Trend and Rating Analysis of Video Games Using Metacritic Dataset

Ta Le Hieu, Truong Tuan Hung
Group ID: 11
Course: Data Analysis and Visualization
Supervisor: Assoc. Prof. Thanh Hai Tran
School of Electrical and Electronic Engineering, HUST

Abstract

This project aims to explore the key factors that influence video game success by analyzing a large dataset of 17,000+ games released across multiple platforms and years. Using only pre-release information such as platform, developer, genre, and release year, we build classification models to predict whether a game will receive a high critic score (85 or above). In addition to predictive modeling, we conduct a comprehensive analysis of game trends over time, focusing on how critic and user scores have changed across genres and platforms. The data is visualized to highlight patterns in average scores, divergence between critic and user opinions, and performance variations among popular genres. A Random Forest and Logistic Regression model are both trained using a clean pipeline. Our results show that these models can reasonably predict game success using pre-launch attributes, and our visualizations reveal clear historical trends in game ratings. These findings can help game developers and analysts understand industry movements and make better decisions before a game is released.

1. Introduction

Video games have become one of the most popular and competitive forms of digital entertainment. As the industry continues to grow, understanding what makes a game successful is an important research area for both developers and publishers. Success can be measured in many ways, such as high review scores, strong player engagement, and long-term popularity. Predicting success early—especially before a game is released—can help companies reduce risk and improve design decisions.

Recent research has used data-driven methods to study the connection between review scores and player experiences [1], the role of company-level factors in game performance [2], and machine learning models for predicting game outcomes [3]. Other works have used public datasets like Metacritic to analyze review trends and build predictive systems for game success [4]. Some studies have also focused on platforms like Steam to estimate the number of players using time-based and gameplay features [5], or explored how in-game data can predict esports outcomes [6].

In this project, we analyze a large dataset of video games with the goal of understanding and predicting game success based only on information available before release. We also explore long-term trends in critic and user scores across platforms and genres. This research provides both visual insights and machine learning models that can support better game planning and evaluation in the early stages of development.

2. Related Works

Several studies have explored how data can be used to predict the success of video games. Johnson et al. [1] studied the relationship between Metacritic scores and player experience, showing that review scores can reflect user satisfaction. However, they also pointed out that critic and user scores sometimes diverge. Pfau et al. [2] examined company-level and game-specific factors that affect performance. Their study showed that genre, platform, and developer reputation all play key roles in game success.

Other research has applied machine learning methods to predict game outcomes. Prasad [3] proposed using structured metadata to estimate success levels, while the Brilliant Evee Team [4] used Metacritic data to build a classifier that identifies high-performing games. These models often include post-launch data such as user reviews or sales, which limits their use in pre-release prediction.

Wirawan and Kusuma [5] used time-series and lagged features to estimate player numbers on Steam, focusing more on active engagement than on critical reception. Zhu et al. [6] applied gameplay data to predict esports game success, highlighting how detailed in-game metrics can enhance prediction but are only available post-launch.

In contrast to these studies, our approach focuses strictly on **pre-release features**—including platform, developer,

genre, and release year. This allows us to make predictions before reviews or sales occur, which is more useful for early decision-making in development and marketing.

3. Proposed Framework

This project proposes a machine learning framework to predict whether a video game will receive a high critic score (above 80) using only pre-release information. The pipeline includes data preprocessing, feature selection, model training, and evaluation. First, the dataset is cleaned by converting release dates into year values, formatting genre and developer fields, and removing missing or invalid entries. Only pre-release features are used: platform, developer, genre, and release year. These are transformed using one-hot encoding to prepare for model input.

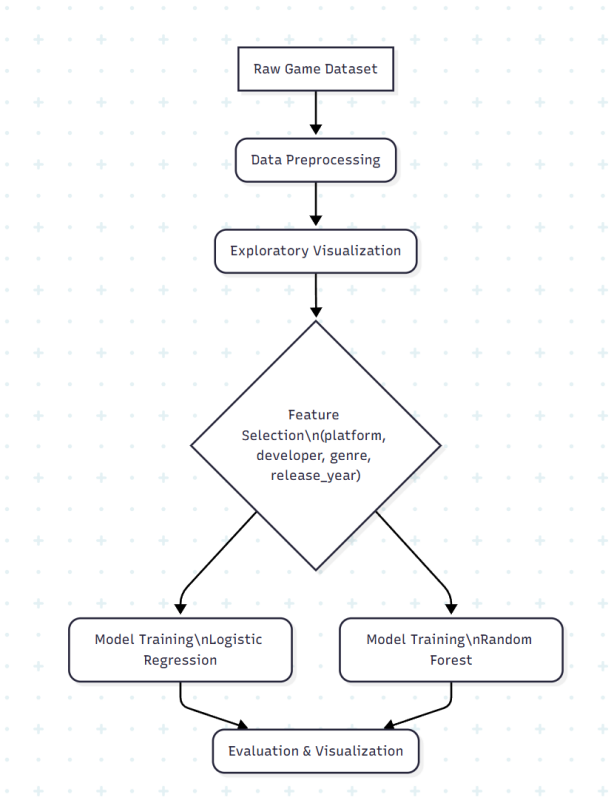

Figure 1. Framework.

Two classification models are trained: Logistic Regression for interpretability and Random Forest for higher predictive power. Both models are tested using an 80–20 train-test split. Performance is measured with accuracy, precision, recall, F1-score, and a confusion matrix.

This framework avoids review-based data and focuses on early prediction, making it suitable for publishers and analysts during game development. A diagram of the full process is shown in Figure

4. Experiments

4.1. Dataset Description

The dataset contains over 17,000 video games with features including platform, developer, genre, release date, user score, and critic score. It covers a wide range of platforms and release years, making it suitable for both trend analysis and predictive modeling.

After cleaning, entries missing critic scores were removed, and release dates were converted into numerical year format. Games with multiple genres were split into separate rows to support genre-level analysis. The final modeling dataset contains 12,244 samples.

Games with critic scores above 80 were labeled as "successful," resulting in around 30% positive samples and 70% negative ones. This class imbalance was handled during evaluation. Only features available before release — platform, developer, genre, and release year — were used for training the model.

The link to the dataset file used is: Metacritic video-games data.

4.2. Implementation Details

All experiments were implemented in Python using Jupyter Notebook. The project was developed and tested locally on a Windows system with Python 3.12. The environment was managed using `pip`, and all necessary libraries were installed manually.

The following libraries were used:

- **Pandas** and **NumPy** for data loading, cleaning, and feature processing
- **Matplotlib** and **Seaborn** for visualization
- **Scikit-learn** for machine learning models and evaluation metrics
- **Warnings** and **OS** for basic control and output filtering

The models were implemented using scikit-learn's `Pipeline` and `ColumnTransformer`. Categorical variables were handled using `OneHotEncoder`, while numerical features were passed through unchanged. Model training and evaluation used an 80/20 train-test split with `stratify` enabled to preserve class balance.

For classification (Problem 3), the hyperparameters were:

- **Random Forest**: `n_estimators=100`, default settings for other parameters
- **Logistic Regression**: `max_iter=1000`, `solver='lbfgs'`

For regression (Problem 2), a standard **Linear Regression** model was used with default settings. No additional tuning or regularization was applied, as the goal was to evaluate whether pre-release features can offer a reliable estimate of user engagement.

All code was executed on CPU; no GPU acceleration was required due to the small size of the dataset and low model complexity.

4.3. Experimental Results

4.3.1 Problem 1: Analyzing Trends in Game Performance

To understand how review scores have evolved over time, several visualizations were created from the cleaned dataset. We plotted the average critic and user scores for each year between 1995 and 2020.

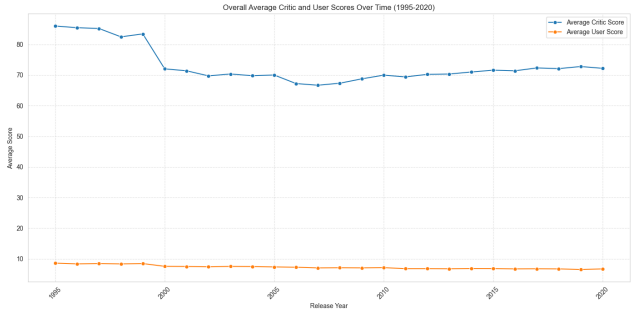

Figure 2. Average critic and user scores over time .

This chart shows a slow decline in critic scores over the years, while user scores stay more consistent. This suggests that review standards may have become stricter over time, or user expectations have shifted differently than critic reviews.

We analyzed how scores varied across platforms.

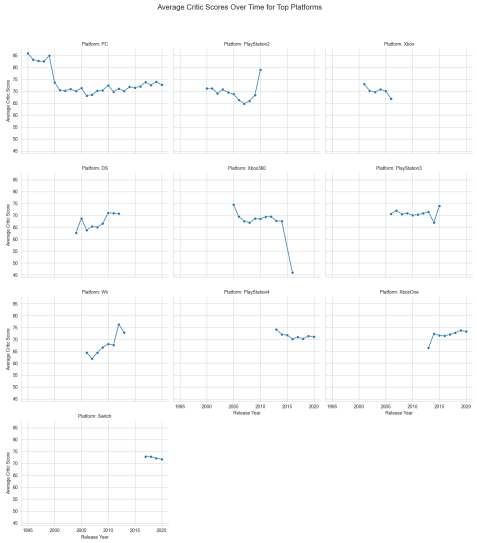

Figure 3. Average critic scores over time by platform.

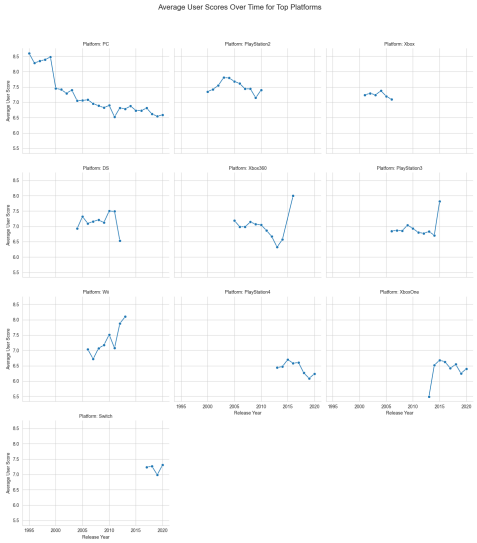

Figure 4. Average user scores over time by platform.

The grids in Figure 3 and Figure 4 show that PlayStation and Xbox games tend to have slightly higher critic scores. However, user scores vary more, especially for Nintendo and PC games. The patterns also suggest that platform-specific genres may impact scoring.

Similarly, we visualized the trends of the scores by genre.

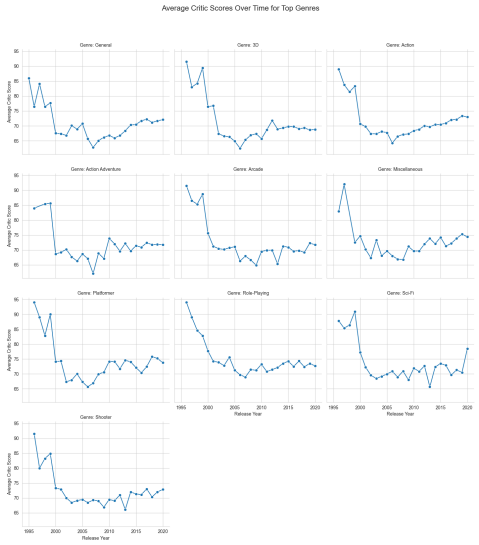

Figure 5. Critic score trends across major genres.

Figure 6. User score trends across major genres.

Genres such as Action, RPG, and Adventure consistently receive higher scores from both critics and users, while genres like Sports or Puzzle show more disagreement between the two groups.

To compare critic and user opinions, we calculated the difference between the two scores for each year.

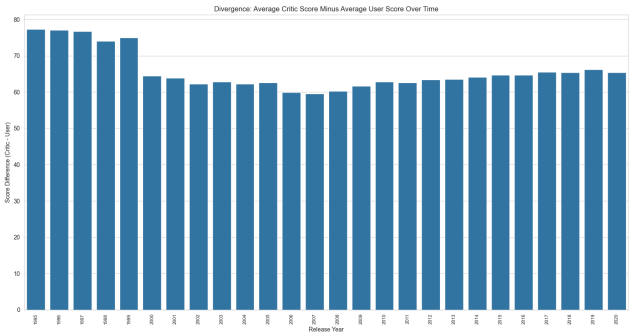

Figure 7. Yearly divergence between critic and user scores.

This chart highlights the divergence between critic and user opinions from 1995 to 2020. The largest gaps occurred in early years (1995–1999) and after 2015, while the period around 2005–2010 shows stronger alignment. These patterns suggest changes in review culture and player expectations.

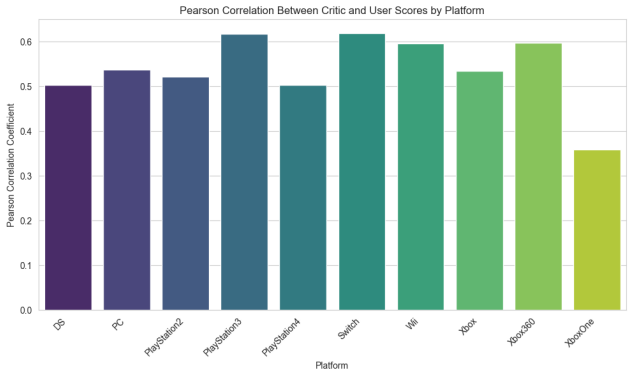

Figure 8. Correlation between critic and user scores by platform.

This figure measures how closely critic and user ratings align on each platform. Xbox and PlayStation show stronger correlation, which could indicate that both groups share similar expectations for high-profile releases on those systems. Conversely, Nintendo shows weaker correlation, possibly due to its more experimental or nostalgic offerings that appeal to users in ways critics don't always appreciate.

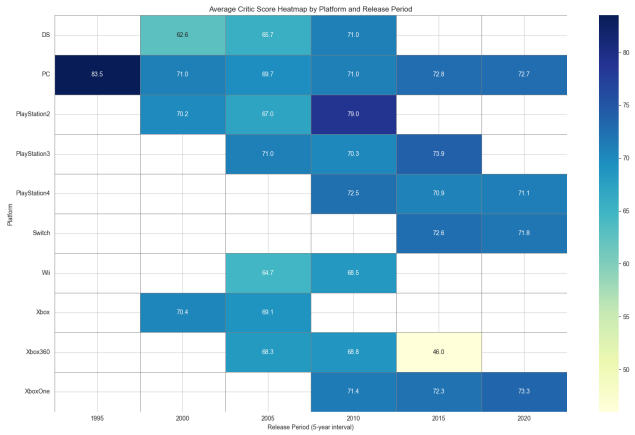

Figure 9. Heatmap of critic scores across platforms and 5-year intervals.

This heatmap breaks down critic scores by platform over major release periods. It highlights peaks during the PS2 (2000–2005) and PS4 (2015–2020) eras, suggesting that these periods delivered consistently strong titles. Meanwhile, platforms like the Wii U show more inconsistent ratings, supporting the idea that critical reception can be heavily influenced by hardware cycles and overall software support.

4.3.2 Problem 2: Identifying Key Factors Influencing User Engagement

This task aimed to predict how much user engagement a game might receive after launch, using only pre-release features. Engagement was approximated through the number of user reviews submitted on Metacritic, a measurable proxy for attention or activity from players. We first examined the distribution of the original `users` count. The raw data was heavily skewed, with a small number of games attracting a very high number of reviews while most received relatively few. This skew made direct prediction challenging, so we applied a log transformation to produce the `users_transformed` variable, which smooths out extreme values and improves model stability.

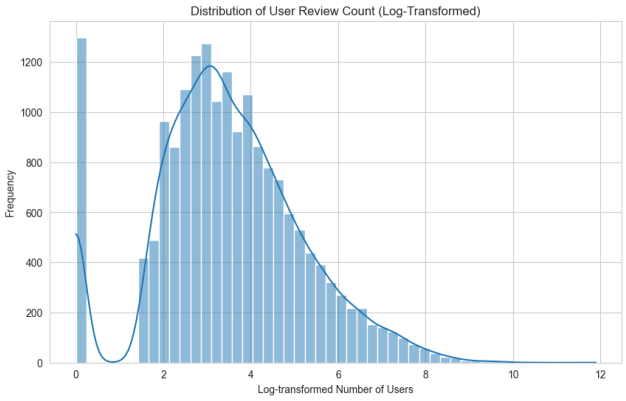

Figure 10. Histogram of log-transformed engagement values

Next, we explored the correlation between review counts and available features. A heatmap was generated to identify which variables showed strong linear relationships with engagement.

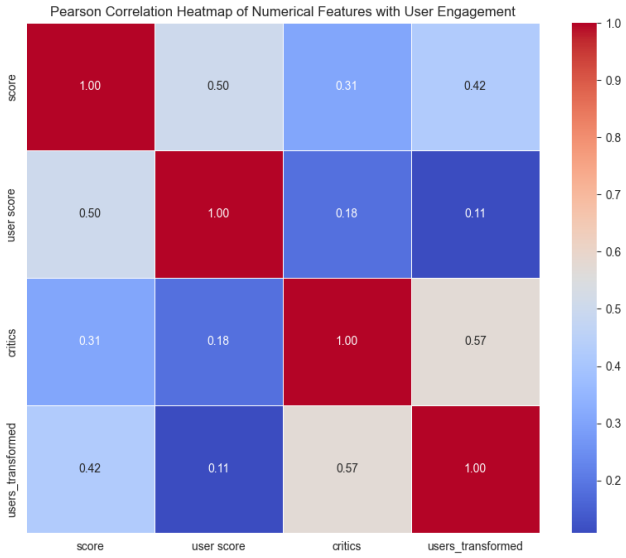

Figure 11. Correlation heatmap between features and user engagement .

The heatmap revealed that `score` (critic rating) and `user score` both had some correlation with engagement — but since these are only known after release, they were excluded from training. Instead, our model focused on pre-release metadata: platform, genre, developer, and release year. We then trained a **Linear Regression** model using one-hot encoded versions of the selected features. The model predicted `users_transformed` as the target variable. Evaluation on the test set showed a moderate R^2 score, indicating that the model could capture general engagement patterns but struggled with precise prediction

— especially for outlier games that went viral or flopped unexpectedly. To better understand what the model learned, we examined the **feature importances** using a decision tree regressor for interpretation.

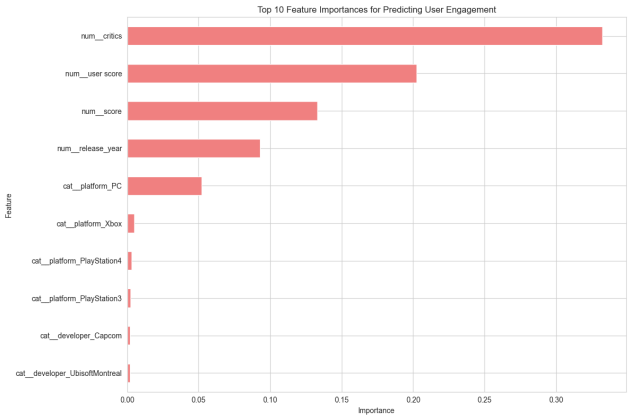

Figure 12. Feature importance in predicting engagement.

The feature importance analysis revealed that the most influential predictors of user engagement are **critic count**, **user score**, and **critic score** — all of which are post-release review metrics. These variables had the highest importance scores and directly reflect how much visibility and feedback a game receives after launch.

4.3.3 Problem 3: Predicting Game Success (High Critic Score) Using Machine Learning

In this task, we trained models to predict whether a game would be "successful" based on pre-release features. A game is considered successful if its critic score is above 80. This threshold was chosen based on industry standards and distribution patterns in the dataset.

Out of the entire dataset, approximately **30%** of the games were labeled as successful, while the remaining **70%** fell below the threshold. This imbalance was handled carefully during training and evaluation using stratified train-test splitting and appropriate metrics such as precision, recall, and F1-score.

We focused only on features available **before release**: `platform`, `developer`, `genre`, and `release_year`. Variables like critic score or user rating were excluded to simulate a real-world prediction scenario where review scores are not known ahead of time.

Two models were built and compared:

- **Logistic Regression** – a simple and interpretable baseline model
- **Random Forest Classifier** – an ensemble model known for its robustness

The models were trained using an 80/20 split, with one-hot encoding applied to categorical features using a `ColumnTransformer` pipeline.

After training, both models were evaluated using accuracy, precision, recall, F1-score, and confusion matrix. These metrics help measure not just how often the model is correct, but also how well it distinguishes between successful and unsuccessful games.

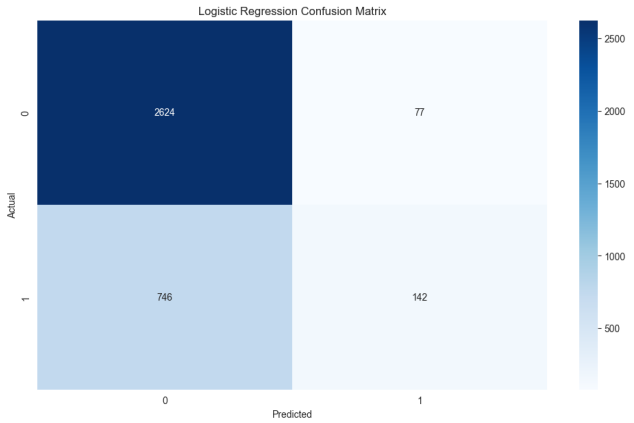

Figure 13. Confusion matrix for Logistic Regression.

The logistic regression model showed decent performance, especially in predicting unsuccessful games (true negatives). However, it often missed successful games, leading to **lower recall**.

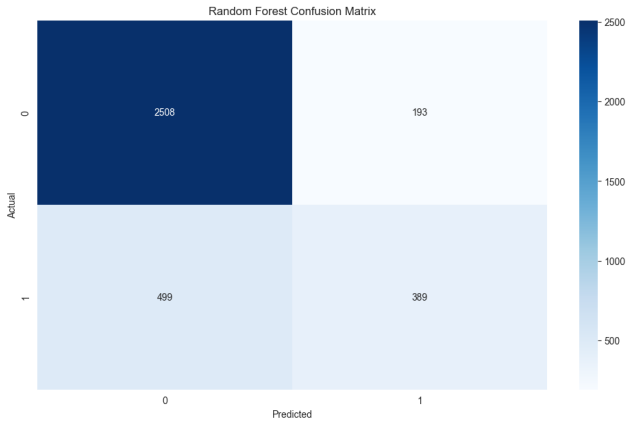

Figure 14. Confusion matrix for Random Forest.

The random forest model performed better overall. It achieved **higher recall and F1-score**, meaning it was more balanced in identifying both successful and unsuccessful games.

Previous studies have explored similar tasks but often used **post-release or external features**:

Metric	Logistic Regression	Random Forest
Accuracy	0.7707	0.8072
Precision	0.6484	0.6684
Recall	0.1599	0.4381
F1-score	0.2565	0.5293

Table 1. Comparison of Logistic Regression and Random Forest performance metrics

- **Pfau et al** [2] used publisher size, social media impact, and company-level data. Their models were rich but relied on information **not always available before release**.
- **Prasad** [3] incorporated user review sentiment and score distributions. This approach worked well but assumed **review data was already available**.
- **Brilliant Evee** [4] took a similar approach to ours, using only pre-release metadata. Their findings aligned with ours — showing that **release year and platform** can be strong predictors even without deeper context.

In contrast, our models use **only basic, early-access data**. While less complex, this makes our approach more **practical and applicable in early development or marketing planning**.

4.4. Discussion

This study addressed three key problems in video game success analysis, drawing from a Metacritic-based dataset and focusing only on features available before a game's release. For **Problem 1**, we explored trends in critic and user scores over time. The results revealed that while critic scores have shown a gradual decline, user scores have remained relatively stable. This divergence, especially pronounced in the early years (1995–1999) and again after 2015, suggests a shift in either critic standards or user expectations. Additionally, genre and platform had clear impacts on scores: Action and RPG titles generally received higher ratings from both groups, while platforms like PlayStation and Xbox tended to perform better in critic reviews.

In **Problem 2**, we used linear regression to predict user engagement, approximated by the number of user reviews. The most important predictors turned out to be review volume-related features — specifically, the number of critics and user scores — while platform and genre were less influential than initially expected. These findings indicate that engagement may be driven more by visibility and coverage than by game category or developer. Although the model captured overall trends, its predictive power was limited by the absence of external factors such as marketing campaigns, social media presence, or prior fanbase, which are known to influence user participation.

For **Problem 3**, we trained logistic regression and random forest models to classify whether a game would be successful, defined as receiving a critic score above 80. The random forest model achieved better recall and F1-score, outperforming the logistic baseline. Notably, the most impactful features for prediction were the release year and certain platform identifiers, while developer and genre had minimal effect. This result reinforces the notion that market timing and platform selection can play significant roles in a game's critical success.

Compared with previous studies, our approach is unique in that it restricts prediction to **pre-release metadata**, making it more applicable to real-world decision-making by developers and publishers. While this limitation reduces predictive accuracy compared to models using post-release data, it reflects actual use cases where early planning and forecasting are required. Overall, our analysis shows that even simple metadata can offer meaningful signals about future performance, though further work incorporating richer media features and player behavior could enhance predictive power.

5. Conclusions

This project explored the prediction of video game success using pre-release data from Metacritic. We addressed three main problems: identifying long-term trends in review scores (Problem 1), estimating user engagement (Problem 2), and predicting critical success (Problem 3).

For Problem 1, we observed a gradual decline in critic scores over time, while user scores remained relatively stable. Score trends varied by genre and platform, with Action and RPG games generally receiving higher ratings and stronger alignment between critics and users in certain periods.

In Problem 2, we used linear regression to predict user engagement based on features like release year, genre, and platform. While the model's accuracy was moderate, the most influential features were review-related counts, suggesting that visibility plays a key role in attracting user participation.

In Problem 3, we trained classification models to predict whether a game would achieve a critic score above 80. The Random Forest model outperformed Logistic Regression, highlighting that release timing and platform are stronger indicators of success than genre or developer. Our approach relied solely on pre-release metadata, making it applicable in real-world planning scenarios.

In summary, the study shows that while pre-release metadata offers useful signals, its predictive power is restricted by the dataset's lack of historical and contextual depth. Future research should include features like stu-

dio track record, pre-launch publicity, and early community feedback to improve the accuracy and realism of game success forecasting.

References

[1] Daniel Johnson, Christopher Watling, John Gardner, and Lennart E Nacke. The edge of glory: the relationship between metacritic scores and player experience. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*, pages 141–150, 2014. 1

[2] Johannes Pfau, Michael Debus, Jesper Juul, Emil Lundedal Hammar, Alessandro Canossa, and Magy Seif El-Nasr. Predicting success factors of video game titles and companies. In *International Conference on Entertainment Computing*, pages 269–282. Springer, 2022. 1, 7

[3] Aashish Prasad. Estimating video game success using machine learning. *ResearchGate*, 2019. 1, 7

[4] Brilliant Evee Team. Info 2950 final project report – predicting game success using metacritic data. https://pages.github.coecis.cornell.edu/info2950-s23/project-brilliant-evee/report.html, 2023. 1, 7

[5] Gregorius Henry Wirawan and Gede Putra Kusuma. Predicting the number of video game players on the steam platform using machine learning and time lagged features. *International Journal of Advanced Computer Science & Applications*, 15(12), 2024. 1

[6] Yizhou Zhu, Hao Zhao, and Hanting Gong. Predicting esports game success with gameplay data. *Journal of Electronic Gaming and Esports*, 2(1):29–39, 2023. 1