**Data Mining in Videogame Sales Dataset**

Samuel Ramirez

Data Mining/Predictive Model 1 CUS 610

**ABSTRACT**

The main goal of my analysis is to find hidden patterns and relationships between the attributes using data mining techniques such as decision tree classifiers, k-means clustering, and the Apriori algorithm. The outcomes of these different algorithms work hand in hand with each other in explaining the intricacies and the connections in the data. For the decision tree classifiers using the Genre Action, it was found that Genre was the best attribute to test on the dataset for accuracy. For the k-means clustering we found the dataset fit into three clusters: cluster 2 with the overwhelming majority of values that were on the bottom of videogame sales, cluster 0 with values that sold better than cluster 2 and cluster 1, which was the top 1% of videogames sold in the dataset. Using cluster 1 and comparing it with other attributes, I was able to make interesting observations on the data. Finally, using the Apriori algorithm, I was able to delve deep into the relationships between the attributes and found very specific connections that I was able to apply to the dataset as a whole.

**RELATED WORK**

In the field of videogames several different data mining techniques are used extensively to gain insight into both the videogame itself as well as the people playing the videogame. In an experiment conducted in UC Santa Cruz, researchers used a regression model called RUnEA to see which videogame features kept players coming back to the videogame over an extended period of time. One of the games they tested this model on was called Infinite Mario and they found that the most important feature in regards to player retention was difficulty level proportional to the skill level of the player [6]. Another data mining technique that was applied to videogames was clustering in World of Warcraft. Researchers wanted to classify players based off their behavior into four different clusters: Explorers, Achievers, Killers and Socializers. The numerical data that they used for the k-means clustering was the time spent in different regions. The conclusion they came too was the majority of the players were socializers. The exact breakdown of the clusters were 66% socializers, 26% explorers, 5% killers, and 4% achievers. These results are useful because this allows for the game to cater more to the most popular cluster of players [7]. The k-means clustering algorithm is used to classify players in other games as well such as Tomb Raider: Underworld, Tera, and Battlefield 2: Bad Company 2. For example, data was collected on different ways players died in game to classify them into 4 clusters: Veterans, Solvers, Pacifists, and Runners. This type of data mining on gaming metrics is important because it gives companies insight into the player experience and allows them to improve their game accordingly [8].

In another experiment both decision trees and K-means is used to classify and cluster players. This experiment combines Bartle’s taxonomy (Achievers, Killers, Socializers, and Explorers), with two more classifications: hardcore and casual players. The 4 new classifications created are Hardcore Achiever, Casual Achiever, Hardcore Killer, and Casual Killer. A list of 12 attributes were collected such as number of items collected, number of coins collected, number of enemies defeated, etc. The K-means algorithm then reads the training set and creates four clusters without their classification labeled. These clusters are passed into the decision tree classifier and a new tree is created using this data. The training data is then labeled with its classifiers. Once this is done this allows for the decision tree classifier to create a new tree to choose between the different classifications. The player’s individual attributes that were stored in game then gets passed through the decision tree and a classification is made for that specific player. This method improves upon earlier usage of Bartles taxonomy and clustering since it uses a combination of K-means clustering and decision tree for the algorithm, and expands upon the different classifications by adding another layer, hardcore and casual [9]. As you can see, clustering and classification are popular data mining techniques in the field of data science. In my dataset I also used decision trees and clustering. However, my approach in using k-means clustering was different from their approach in that while they used the clusters to classify the players based off their behavior in game, I used the clusters for strictly sales between each region. Because of the nature of my dataset I could only classify my clusters into three different categories: critical success, high success, and moderate success.

**DATA**

The dataset that I used is called vgsales and has information on 16,600 of the top videogames sold over the past 40 years. The attributes in the data set are Rank, Name, Platform, Year, Genre, Publisher, NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales, and Global\_Sales. For pre-processing the first step I took was to remove any classifying attributes, which were Rank and Name. Once I did that I was left with three nominal attributes and five numerical attributes. Next I checked for missing data and found missing values in the Year and Publisher attributes. I filled in the missing values for year with the mean of its values and I filled in Publisher with blank values. Lastly, I checked for duplicate data and found and removed 258 duplicate rows. For the decision tree classifier, I only used the three nominal attributes. For this algorithm to work, I converted the three attributes from nominal to binary. Also, since there were severe class imbalances for each decision tree, I used random under-sampling in order to even out the classes. Class imbalance is a common problem in the field of data science that can be remedied by either randomly undersampling or randomly oversampling [1]. For k-means clustering, I only used the numerical attributes that corresponded to each regions sales. I did this in order for the algorithm to work and also because I wanted to see and compare the kind of clusters that would form with respect to each individual regions sales. Finally, for the Apriori algorithm, I chose only nominal attributes since Apriori does not accept continuous data such as sales from each region. The only preprocessing step I took for this algorithm is I converted the dataset into a list in order for the Apriori algorithm to run through it.

**METHODOLOGY**

The three main techniques I applied in my analysis were the decision tree classifier, k-means clustering, and the Apriori algorithm. For the decision tree classifier, I used a training and test set. I split it so that 70% of the data went to the training set and 30% of the data went to the test set. The reason for training and test sets in the first place is because the training set is used to build the model for the decision tree. Once the decision tree model is built using the training set, the test set is used to test the model for accuracy. The goal of decision tree classifiers is to predict classifications of data based on a set with those classifications. Decision trees are created using attribute-value vectors and utilize a divide-and-conquer method [2]. Typically, decision tree classification follows two phases, a tree-growing phase and a tree-pruning phase. In the tree-growing phase the entire data set starts at the top where the root node is and then gradually starts to split based on a specific criterion until the data is divided into separate clusters and cannot be split further. In the tree-pruning phase, certain sections of the decision tree are removed in order to increase accuracy and prevent overfitting [3].

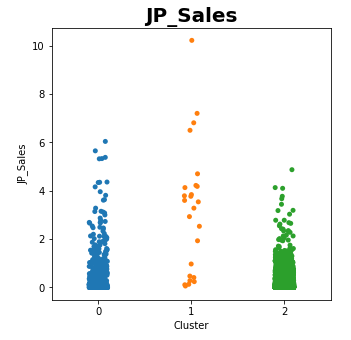
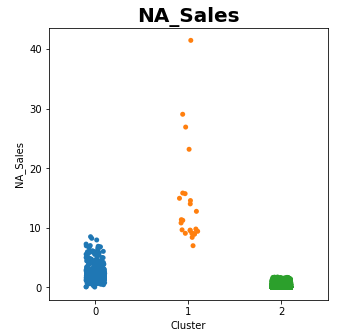
The second data mining technique that I used was k-means clustering. Using videogame sales in each region as the numerical data, I first randomly assigned the data into 3 clusters. After finding the centroid (average) of each cluster I reassigned the data into the cluster with the closest centroid. This procedure was repeated until there were no change in the clusters. An experiment was conducted by researchers at Purdue University to see which approach was better for k-means clustering, interactive or non-interactive. While the interactive approach tried to create its own unique algorithm for data mining, the non-interactive approach creates algorithms that outputs a summation of the dataset so that data mining techniques can be applied to it. Researchers concluded that the non-interactive approach was better since more analysis could be made on it [4]. One challenge that k-means clustering has to deal with is when it is first creating its centroids, it has high sensitivity. An innovative approach to try to combat this issue that was thought of by researchers at the South China Agricultural University is to use ensemble learning for centroid creation. How this works is first an ensemble of base clusters are created using a random amount of clusters and initializations using k-means. These base clusters are then used to create a coassociation matrix which in turn is used to get pre-clustering results. Finally, from these pre-clustering results, the first centroids are created for the k-means algorithm [10].

The third data mining technique that I applied in my analysis was the Apriori algorithm. I used this algorithm on a list of data for nominal attributes and it returned association rules that linked specific values in each attribute together. How the Apriori algorithm works is it first scans the dataset in order to find frequent 1-itemsets. Using those k-frequent itemsets, it then generates (k+1) itemsets and tests it against the dataset. This procedure keeps on repeating until no more frequent itemsets can be generated. This algorithm was one of the first frequent itemset mining algorithms to come out and is still popular today due to its simplicity. It is best known for being used in market basket analysis where companies tried to find products that were being sold frequently together. A problem that the Apriori algorithm has is when generating candidate sets from frequent 1-item sets, it takes a long time and a lot of computing. Researchers found a new innovative approach by skipping the candidate generation set called R-Apriori. This algorithm improved in speed, efficiency, and performance over Apriori. As the dataset size and item number increased, the gap between R-Apriori and Apriori widened, with R-Apriori being the preferred algorithm [5].

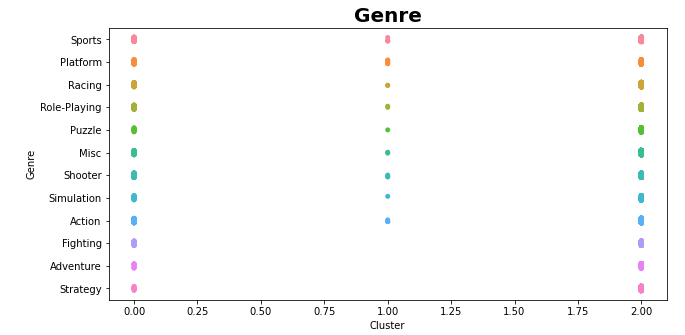
**RESULTS**

I ran a decision tree classifier on the top value from three attributes: Genre, Platform, and Publisher to see which attribute was the best predictor. For Genre I used the Action value, Platform I used the Nintendo DS, and Publisher I used Electronic Arts. After running the decision tree on all three values, I found that the Genre attribute was the best option for predicting the data since its accuracy on the test data was the highest compared to the other two value's accuracy. Action had approximately 79% accuracy, the Nintendo DS had approximately 75% accuracy, and Electronic Arts had approximately 66% accuracy.

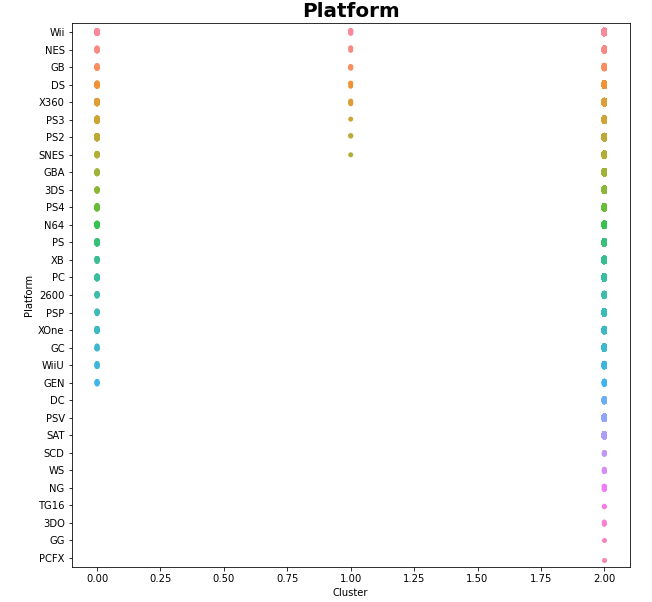
The results of the k-means clustering shows that of the three clusters, the vast majority of the data lies in cluster 2 while a tiny fraction of the data lies in the other 2 clusters, with cluster 1 only containing .15% of the data. With the aid of data visualization, I see that cluster 2, which has almost all the data values, are the sales that did comparably the least well while cluster 1 are those that did the best. This shows me that it is very rare for a videogame to do extremely well and out of the thousands of videogames they are far and few in between.



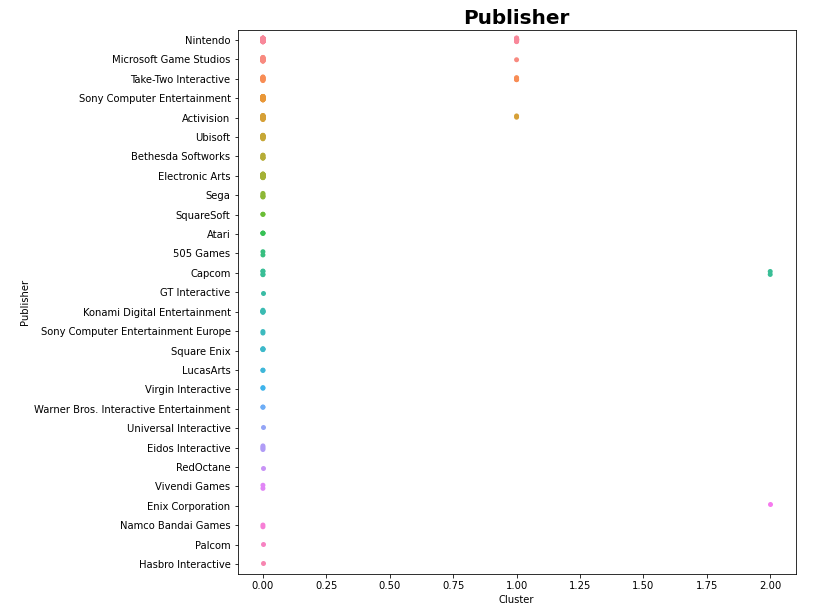
From looking at the scatterplots between the different regions and their clusters I learned a lot. By looking at the scatterplot for North American Sales I see that the data points are almost perfectly aligned with their clusters. This tell me that the North American market is the most important for videogames since they were the main determinant in deciding which cluster a value will belong to. This is in direct contrast to the scatterplot for Japan where each cluster is nearly identical to each other, with only 4 values in cluster 1 being higher than the others. I was surprised to see more points in cluster 2 being higher up than the points in cluster 1. This shows that many games that are popular in Japan are not as popular in other regions and games that sell very well in other regions may not sell well in Japan. This trend also follows in other regions as shown by the Other\_Sales plot.



This plot shows that out of all the genres, the only ones that did not make it into cluster 1 was Fighting, Adventure, and Strategy. This tells me that, since cluster 1 is the cluster with the highest videogame sales, games in these genre are hard to have massive success compared to other genres. This may be because many games in these genres have a dedicated and loyal fanbase who will keep returning to these types of games and so it is harder for new audiences to pick these types of games up. Another reason why is because games in these genres may be harder to market compared to other genres. For example, sports games and racing games will by nature sell more than strategy and adventure games since they appeal to a wider market. Another reason why fighting and strategy did not make it into cluster 1 is because not a lot of games in the dataset have that genre compared to others. However, I find it fascinating that the puzzle genre broke into cluster 1 even though it had the least amount of games in this dataset and I wouldn't think that genre of game would have such massive success.



This plot shows which clusters each platform is in. This tells me that Nintendo consistently has platforms with high videogame sales since the majority of the platforms in cluster 1 belong to Nintendo. Compared to Nintendo's 5 platforms, Sony had two platforms in cluster 1, the PS3 and PS2. Microsoft had only 1 platform in cluster 1, which was the Xbox 360. One thing that I find fascinating is the NES, the Gameboy, and the Super Nintendo (SNES), were all able to break into cluster 1 even though there were other platforms with hundreds of more games than them. For example, the NES and Gameboy only had 98 games on this list and SNES had 239 games, while platforms like the PSP had 1213 games and PS had 1196 games. I believe this shows the strategy for these kinds of platforms and their respective companies. While Nintendo focused on solely quality for its sales, Sony's strategy was to focus on putting out as much games as possible. Sony has continued this tactic as evidenced by the high amount of games on the PS2 and PS3 and were even able to break into cluster 1. However, Nintendo switched tactics and started to output many games as well, as seen by the DS and Wii and were able to get those systems into cluster 1.



This plot shows me that out of the many videogame publishers, only 4 have managed to get into cluster 1, which is the highest videogame sales in the dataset. This tells me that although there are hundreds of videogame publishers, only a handful of them have achieved massive success. This shows that there is a wide disparity between videogame publishers in terms of success, appeal, and quality and that when it comes to videogame publishers, there is a top 1%. One very interesting thing that I noticed is Electronic Arts has the most videogames on this list with a number of 1351, nearly 400 more than the publisher with the second most videogames, Activision. However, even though they have made so many videogames that sell well, they have never over the span of 40 years gotten into cluster 1 and achieved critical success compared to its competitors such as Nintendo or Activision. I feel like Electronic Arts could benefit from better quality control of its videogames, by not focusing on the quantity of games created, but the quality.

The Apriori results give a lot of information about how the nominal attributes of the dataset are connected to each other and shows the many different relationships they have. For example, I found it very interesting that all of the Gameboy puzzle games on this list of videogame sales were created by Nintendo since that association rule had a confidence of 1. I also found it interesting that the majority of Gameboy platformers on this list were created by Nintendo. These two association rules indicate to me that Nintendo did not have enough third party support for the Gameboy. This changes in later years as evidenced by the association rules with Nordic, Misc, and Wii as well as by MTV Games, Wii, and Misc. It made sense that SquareSoft and Square Enix both had multiple association rules with Role-Playing since SquareSoft primarily made Role-Playing games and later became Square Enix.

In this videogame dataset I used three data mining techniques that each brought its own unique perspective and insight into the data. With the decision tree classifier, I learned that out of the three nominal attributes, Genre, Publisher, and Platform, the one that had the highest accuracy on the test data was the Genre attribute. Since Genre was the best predictor, this told me that each publisher and platform had its own tendencies towards certain genres. This is further explained by running the Apriori algorithm on the dataset, where I found certain companies and systems were linked to specific genres. For example, on the Gameboy Nintendo focused on making platformers and puzzle games, while publishers like Electronic Arts focused on making sports games and Squaresoft and later Square Enix specialized in Role-Playing games. From these association rules, I learned that the vast majority of publishers focus on a specific genre and only a handful of companies, such as Nintendo and Microsoft, publish multiple genres. I liked the Apriori Data Mining technique because you get an in-depth look into the data and get to see all the different connections and relationships. I also used K-Means clustering which differed from the other two data mining techniques as this focused on the numerical data of videogame sales and so I was able to analyze the dataset and attributes from a different perspective. For example, by seeing which Publishers, Genres, and Platforms belonged to cluster 1, the top 1% of videogames sold on this dataset, I learned a lot about each attribute. For example, I learned that even though Electronic Arts have the most games on this dataset, they did not break into cluster 1 and thus focus more on just putting out games unlike Nintendo, who have almost half the amount of games as them but were able to break into cluster 1 several times. In conclusion, each data mining approach has its own unique method, gives its own perspective into the dataset, and works in unison with each other to give a clearer and whole picture of the dataset.

**REFERENCES**

[1] Ümit Yılmaz, Cengiz Gezer, Zafer Aydın, and V. Çağrı Güngör. 2021. Data Mining Techniques in Direct Marketing on Imbalanced Data using Tomek Link Combined with Random Under-sampling. 2021 the 5th International Conference on Information System and Data Mining. Association for Computing Machinery, New York, NY, USA, 67–73. DOI:<https://doi-org.jerome.stjohns.edu/10.1145/3471287.3471299>

[2] J. R. Quinlan. 1996. Learning decision tree classifiers. ACM Comput. Surv. 28, 1 (March 1996), 71–72. DOI:<https://doi-org.jerome.stjohns.edu/10.1145/234313.234346>

[3] Kai-Uwe Sattler and Oliver Dunemann. 2001. SQL database primitives for decision tree classifiers. In Proceedings of the tenth international conference on Information and knowledge management (CIKM '01). Association for Computing Machinery, New York, NY, USA, 379–386. DOI:<https://doi-org.jerome.stjohns.edu/10.1145/502585.502650>

[4] Dong Su, Jianneng Cao, Ninghui Li, Elisa Bertino, and Hongxia Jin. 2016. Differentially Private K-Means Clustering. In Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy (CODASPY '16). Association for Computing Machinery, New York, NY, USA, 26–37. DOI:<https://doi-org.jerome.stjohns.edu/10.1145/2857705.2857708>

[5] Sanjay Rathee, Manohar Kaul, and Arti Kashyap. 2015. R-Apriori: An Efficient Apriori based Algorithm on Spark. In Proceedings of the 8th Workshop on Ph.D. Workshop in Information and Knowledge Management (PIKM '15). Association for Computing Machinery, New York, NY, USA, 27–34. DOI:<https://doi-org.jerome.stjohns.edu/10.1145/2809890.2809893>

[6] Weber, Ben George, Michael Mateas and A. Jhala. “Using Data Mining to Model Player Experience.” (2011). <http://alumni.soe.ucsc.edu/~bweber/pubs/submission_3_epex_final.pdf>

[7] Odierna, Bruno Almeida and Ismar Frango Silveira. “PLAYER GAME DATA MINING FOR PLAYER CLASSIFICATION.” A Produção do Conhecimento na Engenharia da Computação (2019): n. pag. <http://alumni.soe.ucsc.edu/~bweber/pubs/submission_3_epex_final.pdf>

[8] Odierna B.A., Silveira I.F. (2020) MMORPG Player Classification Using Game Data Mining and K-means. In: Arai K., Bhatia R. (eds) Advances in Information and Communication. FICC 2019. Lecture Notes in Networks and Systems, vol 69. Springer, Cham. <https://www.researchgate.net/publication/330832073_MMORPG_player_classification_using_game_data_mining_and_K-means>

[9] L. F. Bicalho, A. Baffa and B. Feijó, "A Game Analytics Model to Identify Player Profiles in Singleplayer Games," 2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames), 2019, pp. 11-20, doi: <https://www.researchgate.net/publication/340572670_A_Game_Analytics_Model_to_Identify_Player_Profiles_in_Singleplayer_Games10.1109/SBGames.2019.00013>.

[10] Shasha Xu and Dong Huang. 2019. Ensemble-Initialized k-Means Clustering. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing (ICMLC '19). Association for Computing Machinery, New York, NY, USA, 65–69. DOI:<https://doi-org.jerome.stjohns.edu/10.1145/3318299.3318308>