Predicting Gaming Industry Success with Machine Learning

Nicholas Katrantzidis

500655049

Ryerson University

Table of Contents

1. Revised Abstract

The rapidly growing gaming industry has resulted in thousands of new games, evolving hardware, new genres and much more. A PwC thought leadership piece confirms this industry growth in its latest report stating, “The gaming industry is tipped to maintaining its recent rapid growth, and could be worth 321 billion by 2026” [1]. With new games created each day, many gaming enthusiasts can be overloaded with information and find it difficult to understand which factors play a role in future purchasing decisions, which games are most popular by region, and the age-old question of which console is the best. This report, attempts to provide clarity for the below questions using the Jupyter Notebook Python programming language tool and classification and regression machine learning techniques.

Revised Research Questions

1. Given the full enriched data set and based on related research, which regression algorithm best predicts Global Sales when comparing Decision Tree vs. Random Forest vs. Linear Regression?
2. Does critic-score data prove to be a better indicator for video game success compared to user-score data? How do critic enriched data sets compare to a user enriched data set when using regression algorithms to predict Global Sales?
3. Can we effectively predict the video game platform (aka console) based on the enriched data set? Which attributes in the enriched data set have stronger correlations to predicting gaming platforms?

Research question 1 will compare various regression model techniques, specifically decision tree, random forest, and linear regression and their respective evaluation metrics such as R2 to determine the best model for predicting global sales. Question 2 will compare regression the evaluation metrics of the best regression model (determined in question 1) with 3 types of data sets. The first data set will be baseline with no user or critic data, the second data set will be with a user enriched data set and the third data set will be with critic enriched data set. All 3 regression models will be compared using regression evaluation metrics to determine the impact of user-data vs. critic data. In question 3, this report will use classification machine learning techniques to understand which attributes best predict the platform (aka console) the game is played on. Question 3 will use Scikit learn library and One-Hot encoding on each data set to ensure accurate implementation of classification algorithms. Pandas library will also be used for cleaning the data and preparing the data set appropriately for machine learning. In conclusion, this report will use knowledge gained from CIND820 and related research to answer the proposed research questions. Ultimately the reader should gain a better understanding of how to predict video game sales, the impact of critic vs user scores on video game sales, and how classification can be used to predict video game consoles.

The source of Data used for the proposed research is Kaggle and Meta Critic, which are open source platforms.

2. Literature Review

2.1. Introduction

The gaming industry has grown exponentially leading to a variety of different publishers, genres, platforms, developers, ratings etc. Today we have enough data in the gaming industry to use machine learning techniques and create various models that predict success in terms of sales and ratings.

With most of the gaming industry data openly available to the public today, many analysts have explored the idea of predicting success in gaming. As a supplement to this research, I've outlined 5 research reports that have inspired and informed this paper.

2.2. Comparable Works

Research Paper 1: Predictive Analysis on Commercial Success of Game [2]

In this research paper, the author sets out to analyze sales data from released video games to identify industry trends and develops a prediction model to forecast the probability of a game being successful based on Global Sales. The data set used in this research report is from a variety of sources including Kaggle, Meta Critics, and Gamespot. The analysis used was Random Forest Classifier and Logistic regression, which required the author to use One Hot Encoding method to convert string values to numerical values. The finding in this report suggest that certain genres, ESRB ratings, and publishers are more likely to result in successful games. In addition, the author found critic scores to be strongly correlated with

global sales and found the LR model provided higher accuracy and less loss compared to the RFC model.

Research Paper 2: Predicting Video Game Sales Using an Analysis of Internet Message Board Discussions [3]

In this research paper the author of "Predicting Video Game Sales Using an Analysis of Internet Message Board Discussions" explores the use of NLP and analysis of internet message board discussions to predict video game sales. The methodology used in this report is to generate a weekly corpora by downloading and processing text from the video game community on the internet and uses support vector regression to create a model that is able to predict future sales figures of video games. The findings of the report suggest that in order to create a better prediction of both old and new video games, discussion data is needed on top of sales data. In addition, the online discussion boards consist of a specific niche audience, making the model predictions limited to that audience. Therefore, the online discussion data and model does not represent the broader video game population.

Research Paper 3: Estimating Video Game Success Using Machine Learning [4]

This research does not present any definitive conclusion; however, it lays the groundwork and various approaches to predicting video game success based on descriptive features. The author intends to use Word2vec NLP methods as well as Support Vector Machine, Artificial

Neural Network, K-Nearest Neighbor, and Random Forest algorithms. In addition, these methods will be evaluated and compared by calculating accuracy, precision, and F-score.

Research Paper 4: Machine Learning for Predicting Success of Video Games [5]

This report examines the prediction of video game success. The author creates a database of PC platform games and visually presents the collected data. Finally, machine learning methods such as Support Vector Machine, Artificial Neural Network, K-Nearest Neighbor, Random Forest are employed in experiments to determine the extent to which the post-release success of PC games can be predicted prior to their release. The research revealed a strong correlation between core features known before release and the average number of concurrent players in the first two months after release. Experience as a developer or publisher was found to be the most influential factor in prediction accuracy.

Research Paper 5: Predicting Steam Games Rating with Regression [6]

In this study, the author focuses on predicting game ratings using regression models. To feed the models, data is collected from Steam, Meta Score, User reviews, and more. The findings reveal that tree-based regression algorithms perform better compared to other regression methods. This is shown through the evaluation section where the Random Forest outperformed other methods with an R2 score over 0.9. To enhance the study, the author suggests gathering additional information related to games, such as total playtime, difficulty,

and art style. Additionally, incorporating text classification could further improve the research by encoding variables like game name, summary, and user reviews.

2.3. Literature Review Conclusion and Key Findings

I believe this report will further extrapolate on the related research cited. For full transparency, outlined here are key findings for each report and how they impact this reports research questions.

1. **Research Paper 1**: Using One-Hot encoding to convert categorical values to numerical values. This method will be used to ensure higher accuracy for the classification model needed for research question 3.

2. **Research Paper 2:** Although message board data can improve accuracy for a niche sample, it does not do a good job of representing the broader gaming audience. Therefore, message board data will not be used in answering this reports research questions.

3. **Research Paper 3:** The research sets the groundwork for how to deploy various machine learning techniques and which metrics can be used to evaluate and compare them. Accuracy, precision and f-score will be used to evaluate the classification models in research question 3.

4. **Research Paper 4:** This research shows that average number of concurrent players predicts post-release success. Similarly, in this reports data set we have the "User_Count" attribute. For research question 3, this report will attempt to further extrapolate by confirming if User data has a greater impact than critic data.

5. **Research Paper 5:** Based on this analysis decision trees produce higher R2 scores and therefore more accurate results. This report attempts replicate this analysis with the enriched data set and confirm or deny this conclusion.

2.4. Data Description and Data Cleaning

Table 1: Pre-Data Cleaning Snapshot

Attribute	Description	Data Type	Non-Null Values	Missing Values
Name	Name of the game	object	16717	2
Platform	Console on which the game is running	object	16719	0
Year_of_Release	Year the game was released	float	16450	269
Genre	Game's Category	object	16717	2
Publisher	Publisher	object	16665	54
NA_Sales	Game sales in North America (in millions of units)	float	16719	0
EU_Sales	Game sales in European Union (in millions of units)	float	16719	0
JP_Sales	Game sales in Japan (in millions of units)	float	16719	0

Other_Sales	Game sales in the rest of the world (Africa, Asia excluding Japan etc.)	float	16719	0
Global_Sales	Total sales in the world (in millions of units)	float	16719	0
Critic_Score	Aggregate score compiled by Metacritic staff	float	8137	8582
Critic_Count	Number of users who gave the critic_score	float	8137	8582
User_Score	Score by Metacritic's subscribers	object	10015	6704
User_Count	Number of users who gave the user_score	float	7590	9129
Developer	Party responsible for creating the game	object	10096	6623
Rating	The ESRB Ratings (E = Everyone, T = Teen, A = Adults etc.)	object	9950	6769

Key Findings Addressed for Data Cleaning

The following code and pre-processing steps have been replicated for this research study as they adequately meet the needs required to run machine learning algorithms outlined in this report. The Kaggle user referenced is “Phoenixkit” and their code was made public on Kaggle here: https://www.kaggle.com/code/phoenixkit/lab1-preprocess

The major exception in this research study is not dropping the Publisher, Developer, and Rating attributes. These attributes could impact the classification model needed to answer research question 3.

1. Remove attributes that will not be needed based on research questions
    - All attributes have been identified as relevant for at least 1 research question, therefore non are removed

2. Remove Key Attributes with Missing Values
    - 2 records missing Name values should be removed
    - 269 records missing Year_of_Release values should be removed
    - 54 records missing Publisher values should be removed

3. Review Duplicate Values
    - The Name attribute has 5019 duplicate values. Considering the same games can be released on multiple platforms we will also need to check which records have the same Game Name and Platform and then remove a duplicate
        - Data shows 6 duplicate records with same name and platform
        - Data manually reviewed and duplicate record removed for each pair

4. Clean and standardize Critic_Score, Critic_Count, User_Score, User_Count
    - Create an array of only unique values in the User and Critic score columns
    - Convert all “tbd” values to null in both columns and define as type float
    - Standardize Critic Score and User Score values
        - Drop records with null user and critic scores
        - Multiply User score values by 10 to ensure same scale as critic scores
        - Use critic score values for missing user score values

- Use user score values for missing critic score values

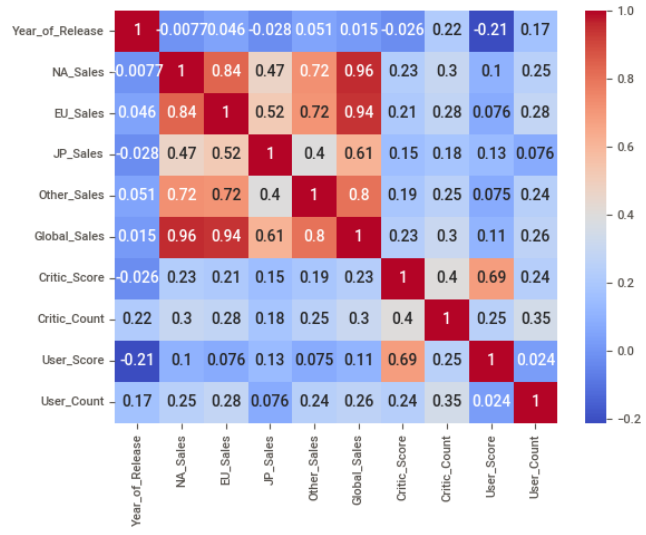
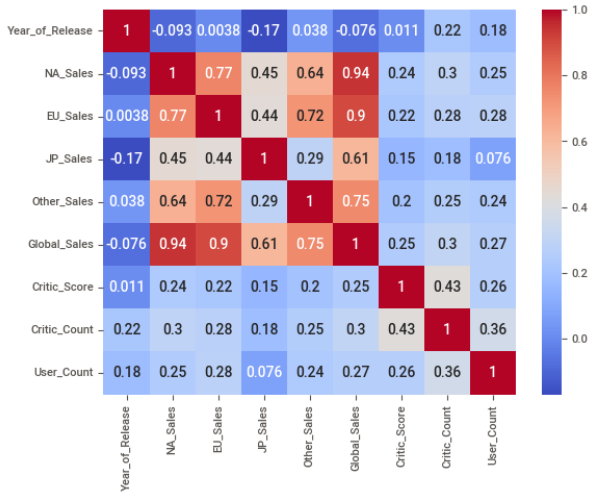
5. Standardize Critic and User attribute values
    a. Replace missing user count values with the mean user count
    b. Replace missing critic count values with the mean critic count
    c. Convert the Critic Count, Critic Score, User Count, User Score, and Year of Release to integer instead of object.

Table 2: Post Data Cleaning Snapshot

Attribute	Description	Data Type	Non-Null Values	Missing Values
Name	Name of the game	object	8451	0
Platform	Console on which the game is running	object	8451	0
Year_of_Release	Year the game was released	int64	8451	0
Genre	Game's Category	object	8451	0
Publisher	Publisher	object	8451	0
NA_Sales	Game sales in North America (in millions)	float	8451	0
EU_Sales	Game sales in European Union (in millions)	float	8451	0
JP_Sales	Game sales in Japan (in millions)	float	8451	0
Other_Sales	Game sales in the rest of the world (Africa, Asia excluding Japan etc.)	float	8451	0
Global_Sales	Total sales in the world (in millions)	float	8451	0

Critic_Score	Aggregate score compiled by Metacritic staff	int64	8451	0
Critic_Count	Number of users who gave the critic_score	int64	8451	0
User_Score	Score by Metacritic's subscribers	int64	8451	0
User_Count	Number of users who gave the user_score	int64	8451	0
Developer	Party responsible for creating the game	object	8451	0
Rating	The ESRB Ratings (E, T, A, etc.)	object	8451	0

3. Data Visualization

Clean Data Correlation Heat Map	Raw Data Correlation Heat Map
	

- NA Sales and EU Sales have a strong correlation with Global Sales.
- Critic Scores and User scores don't seem to have a strong correlation to sales

4. Data Source and Github Repository

- Github Repository: https://github.com/nickkzds/vg-success-pred
- Data Source: https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings

5. Project Methodology

1. **Research Questions:** Topic chosen based on personal interest and curiosity for the gaming industry. Research questions refined and re-stated based on available data sets and early data source exploration.

2. **Data Collection:** Once a viable public data source is found with the right attributes, begin early data scraping. Use Python and pandas library to load the data set and produce simple descriptive reports outlining potential gaps that need to be addressed for data cleaning.

3. **Data Cleaning:** Begin documenting potential data issues and research creative python driven solutions to remove missing values, drop features, normalize data and standardize data types.

4. **Data Pre-Processing:** Begin preparing the data for specific machine learning algorithms described in the research question phase. For classification algorithms use one-hot encoding when dealing with categorical data. For regression algorithms ensure each feature is scaled properly.

5. **Data Modelling & Evaluation:** Feed the cleaned and processed data into the appropriate algorithms and compare evaluation results for each. For example comparing R2 scores for regression, or accuracy, precision and recall for classification.

6. **Communicate Key Findings:** Record key findings and clearly communicate how the results have answered or come close to answering the proposed research questions.

6. References in APA Citation Format

1. Ballhaus, W., Chow, W., & Rivet, E. (2022, June 12). *Global Entertainment & Media Outlook 2022–2026 perspectives report*. PwC. https://www.pwc.com/gx/en/industries/tmt/media/outlook/outlook-perspectives.html

2. Shovo, Sheikh & ShafiulAlam, ShafiulAlam & Habib, Shaikh. (2023). *Predictive Analysis on Commercial Success of Game*. IARJSET. 10. 10.17148/IARJSET.2023.10536.

3. Ehrenfeld, S. E. (2011). *Predicting Video Game Sales Using An Analysis Of Internet Message Board Discussion* (thesis).

4. Prasad, Aashish. (2019). *Estimating Video Game Success using Machine Learning*. 10.13140/RG.2.2.14389.01767.

5. Trněný, M. (2017). *Machine Learning for Predicting Success of Video Games* (thesis).

6. Teja, A. S., Hanafi, M. L., & Qomariyah, N. N. (2023). Predicting steam games rating with regression. *E3S Web of Conferences*, *388*, 02001. https://doi.org/10.1051/e3sconf/202338802001