Video Game Sales Analysis

Harish Yelamarthy
Masters in Business Analytics,
at Northwood University
yelamarthyh50@northwood.edu

Nipun Siddineni
Masters in Business Analytics,
at Northwood University
siddinenin91@northwood.edu

Nitish Bhasker Chetla
Masters in Business Analytics,
at Northwood University
chetlab45@northwood.edu

***Abstract*— The main factors impacting the success of video game sales are examined in this study. We examine sales information from an extensive library of almost 17,000 video games to determine the features most frequently linked to "hit" games—those that have sold more than a million copies. We create and apply a predictive algorithm to help identify possible hit games by utilizing this data. For publishers and developers of video games looking to maximize the financial potential of their creations, this study provides insightful information.**

Keywords—DataAnalysis, Hitgames, sales prediction, Game publishing,Factors influencing sales

I. INTRODUCTION (*HEADING 1*)

Although the video game industry is a vibrant and changing one, making a profit can be difficult. Publishers and developers are always trying to figure out what makes a blockbuster "hit" game different from the others. In order to close this significant gap, a large collection of sales data from about 17,000 video games is analyzed in this study. We will determine the essential traits most closely linked to games that surpass the one-million-seller threshold—a measure of commercial success in the industry—by means of this analysis.

Furthermore, we will create and put into use a useful prediction model by utilizing the insights obtained from the data. This approach will be created to help publishers and developers find games that have the potential to be popular so they can maximize their resources and increase their prospects of success in this cutthroat industry.

The research being conducted explores how machine learning models may be used to pinpoint the elements that contribute to video game success. In a competitive market, developers and publishers stand to gain a great deal by being able to forecast "hit" games, which are those that sell more than one million copies.Maintaining the Integrity of the Specifications

II. DATA COLLECTION

Source of data kaggle

The carefully selected dataset that includes sales data for about 17,000 video games is the foundation of this study. The foundation for determining the critical elements involved in reaching "hit" status—which is defined as selling more than one million copies—is provided by this data., To comprehensively assess the potential drivers of sales success, the dataset incorporates a diverse range of features. Genre classification plays a crucial role, as player preferences can vary significantly between categories like role-playing games (RPGs), first-person shooters (FPS), and simulation games. Release year is another important factor, as market trends and technological advancements can influence player reception over time. Additionally, the platform on which a game is released (e.g., console, PC, mobile) can significantly impact its reach and sales potential.

The dataset goes beyond these fundamentals to provide information on development budget, highly regarded ratings from credible sources, and verified marketing expenses. By examining these characteristics, we can investigate the monetary commitment required to create a game, how player perception is affected by reviews, and how marketing tactics influence sales.

The data acquisition process involved procuring information from reputable industry sources known for their data accuracy. To ensure the validity of our analysis, a rigorous data cleaning process was implemented. This process involved identifying and rectifying any missing values, inconsistencies, or errors within the dataset. By meticulously cleaning the data, we aimed to minimize the impact of outliers and ensure the robustness of our subsequent analysis using machine learning models.

III. EXPLORING THE DATASET

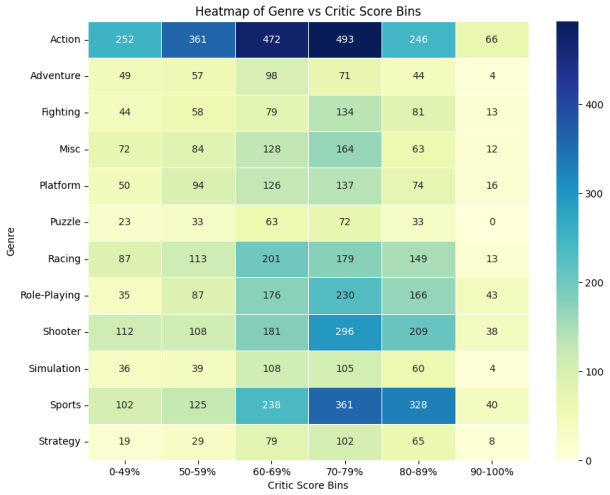

Figure 1 – Genre vs Critic Score

Using a heatmap display, we investigated how reviewer scores varied among different genres of video games (Figure 1). The heatmap indicates that some genres, like "Action" and "Adventure," have a higher concentration of games that score between 50 and 79% in the center, while other genres, like "Strategy" and "Role-Playing," have a higher tendency toward higher reviewer ratings (70–100%). These preliminary findings point to possible differences in the assessments of different genres of games by critics.

Figure 2 – Platform vs Critic Score

Using a heatmap representation, we also investigated how critic ratings were distributed throughout various video game platforms (Figure 2). The heatmap indicates that critic scores might be influenced by platform choices. For example, games on the "PS2" have a tendency to score higher in the 70–79% area, but games on the "PC" seem to favor the 80–88% range. These results call for more research into the possible causes of these platform-specific score distributions.

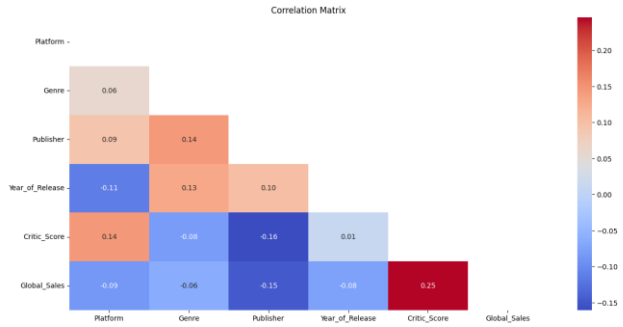

Figure 3 – Correlation matrix

We used a correlation matrix (Figure 3) to examine the dataset. The correlation coefficients between a number of variables were displayed in this complex table, showing the direction and strength of the links between them. The findings indicate that although critically acclaimed games may have somewhat greater sales, critic score is not the only factor that determines a video game's economic success. The association between "Critic Score" and "Global Sales" is small yet favorable. The association between "Year of Release" and "Global Sales" was surprisingly weak, suggesting that games from various eras may have had comparable sales statistics..

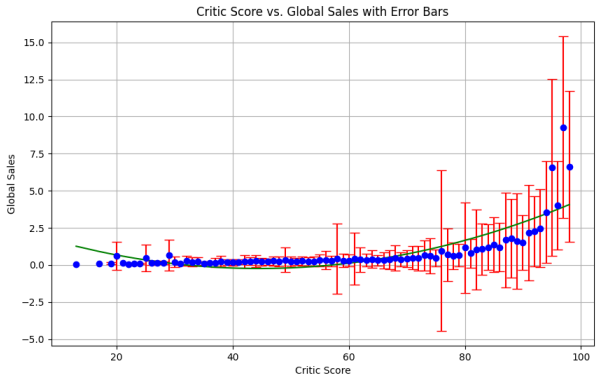

Figure 4 – fitting a polynomial of degree 2 for data points

Using a scatter plot, we looked at the connection between the reviewer score and worldwide sales (Figure 4). The scatter plot indicates a small positive connection between the two variables, suggesting that a video game's commercial success is not solely determined by its critic score, however there is a minor trend for games with higher critic scores to also have greater sales. On the other hand, some games with higher scores seem to have sold comparatively little, while some games with lower scores seem to have sold well.

IV. THEORETICAL BACKGROUND

Machine learning, or ML, is becoming a disruptive force in many fields. Machine learning (ML), a branch of artificial intelligence (AI), enables algorithms to learn from data without the need for human programming. They are able to recognize intricate patterns, formulate data-driven forecasts, and improve their performance over time as a result. The capacity to learn from examples, which promotes adaptability to novel conditions and independent task completion, is at the core of machine learning.

A. Supervised and Unsupervised Learning Paradigms

According to the kind of training data, supervised and unsupervised learning represent the two main paradigms in machine learning:

Supervised learning makes use of labeled data, in which every data point has an associated output or goal variable. In the end, the algorithm is able to predict future outcomes for unseen data by exploring the complex link between the input attributes and the intended output. Common supervised learning applications include classification tasks (like spam email filtering) and regression issues (like stock price predictions).

Unsupervised Learning: On the other hand, unsupervised learning works with data that is not labeled and does not include goal variables or pre-established categories for the data points. Finding buried patterns or innate structures in the data itself is the goal here. Applications for unsupervised learning include principal component analysis and dimensionality reduction (e.g., grouping clients with similar buying habits).
Our model generation approach utilized the advantages of two different techniques, namely logistic regression, and random forest classifier, in the context of supervised learning.

Logistic regression

When it comes to classification issues with binary outcomes, logistic regression is a potent statistical technique (0 or 1). It builds a linear model to calculate the likelihood that an instance falls into a particular class. The logistic function, sometimes known as the sigmoid function, is the foundation of this method. The output of the linear model is converted by this function into a probability value that is limited between 0 and 1. Among the many benefits of logistic regression are its interpretability, which makes it easier to comprehend how characteristics affect the model's predictions, and its computational efficiency

B. Random Forest Classifier

An ensemble learning technique that uses the combined strength of several decision trees to improve overall accuracy and resilience is embodied in a random forest classifier. To encourage a varied group of learners, each decision tree is built using a random selection of characteristics and data points. The random forest combines each tree's individual forecasts during the prediction step, frequently by using a majority vote for classification problems. Random forests are highly praised for their robust performance in a variety of classification tasks, their capacity to handle high-dimensional data, and their ability to reduce overfitting.

Through the strategic use of both logistic regression and random forest, we extract important insights:

Interpretability is provided by logistic regression, which helps us understand which factors have a major impact on the model's predictions.

Random forests use the combined power of several decision trees to provide a strong and accurate overall forecast.

V. RESULTS AND DISCUSSION

The logistic regression model's accuracy is 83.48%, meaning that for around 83.48% of the instances in the dataset, it accurately predicts the result.

Regarding accuracy, the model was able to predict class 0 (negative class) with 84% precision and class 1 (positive class) with 73% precision. Precision is defined as the percentage of all cases categorized as belonging to a specific class that were properly predicted.

The model successfully recognizes the majority of examples that belong to class 0, as evidenced by the high recall, or true positive rate, of 99% for class 0. Class 1 recall, on the other hand, is significantly lower at 9%, suggesting that the model has difficulty accurately classifying cases.

For class 0 and class 1, the F1-score—the harmonic mean of accuracy and recall—is 0.91 and 0.17, respectively. A higher F1-score indicates greater performance because it strikes a balance between recall and accuracy.

The model's predictions are broken out in great depth in the confusion matrix. It demonstrates that the model accurately predicted 1332 cases out of 1342 instances that belong to class 0, and mistakenly forecasted 10 instances that belong to class 1. Likewise, of the 286 examples that belonged to class 1, the model predicted 27 occurrences correctly and 259 instances wrongly as class 0.

ith an accuracy of 83.54%, the random forest model can accurately predict the result for around 83.54% of the dataset's instances.

Regarding accuracy, the model was able to predict class 0 (negative class) with 84% precision and class 1 (positive class) with 68% precision. Precision is defined as the percentage of all cases categorized as belonging to a specific class that were properly predicted.

The model successfully recognizes the majority of examples that belong to class 0, as evidenced by the high recall, or true positive rate, of 99% for class 0. Nevertheless, class 1's recall is just 12%, indicating that the model has difficulty accurately identifying cases that belong to class 1.

For class 0 and class 1, the F1-score—the harmonic mean of accuracy and recall—is 0.91 and 0.20, respectively. A higher F1-score indicates greater performance because it strikes a balance between recall and accuracy.

The model's predictions are broken out in great depth in the confusion matrix. It demonstrates that the model accurately predicted 1326 cases out of 1342 instances that belong to class 0, and mistakenly forecasted 16 instances that belong to class 1. In a similar vein, of the 286 occurrences that belonged to class 1, the model predicted 34 instances correctly and 252 instances wrongly as class 0.

VI. Conclusion

In conclusion, this study delved into the factors influencing the success of video game sales by analyzing a comprehensive dataset encompassing nearly 17,000 video games. Through exploratory data analysis and the application of machine learning techniques, we gained valuable insights into the features associated with "hit" games, defined as those selling over a million copies. Our research identified critical elements such as genre classification, critic scores, and platform preferences as significant contributors to a game's commercial success.

Moreover, the predictive algorithms developed in this study demonstrated promising performance in identifying potential hit games, with logistic regression and random forest models achieving accuracies of approximately 83.48% and 83.54%, respectively. These models offer valuable tools for publishers and developers seeking to maximize the financial potential of their creations by identifying games with the highest likelihood of success.

VII. Future Scope

Future research might go in a number of directions, even while this study delivers insightful information about the variables affecting video game sales and prediction models to help industry decision-makers:

1. include other information: To improve the models' forecast accuracy, future research may look into include extra data including player demographics, in-game mechanics, and marketing tactics.

2. Temporal analysis: Player preferences and trends change over time in the dynamic video game business. Future studies might look at how success criteria and sales trends have evolved over time and throughout different gaming generations.

33. Cross-validation and model comparison: Future research might make use of cross-validation strategies to assess the effectiveness of other machine learning methods outside of logistic regression and random forest classifiers in order to guarantee robustness and generalizability.

4. Qualitative analysis: Along with quantitative analysis, qualitative research techniques like surveys or interviews with gamers and industry professionals may provide further light on the individualized elements of player preferences and game success.

5. Adaptation to changing technologies: Future studies may examine how game sales and success variables are impacted by new platforms like virtual reality (VR) and augmented reality (AR).

All things considered, more study in this field has the potential to improve our comprehension of the dynamics of the video game business and offer insightful advice to stakeholders hoping to successfully negotiate this competitive environment.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.