Team 10 Final Report - Predicting features of Blockbuster Games.

Background

A video game is an electronic game that can be played on a computing device [1-2]. Under the premise of playing in moderation, video gaming can have many beneficial effects, including improving physical fitness, social behavior, and cognition and brain functions [2-5]. The global video game market size was valued at USD 151.06 billion in 2019 and is expected to grow at a compound annual growth rate of 12.9% from 2020 to 2027[1, 6, 7].

Motivation

Game companies around the world are interested in retaining satisfying profit. Thus, game companies are extremely careful when making new products because success of new products can be directly related to the survival of business organizations [6, 7, 8, 19] owing to the high budget associated with producing high-quality games and the growing size of development team. Besides that, it is never an easy task to estimate the chances of video game success due to the increasing dynamic market [2, 5, 6, 7]. Our project targets benefiting game companies in increasing their profits by leveraging machine learning techniques on the vast amount of data available.

Related work

Using data-driven approaches to perform prediction in game industry has come to existence for quite a while. Unsupervised machine learning algorithms such as K-means clustering and Principal Component Analysis (PCA) have been used for segmenting user types [8, 9, 13]. Supervised learning techniques have also been widely used for predicting game ratings and sales, in which regression [14, 24] and neural networks [18] are the most popular models and have decent prediction accuracy. Usually, researchers use numerical and categorical values such as game genre and platform as labels for training models [14, 18]. However, some pioneers have brought textual data into play. For instance, [7] has used game description data to create machine-learning-based game genres and [20, 23] has performed sentiment analysis based on Gaussian Naive Bayes and Decision tree algorithms by using review data. This has opened a new door to data-driven research in game industry since textual data such as game storylines and user reviews are much more abundant compared to numerical/categorical data. Machine learning has also been applied in the industry for other purposes such as predicting game currency value and recommending in-game purchase items [26, 27]. Researchers have also made various visualizations regarding game data [15] and concluded that visualization techniques such as charts, heatmaps, movement visualizations, self-organizing maps and node-link representations are useful for answering specific data analytical questions.

Based on the literature review, it is identified that researchers in the game industry have been conducting research based on numerical/categorical data or textual data independently, while the two can be combined to provide more comprehensive results. Besides, most game data visualizations that have been done so far are static plots, although an interactive visualization will provide the users with much more details and better experience.

Project Objective

Our goal is to use machine learning models to predict new games' success rate in the current market. We will present our results in an interactive interface through visualization. To officially define the objective of this project, we have proposed a methodology that utilizes the historical data available at main-stream game databases to establish various machine learning models, which are used to estimate the success rate of new products and design new games that can achieve high investment returns in the current market. We also analyze user reviews from textual data and provide visualization of user feedback to gain more insights on popular game features. An interactive user interface has been created to give audience more convenience to access the research outputs.

Methodology

Overview

The methodology built in this project can be adopted by any video game company to provide advice for new product designs from a data-driven point of view. Using VGChartz[8-11] and Steam API Datasets [12, 13], our approach aims to combine numeric and textual data to estimate potential blockbuster games [18] and analyze the user sentiments of these games. Our AI models utilize regression techniques like decision trees, random forest, multiple linear regression to find the best model for predicting sales. Additionally, sentiment analysis of user reviews will be performed for opinion mining [12, 13]. Finally, an interactive user interface would visualize detailed insights on the factors contributing to the potential blockbuster game in addition to user's opinion on that game genre. The idea is to apply explainable AI to help users explore the game sales data, understand weights of each input data feature in the machine learning model in predicting the sales, and interactively showing key features contributing to success and user opinion using word cloud. To overcome the limits of current practices which are restricted to analyzing specific features, we aim to combine all key features like Genre, Rating, Game categories,

Pricing, play time, and user sentiment in determining game success [16]. Our interactive interface will guide users on the criteria to create and market the next blockbuster game [14-15].

Data Preparation

Source: Using VGChartz[8-11] and Steam API Datasets [12, 13], our approach aims to combine numeric and textual data to estimate potential blockbuster video games [18] and analyze user sentiments on these games. We found a dataset on Kaggle that contains playtime data [28]. The dataset is called "steam-200k.csv" and contains 200,000 rows, each representing a single record of playtime of a player (in the form of the player's ID) for a specific game. To obtain the "regional playtime" for the games, the team decoded the player ID into the Steam ID and used the Steam API to get country information of the players, and then merged countries of the same region into individual regions, namely North America (NA), Europe (EU), Japan (JP), and Other (Other). Thus, the original steam-200k.csv has been converted into regional_playtime.csv, with each entry representing the playtime of a game in a certain region. We have combined this file with the file that contains the basic information about steam games, we can perform supervised machine learning to predict regional playtime with respect to game features such as genre, rating, platform, and price. For sentiment analysis, we downloaded data from Steam API and prepared a data set with over 19 columns and 996 games. Key attributes that were used were Steam AppID, game name, genre, playtime, reviews ("steam_data_final.csv"). The review dataset obtained from Kaggle "steam_reviews.csv" was a 7GB file which had 21 million user reviews for over 300 games in multiple languages. The two datasets were cleaned and merged to aggregate and generate insights of user reviews at the genre level.

Machine Learning for Prediction

Predicting regional and global sales

For this project we implement Gaussian Linear Model to predict game sales from game releases duration, ratings, publisher, genre, and platforms. Our Intuition on trying to predict regional and global sales using Gaussian Linear Model (GLM) compared to state-of-the-art regression model like kernel ridge and lasso regression is because we want to leverage the generative regression ability of GLM to predict sales. Both Gaussian Linear model and kernel/lasso regression employ a kernel trick to process the multidimension data and predict the next point regression. However, GLM can choose the kernel's hyperparameters based on gradient-ascent on the marginal likelihood function. In normal kernel ridge we need to find our own parameter by experiment or by performing a grid search on a cross-validated loss function. Another difference is that GLM can learns a generative, probabilistic model of the target function. This means while kernel ridge is mostly only providing predictions.

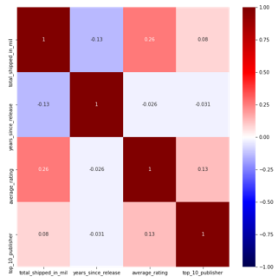
Our project to predict sales is based on the VGChartz[8-11]. The dataset contains 55,792 game sales data. In order to predict the global sales, we tried to use the relation of ratings, Platform, Genre, years from release, and top publisher to predict global sales. In order to get this relation, we first run data preparation using assumption on how to calculate and define features. The ratings are an average from VGChartz ratings, User Rating, and Critics Rating. Our assumption is higher rating means the game is more successful. The year from release is counted from the year game is published until the time we got on 2020. Our assumption is the longer the game is from year of release, the more popular the game since it manages to survive for a long time in the market. The top publisher is a binary column that shows whether the game is from top ten publisher or not. Our assumption is top publisher has more money to advertise game and give it higher chance to become blockbuster. Genre and Platform is used to be a learning parameter to understand a characteristic of blockbuster game. The sales will be represented in number of copies sold in thousands. Our assumption is a blockbuster game has be popular and the number of copies getting sold and shipped is more important than the profit it generates. The dataset also doesn't contain real game profit. After this preparation we find first try to visualize and learn some covariance from our datasets. We run the regression model on each specific region dataset based on four regions, i.e. North America, Europe, Japan, and Other. The results can be seen in **Figure 1**.

This correlation features offer us an insight on how each regional market is different. For example, in North America year's since release has negative correlation to sales, while in Japan this features is the highest correlation to sales. We infer that this might be due to the fact in the Japan Market the game with longer year since release has bigger fanbase. The fanbase in Japan is loyal to the title of the game and that's why the year since release correlate well with sales. Another interesting observation in Japan market we found is years since release has positive correlation to average rating, while in North America, Europe, and Other market it has negative or neutral values. This reinforces our inference of Japan Market loyalty of certain title is strong.

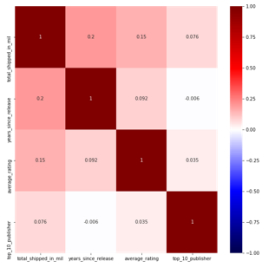
We observe in North America market the year since release feature is the lowest while the feature with highest correlation is average rating. We infer that this might be due to the fact in the North America Market the customer prefer to have a new highly rated game, compared to staying loyal to certain game title. Another interesting observation in North America, Europe,

and Other markets is that “top 10 publisher” feature has the most positive correlation to average rating, while in Japan it has neutral values. This means marketing from top 10 publisher to make a game popular is more effective outside of Japan to increase the possibility of getting a higher average rating.
Just looking at the correlation market we found that the characteristic of sales on Japan Market is very different with the rest of the world’s market. Games coming from a top publisher an get more exposure and marketing doesn’t correlate that well in rating compared to other’s market, while older game that has a long release year has a high correlation with number of sales. On the opposite, other world’s market seems to value ingenuity and new games that published from top publisher.

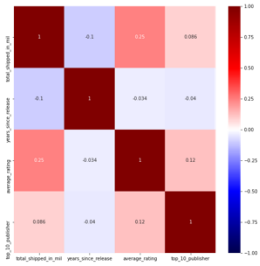
North America Market Feature Correlation to Sales



Japan Market Feature Correlation to Sales



Europe Market Feature Correlation to Sales



Other Market Feature Correlation to Sales

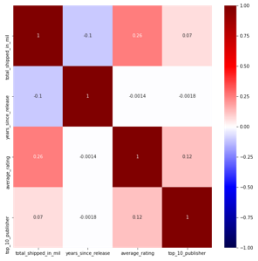


Figure 1: Market Features correlation to sales

Predicting regional playtime
Based on the literature survey, the team has found that researchers in the game industry have mainly focused on the prediction of sales. However, we consider playtime as another significant indicator of a game’s success, since a long playtime means that players tend to spend more time on the game and will consequently tend to engage with more in-game purchases and to purchase the sequels of the game. The team will perform a playtime prediction in different regions based on several game features, including genre, rating, platform, and price. This can be an improvement that we make with respect to the state-of-the-art research.
The steam-200k dataset (pre-processed as described earlier) will be used for this prediction. As mentioned earlier, the dataset has been processed and the regional playtime for each game has been obtained. Through basic analysis, we have identified that game genre, rating, platform, and price would be the most significant factors regarding the playtime of games. As a result, we extract these four metrics in the dataset as the features for the machine learning models. Specifically, the three categorical metrics, i.e. genre, rating, and platform, will be pre-processed through one-hot encoding to be used in the machine learning models. To predict games’ regional playtime, we will try various regression models, compare their performance, and select the optimal one for our prediction in the end. Since we will use categorical metrics in this prediction, we select decision tree, random forest, and XGBoost as candidates for the supervised machine learning models. For the random forest model, we will use grid search method to determine the optimal hyperparameter setting for it. When evaluating the model performance, we will use R2 score as the main criterion. Last but not the least, we will have four models (of the same kind) four for predicting the four regions, i.e. North America, Europe, Japan, and Other, respectively. In other words, we will start with a model for predicting the game playtime in North America, and use similar approaches to build three models for predicting playtime in the other regions.

Sentiment analysis

A dataset containing 21 million user reviews of around 300 different games from Steam was used for the sentiment analysis of user reviews on games and a visual representation of sentiments using a word cloud. Our goal was to analyze user sentiments by grouping games based on their genres and interpret the results. To accomplish this, we collected steam games and genre related information from Steam API and processed the data in the format we required. This dataset includes information on Steam AppID (unique identifier for every game in steam), game name, genre, player count, price, languages etc. Additionally, we sourced a 2021 Kaggle dataset that contained consolidated user review data for every game. The dataset contained other information such as Steam App ID, Game Name, Review Language, Review Time Stamp, Recommended Game, etc. For our project, we used only English reviews from the dataset. After pre-processing, cleaning, removing null values and duplicates, we had a dataset with around 10 million English reviews. For each of the reviews, a

sentiment polarity was produced and based on the polarity, each user review was labeled as positive, neutral and negative. **Figure 2** shows review counts for games that were analyzed and counts per genre, while **Figure 3** shows the sentiment polarity per genre.

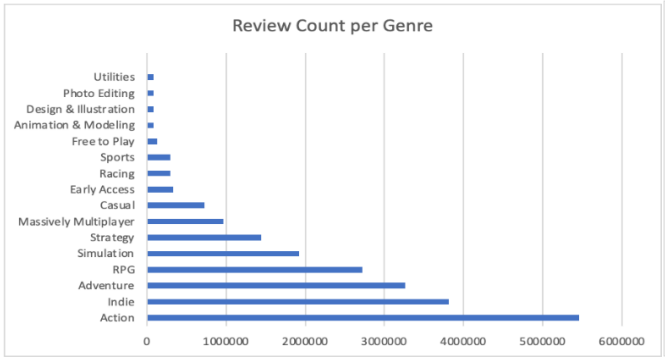


Figure 2: Review count per Game

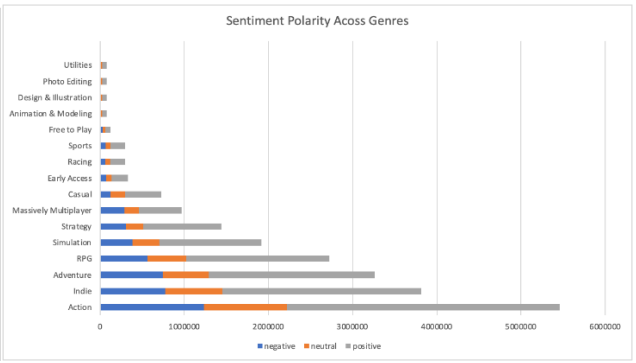


Figure 3: Sentiment Polarity per Genre

Using the sentiment polarity label that was generated, we analyzed text reviews for both positive and negative sentiments. After removing stop words, and performing stemming, word clouds were generated from positive and negative user reviews. The word clouds will act as visual representations of game elements that users like or don't like about the game. This will provide vital information and feedback for developing specific game features in each genre of game development. From the charts, we see that "Action" seems to be a popular game genre.

Figure 4 is an example of a word cloud with positive user sentiments on games in the "Racing" genre. From the word cloud, we can interpret positive reviews and some insights on specific games and features that will be very useful. For instance, words like "competitive" and "friend" indicate that there is a demand for multi-player games that are more preferred in this genre.

Figure 4: Word cloud for Positive sentiments for Racing Genre

On analyzing negative sentiments for games, we were not able to make out specific characteristics of the game that caused the dislike. As a result, we have decided to visualize only positive user sentiments which is more insightful. Further, results of our analysis will be incorporated in the visualization tool, that will enable users to see these word clouds for every genre and understand trends and user sentiments in addition to the machine learning model predictions. This will enable game producers to analyze both results in developing future game features.

Visualization

The visualization system consists of two main parts, a data input panel on the left and a charting area on the right. Above the two is a dropdown that allows users to toggle between three different visualizations. The three visualizations include: a prediction of game sales over the years since release, a prediction of game play time for each market region and a positive sentimental analysis to inform game analyzers about positive insights depending on game genres. Each visualization has a different set of parameters for users to select via either a dropdown (for categorical data) or a slider (for continuous data). **Figure 5** shows an example of the prediction of game play time visualization.

For optimal performance, machine learning models are pretrained and word clouds are generated for each genre. The visualization system uses JavaScript “fetch” function to pull data from a Flask backend via RESTful calls. There are in total 2 GET APIs, one for each of the two machine learning models. There is one machine learning model for each market region. Upon user input in the visualization parameter panel, a GET call with the set of selected parameters of the UI components are passed to the corresponding API for that machine learning model. The Flask backend then calls a python module that loads the machine learning model using Pickle model manager and predicts an output with the parameters passed. The prediction result is passed back to the JavaScript to generate visualization in real time. For the word clouds, a word cloud in the form of a jpg file is generated for each genre. The JavaScript pulls the image from its static folder via a local relative path. Despite having simple chart types, the strength of the visualization lies in its integrated pipeline with machine learning models from the backend. With a fast model training backend and possibly real time online learning, we can easily scale to make the pipeline truly real time by updating the model files whenever needed.

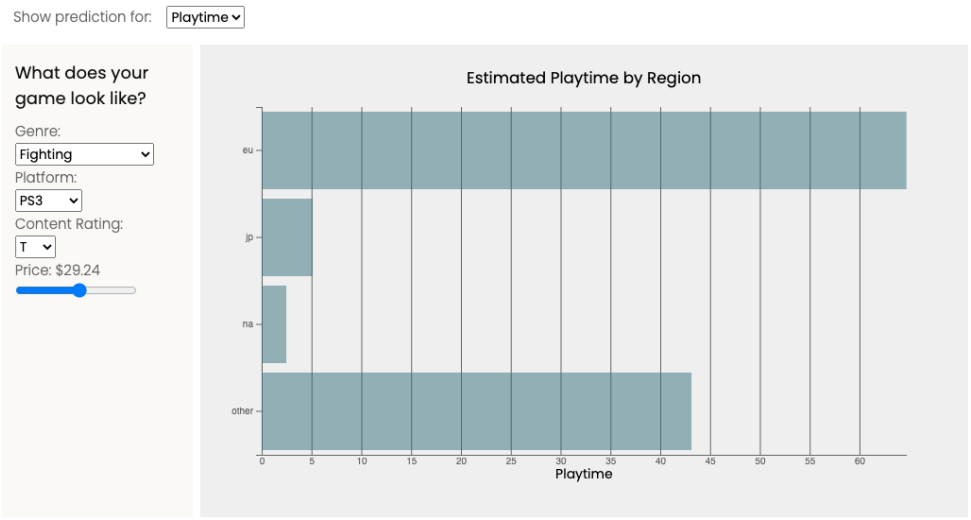


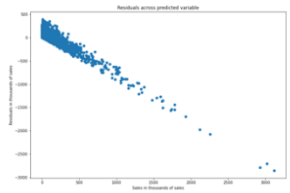
Figure 5: Visualization of Playtime Prediction

Experiments and Results

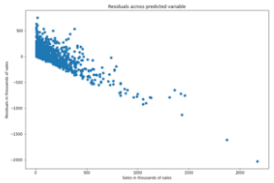
Machine Learning for Prediction

Predicting regional and global sales

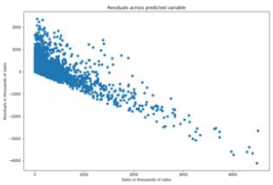
The data we got from preprocessing is then get trained inside or GLM models. We used Scaler to make sure that each feature has a fair contribution. From our early experiment we use Gaussian Distribution to predict the model. However, we got a very high RMSE of 1.35 million sales out of the scale of 3-4 million. We then change our model to use a Poisson Distribution to have a better result. We assume that the way a game is sold follows a Poisson distribution better. We assume most game will get sold about when the game is released, then slowly get more sales as it goes along. The results for our Machine Learning model can be seen in **Figure 2**. This figure shows the error residual of our model prediction with actual dataset sales. It’s a regression model visualization to show how our model predict the sales. On average we do see an error of plus minus 500 thousand sales (RMSE actual value is 585.06) from the total of almost 3-4 million scale. Since regression model doesn’t have accuracy the way we think this model has worked well if as many points lies in the y axis of delta error within 500 thousand range. From the model visualization we can see that most point indeed lies in this range. However, we do see a long tail for some outlier points, especially in the Europe Market.



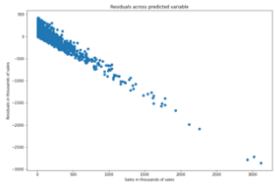
North America Market Prediction to Sales



Japan Market Prediction to Sales



Europe Market Prediction to Sales



Other Market Prediction to Sales

Figure 6: Final Model Residual to Sales

Predicting regional playtime

Through our experiments, we have identified that the random forest model outperforming the other two models. The best number of trees for it to have is found to be 128. With this number of trees, the R2 value of random forest is larger than both the decision tree's and the XGBoost's. Consequently, we will use the random forest model with 128 trees as our final regression model for predicting regional playtime of games. One thing worth mentioning here is, even the best performing model, random forest, does not have a very high R2 score in the end. We consider the reason being that the playtime dataset we use has a limited size. As mentioned earlier, although we started off with about 100,000 rows of playtime records, most of the players in this dataset have chosen to hide their country/region information, so we ended up with only less than 10,000 rows of playtime records. The limitation of the data size may be the primary reason for our regression model not having an excellent performance. However, given the size of the training and testing data, we consider the model still has a decent performance in predicting regional playtime, which means the methodology we built is still valid. In the real world, if provided with more playtime data, this model will have much better predicting performance.

After building the prediction model and the visualization tool, the team performed some experiments and have found some interesting results which are summarized below:

First, the game playtime does not necessarily have a negative linear relationship with its price. For instance, it is observed that for action games on PC platform with content rating as E, the playtime in North America and Other Places is high when the price is low and the playtime would decrease as the price gets higher. However, it is opposite for European and Japanese players. They tend to spend less time on the action games with lower prices and more time on those with higher prices. The observation indicates that the players in Europe and Japan tend to spend more time on expensive action games, and this is likely because that an action game with a higher price can have better designs, details, and graphics, which are attractive to the players at these regions.

The other observation is that players in different regions may prefer different platforms. For example, for sports games on the PC platform, it is found that players in Europe would spend the longest time. Meanwhile, for sports games on the X360 platform, it can be seen that players in North America spend significantly longer time than those from other regions. This can be explained by the fact that X360 is a product of Microsoft, which is based in the US, which would lead to its high popularity in North America.

Conclusion

Our project proposed a way to provide advice for new product designs from a data-driven point of view. Using VGChartz and Steam API Datasets, our approach combines numeric and textual data to estimate potential blockbuster games and to analyze user sentiments on these games. Our modelling uses multiple linear regression to find the best model for predicting sales and tree method to find the prediction of playtime. We also perform sentiment analysis of user reviews for opinion mining. Finally, an interactive user interface would visualize detailed insights on the factors contributing to the potential blockbuster game in addition to user's opinion on that game genre.

Our project aims to have explainable AI to help users explore the game sales data, understand weights of each input data feature in the machine learning model in predicting the sales, and interactively showing key features contributing to success and user opinion using word cloud. Current practices are mostly restricted to analyzing specific features and we aim to combine all key features like Genre, Rating, Game categories, Pricing, play time, and user sentiment in determining game success. Our interactive interface would hopefully help to guide users on how to create and market the next blockbuster game.

Distribution of Effort

Every team member has equal contribution in the distribution of activities.

Reference

1. Statista Research Department, “Video Game Industry - Statistics & Facts,” Statista, Jan 18, 2021.
2. Grand View Research Group, “Video Game Market Size, Share & Trends Analysis Report By Device (Console, Mobile, Computer), By Type (Online, Offline), By Region, And Segment Forecasts, 2020 - 2027,” Grand View Research, May, 2020.
3. McDougall J., Duncan M.J. Children, “video games and physical activity: An exploratory study,” Int. J. Disabil. Hum. Dev., 2008, doi: 10.1515/IJDHD.2008.7.1.89.
4. Cole H, Griffiths MD, “Social interactions in massively multiplayer online role-playing gamers,” Cyberpsychol Behav, Aug, 2007.
5. Toril P, Reales JM, Ballesteros S, “Video game training enhances cognition of older adults: a meta-analytic study,” Psychol Aging, Sep, 2014.
6. Cabras, I., Goumagias, N., Fernandes, K., et al. “Exploring survival rates of companies in the UK video-games industry: An empirical study,” Technological Forecasting & Social Change, Oct 3, 2016, doi: https://doi.org/10.1016/j.techfore.2016.10.073.
7. Prasad A., “Estimating Video Game Success using Machine Learning,” School of Computing National College of Ireland, Aug, 2019, doi: 10.13140/RG.2.2.14389.01767.
8. Aziz, Amar & Ismail, Shuhaida & Othman, Muhammad & Mustapha, Aida. (2018). Empirical Analysis on Sales of Video Games: A Data Mining Approach. Journal of Physics: Conference Series. 1049. 012086. 10.1088/1742-6596/1049/1/012086.
9. Julie Marcous and Sid-Ahmed Selouani, “A hybrid subspace-connectionist data mining approach for sales forecasting in video game sales industry”, 2008, 978-0-7695-3507-4/08, IEEE.
10. KEERTHANA , BODDURU, and Rao, K.VENKATA, SALES PREDICTION ON VIDEO GAMES USING MACHINE LEARNING,2019, JETIR June 2019, Volume 6, Issue 6 ISSN-2349-5162.
11. TM Geethanjali, Ranjan D,Swaraj HY, ThejaskumarMV,Chandana HP, Video Games Sales Analysis: A Data Science Approach, 2020,IJCRT,Volume8,Issue 5 May 2020, ISSN:2320-2882.
12. Toy, E. J., Kummaragunta, J. V., & Yoo, J. S. (2018). Large-scale cross-country analysis of steam popularity. 2018 International Conference on Computational Science and Computational Intelligence (CSCI). doi:10.1109/csci46756.2018.00205.
13. Zuo, Zhen, Sentiment Analysis of Steam Review Datasets using Naive Bayes and Decision Tree Classifier, https://core.ac.uk/download/pdf/159108993.pdf, March 2021.
14. Dheandhanoo, T., Theppaitoon, S., & Setthawong, P. (2016). Game play analytics to measure the effect of marketing on mobile free-to-play games. 2016 2nd International Conference on Science in Information Technology (ICSITech). doi:10.1109/icsitech.2016.7852620.
15. Wallner, G., & Kriglstein, S. (2013). Visualization-based analysis of gameplay data – a review of literature. Entertainment Computing, 4(3), 143-155. doi:10.1016/j.entcom.2013.02.002.
16. Gil, R., & Warzynski, F. (2015). Vertical integration, exclusivity, and game sales performance in the US video game industry. The journal of law, economics, and organization, 31(suppl_1), i143-i168.
17. Faisal A., Peltoniemi M., “Establishing Video Game Genres Using Data-Driven Modeling and Product Databases,” Sage Journals, 2015, doi: https://doi.org/10.1177/1555412015601541.
18. Cox, J. (2014). What makes a blockbuster video game? An empirical analysis of US sales data. Managerial and decision economics, 35(3), 189-198.
19. Marchand, André, and Thorsten Hennig-Thurau. "Value creation in the video game industry: Industry economics, consumer benefits, and research opportunities." Journal of interactive marketing 27.3 (2013): 141-157.
20. Zhen Zuo. “Sentiment Analysis of Steam Review Datasets using Naive Bayes and Decision Tree Classifier” (2018).
21. Clarke, Rachel Ivy, Jin Ha Lee, and Neils Clark. "Why video game genres fail: A classificatory analysis." Games and Culture 12.5 (2017): 445-465.
22. Strååt, Björn, and Harko Verhagen. "Using User Created Game Reviews for Sentiment Analysis: A Method for Researching User Attitudes." GHITALY@ CHItaly. 2017.
23. S. Chakraborty, I. Mobin, A. Roy and M. H. Khan, "Rating Generation of Video Games using Sentiment Analysis and Contextual Polarity from Microblog," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 157-161, doi: 10.1109/CTEMS.2018.8769149.

24. A. Krishna, A. V, A. Aich and C. Hegde, "Sales-forecasting of Retail Stores using Machine Learning Techniques," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2018, pp. 160-166, doi: 10.1109/CSITSS.2018.8768765.
25. Paul Bertens, Anna Guitart, and Pei Pei Chen, “A Machine-Learning Item Recommendation System for Video Games,” 2018 IEEE Conference on Computational Intelligence and Games (CIG), Oct. 15, 2018, DOI: 10.1109/CIG.2018.8490456
26. Young Bin Kim, Kyeongpil Kang, Jaegul Choo, Shin Jin Kang, TaeHyeong Kim, JaeHo Im, Jong-Hyun Kim, Chang Hun Kim, "Predicting the Currency Market in Online Gaming via Lexicon-Based Analysis on Its Online Forum", Complexity, vol. 2017, Article ID 4152705, 10 pages, 2017. https://doi.org/10.1155/2017/4152705
27. Paul Bertens, Anna Guitart, and Pei Pei Chen, “A Machine-Learning Item Recommendation System for Video Games,” 2018 IEEE Conference on Computational Intelligence and Games (CIG), Oct. 15, 2018, DOI: 10.1109/CIG.2018.8490456
28. Kaggle – Steam Video Game Dataset, https://www.kaggle.com/tamber/steam-video-games, Accessed Date: 3 March 2021
29. Juho Hamari, Kati Alha, Simo Jarvela et al, “Why do players buy in-game content? An empirical study on concrete purchase motivations,” Computers in Human Behaviour, vol. 68, pages 538-546, March 2017.
30. Olli I. Heimo, J. Tuomas Harviainen, Kai K. Kimppa & Tuomas Makila, “Virtual to Virtuous Money: A Virtue Ethics Perspective on Video Game Business Logic,” Journal of Business Ethics, 153, 95-103, 10 December 2016.
31. Xudie Ren, Haonan Guo, Shenghong Li et al, “A Novel Image Classification Method with CNN-XGBoost Model,” IWDW 2017: Digital Forensics and Watermarking pp 378-390, 26 July 2017.
32. Tianqi Chen, Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System,” KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA, August 2016.
33. Rory, Mitchell, Eibe Frank, “Accelerating the XGBoost algorithm using GPU computing,” PeerJ Computer Science, doi.org/10.7717/peerj-cs.127, 24 July 2017.