DOI: 10.54254/2754-1169/2026.NJ31379

Beyond Data Leakage: Robust Pre-Release Forecasting of North American Video Game Sales

Zichuan Zhan

NYU Tandon School of Engineering, New York University, New York, USA
zz5283@nyu.edu

Abstract. The video game market is large yet volatile, making reliable sales forecasts critical for publishers and retailers. Many prior approaches overlook temporal dependencies and risk data leakage, inflating accuracy. This study presents a leakage-free ensemble framework for forecasting North American video game sales. Time-aware features, specifically lagged “previous-title” sales and three-year rolling trends by genre, platform, and publisher, are engineered strictly from past information. Using VGChartz-derived historical data, Random Forest (RF), Gradient Boosting (GB), and histogram-based Gradient Boosting (HGB) are evaluated under a time-based train/test split and time-series cross-validation. Results show that the HGB consistently outperforms linear baselines, achieving a Coefficient of Determination (R^2) of approximately 0.384, along with the lowest Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Feature-importance analyses highlight prior-title sales and publisher trends as dominant predictors. These findings demonstrate that avoiding leakage and incorporating temporal structure significantly improve forecast reliability, offering valuable insights for sales planning and strategy.

Keywords: video game sales, feature engineering

1. Introduction

Predicting video game sales is critical due to the industry’s rapid growth and high stakes. The global video game market generates billions in revenue, and anticipating demand helps developers and publishers plan production, marketing, and distribution. Sales forecasts guide strategic decisions in pricing and inventory management. For example, Wu notes that different gaming platforms historically have produced markedly different sales outcomes, underlining the need for nuanced predictive modeling [1]. Similarly, Huang emphasizes the complex mix of factors such as genre, platform, marketing, and other influences that affect sales and the challenges they pose for forecasting [2]. Given this importance, robust forecasting methods that leverage rich game data while avoiding pitfalls are essential.

Prior studies have applied various machine learning approaches to game sales prediction, with ensemble methods showing strong performance. Wu evaluates classical algorithms including linear regression (LR), support vector machine (SVM), random forest (RF) and gradient boosting (GB) on a 16,719 game global dataset and reports that gradient boosted trees yield the most accurate predictions [1]. Huang compares neural networks, extreme gradient boosting (XGBoost) and Light

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.NJ31379

Gradient Boosting Machine (LightGBM) on similar data and finds that tree based ensembles can achieve lower RMSE than neural models [2]. Affana et al. also use Gradient Boosting Regressor (GBR), Ridge, and other regressors on the same Kaggle VGChartz data, confirming that ensemble regressors reliably predict game sales [3]. Li and co-authors study an eight-week-ahead game sales forecasting task based on VGChartz data and show that a hybrid feature selection scheme which combines correlation based screening with random forest importance can substantially improve the accuracy of tree based models such as LightGBM [4]. A recent literature review by Saptadi et al. notes that ensemble techniques, in particular XGBoost and RF, consistently outperform simple regressors in sales forecasting tasks [5]. These studies highlight the advantages of ensembles, which can capture nonlinear effects and interactions among features.

However, several recent studies rely on post-release signals and random splits, which can inflate accuracy. Huang includes ranking-style outcomes observed only after launch [2]. Affana uses critic and user ratings in work published in 2025 that are unavailable pre-release [3]. Wu reports results under a conventional 80/20 random split rather than a release-year split, ignoring intuitively important predictive features such as the sales volume of the previous work and the trend of the publisher [1]. By contrast, this work builds features strictly from pre-release information, including lagged previous-title sales and three-year rolling trends, and evaluates with release-time splits and time-series cross-validation, thereby preventing temporal leakage and matching real-world forecasting.

2. Method

2.1. Dataset

A publicly available VGChartz-derived video game sales dataset from Kaggle is used in this study. The dataset covers approximately 16.7 thousand titles and reports release year, genre, platform, publisher, and regional sales, including North America [6]. The analysis is restricted to games released between 1980 and the last year represented in the dataset. Records with missing release year or missing North American sales are removed, and all quantitative fields are converted to numeric types. North American sales are measured in millions of units. To stabilise variance and mitigate the right skew caused by blockbuster titles, the target variable is transformed using the natural logarithm of one plus the North American sales.

2.2. Model

Three ensemble regression models are employed in the analysis: RF, GB, and HGB. RF is a bagging-based ensemble of decision trees that reduces variance and is robust to noisy features. It provides built-in measures of feature importance and handles high-dimensional data effectively. In this study, RF consists of 300 trees with a maximum depth of 20. GB builds trees sequentially in order to minimise residual errors and is known for high predictive accuracy on tabular data, although it may overfit if not properly tuned. GB is configured with 200 trees, a maximum depth of 6, and a learning rate of 0.1. HGB is a more recent algorithm that also builds additive trees but relies on histogram-based splits to improve efficiency, which makes it faster on large datasets. In the configuration used here, HGB runs for 200 boosting iterations and uses a maximum depth of 6 and a learning rate of 0.1, matching the depth and learning rate used for GB. HGB can capture non-linear interactions among features like GB. Previous research, including work by Wu and by Saptadi and

DOI: 10.54254/2754-1169/2026.NJ31379

co-authors, suggests that ensemble models typically outperform simpler linear regressors for sales forecasting tasks [1,5].

2.3. Feature engineering

Each model takes as input a set of engineered features. Let y_i^N denote the North American sales of game i and F_i and S_i denote its franchise and platform, and let t_i be its release year. The first feature, previous game sales, is defined as the North American sales of the most recent eitle in the same franchise and on the same platform. Formally, for each game i the previous-title feature P_i is given by

$$P_i = y_j^N \tag{1}$$

where game j satisfies $F_j = F_i$, $S_j = S_i$ and $t_j < t_i$, with t_j as large as possible. If no such title exists on the same platform, the search is extended to earlier titles in the same franchise on any platform; if no earlier title exists at all, P_i isis set to zero.

Trend features summarise recent market history at the publisher, platform and genre levels. For each publisher p and year t, let $\bar{y}_{p,t}$ be the mean North American sales of all games released by ppp in year t. The three-year publisher trend is a backward-looking moving average over past years,

$$T_{p,t}^{pub} = \frac{1}{m_{p,t}} \sum_{k=1}^{m_{p,t}} \bar{y}_{,t-k} \tag{2}$$

where $m_{p,t}$ is the number of available years before t, capped at three. Analogous trends $T_{s,t}^{plat}$ and $T_{g,t}^{gen}$ are defined for each platform s and genre g.

In addition to sales-based trends, three-year rolling counts of releases are constructed. For example, let $N_{s,t}^{plat}$ denote the number of titles released on platform sss in year ttt. The corresponding rolling count is

$$C_{p,t}^{plat} = \frac{1}{m_{s,t}} \sum_{k=1}^{m_{s,t}} N_{s,t}^{plat} \tag{3}$$

with $m_{s,t}$ defined in the same way as above. Similar rolling counts are computed for publishers and genres.

Share features normalise these rolling counts by the overall activity in the market. Let C_t^{total} denote the three-year rolling count of all titles released in year ttt. The platform, genre and publisher market-share features are then

$$M_{s,t}^{plat} = \frac{C_{s,t}^{plat}}{C_t^{total}},\ M_{g,t}^{gen} = \frac{C_{g,t}^{gen}}{C_t^{total}},\ M_{p,t}^{pub} = \frac{C_{p,t}^{pub}}{C_t^{total}} \tag{4}$$

Categorical variables are encoded in a way that preserves interpretability. The twenty most frequent publishers are represented as separate categories and all remaining publishers are grouped into a single other category. Platforms and genres are label encoded, and any category that does not appear in the training period is mapped to a dedicated unknown class. The release year is retained as a numeric feature indicating temporal position. No aggregated or time-derived feature uses information from the game’s own release year or from any future year.

DOI: 10.54254/2754-1169/2026.NJ31379

2.4. Experimental design

The data are split chronologically to avoid data leakage. Games released in 2014 or earlier form the training set, and games released from 2015 to 2022 form the test set. This setting mimics a realistic forecasting scenario in which only past data are available when predicting future titles. As Sun et al. highlight, such expanding-window cross-validation prevents “information leakage across temporal boundaries” and yields more reliable error estimates [7]. After splitting, missing feature values are imputed as zero, and continuous features are standardised where appropriate. Categorical encoders for platform, genre and the set of high frequency publishers are fitted on the training data only, and any category that appears later in validation or test years is mapped to a dedicated unknown class. All models are trained on sales after a logarithmic transformation. Performance is then measured on the test set on the original sales scale by applying the inverse transformation to the predictions. The reported metrics are R squared, root mean square error RMSE and mean absolute error MAE for North American sales, which allows comparison with earlier studies. In particular, RMSE in millions of units is the main error measure reported by Wu and by Huang [1,2]. Experiments are run for the ensemble models RF, GB and HGB.

3. Result

Figure 1. Feature correlation heatmap (photo/picture credit: original)

The engineered features show plausible relationships with game sales. The correlation matrix in Figure 1 reports Pearson correlations between each feature and the logarithm of North American sales. The three year publisher trend feature has the strongest positive correlation with the log sales target, approximately 0.38, followed by the previous game sales feature at about 0.34. The platform trend feature and the genre and publisher market share features have mild to moderate correlations with the target, with values between roughly 0.08 and 0.23, which still indicates useful predictive signal. The static categorical encodings for publisher, platform and genre are not displayed as numeric variables in the matrix but contribute indirectly through the ensemble models. No pair of features exhibits very high mutual correlation, with all pairwise coefficients below 0.8, which suggests limited multicollinearity among the engineered features. The moderately correlated trend and share features support the design choice of summarising multi year history because they capture gradual shifts in demand without introducing redundant information.

DOI: 10.54254/2754-1169/2026.NJ31379

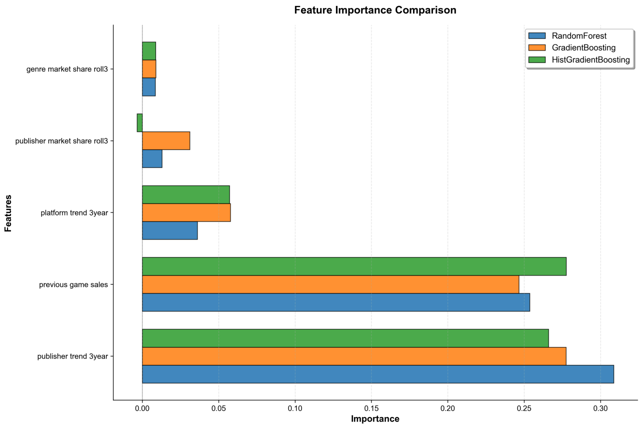

Figure 2. Permutation importance of ensemble models (photo/picture credit: original)

Figure 2 compares permutation importance scores from the three ensemble regressors, averaged over trees. Across all models the publisher three year trend and the previous game sales feature dominate, each receiving importance values in the range 0.25 to 0.30, whereas all other features remain well below 0.10. HGB assigns the highest weight to previous game sales, while RF and GB give slightly more weight to the publisher trend, but in every case these two temporal signals are clearly the primary drivers. The platform trend feature forms a second tier of predictors, followed by the genre and publisher market share features, whose contributions are small but non-zero. This ranking is consistent with the correlation analysis in Figure 1 and indicates that both franchise momentum and publisher-level history carry much stronger predictive information than broader market share effects. Overall, the feature importance profiles confirm that the engineered temporal structure, in particular lagged sales and multi year trends, is central to the ensembles' forecasting performance, in line with the trend based findings reported by Gray and co authors [8].

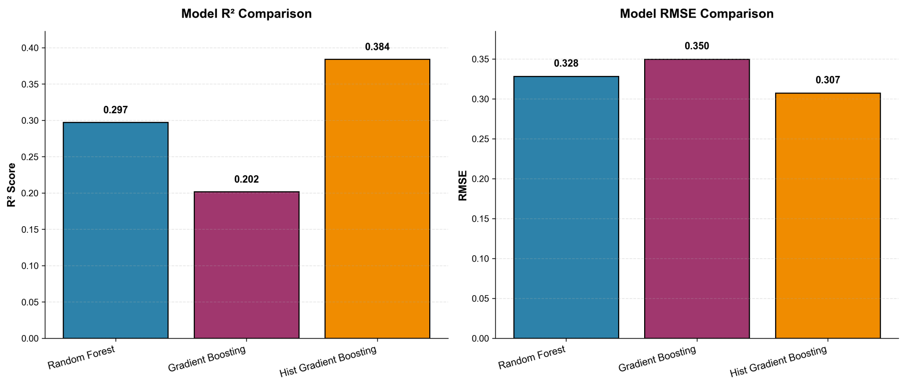

Figure 3. Performance comparison of ensemble models (photo/picture credit: original)

In terms of forecasting accuracy, the three ensemble models show clear differences in Figure 3. HGB achieves the highest test R-squared, about 0.384, and the lowest RMSE, about 0.307. RF attains a test R-squared of roughly 0.297 with RMSE around 0.328, while GB performs worst with a test R squared near 0.202 and the largest RMSE of about 0.350. These values indicate that even the best ensemble leaves a substantial amount of variance unexplained, which is typical for sales prediction tasks, yet HGB delivers a noticeable improvement over RF and especially over GB. This performance ranking is consistent with earlier findings by Wu and by Huang, where boosted tree ensembles tend to provide more accurate sales forecasts than simpler methods.

DOI: 10.54254/2754-1169/2026.NJ31379

The results highlight the value of carefully designed temporal features and ensemble learning for game sales forecasting. Among all predictors, the three year publisher trend and the previous game sales feature are consistently dominant, and the platform trend feature forms a clear second tier. These patterns indicate that franchise momentum and publisher level history carry the most informative signals for future demand, with broader platform and genre trends playing supporting roles. At the same time, the best model, HGB, reaches a test R squared of only about 0.38, which underlines that video game sales remain intrinsically hard to forecast.

4. Result

A central contribution of this study is the explicit control of temporal leakage. As Borycki and Mucci notes, even subtle use of future information can inflate reported accuracy by a large margin [9,10]. The experimental design in this work keeps training and evaluation strictly separated in time and constructs every temporal feature from past years only. Aggregations are shifted back by one year and time series cross validation respects the chronological order of releases. The moderate but realistic error levels observed in the test results contrast with the overly optimistic scores often reported under random splits or when post release signals are used. This suggests that much of the apparent performance in earlier studies may reflect leakage rather than genuine predictive power.

The feature importance profiles offer additional interpretability. They show that strong franchise histories and favourable publisher trends materially raise predicted sales, whereas changes in platform and genre level market share have smaller but still meaningful effects. For decision makers, this implies that investment decisions around sequels and publisher portfolios are likely to have greater impact than marginal shifts between platforms or genres. The analysis therefore complements market analytics work that emphasises the role of long term catalogue management and publisher branding in entertainment industries.

Several limitations s hould be acknowledged. The data are restricted to North American sales, so regional differences in consumer behaviour are not captured. Important drivers such as marketing expenditure, competitive releases and social media attention are absent from the feature set, which likely explains part of the remaining variance. The VGChartz derived dataset may also contain measurement noise and missing titles. In addition, the trend window is fixed at three years; alternative horizons such as five year or adaptive windows could change the relative importance of recent and older history. A further limitation is the heavy concentration of games with very low demand. Trněný reports in a Steam based study that about one third of titles have fewer than one concurrent player on average during the first two months after release and that a naive baseline which always predicts a very low player count can match more sophisticated models on these games [11]. The sales data analysed here exhibit a similar long tail, so the proposed models are expected to be more informative for mid and high selling releases than for very small games, where outcomes remain difficult to distinguish from random noise.

These limitations point to avenues for future work. One direction is to incorporate external signals, such as web search interest, pre release coverage or social media activity, while preserving the same strict temporal separation. Another is to explore richer modelling families, including deep learning architectures, again under leakage free evaluation. From an applied perspective, the present models are most suitable for supporting medium term planning decisions, such as production quantities and marketing budgets for upcoming releases, rather than for precise point forecasts at the level of individual titles. By combining transparent ensemble methods with principled temporal design, the study provides a cautious but practically useful framework for forecasting game sales without overstating predictive certainty.

DOI: 10.54254/2754-1169/2026.NJ31379

5. Conclusion

This study introduced a leakage free ensemble framework for forecasting video game sales based on explicitly temporal features. Time aware predictors such as lagged franchise sales and multi year trends at the publisher, platform and genre levels were constructed using only information that would have been available before release and were evaluated under strictly chronological data splits. The ensemble models RF, GB and HGB all benefited from these features, with HGB delivering the strongest test performance and RF ranking second. Feature importance analyses consistently highlighted prior sales and publisher level trends as the most influential drivers, which confirms the central role of franchise momentum and market trajectory in game sales forecasting. Taken together, the results show that rigorous temporal validation combined with transparent ensemble models can produce more trustworthy forecasts than approaches that ignore the time dimension.

The framework also has limitations. It relies on a VGChartz derived dataset that may omit digital only releases and may contain measurement noise, and it does not incorporate external drivers such as marketing expenditure, competing releases or social media attention. Overall, the models provide informative forecasts and useful relative rankings for mid and high selling titles, whereas predictions for small scale or independently developed games should be interpreted as coarse signals rather than precise estimates. Under this interpretation, the proposed framework is best viewed as a decision support tool for publishers and larger studios when managing portfolios, allocating resources and assessing risk, rather than as a mechanism for making fine grained revenue commitments for individual low volume projects.

References

[1] Wu, K. (2025). Machine learning models-based video game sales prediction. In Proc. of the 1st Int. Conf. on E-commerce and Artificial Intelligence (ECAI 2024) (pp. 45–50).
[2] Huang, W. (2023). Research on the prediction on the sales of electronic games. In Highlights in Business, Economics and Management (GEFHR 2023) (Vol. 21, pp. 13–21). Springer.
[3] Affana, A., Vishwakarma, S., & Kumar, S. (2023). Video game sales prediction model using regression model. SSRN Electronic Journal.
[4] Li, J., Zheng, Y, Hu, H., Lu, J., & Zhan, C. (2021). Predicting video game sales based on machine learning and hybrid feature selection method. In Proceedings of the 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE) (pp. 1–8).
[5] Saptadi, S., Budiawan, W., Fadhila, R. N., & Saputra, D. W. N. (2025). Machine learning applied to sales prediction modeling: A systematic literature review. International Journal of Intelligent Systems & Knowledge Engineering, 15(3), 45–60.
[6] rush4ratio. (2016). Video Game Sales With Ratings. Kaggle. https: //www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings
[7] Sun, Y., Qu, Z., Zhang, T., & Li, X. (2025). Adaptive ensemble learning for financial time-series forecasting: A hypernetwork-enhanced reservoir computing framework with multi-scale temporal modeling. Axioms, 14(8), 597.
[8] Gray, D., Shalan, A. M., & Kadlec, C. (2024). Sales predictions for video games using predictive analytics of market data. In Proceedings of the 2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 1510–1516). IEEE.
[9] Borycki, B. (2025, August 7). Data leakage in time-dependent feature engineering. BitPeak Blog. Retrieved from https: //bitpeak.com/data-leakage-in-time-dependent-feature-engineering/
[10] Mucci, T. (2024). What is data leakage in machine learning? IBM. Retrieved from https: //www.ibm.com/think/ai-impact/data-leakage-in-machine-learning
[11] Trněný, M. (2017). Machine learning for predicting success of video games (Master's thesis, Masaryk University, Faculty of Informatics).