Ricky Gajeski, Cheyann Odle

Prof. Beagley

DATA-151-A

5/07/23

Video Game Sales

**Introduction**

In 1983, after the release of E.T. the Extra-Terrestrial video game for the Atari 2600, the video game market fell into a recession as a response to the game’s (and many others) poor quality. For the next two years, the video game struggled to survive as consumers around the globe rejected video games. It was not until the release of the Nintendo Entertainment System in 1985 did consumers regain trust in the video game market and started to purchase and play them again. As with any business, it is best that an entire industry does not collapse, especially if is avoidable by your own control. Moving forward, it became important for video game companies to reflect as to why or what makes a game successful. As data analysts, we want to discover the most common variables within best-selling video games and create a model to predict what games with certain variables would be best sellers. This is an important question to answer as it would keep the video game industry thriving and overall help improve consumer reception to video games as the product would hopefully come out better and make more tailored to trends of the industry at the time. So far, the video game market has been successful in staying profitable. In fact, according to Paul Tozour of *Euronews*, the video game industry should reach a value of $326 billion by 2026, and its success during the financial crisis of 2008 has shown that the video game industry is recession-proof (*Euronews*). But Tozour also points out a potential industry problem: consumer interest is at threat if another recession occurs while the industry’s reputation stays in a negative light. With a potentially shaky future ahead, it would be important for video game companies to understand what common trends appear in the best-selling video games. Therefore, our data analysis becomes important as it would help provide video game companies with a list of trends that will help create a successful video game rather than another E.T.

A good way to begin is to start just start comparing features. “What Makes a Blockbuster Video Game?” is the title of a peer reviewed article by Joe Cox questioning what quality combination possibly make up a blockbuster video game using US sales data. Joe Cox used ordinary least squares and logistic regression models to run on his data set. He used an estimation of those models to counter the long tail of his data. The importance of a video game’s profitable success is emphasized by how much risk goes into making a videogame, by explaining how the company is in charge of all the costs but may not be able to make a profit due to the game flopping. The game flopping then bankrupts the company. This can be avoided if companies knew what qualities to focus in on. Joe Cox found, by his data, “handheld platforms are found, on average, to sell significantly fewer copies than for home platforms “ (Cox). He concluded that sales are up further with three qualities: the more major a publisher, a home platform, and good quality.

While qualities are necessary to determine relationships between features it is necessary to keep an eye on the qualities and the direction of impact or lack of impact, they have on the features you are comparing. Hoon S. Choi and coauthors wrote a peer reviewed article covering a how different qualities can correlate to the overall outcome of video game sales. “According to cue utilization theory, products consist of wide range of of quality cues such as price, brand name, packaging, and color, which indicate their potential quality to consumers” (Cho et al). This gives some examples as to what counts as a quality to include or consider comparing to overall sales. The most unexpected thing could have the deciding factor whether someone purchases something or not, especially in terms of video games. With so many games to choose from, it becomes even more critical to maximize the chances for a sale.

Can a video game be classified a selling a High amount of units or a Low amount of units ?

**Data and Methods**

After searching for a data set, we found one from Kaggle named Global Video Game Sales. In this data set, data entries include the video game’s title, publisher, platform, release year, genre, and units sold (both regional and global). With this data set in hand, we decided to use google collab for the time being to code our model, then import it to GitHub due to our familiarity with coding on Collab and ease of access for both of us to work on the code. The objective for this dataset was to run a classification model to determine whether a game will be a high or low unit seller. To start, we first cleaned our data by removing entries that had null in any of the variables. With the nulls dropped, we then created a few graphs to better visualize our data. Which look like the following:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

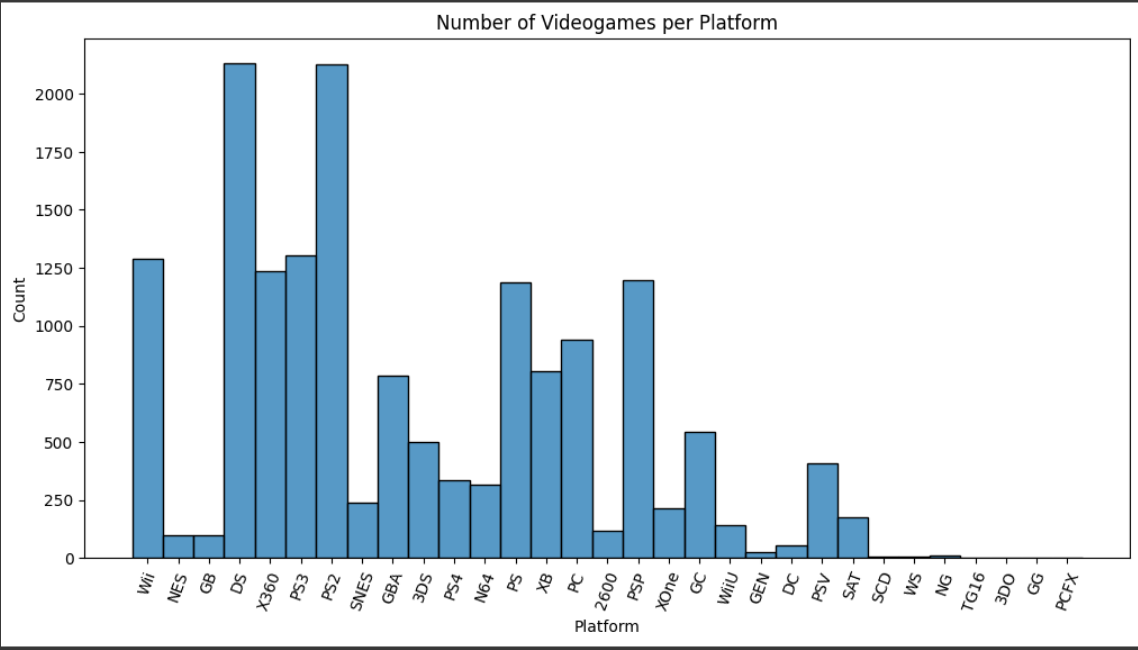
Description automatically generated Now having a good idea of how our data looked, we proceeded to create a heat map of the correlation values between the variables, which produced the following:

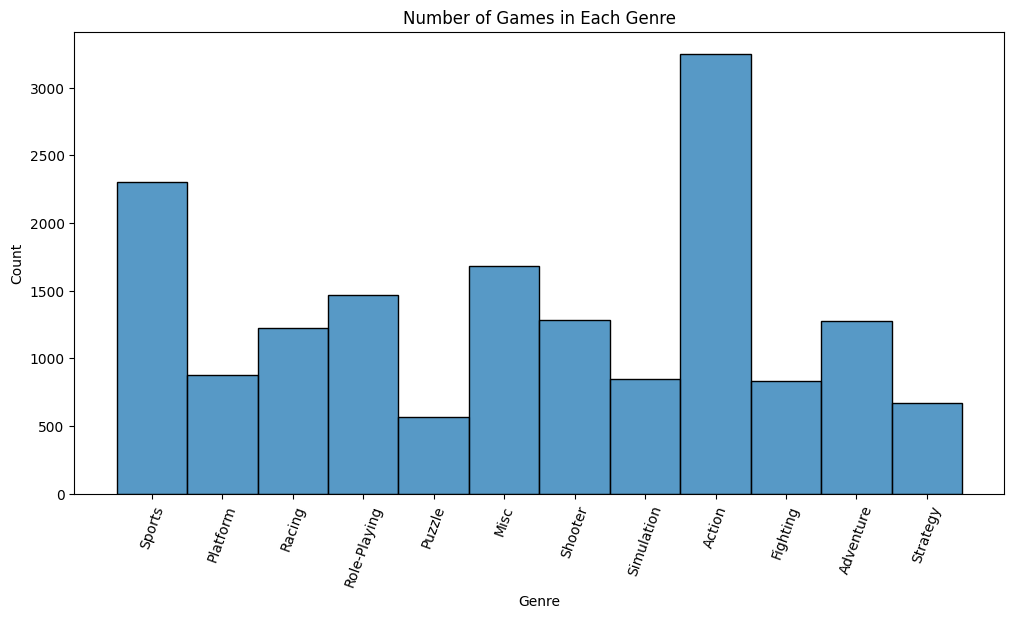


From this map it was noted that the regional units sold had a high correlation with the Global Sales. To make sure they were viable to use as qualities, the sum of all the regional sales were equalized and compared to the global sales. The values were equal, so the regional sale columns were dropped. Genre, Platform, and Publisher were transformed from a string to an integer using the one-hot method. One-hot means that each category option becomes a column and a 1 means that this value has that category or a 0 means that it does not have this category. An additional categorical column was created titled “High/Low”. This was column was created by labeling any Global Sales over the median (0.17) as “High” and any sale equal to or less than the median as “Low”. Four models were chosen: Gaussian Naive Bayes, Decision Tree Classifier, Random Forest Classifier, and an AdaBoost Model. The feature matrix was defined as X= Platform, Genre, Publisher, and Year. The target matrix was defined as y= High/Low. The Gaussian Naïve Bayes Model has a F1-score for High as 0.46 and Low as 0.62. The Decision Tree Model had an F1-score for High as 0.68 and Low as 0.65. The Random Forest Model had an F1-score for High as 0.69 and Low as 0.68. The AdaBoost Model had an F1-score for High as 0.33 and Low as 0.66. Naïve Bayes was chosen as this model works good for training datasets with independent parameters and isn’t likely to overfit. Decision Tree modeling was chosen as it provides a good visual on how it obtains the High/Low answer. The Random Forest is a summation of several decision trees (however many decided upon) and is good for decreasing variance. Adaboost is good for lowering bias and does prone to overfitting, but is severely weak to outliers.

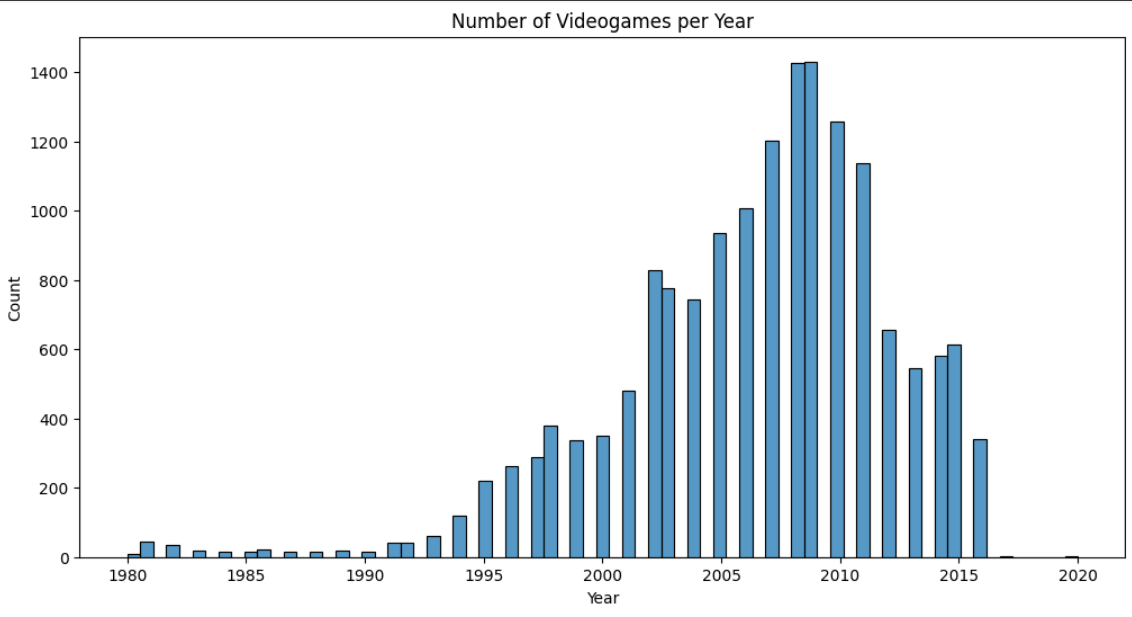
**Results**

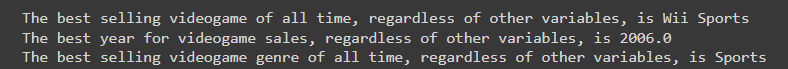
The bar graph titled: Number of Video Games per Platform shows DS and PS2 as holding the highest values of games for a platform. This bar graph has a right skew.



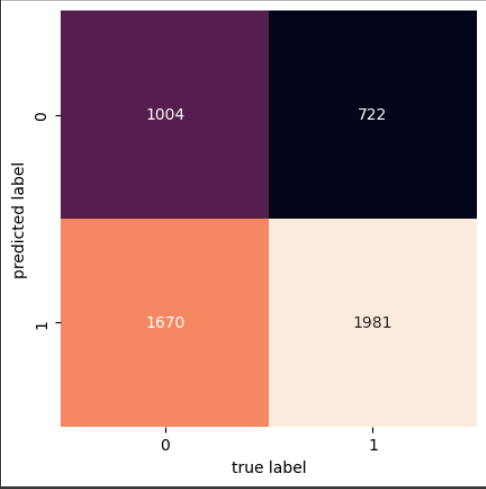
The bar graph titled: Number of Games in Each Genre has action as having the most games and puzzle games having the least amount. This bar graph is normally distributed.T

The bar graph titled: Number of Videogames per Year has 2006 with the most units of games sold and the highest selling unit game of all time is Wii Sports.

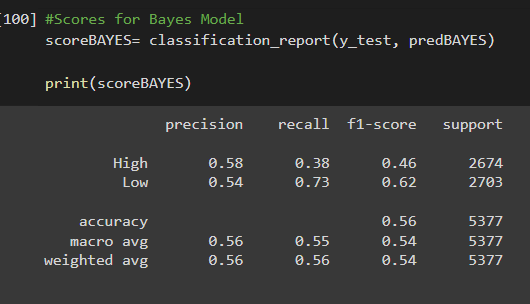




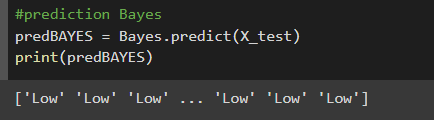
The Naïve Bayes Confusion Matrix: The 0 on the side is a place holder for “High” and 1 is a place holder for “Low”. The dark blue box has 722 as mislabeled and the orange box has 1670 as mislabeled. While the purple has 1004 labeled “High” correctly and the tan box has 1981 labeled “Low” correctly.



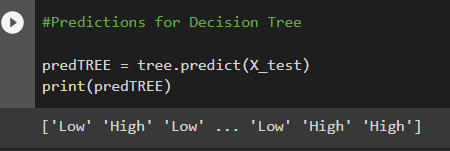
The scores for the Naives Bayes Model a F1-score for High as 0.46 and Low as 0.62.



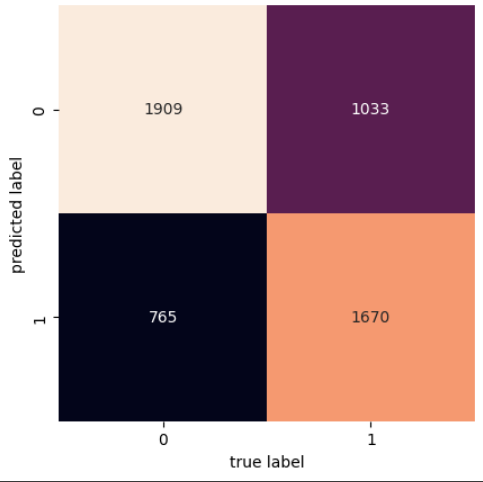
The prediction for Naives Bayes Model.



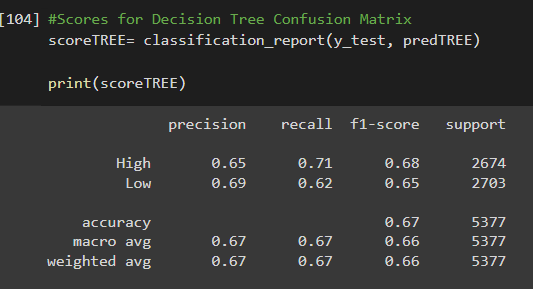
The prediction for the Decision Tree Model



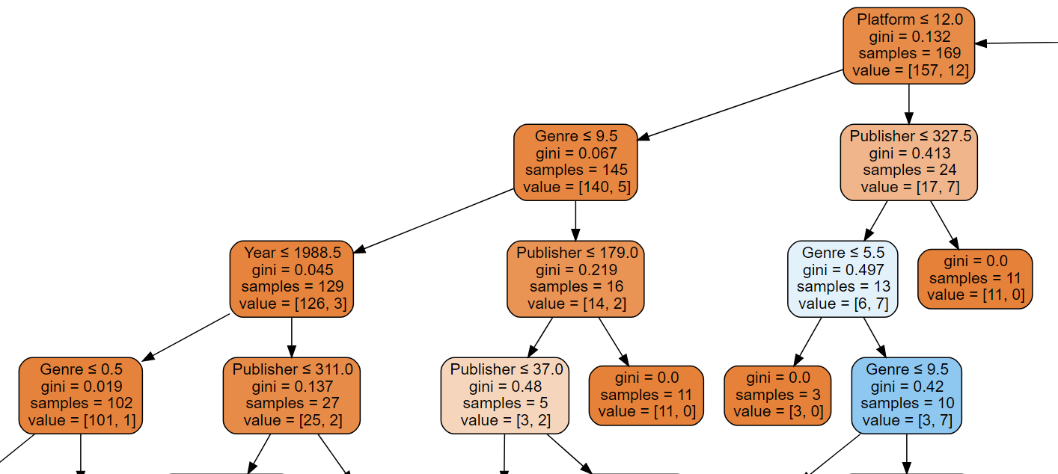
The confusion matrix for the Decision Tree. The purple box has 1033 as mislabeled and the dark blue box has 765 as mislabeled. While the tan has 1909 labeled “High” correctly and the orange box has 1670 labeled “Low” correctly. This matrix is almost the reverse of the Naives Bayes matrix. This matrix is labeling more as “High” incorrectly as compared to the Naives Bayes confusion matrix.



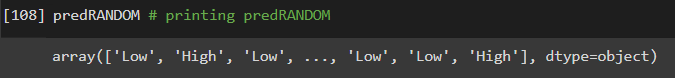
The Decision Tree Model had an F1-score for High as 0.68 and Low as 0.65.



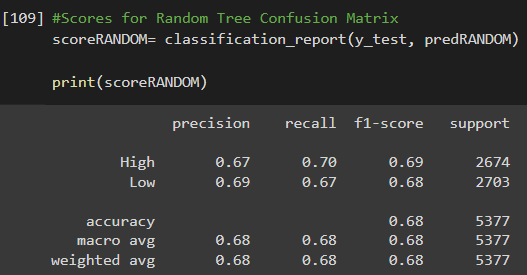
This is a small portion of the Decision Tree and how it is grouping the variables.



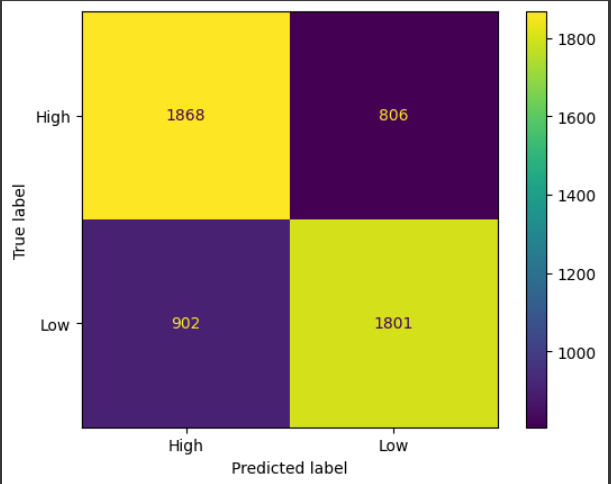
Random Forest Model predictions.



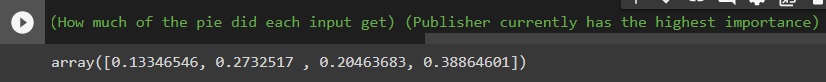
The Random Forest Model had an F1-score for High as 0.69 and Low as 0.68.



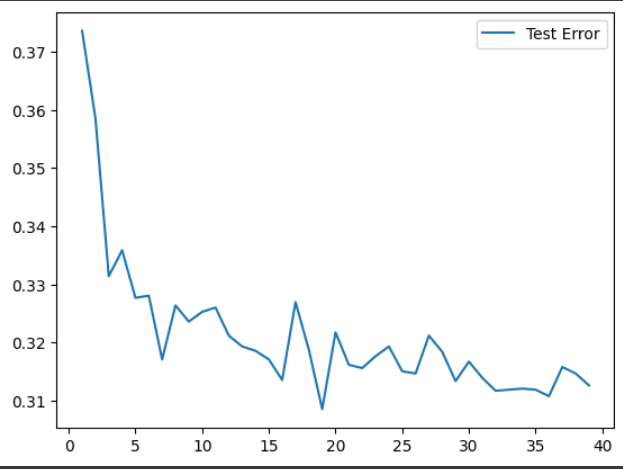
The top right purple box has 806 as mislabeled and the bottom left purple box has 902 as mislabeled. While the top left yellow box has 1868 labeled “High” correctly and the bottom right yello/green box has 1801 labeled “Low” correctly.



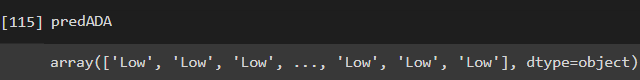
The order of importance for the Random Forest Model. In order of importance Publisher, year, genre, and platform.



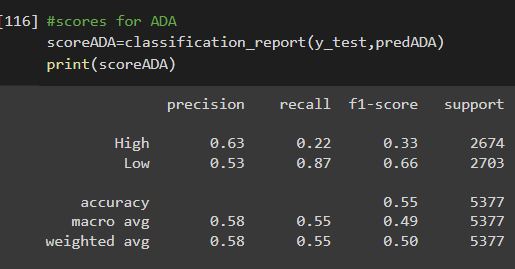
This graph is to see a visual if the Random Forest is over fitting, as the graph is going towards a zero slope without spiking upwards it means that there currently is no overfitting.



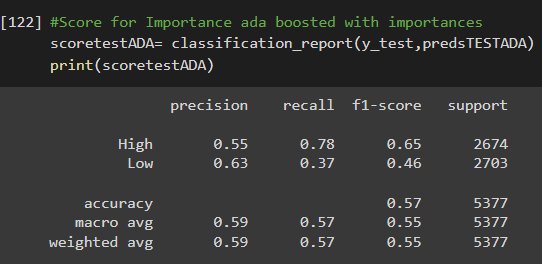
Predictions for the AdaBoost Model.



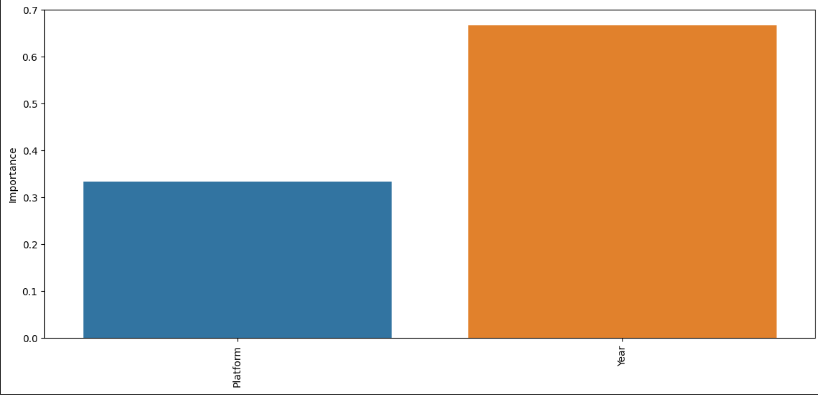
The AdaBoost Model had an F1-score for High as 0.33 and Low as 0.66 before the importance features.



The AdaBoost Model had an F1-score for High as 0.65 and Low as 0.46 after focusing on important features.



These are the two features that Adaboost identified as most important, year and platform.



**Discussion**

The data models were switched from regression models to a classification models, as predicting a general High/Low would be more straightforward to determine from other categorical values. From preliminary graphs some basic insight was given into the dataset. PS2 and DS have the most in count. Action has the most games in terms of genre. Wii sports has the most units sold in terms of Global Sales. The best year for units sold was 2006. Future running of the models may include one-hot coding the year. This is to prevent the year as being registered as a number and to have it be registered a category. This could alter the Adaboost model as it read Year as having the most importance. Further running models on the dummies may yield more accurate results as well. To help add to the models more columns should be incorporated, the models added should extend on the other columns. Hoon Choi and coauthors brought up an interesting point between publishers and game developers. Publishers deal more with the advertisement aspect of game sales while game developers make the game. The idea is that there may be games that sell more units may be developed by the same game developer as opposed to being advertised by the same publisher. The columns added to the dataset due to this point will be game developers and advertisement budget. To further add on to the dataset the type of console should be added. By giving the models more features they can yield better results by possibly having better combinations of features that can predict video game units sold.

**Conclusion**

Four classification models were run to determine whether a video game will sell a high or low number of units. The models ran are Gaussian Naïve Bayes Model, Decision Tree Model, Random Forest Model, and an AdaBoost Model. Out of the four models ran, The Random Forest Model has the current best F1 score with a weighted average of 0.68. Next highest weighted average F1-score is 0.66 using the Decision Tree Model. The third highest F1-score is AdaBoost modeling with 0.55. And the Gaussian Naïve Bayes Model has the lowest F1-score at 0.54. The Random Forest Model has an okay F1 score, this could be better. What that means is that over half the data generated from the target matrix when compared to the actual “High/Low” is being labeled correctly at a score of 0.66. For the Random Forest Model there was an area where the importance feature was run and publishers were found to be the most important feature. If publishers are the most important feature in determining a high unit selling video game, this can create less promotion for smaller publishers. This can happen as those in the video game industry do want their game to do well in the market, so if a publisher were to be the biggest factor that publisher will then get the most opportunities. This skew in opportunities could eventually shrink the publisher competition market. Another thing that may be tested is the number of platforms to a publisher, as the more platforms a publisher has the more opportunities, they are going to have to sell units.

Work Cited

Choi, Hoon S., et al. “The Effect of Intrinsic and Extrinsic Quality Cues of Digital Video Games on Sales: An Empirical Investigation.” *Decision Support Systems*, vol. 106, 2018, pp. 86–96., <https://doi.org/10.1016/j.dss.2017.12.005>. Accessed 21 Apr. 2023.

Cox, Joe. “What Makes a Blockbuster Video Game? An Empirical Analysis of US Sales Data.” *Managerial and Decision Economics*, vol. 35, no. 3, 2014, pp. 189–98. *JSTOR*, <https://www.jstor.org/stable/26607770>. Accessed 20 Apr. 2023.

Tozour, Paul. “Video Game Industry Is Incredibly Profitable. It’s Also Highly Toxic.” *Euronews*, 5 Apr. 2023, www.euronews.com/2023/04/05/video-game-industry-is-so-profitable-its-recession-proof-its-also-highly-toxic#:~:text=Video%20game%20industry%20is%20so%20profitable%20it. Accessed 11 Apr. 2023.