ST4248
STATISTICAL LEARNING II

Video Games Sales Prediction using Machine Learning Techniques

With the gaming industry experiencing significant growth in recent years, it has become a multi-billion dollar industry with a global audience. With global revenue reaching $180 billion in 2020, identifying the characteristics that make a game successful and the factors that influence its sales has become increasingly important for developers, publishers and investors. With this knowledge, developers will know how to make games that are better received by the target audience, and stakeholders will be better informed about which games to invest in. In this report, we will explore the relationship between game characteristics and sales, whether ratings by critics or users are good predictors of sales, and the impact of publishers on sales, among other questions. To achieve this, we will use machine learning models such as XGBoost and KMeans clustering, as well as linear regression to analyze the dataset and provide insights into the gaming industry.

Prepared by:
A0216305R

I.Data Collection and Preprocessing

We begin our analysis with a dataset obtained from Kaggle from user gregorysmith. BeautifulSoup, a library in Python was used to gather the data from a web scrape of vgchartz.com. The dataset includes 16,598 records of video games released between 1980 and 2020, with sales data broken down by region (Europe, Japan, North America, Other) as well as Global sales in millions of copies sold(Fig,A). However, this dataset lacked information on critics reviews and user ratings, which are important factors in determining a game's success. Hence, I merged the data with another dataset by user lombardoparedes containing games with metacritic scores since 2000. I grouped the data by the game title and the platform the game was on to ensure consistency in merging. The resulting merged dataset has 14 variables and 16,598 entries. Next, we preprocess the data. Since many games in the original dataset did not have review scores, the resulting dataset contained many missing values. After removing these values, the dataset had 5514 entries. To ensure that the critic scores were on the same scale as the user scores, I divided the metascore by 10. As the rank variable is dependent on sales and may cause multicollinearity issues, I removed the rank and name columns from the model training process. Categorical features such as year, genre and platform were converted to factors, and sales data across regions were normalized to improve model performance. Lastly, the dataset was split into training and test sets to avoid overfitting and to estimate model performance.

II. Exploratory Data Analysis

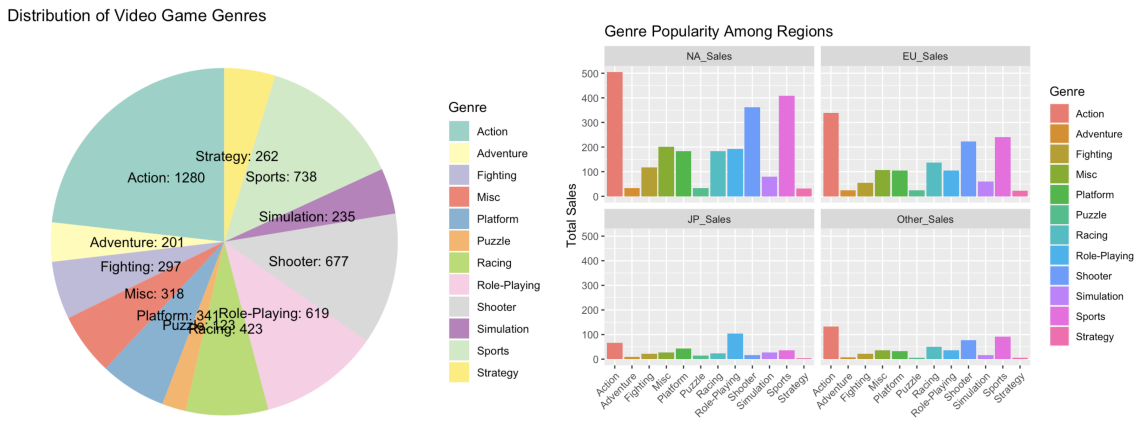


Fig.1 EDA By Genre

Firstly, I conducted EDA to find out the breakdown of genres of games produced and which genre has the highest sales in each region. The results above (Fig. 1) show that Action Games seem to be the most common

type of game produced. However, from the barplot, it can be seen that different regions consumers have a preference for different types of games. For example, while action games are popular in North America, Role-Playing Games are most popular in Japan. We will explore these further in statistical analysis.

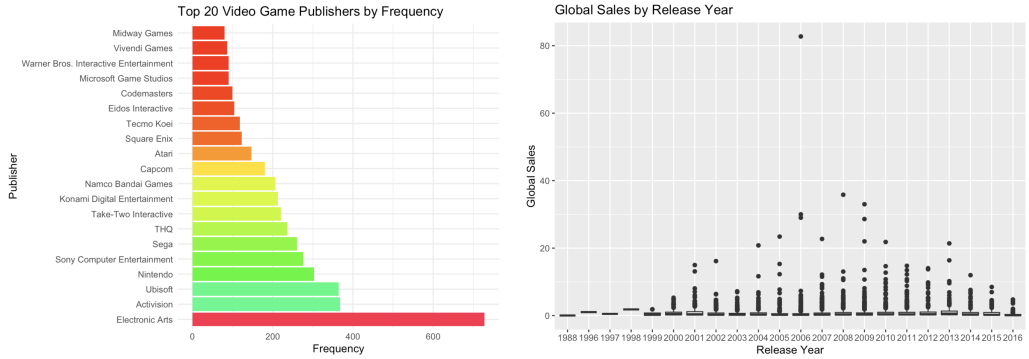


Fig.2 EDA for top 20 Video Game Publishers and Global Sales by Release Year

To find out if a relationship existed between the release year and sales, I created a boxplot using year as a categorical variable and plotted it against global sales of each that year. The data shows not much deviation in sales by year, with some years having outliers where games sold exceptionally well(2008). The top 20 publishers were also visualized to find out who are the most popular publishers(Fig.2).

Lastly, I computed the correlation matrix and plot to find out if any collinearity existed among features.

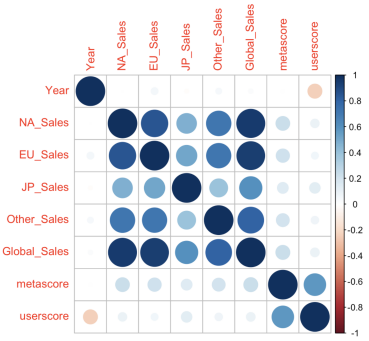


Fig.3 Correlation Plot of features

The plot above shows a strong positive correlation between many of the sales data. This makes sense for global sales as it is dependent on sales in other regions. With the insights gathered from the correlation plot, I will split the dataset further into each region and train the model with each regions' sales as the response variable.

III.Statistical Analysis

Game characteristics that affect Sales

To investigate the relationship between the features and the sales of games, we will determine the most important features via feature selection. Firstly, I have trained a Multiple Linear Regression model to serve as a baseline. I will then use XGBoost, an ensemble learning model that uses multiple decision tree models to make predictions and combines their results to achieve higher accuracy. It will help determine feature importance by calculating how much each feature contributes to the overall prediction accuracy of the model. This information can be used to identify the most important features that affect video game sales.

The XGBoost model was trained separately 5 times, each time with a different region(EU, JP, NA, Other, Global) in order to find out if a different region would result in a different feature importance score. My main hypothesis was that consumers of video games from different regions would value different features of games, one example being that Japanese consumers from the JP market would prefer games made by Japanese publishers. Next, we will evaluate the model.

To assess model accuracy, we will use the Test Root Mean Squared Error (RMSE) and R-squared value, which explains the variance explained by the model. The multiple linear regression model resulted in a test RMSE of 0.002891, which could be attributed to overfitting and the use of collinear sales data during training. We will examine the XGBoost results, where an example output has been included in the appendix(Fig.b)

Region	Test RMSE	Top features in order
Global	1.122	metascore,userscore,Year,PublisherNintendo,Platformwii,GenreShooter,GenreRacing,PublisherTake-TwoInteractive,GenreMisc,GenreSports
North America	1.076	metascore,Year,userscore,Platformwii,PublisherNintendo,GenreShooter,Platformx360,Platformpc,GenreSports,GenreRacing
Europe	1.143	metascore,userscore,Year,Platformwii,GenreRacing,PublisherNintendo,GenreSports,GenreShooter,Platformps3,PublisherTake-TwoInteractive
Japan	1.196	metascore,Year,PublisherNintendo,userscore,Platformpsp,GenreRole-Playing,PublisherSquareEnix,Platformwii,PublisherCapCom,Platformps2
Rest Of World	1.056	PublisherTake-TwoInteractive,Platformps2,userscore,Platformps3,GenreRacing,Platformwii,PublisherNintendo,PublisherActivision

Fig.4 XGBoost Results

From the results, several findings can be made. The results suggest that my initial hypothesis was correct - Japanese publishers tend to perform better in their home market of Japan. Specifically, the analysis shows that two Japanese publishers, Capcom and SquareEnix, are important factors in the Japanese video game market. Another factor is the console on which the game is played. In particular, video games played on Japanese

consoles, such as the ps2 made by Japanese company Sony, tend to sell better in Japan compared to games played on x360 made by Microsoft. This suggests that the region in which a game is played has a significant impact on its sales performance. Additionally, the North American market tends to agree with the global market trends, possibly due to its significant share in the global consumer base of video games.

Determining if ratings are good predictors of sales or vice-versa

Next, we would like to find out if ratings are good predictors of sales or the other way around. I performed simple linear regression. I created a new variable total_score, which was computed by metascore*0.4 + userscore*0.6, where I slightly favour the user scores as user scores may be more important in industries such as gaming, where the target audience is the user. With total_score as the response variable, I ran a simple linear regression using global sales and vice versa. The results are as follows.

	Sales predicted by score	Score predicted by sales
RMSE	1.29132	1.20915
F-statistic	150.6	150.6
R-Squared	0.03757	0.03757
p-value	<2.2e-16	<2.2e-16

Fig.5 Linear Regression Results

Interestingly, both provided a significant test with a high p-value, indicating the model was accurate and there was a strong positive relationship both ways. To examine this relationship further, my research has led me to propensity score matching. Propensity score matching is a causal inference method that can help to determine the causal relationship between game ratings and sales as it can help to control for selection bias and confounding variables.The propensity score is the probability of a game being in the high-rated (or high-selling) group, given its observed covariates. To train the model accurately, I created two binary variables, high_score and high_sales, where a high rating is defined as ratings above 7 and high sales would be above 1 million copies sold. I included the confounding variables such as Genre, Year, Platform and Publisher.

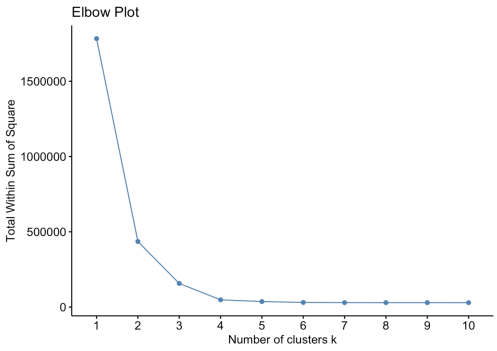
	Sales predicted by score	Score predicted by sales
F-Statistic	3.004	4.427
R-squared	0.1134	0.3319
p-value	<2.2e-16	<2.2e-16
Causal effect	0.5115	0.6568

Fig.6 Propensity Score matching results

The propensity score matching results are interesting as they show that the while both models are statistically significant(p-value <2.2e-16), the score predicted by sales had a much higher R-squared value. This might suggest that higher sales indicate a higher rated game, which was surprising to me as my initial hypothesis was otherwise. Furthermore, propensity score matching is able to tell us that an estimated causal effect of "high_rating" of 0.5115 means that, on average, games with a high score had global sales that were 0.5115 units higher than games with a low score.

Impact of publishers on sales

Lastly, we would like to find out the impact of publishers on sales. K-means clustering was used to cluster the publishers based on the total sales made. From the elbow plot(Fig.7), we will use 5 clusters.



Publisher <fctr>	total_sales <dbl>	cluster <fctr>
Atari	74.53	1
Bethesda Softworks	69.62	1
LucasArts	67.58	1
Deep Silver	18.11	2
Acclaim Entertainment	18.02	2
Sony Computer Entertainment Europe	15.26	2
Nintendo	846.13	3
Electronic Arts	720.53	3
Activision	447.15	4
Take-Two Interactive	325.27	4
Sony Computer Entertainment	323.32	4
Microsoft Game Studios	173.66	5
Sega	167.68	5
THQ	142.76	5

Fig. 7 K-Means Clustering Elbow Plot and top 3 Publishers in each cluster

The results revealed that there were distinct clusters of publishers that had varying degrees of influence on sales. In cluster 1, the publishers grouped together potentially indicate a common target audience or genre preference among their games. In cluster 2, the publishers have lower sales compared to other clusters, suggesting that they may have smaller market share and face tougher competition. In cluster 3, the publishers have the highest total sales, indicating they are the most successful.

IV. Conclusion

In conclusion, our analysis of the video game dataset has revealed that in order of importance, ratings, Year, publisher, and genre are the most important factors that influence a game's success and sales. These findings can guide game developers, publishers, and investors in making informed decisions and contributing to the growth and success of the gaming industry.

V. Appendix

References:

Paredes, L. (n.d.). *Metacritic games*. Kaggle. Retrieved April 15, 2023, from https://www.kaggle.com/datasets/destring/metacritic-reviewed-games-since-2000

R-bloggers.(2022, April 4). Propensity score matching. R-bloggers. https://www.r-bloggers.com/2022/04/propensity-score-matching/#:~:text=This%20code%20tells%20R%20to,out%20using%20the%20summary%20function.

Smith, G. (n.d.). Video Game Sales. Kaggle. Retrieved April 15, 2023, from https://www.kaggle.com/datasets/gregorut/videogamesales

NAME	The name of the video game title
PLATFORM	The gaming platform the game was released on (e.g. Xbox, PlayStation, Nintendo, PC, etc.).
RANK	The ranking of the game based on its total sales.
YEAR	The year the game was released.
GENRE	The genre of the game (e.g. Action, Sports, Shooter, Role-Playing, etc.).
PUBLISHER	The publisher of the game.
NA_SALES	The sales of the game in North America, measured in millions of copies.
EU_SALES	The sales of the game in Europe, measured in millions of copies.
JP_SALES	The sales of the game in Japan, measured in millions of copies.
OTHER_SALES	The sales of the game in the rest of the world, measured in millions of copies.
GLOBAL_SALES	The total sales of the game worldwide, measured in millions of copies.
METASCORE	The Metascore given to the game by Metacritic, which is a weighted average of all the critic reviews for that game.It ranges from 0 to 100
USERSCORE	The User Score given to the game by Metacritic users, which is an average of all user ratings for the game. It ranges from 0 to 10

Fig a. Breakdown of features in merged dataset

Feature <chr>	Gain <dbl>	Cover <dbl>	Frequency <dbl>
metascore	3.045773e-01	2.062204e-01	0.1860111505
userscore	1.471790e-01	8.178334e-02	0.1543335023
Year	1.050694e-01	6.960475e-02	0.1292448049
PublisherNintendo	6.650576e-02	5.247263e-02	0.0192600101
Platformwii	5.486458e-02	2.612409e-02	0.0258489610
GenreShooter	3.645464e-02	1.262153e-02	0.0281297516
GenreRacing	3.214283e-02	9.808635e-03	0.0190065890
PublisherTake-Two Interactive	2.332301e-02	7.698286e-03	0.0169792195
GenreMisc	2.302151e-02	2.241773e-02	0.0210339584
GenreSports	2.159448e-02	1.242413e-02	0.0167257983

Fig b. Top 10 features of Global Sales XGBoost model with Test RMSE: 1.122