Research on Influencing Factors of Video Game Sales using Binary Logistic Regression

Mingze Song*

Jinan Foreign Language School, Shandong, 250000, China

* Corresponding Author Email: mingze1022@stu.sqxy.edu.cn

Abstract. In recent years the game industry welcomed rapid development. As a result, it will be meaningful to figure out what factors influence the sales of games and how they influence them. Based on games from Electronic Arts, this paper intends to use binary logistic regression to analyze the effects of platform, genre, critic score, user score, and rating. The dependent variable is sales in North America. The data is from Kaggle, Metacritic, and the Entertainment Software Ratings Board. According to the regression model, which has an acceptable predicting accuracy of 80.09%, it can be inferred that critic score, platform, and user score represent the significance, while genre and rating don't show a significant relation with sales. Among the three significant variables, critic score has a positive correlation with sales, while user score has a negative correlation. This means games with high sales usually have high critic scores but do not always have high user scores. As a categorical variable, the platform's significance means different platforms may suit different games in general. By contrast, players do not show a preference for a certain game genre or some intense content.

Keywords: Binary logistic regression, game industry, game sales.

1. Introduction

Video games are regarded as the "ninth art" for their cultural influence as an increasing number of people accept them, while their economic achievements are also remarkable. Up to 2022, there have been over 600 million players enjoying different kinds of video games, which account for nearly 46% of the population of China at that time. Such plenty of players generated considerable profits. The online game revenue in 2022 was 266 billion yuan, according to the website Insight and Info. The game industry has become a crucial component of people daily lives.

The crucial position and rapid development have made the game industry a good material for researchers. Early in 2011, researchers found that the game industry could survive an economic crisis because of its low costs and creativity [1]. It can explain why the game industry can rise to be crucial. Some researchers studied the recent facts of the game industry through PCA, such as financing level, user amount, and input into invention. They conclude that the game industry has made much progress, but can develop to be bigger [2]. The effect of the game industry is much more than the economy. Games are more like cultural symbols. As information technology and the Internet develop, online games have combined the advantages of many high-techs [3]. Video games can indeed function as a bridge linking different cultures, which is also comprised of many advanced technologies. As a result, it is a natural thought to study the influencing factors.

In China, few studies directly focus on the influencing factors of game sales, however, many researchers' studies have some relationships with them. On the one hand, researchers try to figure out the meaningful applications of video games. A broader application means more sales. It is widely discussed whether educators can apply games to teaching. For example, kids are generally interested in games, so educators in primary schools can try to use video games to stir up students' passion for studying [4]. In this case, educational games will be more salable. The application of sports games is also a hot topic. Combined with sports activities and cognitive training, it has already proved that sports games have a positive effect on patients. The treatment using sports games has great potential. Further research needs to focus on a more comprehensive and personalized design [5]. Another interesting idea is about food culture. Researchers analyzed the characteristics of interactive games,

proposing that these games can create a world containing different foods. Such a virtual world can give players an immersive experience, which may do better in the spread of food vultures [6]. This proposal shows us the possibility of increasing games consisting of food as an element. The development of video games engaged more people to study it and explore its potential.

On the other hand, some researchers noticed the latest scientific achievements. They aimed at predicting a different future when new technologies are integrated into video games. Metaverse is an advanced concept raised in the 1990s by Neal Stephenson. It described a virtual space where people can experience a different world but feel true. In recent years VR, AR, and MR become better and better, making the metaverse a reality but not just a thought. With an open world for players, there is possibly a boost in video game sales. Many companies have already put forward an initiative to look for some methods that enable video games to approach metaverse games [7, 8]. Another example is artificial intelligence (AI). Researchers believe that AI is a great choice for inventing strategic games. This is because nowadays AI can independently help players make decisions, or it can just be the enemy against players. Researchers also give further explanations of how to build an AI [9]. The use of AI can bring a boost in sales. What's more, 5G is an important advanced technology. Employing 5G, "cloud games" will develop quickly. Since cloud game isn't limited to computers or phones and does not occupy internal storage, it attracts many companies' interests. Thus, researchers studied the cost, customers' experience, and content of games to give some practical advice for the cloud game industry [10]. To make their products sold better, companies must at least perform the 3 perspectives above well. To sum up, 5G, AI, and metaverse are some of the instances of high-tech. Their participation produced more material for researchers to study sales.

The game industry has a bright future according to the essays shown in the former. However, it is even unclear what factors influence game sales and how big their influence is. Based on this, this paper intends to clarify the influencing factors preliminarily by logic regression.

2. Methods

2.1. Data Source

The data is mainly from a dataset on the website Kaggle. It contains a pack of basic information about various video games of Electronic Arts (EA) from 2010 to 2017, with a total sample size of 231. EA is one of the most popular game companies around the world, which has published various games. Therefore, it is typical for the whole game industry. Some scores or ratings are organized from Metacritic and the Entertainment Software Ratings Board (ESRB), which will be introduced in the latter part.

2.2. Indicators Selection

After reading the analyses of the current situation of the game industry, platform, genre, critic score, user score, rating, and release year are chosen as the independent variable, and sales in North America (NA sales) are chosen as the dependent variable (Table 1).

Table 1. Basic Information of the data

variable	type	range	average	medium	variance	SD
platform	categorical	[1,3]	2.64	3	0.54	0.74
genre	categorical	[1,10]	7.94	9	3.91	1.98
critic score	numerical	[0,100]	75.27	78	138.45	11.77
user score	numerical	[0,10]	6.19	6.20	1.77	1.33
rating	categorical	[1,4]	2.33	2	1.61	1.27
NA sales	numerical	[0.01,2.53]	0.53	0.37	0.27	0.52

The critic score is cited from the website Metacritic, an authoritative critic website. It is the average number of many professional critics, thus lots of game producers keep track of the scores. The user

score is also from Metacritic, given out by the subscribers of the website, which represents the comments of players. The rating is evaluated by the ESRB, also known as the ESRB rating. ESRB is an association measuring the level of some intensive content, such as violence, blood, or drugs, and it can decide whether people prefer games with intense content. Most game producers admit the ESRB ratings. Table 1 shows the statistical properties of the data. Sales are recorded in millions of units. All the categorical variables are treated as integers.

2.3. Method Introduction

To gain a deeper understanding of the relationship between independent variable and dependent variable, platform, genre, and rating need to be well-categorized and transformed into numbers.

Table 2. The Mapping between Categorical Variables and Value

Genre	Value	Genre	Value	Rating1	Value	Platform2	Value
Puzzle	1	Action	6	E	1	Personal computers	1
Racing	2	Sports	7	E10+	2	Handheld game consoles	2
Adventure	3	Shooter	8	T	3	Home video game consoles	3
Fighting	4	Strategy	9	M	4		
Platform	5	Simulating	10				

Note: E refers to everyone. E10+ refers to everyone above ten. T refers to teenager. M refers to maturity. The ratings stand for the lowest age suitable for a game. Handheld game consoles include 3DS, DS, PSP, and PSV. Home video game consoles include PS3, PS4, Wii, WiiU, X360, and XOne.

Table 2 explains how the categorical variables can be transformed into numbers. With this data, the model can be better established. This research tended to adopt binary logistic regression as the first choice. Every independent variable will be analyzed by the model to figure out whether the sales will be high (above the third quartile) or low (below the third quartile). Products with high sales will be shown as 1, or they will be shown as 0. The final formula will be:

$$\ln\left(\frac{p}{1-p}\right) = \sum_{i=1}^{n} \beta_i x_i + b \quad (1)$$

This way, it is possible to understand the influencing factor of sales. There is also a possibility that the independent variable has no relation to the dependent variable. Everything will be clarified after the results.

3. Results and Discussion

3.1. Basic Information

Considering that the dataset may contain incomplete samples and extreme samples, the actual data size involved in the analysis is 226.

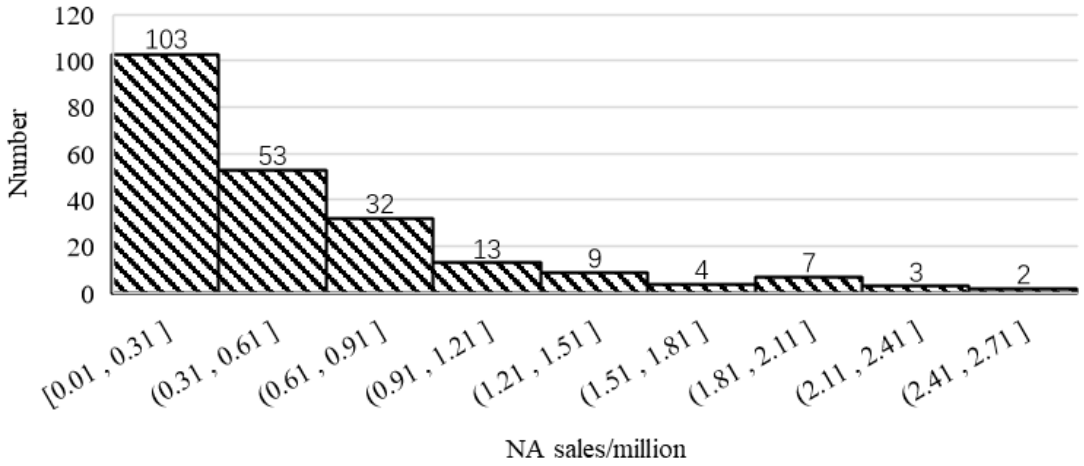


Fig. 1 Histogram of NA sales

Fig 1 shows the distribution of NA sales. The data mainly concentrates on the lower parts and the amount generally decreases with the increase in sales. The distribution shown is consistent with ordinary views of sales. However, the concentration of the data is a disadvantage for analysis.

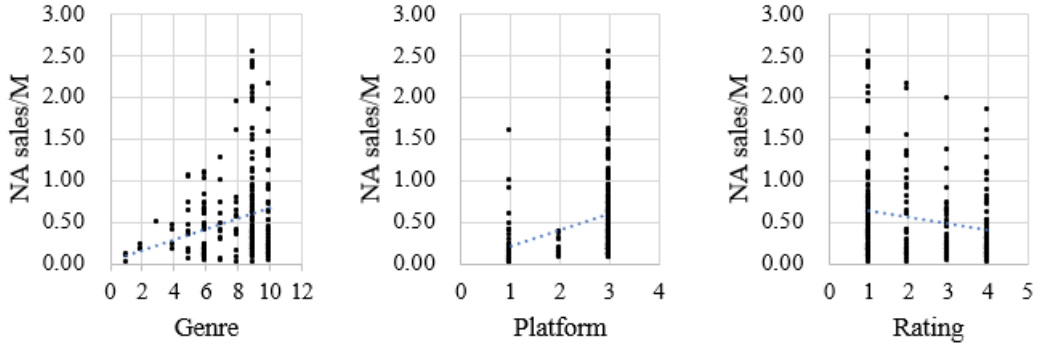


Fig. 2 Distribution of categorical variables

As shown in Fig 2, the data obviously concentrates between 0 to 0.5 when genre is 9, 10, platform is 1, 3, and all values of rating. What's more, other values are also under the circumstances that though the maximums have distinct variations, the minimums are almost the same, around 0.1. The general tendency is also shown in the figure. It is easy to clarify that the tendency of genre and platform is upward while the tendency of rating is downward. s shown in Fig 2, the data obviously concentrates between 0 to 0.5 when genre is 9, 10, platform is 1, 3, and all values of rating. What's more, other values are also under the circumstances that though the maximums have distinct variations, the minimums are almost the same, around 0.1. The general tendency is also shown in the figure. It is easy to clarify that the tendency of genre and platform is upward while the tendency of rating is downward.

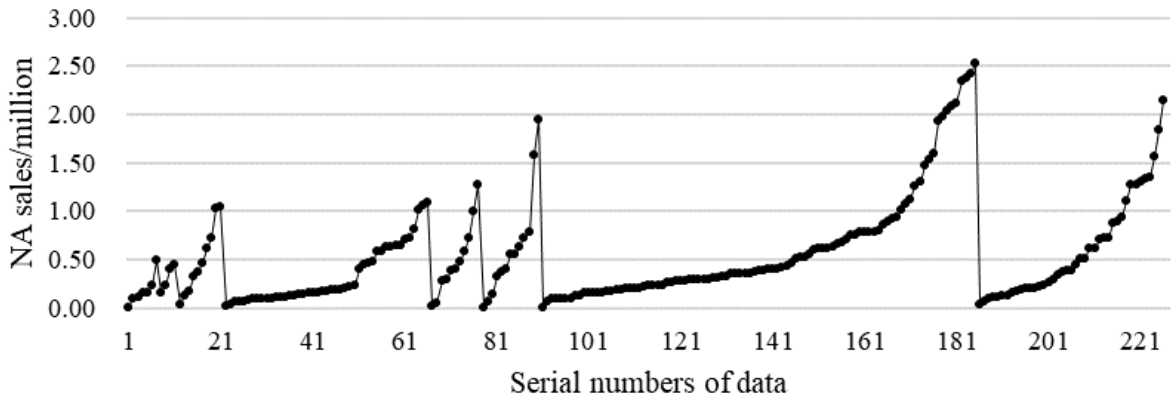


Fig. 3 NA sales divided into different groups according to genre

Fig 3 shows the change in sales in different game genres more clearly. The sales are ranked in order of its value, and separated according to genre from 1 to 10. From the variance of each group, the general tendency which cross groups is unclear. According to the general tendency, using binary logistic regression is a better choice, as it is actually a kind of grouping that lowers the disadvantage of the data.

3.2. Model Results

After figuring out the general tendency of each independent variable, it will be more predictable and convincing to fit the model better.

Table 3. Predicting Accuracy of Binary Logistic Regression

		Predicted Value		Forecast Accuracy	Predicting Error Rate
		0	1		
Ture Value	0	152	17	89.94%	10.06%
	1	28	29	50.88%	49.12%
Total				80.09%	19.91%

Table 3 shows the assessment of the binary logistic regression. The model represents an acceptable forecast accuracy, with a total accuracy of 80%. However, the forecast accuracy of 1 is just 51%, which needs further research on deeper influencing factors.

Table 4. Results of Binary Logistic Regression

Sum	Notation	Regression coefficient	Standard error	Wald χ^2	p	OR	OR 95% CI
Rating	X_1	0.084	0.153	0.298	0.585	1.087	0.805 ~ 1.468
Critic Score	X_2	0.178	0.035	26.539	0.000	1.195	1.117 ~ 1.279
User Score	X_3	-0.567	0.182	9.768	0.002	0.567	0.397 ~ 0.809
Platform	X_4	1.156	0.351	10.856	0.001	3.177	1.597 ~ 6.320
Genre	X_5	0.194	0.116	2.810	0.094	1.214	0.968 ~ 1.522
intercept	-	-16.600	2.951	31.644	0.000	0.000	0.000 ~ 0.000

Table 4 gives out the result of binary logistic regression. According to table 4, the equation can be formed as:

$$\ln\left(\frac{p}{1-p}\right) = -16.600 + 0.084 * x_1 + 0.178 * x_2 - 0.567 * x_3 + 1.156 * x_4 + 0.194 * x_5 \quad (2)$$

Where p refers to the probability of high sales. The regression coefficients and OR values of rating, critic score, platform, and genre are respectively above 0 and 1, while the two values of user score are below 0 and 1. However, both the critic score and user score represent the comments on games. The different results of them are quite unusual.

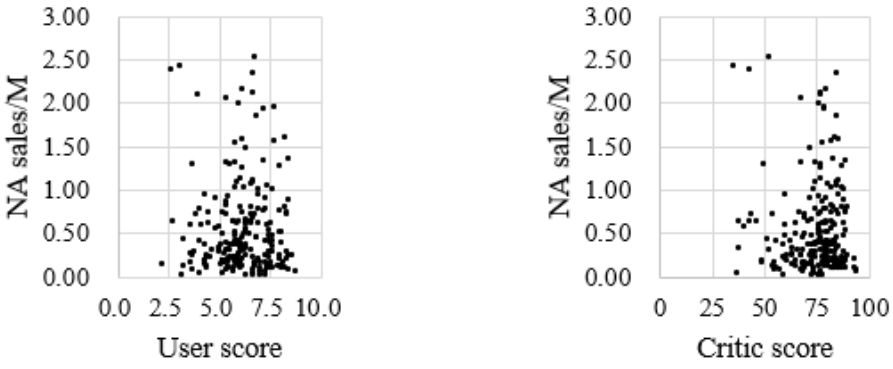


Fig. 4 Scatter diagram of user score and critic score

From Fig 4, the general tendency can be found in an obvious form. The NA sales go higher when the critic scores increase, and the maximum appears when the critic score is almost the highest. By

comparison, the maximum appears when the user score is around the medium value, and its shape is more like a quadratic function. That's why the user score seems to have a negative correlation with NA sales. It can also be inferred that the critic score is more authoritative.

The inference can be proven with the value of p and the standard error value. The critic score has the lowest value of p, below 0.001, of all the variables. The value of p of the user score and the platform are also small, with 0.002 for the user score and 0.001 for the critic score. The genre has a value of p of 0.094, and the value of p of the rating is the largest, about 0.585. The two variables have no significance to the change in NA sales. The standard error generally varies with the value of p, which also describes the relevance. Also, the critic score has the narrowest interval of OR 95% CI, which is 1.117 ~ 1.279. Besides, the OR 95% CI intervals of rating, genre, and user score are under the acceptable range, while the OR 95% CI interval of platform is too wide.

4. Conclusion

This paper clarified the influencing factors of game sales in North America based on 277 games in EA. Using binary logistic regression, a model is established to predict the probability that a game will have high sales according to its rating, critic score, user score, genre, and platform. The main conclusion is that critic score has a positive correlation with NA sales while user score has a negative correlation. The platform of a game can also affects its sales. However, genre and rating have no significant influence on sales. Since there are only few researches that directly focus on the influencing factors of game sales, this paper can be a try for a kind of unknown products, which preliminarily figures out some influencing factors. As a result, it can be a reference for some further studies, and give some clues for game producers to make better-sold games. There are also some remained questions in the paper that need deeper analysis to raise the accuracy of predictions of high sales. It also needs more research that why user score changes differently from critic score. With all these problems fixed, the result of relevant research will be much more accurate.

References

[1] Li Luqi. On the Development of Online Gaming Industry from the Perspective of Economics. Commodity and Quality, 2011, (8): 72-73.

[2] Shang Jie, Zhang Zhiyuan. Empirical Analysis of China's Online Gaming Industry Development. Consumer Guide, 2010, 5: 52-54.

[3] Shang Hui. Research on the Status of China's Online Gaming Industry Development. Henan University, 2009.

[4] Ma Hongmei. Brief Discussion on the Application Significance of Educational Games in Primary School Mathematics Teaching. New Campus, 2023, 12: 52-53.

[5] Chen Zhuoqi, Xu Linyan, Zou Jihua, et al. Application Status of Exercise Games in Old People with Mild Cognitive Impairment. Journal of Nursing Science, 2024, 39(1): 111-115.

[6] Zhan Songqi. Application Research of Game Interaction Design in Food Culture Communication. Lu Xun Academy of Fine Arts, 2022.

[7] Fang Lingzhi, Shen Huangnan. The Changes of Technology and Civilization: A Conceptual Study of Meta-universe. Industrial Economics Review, 2022, (1): 5-19.

[8] Yu Guoming, Geng Xiaomeng. Meta-universe: A Future Ecological Picture of Media Society. Journal of Xinjiang Normal University, 2022, 43(3): 110-118.

[9] Sun Yuxiang, Peng Yihui, Li Bin, et al. Review of Intelligent Gaming: Enlightenment of Game AI to Operational Deduction. Journal of Intelligent Science and Technology, 2022, 4(2): 157-173.

[10] Li Ruixuan, Wang Ziyi, Liu Yang. Research on 5G Cloud Gaming Service Development Trend and Prediction Method. TD Industry Alliance, China Electronics Technology Group Corporation Seventh Research Institute, Beijing Branch, 2022.