Analyzing Game Genres and Their Success

Te Hsin Kung

Department of Computer Inforation System, Cal Poly Pomona

CIS 4321.02: Data Mining

Dr. Fadi Batarseh

Dec 1, 2021

Introduction

Video games have never been a part of my life before I came to United States. Personally, I start playing the first-person shooting game named “Valorant”, which is developed by Riot Games and the company started in Los Angeles, California in September, 2006. In 2020, League of Legends, made by the same company generated over 1.75 billion dollars. Compare to 2019, the revenue of its game generates an increase around 250 million. (Gulati, 2021)

By this we can say that the video games market has been grew big as the technology improves, video game is attractive among different ages, but have you ever wondered which genre is the most. In the article “12 Most Popular Video Games In 2021 You Can Play Right Away”, Gulati mentioned in the graph that top 5 games are Call of Duty Warzone, Minecraft, Animal Crossing: New Horizon, Grand Theft Auto 5 and League of Legends. We can see that different types of genre have its own players market. () Some of the top biggest companies in the world are Sony, Tencent, Nintendo, Microsoft and Activision Blizzard in 2021. (All Top Everything, 2021) The video games industry is valuated at 65.5 billion dollars currently and US consumers reported spending 11.6 billion in 2020. The hardware used for gaming is a separate market and is currently valuated at 472 million dollars. (Clement)

The reason why the video game industry is so big is caused by games becoming an integral part of people’s daily life such as streaming on Twitch, YouTube and so on. There are also professional players who makes a living by playing video games. One of the biggest video game professional players, Michael “Shroud” Grzesiek has made over 2 million per year from endorsement, tournament winning, sponsorship and more. (Statista, 2021) Therefore, I became more interested in learning how the video game industry works. What I want to know is that how

well the video game sales can reach one million sales. In my future career, I plan to work in the video game industry, so I am willing to learn more about how the industry works.

Problem statement

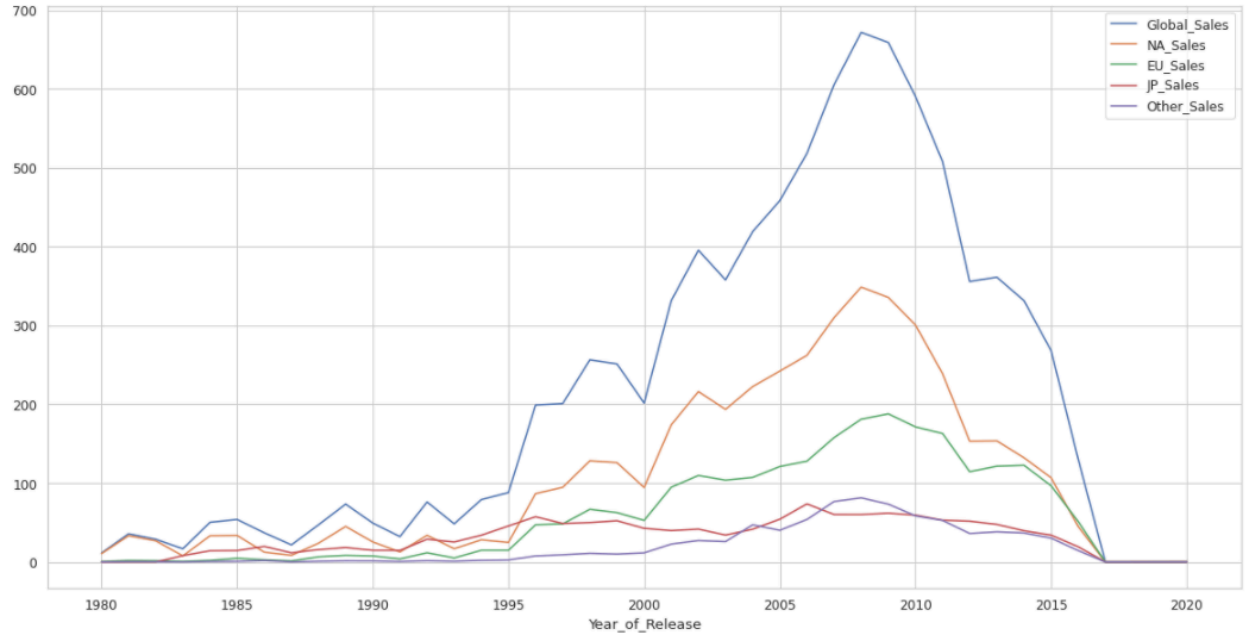
What I am seeking is that what's the game sales performance in each region based on each year? Which game platform release the most games? (Top 5 platforms) Which developer has the most game release? Which publisher has the most published game? What are the most top selling genre? Do these games have the good critical score compare with their sales? In the simplest term, I want to find out how consumers choose what games they want to buy out of hundreds of new games coming out each year.

One approach that I am taking to figure out my problem is by filtering whether a game reaches a certain sales volume or not. Our target in the prediction model will be using the defined hits as with the global sales reach over 1 million units or higher, where I will be predicting if a game will be a hit or not. I will be using decision tree classification and logistic regression model to compare the model result to test the prediction accuracy score. Using logistic regression model to understand the statistical relationship between the dependent variable and independent variables. We can understand the likelihood of the event happening or a decision being made. One reason why I am interested in the global sales is that companies can track whether their current games are selling well or they can even use the previous game sale records to predict whether their future games will succeed or not. This way companies can see how their games is performing in the market and choose genres that give the most success chance. With this kind of data, companies can create the ideal games that will significantly steal market shares. There has been hundreds and thousands of games since the very first video games released and companies need to go through this extensive history to catch the customers attention.

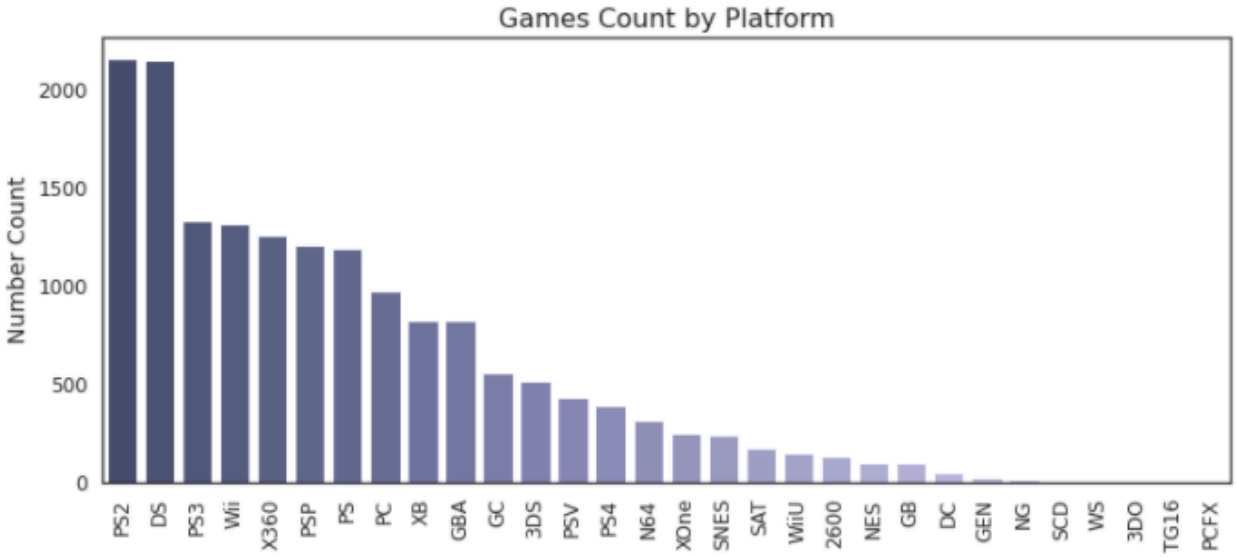
Understanding the dataset

Finding the dataset was challenging because some of the companies do not release their revenue on game sales. I looked through Kaggle and seek for dataset, some of the dataset does not include enough attributes for me to analyze. There was also dataset with messy symbols or too many missing value that will affect the results. In the end, the dataset I chose was called "video games sales 22 Dec 2016" from Kaggle. This is a dataset include different region's sales such as North America, Europe, Japan, etc., name of the game, which platform does it come from, it's publisher, year of release, genre and so on. It has a total of 16 attributes and 16,717 instances. The data type of the attributes includes nine float data types and seven objects data type. The overall dataset does not include messy symbols. The missing value in the dataset include year of release attribute with 269 missing value, genre has 2, publisher has 54, critic score and count both have 8582, users score has 6704, user count has 9129, developer has 623, rating has 6769 When visualizing the missing value shows. My target variable will be aiming at its sales performance which is the attribute called global sales. The main goal is to check and see if each game sales reach one million or higher.

Data analysis and results

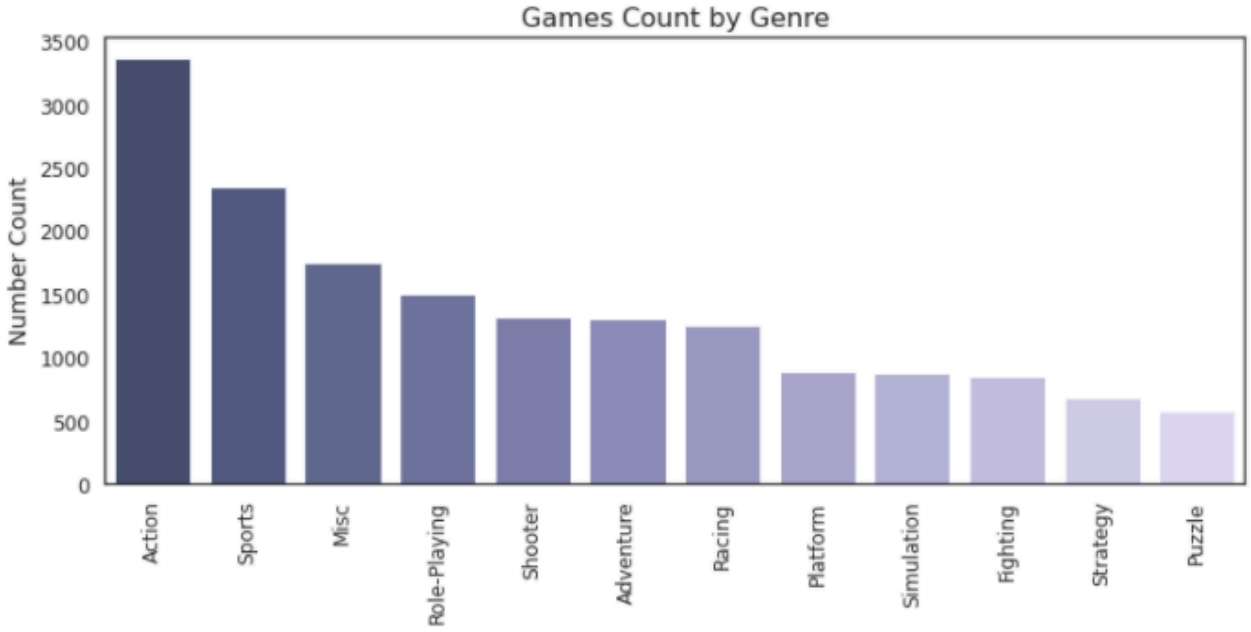


The line graph shows a high peak between 2008 and 2010 where the global sales were the highest and each region's sales also peaked between 2005 and 2010.

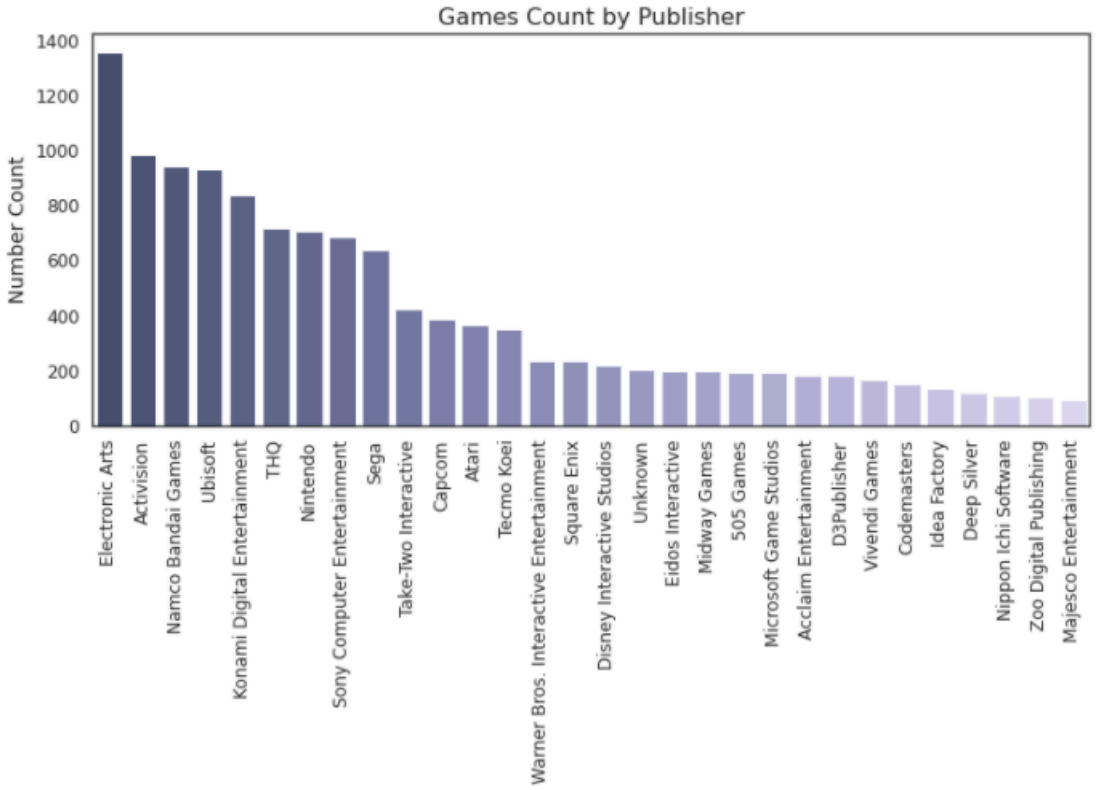


In the EDA, I create a loop to group the games count by platform, developer, publisher and genre. When we take a look at the top five we can see that PS2 has the highest game counts at

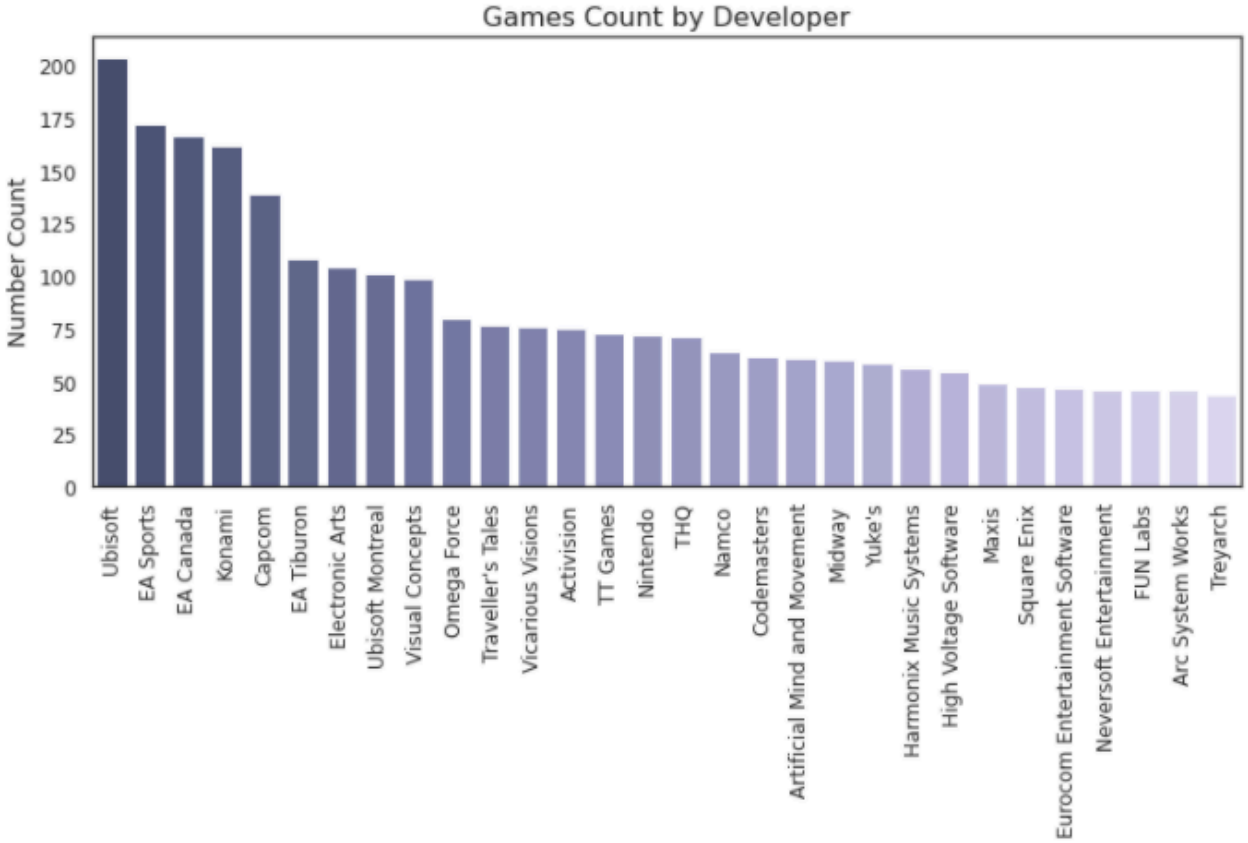
over 2,000 games. DS came close in second place and the bottom three are PS3, Wii and Xbox 360.



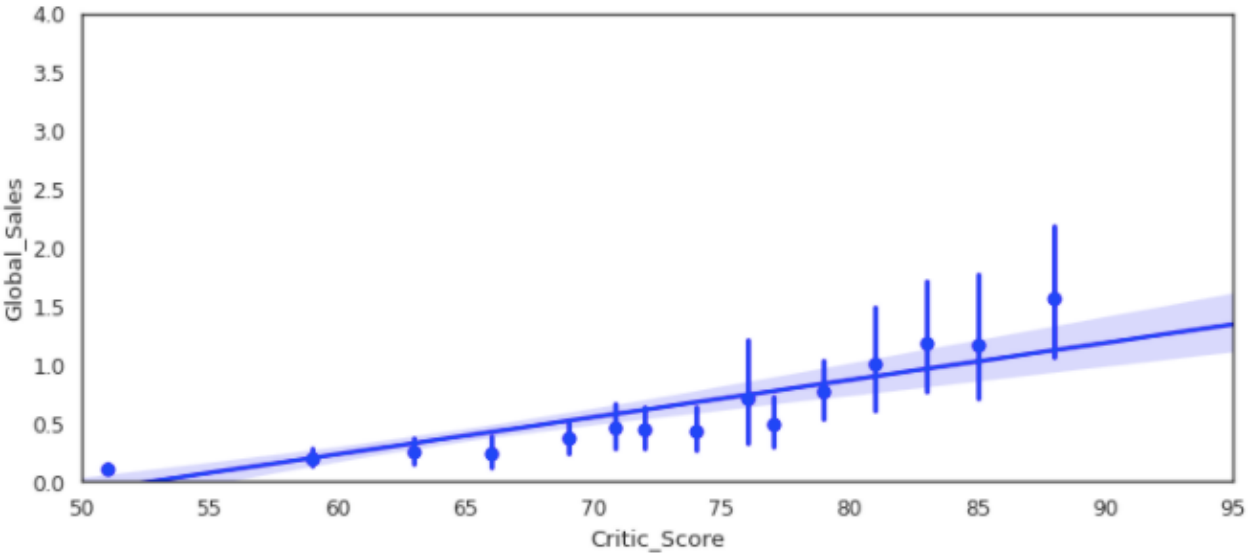
This bar chart shows the number of games in each category. Action has the highest game count sitting at over 3,300 games followed by Sports at over 2400.



This chart shows the number of games each company has published. Electronic Arts sits in the first place at over 1,300 games and Activision comes in second at over 900 games.



This plot chart shows that developer Ubisoft has around 200 game count followed by EA sports with nearly 175 games, EA Canada and Konami have almost the same count, lastly Capcom sits around 150 game counts.



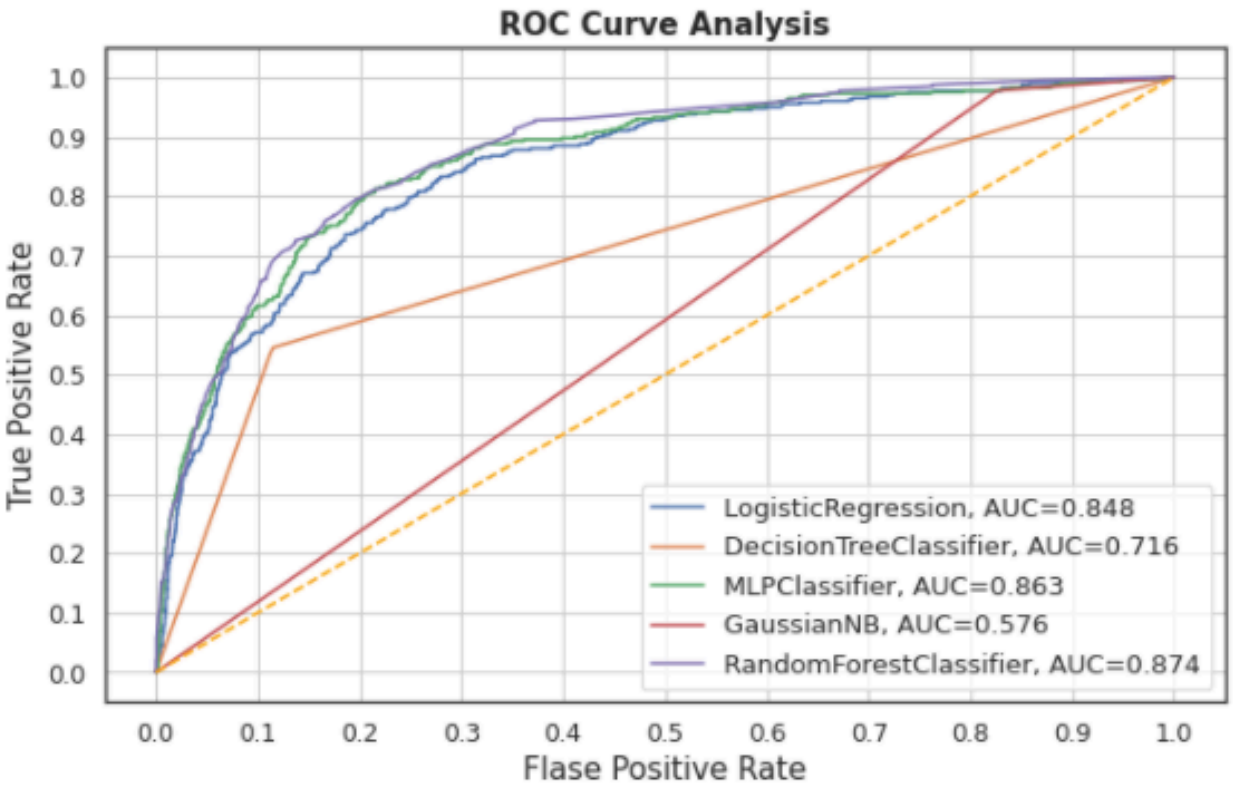
This correlation graph show global sales is proportional to critical score.

	precision	recall	f1-score	support
0	0.88	0.97	0.92	1993
1	0.66	0.33	0.44	402
accuracy			0.86	2395
macro avg	0.77	0.65	0.68	2395
weighted avg	0.84	0.86	0.84	2395

The decision tree model gives an accuracy score of 0.86 which was not overfitting and the precision has 0.88 with a recall score of 0.97. This is better than what I got before using the best model.

	precision	recall	f1-score	support
0	0.89	0.95	0.92	1993
1	0.63	0.40	0.49	402
accuracy			0.86	2395
macro avg	0.76	0.68	0.70	2395
weighted avg	0.84	0.86	0.85	2395

The logistic regression model gives us an accuracy score of 0.86, precision has 0.89 and the recall score is 0.95. Use Decision tree and logistic regression to see the accuracy score and both came out 0.86, 86%.



	LR	DT	NN	NB	RF
accuracy_mean	0.81	0.78	0.82	0.29	0.81
accuracy_stdev	0.29	0.21	0.21	0.29	0.28

Based on the result we got, neutral networks model has the best performance following by logistic regression model then decision tree model has the worst performance. The accuracy mean for logistic regression is 0.848(84.8%), neural network has 0.874(87.4%) and decision tree is 0.716(71.6%). For the AUC score results, both logistic regression classifier and neural networks classifier seem to do a pretty good job, but the neural networks classifier appears to perform slightly better. Comparing Logistic regression model and decision tree classifier, logistic regression has an AUC score of 0.848 slightly close to a perfect classifier, but on the other hand decision tree classifier has the lowest score of 0.716. Therefore, we should

use logistic regression model since overall it has the best performance compare to decision tree model.

In the line graph, we can see that the high peak lands in 2009 and each regions sale were also supporting the results. For the games count by platform, it shows that PS2 has the top sales among all other platforms. Games count by publisher graph shows that Electric Arts release nearly 1,400 of games. For the Games count by developer, Ubisoft has the highest game count. The most purchased genre is action games, which companies can create more this kind of genre. Based on the correlation graph, we can see the performance of global sales is growing as the critic score goes higher.

Conclusion

After going through the dataset, I was able to find the answer that I was looking for. The Action market is already over saturated, and we can see that there are other genres yet to be filled. If game developer creates other genres they may have an easier time competing with other companies as there are less games to choose from. In the graph games count by developers, we can see that EA sports have the second highest game count. However, there are several companies that out rank EA when it comes to most revenue generated.

References

All Top Everything. (2021, October 20). Top 10 biggest video game companies in the world. All Top Everything. Retrieved December 2, 2021, from https://www.alltopeverything.com/top-10-biggest-video-game-companies/.

Clement, J. (n.d.). Topic: Video game industry. Statista. Retrieved December 2, 2021, from https://www.statista.com/topics/868/video-games/#:~:text=Video%20games%20are%20a%20billion,over%2077%20billion%20U.S.%20dollars.

Gulati, S. (2021, June 15). 12 most popular video games in 2021 You can play right away. Fossbytes. Retrieved December 2, 2021, from https://fossbytes.com/most-popular-video-games/.

www.statista.com

statista.com

Zippia. (2020). Working at Riot Games: Employee Reviews and culture - zippia. Retrieved December 2, 2021, from https://www.zippia.com/riot-games-careers-36774/.