Will This Game Be a Hit?

Explaining and Predicting Video Game Review Scores Using Machine Learning

Yuhan Wang; Kaiyue Deng

University of Southern California

DSCI 510: Principles of Programming for Data Science

Dr. Itay Hen

December 15, 2025

Project Title and Team Members

Project Name

Will This Game Be a Hit? Explaining and Predicting Video Game Review Scores Using Machine Learning.

Team Members

Yuhan Wang (USC ID: 8914950372; Email: ywang204@usc.edu)

Kaiyue Deng (USC ID: 7459633091; Email: kaiyuede@usc.edu)

Research Question and Short Description

This project asks: Which factors best explain and predict video game review scores, and how do structured metadata and review-text signals jointly shape ratings - especially across critics vs. users? We analyze a dataset of games and associated reviews to (a) compare region-related effects (publisher vs. developer), (b) extract aspect-based sentiment signals from review text, and (c) integrate both into a regression model to estimate feature importance and predictive performance.

Data Sources, Size and Collection Approach

2a. Data Sources and Approach

Data Sources, Fields and Sizes

We collected video game pages from public web sources -Metacritic, focusing on fields that align with our proposal: title, platform, developer, publisher, critic score, user score, and review text. The final dataset includes 259 games. For text analysis, we collected 1,553 reviews in total (909 critic reviews; 644 user reviews).

Collection Approach

Because the API of the Metacritic workflow required querying by exact game title, we first scraped a list of game names using the web crawler software we wrote ourselves and then used that list to automate data retrieval for the full set of games through the Metacritic API. We stored both raw outputs and cleaned datasets after processing for the following analysis.

2b. Changes from the Original Plan

Compared with the midterm proposal, we have changed the main focus from publisher-region to developer-region. The original analysis focus emphasized whether the publisher region would meaningfully relate to ratings. After completing the publisher-region analysis, we observed that score distributions and medians across publisher regions were relatively similar, suggesting a weak relationship. We then shifted to the developer region, where differences were clearer, and within this, the Japan developer group showed a noticeably higher score distribution compared to other regions.

2c. Key Challenges (and Fixes)

Title Contamination in Sentiment Analysis

Game titles frequently reappear in reviews and may contain words that a general sentiment tool reads as negative (e.g., "dead," "evil"). We addressed this by treating each game title as a custom stopword to reduce false sentiment signals.

Domain Mismatch for VADER-Style Lexicon Rules

Gaming vocabulary can be misread without context (e.g., "kill," "dead" can be neutral or even positive in certain genres), and sometimes VADER cannot comprehend relatively complex grammar structures like clauses or specific expressions by game lovers. So we implemented a custom adjustment function to reweight recurrent domain terms and iteratively checked false positives/negatives, producing a correction table for refinement. After these corrections, the sentiment-score correlation increased from 0.144 to 0.162 (about 12% improvement).

Visualization Formatting and Grouping

Early plots made it clear that publisher-region patterns were weak but developer-region patterns were stronger. This guided us to prioritize developer-region comparisons and to note small-sample limitations for some regions.

Analysis and Visualization Methods

3a. Analysis Techniques and findings

EDA (Structured Factors)

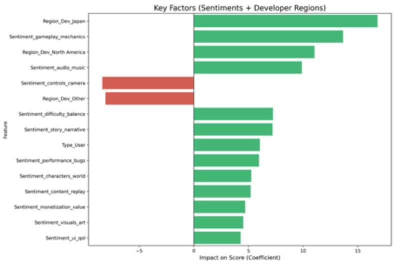
We compared rating distributions across publisher regions and developer regions. The publisher-region boxplot suggested limited differentiation, while the developer-region boxplot showed more separation, especially for Japan-developed games.

NLP (Aspect-Based Sentiment)

We extracted aspect-related sentiment from reviews (e.g., story, gameplay mechanics, audio/music) and computed correlations with overall scores. Story Narrative and Gameplay Mechanics showed the strongest alignment with ratings (highest correlations), indicating that these dimensions act as core "evaluation anchors." We also compared critic vs. user aspect sentiment and found systematic differences (e.g., critics more positive on Multiplayer/Monetization; notable divergence on Controls/Camera).

Regression Modelling

We trained a linear regression model that combines structured metadata with NLP-derived aspect sentiment features. The initial model performance was MSE = 638.84 and $R^2 = 0.1434$. After refining the training set, performance improved substantially to RMSE = 9.12, $R^2 = 0.7350$. Specifically, we (1) removed “uninformative” reviews whose aspect sentiment scores were all ~0, and (2) applied residual-based outlier filtering by dropping extreme-error cases (e.g., potential sarcasm/noise) and retraining on the remaining data.



Most Important Regression Drivers (Top 4)

In the “Feature Importance: What drives game ratings?” output, the top three features are:

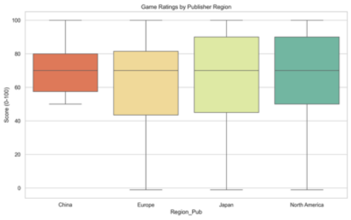
1. **Region_Dev_Japan**: Japan-developed games are associated with higher predicted scores.
2. **Sentiment_gameplay_mechanics**: consistent with our expectation that gameplay is a major determinant of ratings.
3. **Region _Dev_North America**
4. **Sentiment_audio_music**: an unexpected top driver, suggesting that audio/music quality signals in reviews contribute strongly to overall evaluation even when not emphasized in our initial hypothesis.

Interpretation

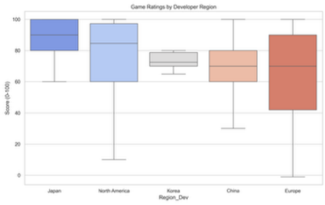
The strong Japan developer effect can be interpreted partly through a “country-of-origin / reputation halo” lens: users and possibly critics may bring prior impressions about Japanese developers or Japanese games that influence judgments of current titles. In addition, our review sampling strategy (e.g., selecting comments from the starting and ending pages of the Metacritic) could amplify perceived differences, so this explanation is suggestive rather than causal. Some factors (e.g., controls/camera sentiment) show negative coefficients, indicating potential penalties on overall score.

3b. Figures

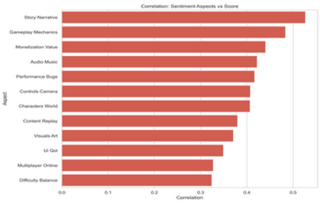
- **Fig.1 Publisher region vs. score, boxplot.** The x-axis groups games by publisher region and the y-axis shows the overall rating; Each box shows the IQR with the median line; whiskers indicate the overall spread. Publisher-region medians are very similar (around ~70), suggesting limited association.



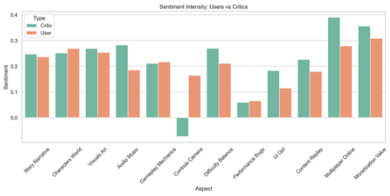
- **Fig.2 Developer region vs. score, boxplot.** The plot uses the same boxplot elements as Fig.1 but groups games by developer region to capture production-side differences. Developer region shows clearer separation than publisher: Japan has a noticeably higher distribution (median close to ~90), while North America and Europe have larger samples but wider spreads and more low-score outliers; Korea/China are exploratory due to smaller counts.



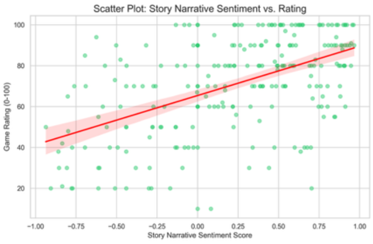
- **Fig.3 Aspect sentiment–score correlation, bar chart.** The x-axis shows the correlation between aspect-level sentiment and the y-axis lists 12 review aspects sorted from highest to lowest correlation. Story Narrative (~0.53) and Gameplay Mechanics (~0.49) rank highest, indicating they function as core "evaluation anchors," whereas Multiplayer Online and Difficulty Balance are lowest, likely because they are more preference-dependent and not relevant to every player.



- **Fig.4 Critic vs. User aspect sentiment comparison, grouped chart.** For each aspect, we compare the average sentiment strength in critic vs. user reviews. Critics are more positive on Multiplayer Online and Monetization Value, while Controls/Camera shows a clear divergence (critics negative, users positive), suggesting critics apply stricter standards to usability/technical feel while users may be more tolerant or prioritize other elements.



- **Fig.5 Story sentiment vs. score, scatterplot with trend line.** Each point represents a game, with story sentiment on the x-axis and score on the y-axis; the upward trend line indicates an overall positive association. At the same time, the points are widely dispersed, implying that even the strongest single aspect cannot predict ratings on its own and that scores are driven by a combination of factors.



3c. Observations and Conclusion

Overall, publisher region alone does not explain ratings well, while developer region - especially Japan - shows a stronger relationship with scores. Gameplay mechanics sentiment is consistently important, and audio/music sentiment emerged as a surprisingly strong driver. Critic and user reviews differ in what they reward or penalize, which helps explain why the two score types can diverge. Finally, the continuously increasing R^2 indicates that our improved combined structured + NLP approach effectively identifies meaningful drivers and produces a transparent, reproducible analysis.

3d. Impact of findings

Practically, our results suggest that ratings are linked more to “who builds the game” than “who publishes it,” and that review text can surface concrete quality dimensions tied to ratings. For product or market analysis, combining metadata with aspect sentiment provides a clearer explanation of *why* scores differ across games and why critics vs. users may disagree.

Future Work

With more time, we would (a) expand the dataset beyond 259 games and rebalance regions to reduce small-sample bias; (b) improve aspect extraction and domain-adapt sentiment scoring to reduce lexicon errors; and (c) test non-linear models such as Random Forest to capture interactions (region × genre × sentiment) that a linear regression may miss. We would also specifically test the robustness of the Japan developer effect under alternative sampling strategies and additional controls.