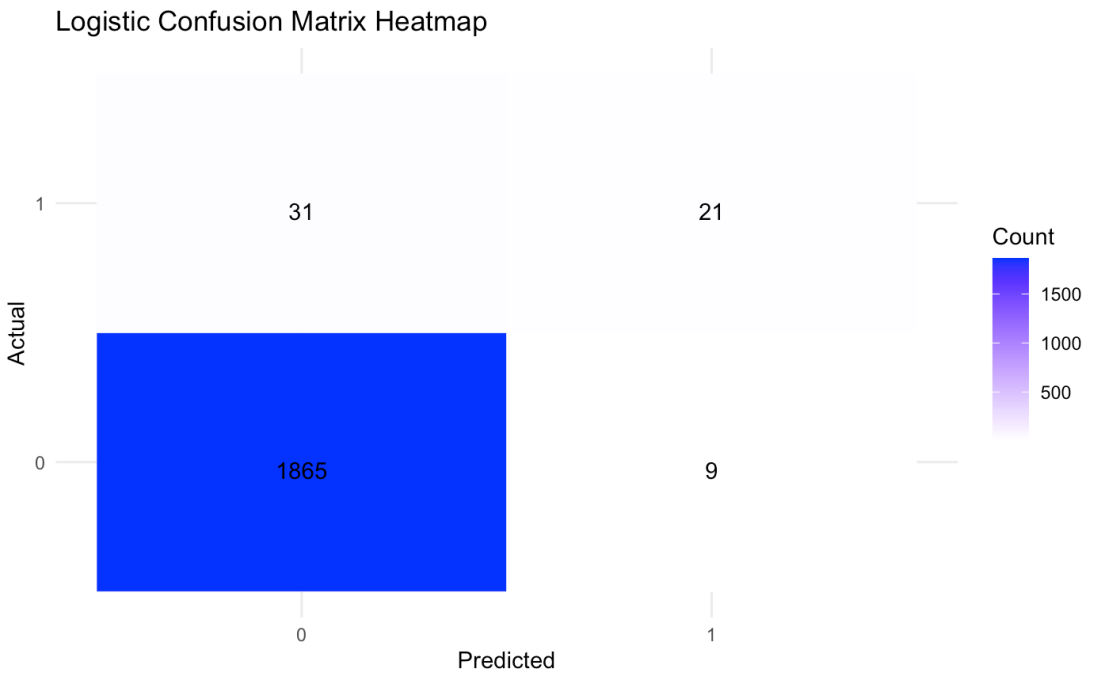
We set family = binomial(link='logit') which means the dependent variable is binary and the logit function to be used as the link function.

$$logit\big(P(Y=1)\big) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$$
$$= a_o + a_1 * Year + a_2 * NOR + \cdots + a_k * Global\_Sales$$

After that, the logistic function is applied to convert log odds to probabilities:

$$P(Y=1) = \frac{1}{(1+e^{-logti(P(Y=1))})}$$

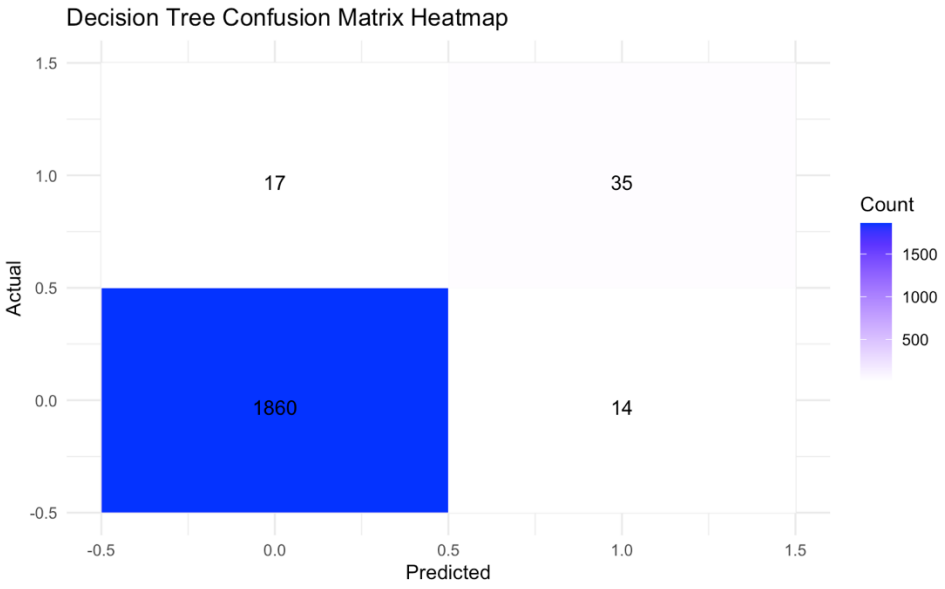
The result is below.



Graph 6: Logistic Confusion Matrix Heatmap

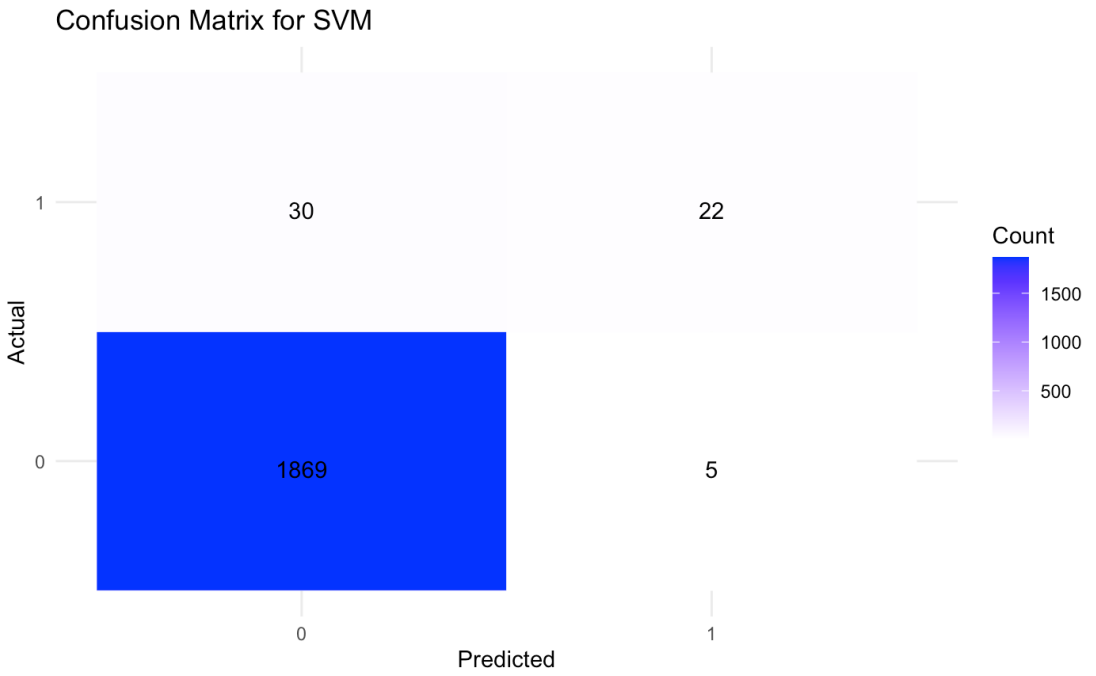
The heatmap shows the relationship between the actual values (Actual) and the model's predicted values (Predicted). The top left and bottom right corners of the matrix show the number of correctly classified by the model, corresponding to True Positives and True Negatives, respectively. The top right and bottom left corners show the number of model misclassifications, corresponding to False Positives and False Negatives, respectively. Darker colors represent higher numbers and lighter colors represent lower numbers. In this model, the model is very accurate in predicting the negative categories, but less accurate in predicting the positive categories. We will discuss this problem after comparing the 3 models. The accuracy for the Logistic Model is 0.9792.

For the second method Decision Tree, we use rpart() (Recursive Partitioning and Regression Trees)function which is designed for the construct the decision tree model. The result is below. The accuracy for the Decision Tree is 0.9839.



Graph 7: Decision Tree Confusion Matrix Heatmap

For the third method Super Vector Machines, svm() function has been used to predict, The result is below. The accuracy for the SVM is 0.9818.



Graph 8: SVM Confusion Matrix Heatmap

Through the above three models, we discovered that the Decision Tree had the highest accuracy, reaching 0.9893. In the prediction process, all three models demonstrated good predictive ability for the Not High Rating class, with the difference in accuracy primarily manifested in the prediction of the High Rating class. This is mainly because, within our data, a significant number of high-rated games had unsatisfactory sales, making it difficult for the models to distinguish between these cases, leading to high-rating games being mis predicted as low-rating.

Conclusion

In this article, we focused on two questions: the feasibility of predicting global sales based on non-regional sales, and if one region were to be included, which would be the most appropriate. We found that using variables such as Rating and NoR (Number of Rating) was insufficient for predicting global sales, with less than ideal results. Utilizing k-cross validation, MLR, and neural networks with sales data from the three major regions, the k-cross validated MLR model yielded the highest fit when including NA Sales. This predictive model is significant for game companies, especially smaller ones, as we know that advertising constitutes a substantial part of game expenses and promotion is a continuous process. For instance, Valorant became popular in the Americas before gradually expanding globally. Therefore, this model can help game companies decide whether to expand internationally and estimate global sales to balance the costs of further promotion. Future research could classify games by genre or publisher—for example, testing Action games in North America, Role-Playing games in Japan, and football games in Europe—and provide theoretical justifications for these

strategies. Incorporating the game's release price could refine the model further. Or, focusing on a specific series like Activision's Call of Duty could yield higher prediction accuracy.

Our second inquiry delved into how to ascertain whether a game would achieve a very high rating. We discovered that using decision trees could achieve the highest degree of fit with an accuracy of 0.9839. This is significant for game companies, as player reviews greatly influence purchase decisions in today's connected world. Predicting a game's overall rating based on sales could inform improvements to content to enhance ratings or preserve elements that players favor for future titles. A case in point is Battlefield 2042, which initially had strong sales but poor ratings, deterring some players from purchasing. However, subsequent updates by EA improved the ratings and sales. Future research could approach ratings not just as a classification issue but aim to predict precise scores, and analyze ratings by platforms such as Xbox, PlayStation, and PC, as the same game can perform differently across platforms, aiding studios in pinpointing areas for improvement.

When predicting game outcomes, pricing inevitably plays a crucial role. If we can gather pricing data for each game across different regions, it will allow for a more nuanced analysis of the market and enable tailored strategies for each one. Additionally, with the Asia-Pacific market's robust purchasing power today, future data collection focused on this region could aid in better data analytics. For future predictive algorithms, we could explore using Polynomial Linear Regression to forecast sales, adjusting the weight of specific parameters. Algorithms like K-means and KNN could be applied to classification problems, and we could consider dividing ratings into more than two categories for prediction purposes.