Predicting Gaming Industry Success with Machine Learning

Nicholas Katrantzidis

Dr. Tamer Abdou

500655049

2023-07-18

Table of Contents

Table of Contents

[Revised Abstract 5](#_Toc140615757)

[Literature Review 7](#_Toc140615758)

[Data Source and GitHub Repository 11](#_Toc140615759)

[Project Methodology 11](#_Toc140615760)

[Data Description and Cleaning 12](#_Toc140615761)

[Feature Engineering 16](#_Toc140615762)

[Modelling and Final Results 18](#_Toc140615763)

[Project Insights & Conclusions 19](#_Toc140615764)

[Limitations of Work and Addressing Gaps 23](#_Toc140615765)

[References in APA Citation Format 23](#_Toc140615766)

## 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 and gaming companies are interested in factors the affect global sales. This report, attempts to provide clarity for the below questions using the Google Collab programming tool and various regression machine learning techniques.

1. Given the enriched data set and based on related research, which regression algorithm best predicts Global Sales when comparing Linear Regression vs. Random Forest vs. SVM vs. Boost Gradient Regression?
2. Does critic-score data prove to be a better indicator for video game success compared to user-score data? Which features are most important for the highest performing models when predicting Global Sales?
3. Which features have the highest importance when using classification to predict global sales over 1 million?

The results for question 1 will compare various regression model techniques, specifically Linear Regression, Random Forest Decision Tree, Support Vector Regression, and Gradient Boost Regression and their respective evaluation metrics with K-fold cross validation, Training Time and Memory Consumption. The results for question 2 will compare the features importance for the highest performing models. The results for question 3 will describe the classification model and its accuracy, precision, and f-score. To achieve these results, this project will use dummy encoding for the categorical variables to ensure accurate implementation of the regression models. Pandas, numpy, sklearn, seaborn, matplotlib, and scipy libraries will be used for cleaning the data, preparing the data, modelling the data, and evaluating the models. Question 3 will use Random Forest Classifier to understand which features predict gaming consoles. In conclusion, this report will use knowledge gained from CIND820 and related research to answer the proposed research questions. The reader should gain a better understanding of how to predict video game sales and the importance of critic vs user features.

## Literature Review

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.

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.

**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 an Encoding method to convert string values to numerical values. The finding in this report suggest that certain genres, ESRB ratings, and publisher are 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 only the niche that participate on message boards.

**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 classification methods were 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.

**Literature Review Conclusion and Key Findings**

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

1. **Research Paper 1**: Using encoding to convert categorical values to numerical values. This method will be required to run both the proposed classification and regression models for question 3 and compare results.
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 classification models. 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 report’s data set we have the “User\_Count” attribute. For research questions 1 and 2, this report will attempt to extrapolate by confirming if user data has a greater impact than critic data.
5. **Research Paper 5:** Based on this paper’s analysis decision trees regression produces higher R2 scores and therefore more accurate results. This report will also use an R2 score to evaluate the classification model for research question 1 along with RMSE and test this hypothesis.

## Data Source and GitHub Repository

GitHub Repository: <https://github.com/nickkzds/vg-success-pred>

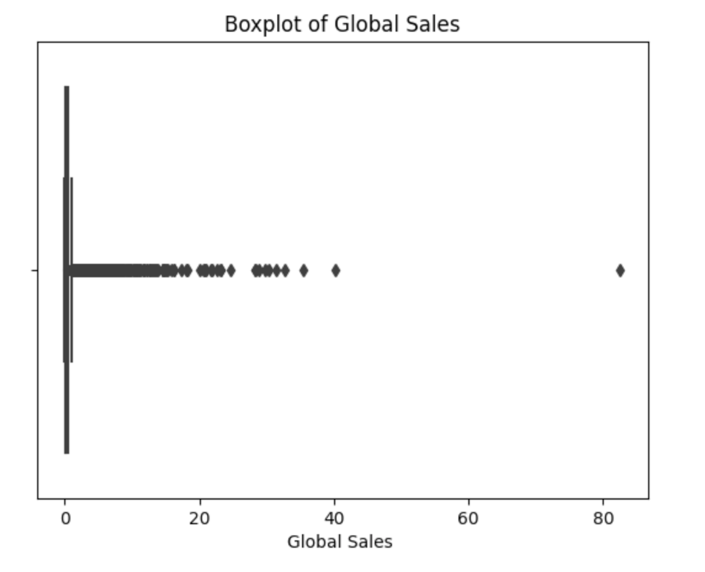
Data Source: <https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings>

## Project Methodology

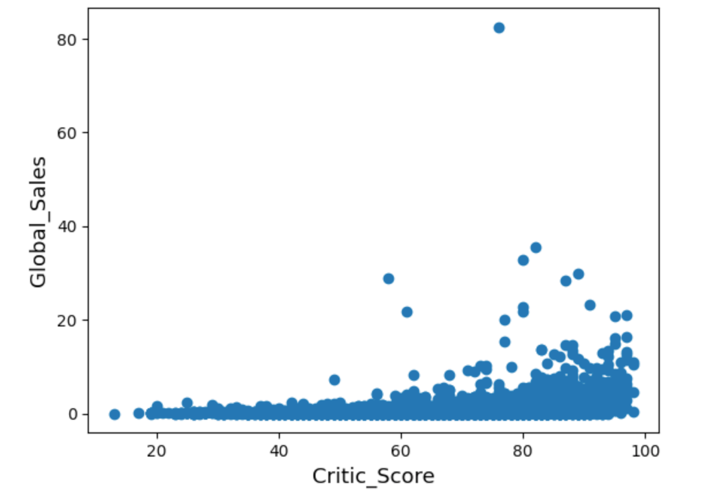
* **Research Questions:**  Topic chosen based on personal interest and curiosity for the gaming industry. Research questions refined and re-stated based on available data sets and early data source exploration.
* **Data Collection:** Once a viable public data source is found with the right attributes, 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.
* **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.
* **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.
* **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.
* **Communicate Key Findings:** Record key findings and clearly communicate how the results have answered or come close to answering the proposed research questions.

## Data Description and Cleaning

1. Utilize Univariate and Bivariate analysis to check for outliers
   1. Univariate analysis: used seaborn library to create a boxplot of the target variable “Global Sales” and discovered 1 clear outlier.

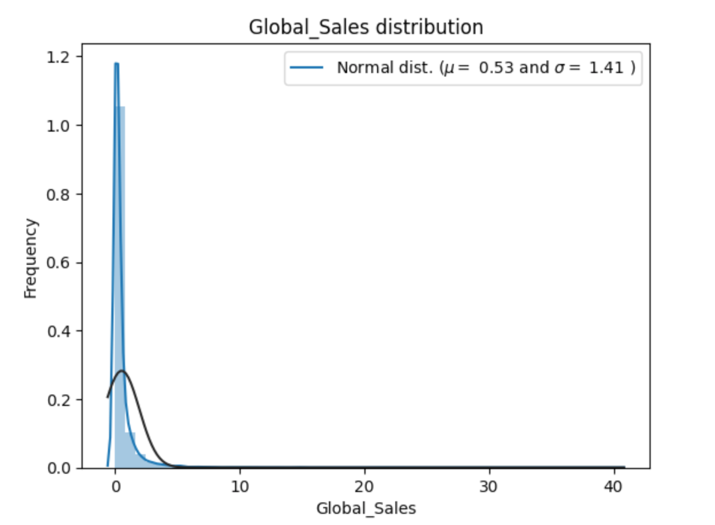


* 1. Bivariate analysis: used matplotlib to create a scatterplot of the target variable and 1 independent variable (Critic\_Score) to confirm the outlier



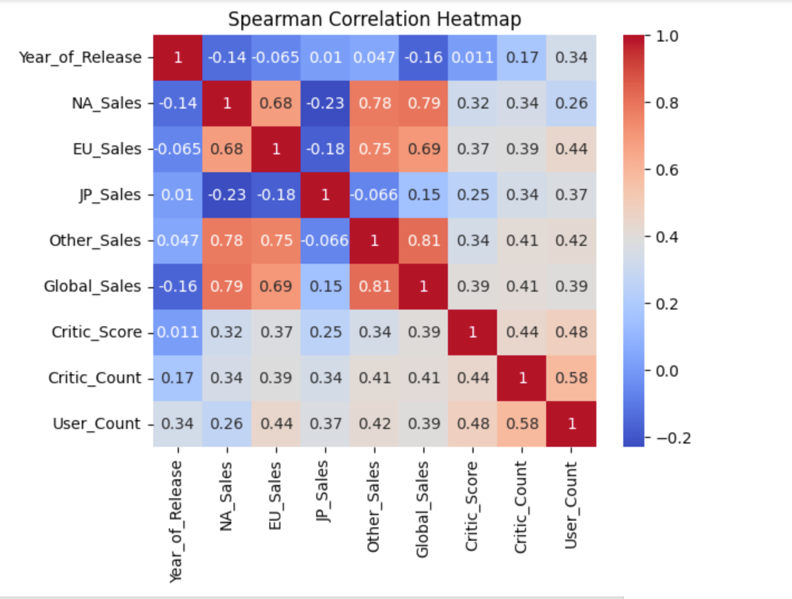
* 1. Removed Outlier with by removing record with critic score & global sales > 60

1. Check Data Distribution
   1. Used seaborn library to create a dist plot and prob plot of global sales



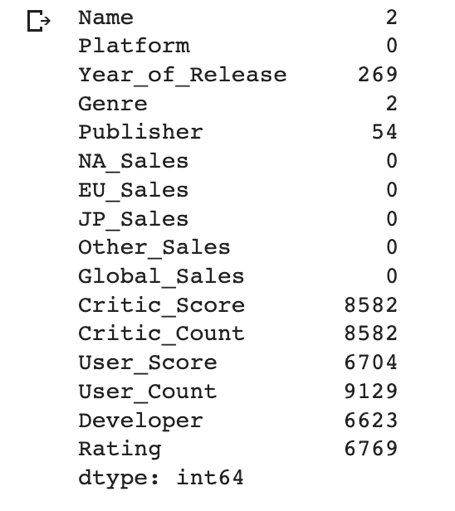
* 1. Distribution of global sales is non-normal
  2. As a result, project modelling will focus on Gradient Boosting Regression, Random Forest Regression, Support Vector Regression, and Linear Regression for question 1&2
  3. As a result, project modelling will focus on Logistic Regression and Random Forest Classifications Models.

1. Checking Correlation and Linearity
   1. Check variable correlation to understand relationships of data
   2. Using Spearman correlation and found Sales data from other countries to be highly correlated with Global Sales. As a result, sales data will be removed from modelling to prevent redundancy and dimensionality issues.

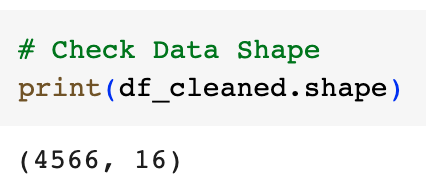


1. Check number of missing values

The Video Game Sales with Ratings data set is a combination of VGChartz video game sales data and Metacritic review data. Combining these two data sets lead to one data set with significant missing values in the review data. Specifically, the Critic Score, Critic Count, User Count columns were missing approximately 50% of their column data.



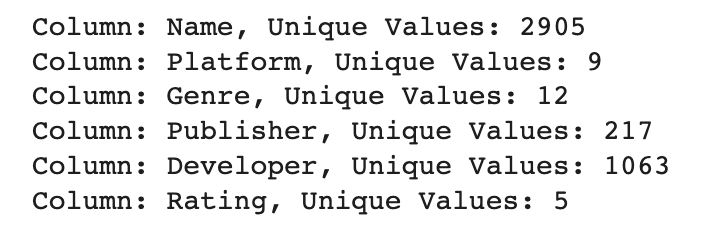
* 1. Removed records in data where Name, Genre, Year of Release, and Publisher have missing values
  2. Reviewed unique values in Platform Column and reduced variables to most recent platforms with a high number of unique values.
     1. Focused on PS3, PS4, X360, XOne, Wii, WiiU, DS, PSP for question 1&2
     2. Included all platforms for question 3 as its easier to run binary classification
  3. Dropped additional values where key feature Critic\_Score has missing values
  4. Imputed data on remaining missing values for User\_Score, User\_Count, Developer, and Rating
  5. Final Data Shape with no missing values and imputed data:



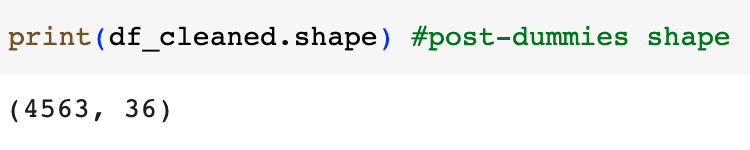
1. Matched Scales for User\_Score and Critic\_Score
   1. User\_Score was in a range of 1-10 and Critic\_Score was in a range of 1-100. Multiplied User\_Score by 10 to match Scale
2. Checked for Duplicate Values
   1. Found 1661 duplicate names data in question 1&2
   2. At first this may seem like an issue; however, many games are released on multiple platforms (especially successful ones). Therefore, duplicate name values can remain in data set.
   3. To further investigate, checked data set for duplicates values with the same Name and Platform to ensure no duplicates that are errors and found 6
   4. Removed suspicious duplicate entries

## Feature Engineering

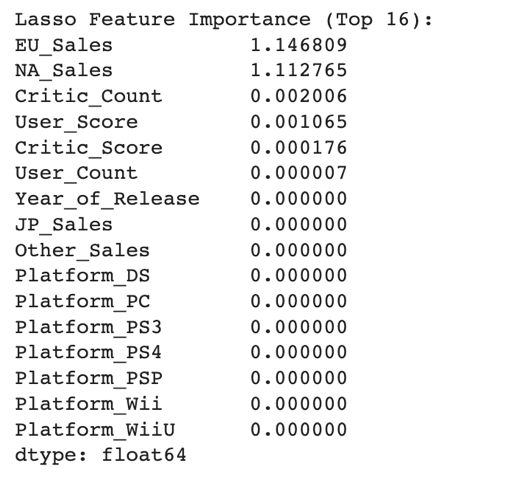
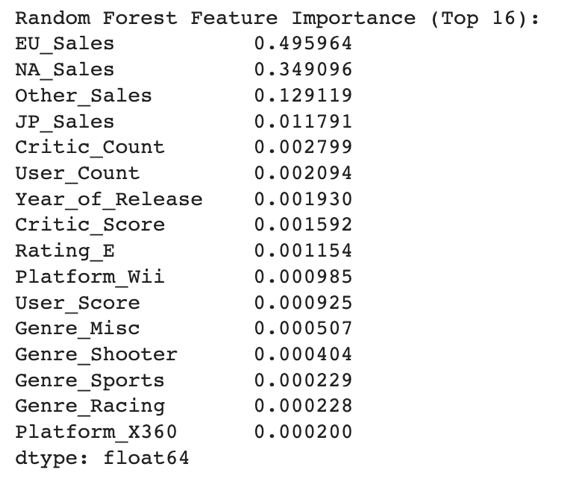
1. Check Unique values in each feature to choose drop candidates for encoding



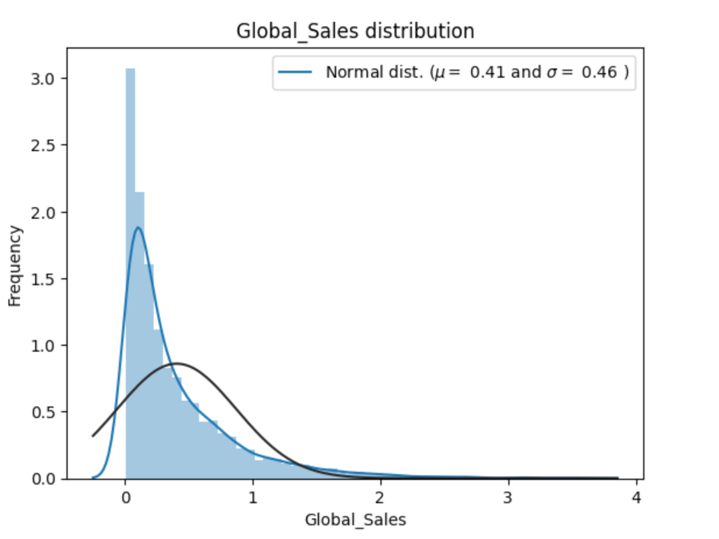
* 1. Name, Developer, and Publisher will be dropped to avoid curse of dimensionality and overfitting for question 1&2
  2. Created dummy data for only Platform, Genre, and Rating and checked updated dummy data shape for question 1&2



1. Use L1 Lasso and Random Forest for Feature selection
   1. Checked Lasso Feature Importance and Random Forest Feature Importance to understand impactful features for question 1&2

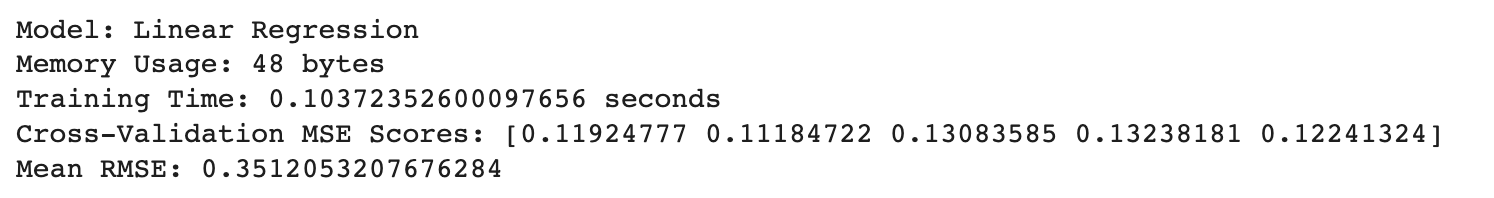


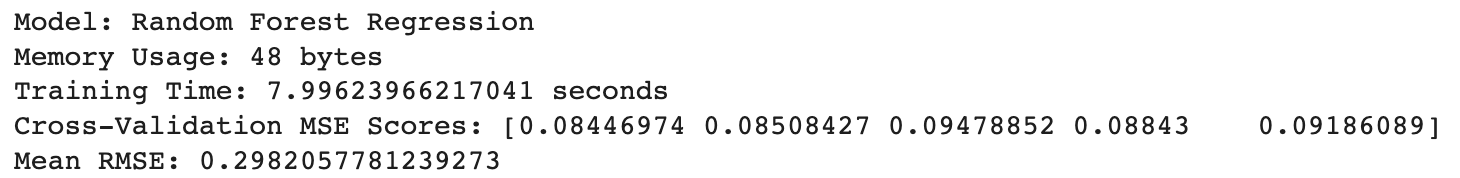
* 1. Sales data from other countries were consistently high and may cause modelling issues. Given the context of focusing on global sales, these features may cause redundancy and therefore will be dropped.

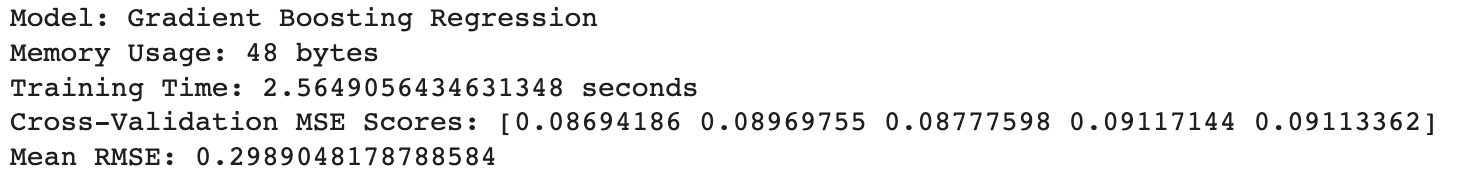
1. Splitting the Data and Scaling
   1. Applied log1p function to all elements of the columns to both the training and test sets. As a result, new distribution is more normal.
   2. Transform all independent variables into a similar range with the MinMax Scaler in train and test sets.
   3. Data is now ready for modelling for question 1&2

## Modelling and Final Results

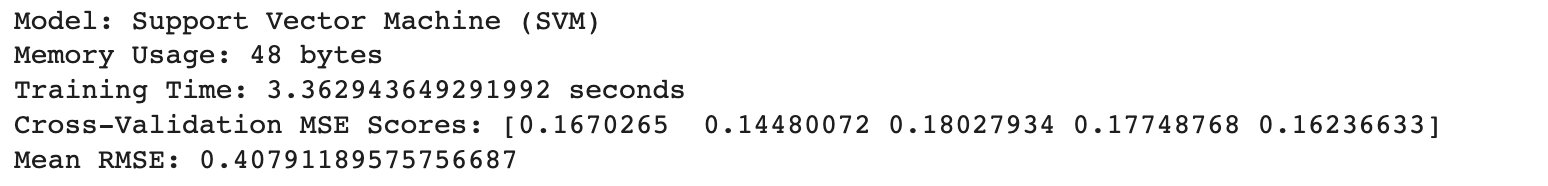
1. Linear Regression



1. Random Forest Regression
2. Gradient Boost Regression



1. Support Vector Machine



1. Random Forest Classification Results

A number on a white background

Description automatically generated

1. Logistic Regression Classification Results

A number on a white background

Description automatically generated

## Project Insights & Conclusions

**Research Question 1:** *Given the enriched data set and based on related research, which regression algorithm best predicts Global Sales when comparing Linear Regression vs. Random Forest vs. SVM vs. Boost Gradient Regression?*

Four regression models and two classification models were used to predict the Global\_Sales target variable in the above analysis and results. Comparing the first four regression models, SVM presented the highest RMSE score and the second highest training time score. The SVM regression model proved to one of the worst performing models given its lower predictive accuracy and second highest training time. Regarding memory consumption, all 4 models used exactly 8 bytes.

The top performing regression models are the Gradient Boosting Regression and the Random Forest Regression with the highest predicative accuracy. Both resulted in RMSE scores of roughly 0.29 compared to SVM at 0.4 and LR at 0.35. The final differentiating factor between the two models is the training time. Although the Random Forest Regression has a slightly better RMSE score, the training time for RF is 7.9 seconds which over double the training time of the Gradient Boosting Regression (2.5) and delivered similar RMSE results.

In conclusion, Gradient Boosting Regression is the more well-rounded model in terms of efficiency and accuracy. If given, similar research question but a much larger data set (with a similar makeup to the one used) the GBR would be the best choice to avoid the longer training times of Random Forest.

**Research Question 2:** *Does critic-score data prove to be a better indicator for video game success compared to user-score data? Which features are most important for the highest performing models when predicting Global Sales?*

Focusing on the best performing regression model a histogram was created to plot the feature importance scores of all the features to identify the highest scoring and most influential feature.

A graph of a number of blue bars

Description automatically generated

User\_Count proves to be the strongest indicator followed closely by Critic\_Count. When comparing the Critic\_Count and User\_Count importance, User proves to be the most influential. This makes logical sense, as the games that appeal to the most users will likely have the highest global sales. Interstingly when comparing Critic\_Score to User\_Score, we can see that Critic\_Score is much more influential and important in the model. This proves that although there may be a less critic’s reviewing games – those fewer that do are a much stronger and more accurate indication of a games sales success. From this analysis we can infer that User\_Score is not as reliable and could even skew the model if there are a significantly higher amount of User reviews than Critic reviews.   
  
In conclusion, the number of users that play a game is the most important feature in predicting global sales and when it comes to user vs. critic reviews; the critic’s are a more reliable source information.

**Research Question 3**: *Which features have the highest importance when using classification to predict global sales over 1 million?*

Two classification models were tested, the random forest classification and the logistic regression classification. Both models scored very similar results in terms of accuracy, precision and f1 scores with the logistic regression model showing slightly better results. The differentiating factor between these two models is the training time which was significantly longer for logistic regression.   
  
Given the context of this project, I would recommend Random Forest Classification model as the best option with its similarly strong results and faster training time. Analyzing the features based on the Random Forest Classifier, we can see once again that User\_Count is the most influential.

A graph of a number of blue bars

Description automatically generated

## Limitations of Work and Addressing Gaps

**Small Dataset**: Given only 4000 records for the regression model and 9000 for the classification model it’s possible the dataset is too small capture the complexity of the present relationships. To address this gap, data scraped from VGChartz and Metacritic could be scraped to enhance the better capture complex relationships.

**Limited Feature Space**: The used feature space is small and could potentially be improved with supplemental data. For example, social media data could be taken into account to understand the sentiment towards specific games. For this given problem, features were removed due to the noise they would have introduced in the encoding phase. Perhaps other ML approaches that don’t require encoding could be used on this enhanced feature space.

## 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