Forecasting Global Video-Game Sales Using NLP and Metadata Features

Team: Eason Lin (ylin5170@usc.edu), Max Clark (maclark@usc.edu), Dazhou Luo (dazhoulu@usc.edu) , Suer Yang (yangsava@usc.edu)

GitHub Repository: https://github.com/maxaclark/forecasting-video-game-sales

1. Problem Definition

This project forecasts the number of video-game copies sold globally using structured metadata and NLP-based title features. We frame the task primarily as regression and secondarily as a three-class sales-tier classification. The overarching objective is to understand whether platform, publisher, year, genre, and title semantics can reliably explain or predict commercial success.

2. Background and Motivation

2.1 Why This Problem Is Meaningful

The global video-game industry surpasses $200B annually, and publishers make multi-million-dollar decisions about release timing, marketing budget, and platform targeting. Accurate sales forecasting enables studios, analysts, and investors to evaluate commercial potential before launch. Understanding the underlying drivers of success - such as IP franchise strength, genre trends, or platform cycles - can support more effective decision-making in product planning and portfolio management.

2.2 What We Aim to Solve

We target two core goals: Predict global sales from metadata + title and identify interpretive factors most associated with high sales outcomes. We address two questions:

- Can structured metadata (genre, year, punisher, regional sales) explain and predict sales outcomes?
- Do NLP-derived title features meaningfully improve prediction accuracy beyond metadata alone?

2.3 Related Work

Past research uses metadata-only linear regression, tree models, or Naive Bayes classifiers. Most prior work does not deeply leverage NLP or contextual embeddings to capture title semantics. Existing studies show moderate accuracy but limited explanatory power.

2.4 Our Contribution

Our main contributions are:

- **Use of BERT embeddings** to encode semantic information in game titles, a technique largely missing from prior literature.
- **Novel FranchiseTag feature**, which detects sequels or recurring franchises using rule-based patterns, improving prediction for blockbuster titles.
- **Platform consolidation strategy** that reduces noise caused by duplicated entries for multi-platform releases.
- **Comprehensive multimodel comparison** (KNN, XGBoost, LightGBM, neural networks, SVM), enabling evaluation across regression and classification frameworks.
- **Interactive Streamlit dashboard** for real-time exploration of predictions and feature importance.

3. Dataset Description

3.1 Dataset Quantity and Statistics

Dataset: Kaggle Video Game Sales; 16,598 records (16,324 after cleaning). Key variables include title, platform, genre, publisher, year, regional sales, and the regression target **Global_Sales**. Global_Sales and the regional sales columns represent millions of copies of each game that was sold worldwide or in the respective global region (ie. North America vs. Europe vs. Japan). The distribution of Global_Sales is highly right-skewed: a few blockbuster IPs dominate overall revenue.

3.2 Dataset Quality & Limitations

The dataset provides strong longitudinal coverage, diverse genres and platforms, and a useful mix of numeric and text features. Its limitations include missing publisher data, inconsistent platform labels, approximate VGChartz sales figures, and the absence of key external factors such as marketing budgets.

3.3 Variables & Data Types

Categorical: Genre, Publisher, Platform (later removed after consolidation). **Numeric:** Year, Regional Sales. **Text:** Name (processed via NLP). **Target:** Global Sales (continuous)

4. Methods and Approach

4.1 Preprocessing & Feature Engineering

Steps performed:

- Removal of invalid or missing rows (274 dropped)
- Normalization and cleaning of title text
- Imputation of missing publishers with “unknown”
- Year→int conversion
- **Platform-consolidation:** identical titles across platforms had sales aggregated to reduce duplicated noise
- **FranchiseTag:** rule-based detection of sequels/series (e.g., “II”, “3”, “2020”)

4.2 NLP Methods

- **BERT embeddings:** Contextual token-level representation; strong for short specialized titles.

4.3 Machine Learning Algorithms

For our baseline model, we started with K-Nearest Neighbors (KNN) regression. We tuned it using grid search cross-validation using the GridSearchCV package from scikit-learn. The best model considered the 12 closest neighbors, weighted by distance, using Euclidean distance as the metric. Next we trained and tuned an XGBoost model over all relevant parameters. The best

XGBoost tree allowed 80% column samples per tree, had a learning rate of 0.1, max depth of 5 per tree, 200 total trees, and allowed 100% of the training samples to each tree in the ensemble. We followed this with a LightGBM. Relevant tuned parameters for the LightGBM include 36 leaves per tree, max depth of 11, learning rate of 0.023, and 228 total trees. For regression, the final model we trained was a feed-forward neural network. It had 4 total layers with ReLu activation in the hidden and input layers and linear activation in the output layer for a regression output. It was trained over 50 epochs using the Adam optimizer. To better understand the results of the neural network we binned the target variable and predictions into “Low”, “Medium”, and “High” categories. “Low” was defined as any game with less than one million global sales, “Medium” was one million to 5 million global sales, and “High” was any game with over 5 million sales, in accordance with the Putra et al. 2025 study. This made the neural network a sort of pseudo-classifier despite giving numerical outputs. To compare these classification results we also built and trained a Support Vector Machine due to its prowess in dealing with high-dimensional data.

4.4 Evaluation Metrics

For Regression, we used MAE as the primary metric as the target variable was heavily skewed. For Classification, we focused on F1-score due to imbalanced classes, with smaller emphasis on accuracy.

5. Experiments

5.1 Experiment Setup

Python 3.11; libraries: pandas, NumPy, Matplotlib, scikit-learn, XGBoost, LightGBM, PyTorch/TensorFlow, BertTokenizer/BertModel. MLflow used for experiment tracking. Train/valid/test split applied temporally to avoid leakage.

5.2 Training Details

Numeric features scaled; categorical features one-hot encoded; NLP embeddings concatenated with metadata. Neural network trained over 50 epochs with early stopping. XGBoost, LightGBM, and SVM tuned via grid search.

5.3 Model Comparisons

The KNN model served as a baseline to show proof of concept. It performed decently well (MAE of 0.723 million sales copies) given its simplicity, but it struggled with overfitting (validation MAE of 0.239). In contrast, the neural network provided the best balance between MAE and stability, showing strong performance on both validation and test sets. The testing MAE for the neural network was 0.639, but, more importantly, it was more effective at identifying "hit" games compared to the other models. The table below displays the 10 highest predicted values of the neural network.

	Name	Actual_Global_Sales	Predicted_Global_Sales
1480	call of duty black ops 3	25.32	21.851910
982	nintendogs	24.76	17.653006
575	fifa soccer 13	16.16	15.350302
175	call of duty advanced warfare	21.90	15.319038
1643	the sims 4	2.97	12.301384
1108	grand theft auto 2	3.42	10.873234
97	wii fit	22.72	10.528362
603	fifa soccer 07	6.38	9.545191
1946	wii sports club	0.40	9.358070
1660	assassins creed iii	13.10	9.312297

The gradient-boosting models, including XGBoost and LightGBM, were competitive with regards to MAE (0.602 and 0.724 respectively), but they achieved this by primarily predicting lower values for global sales – they did not handle the skewness of the data. Below are the 10 highest predicted values for the XGBoost model.

	Name	Actual_Global_Sales	Predicted_Global_Sales
1532	madden nfl 09	7.14	5.404613
1108	grand theft auto 2	3.42	5.195892
743	new super mario bros wii	28.62	4.519134
2072	madden nfl 11	5.94	4.460551
564	the legend of zelda	6.51	4.118402
68	mario party ds	9.02	3.784997
949	madden nfl 99	2.64	3.594989
713	super mario galaxy 2	7.69	3.533603
2025	super mario advance	5.49	3.220605
1643	the sims 4	2.97	3.153491

When applied to classification, the neural network achieved a macro F1-score of 0.574, with "Low", "Medium", and "High"-specific F1-scores of 0.9, 0.34, and 0.48 respectively. In comparison, the SVM outperformed the neural network with a macro F1-score of 0.61, and class-specific F1-scores of 0.91, 0.38, and 0.54 respectively, improving in every class.

5.4 Error Analysis

In our error analysis, we observed that most residual errors were driven by a few specific patterns. Outlier franchises such as Pokémon, Mario, and Wii Sports produced unusually high sales that were difficult for models to predict accurately. Additionally, niche genres often showed unpredictable spikes that increased variance in the predictions. Transitional console-generation releases also contributed to errors, as these periods tend to create irregular shifts in market behavior.

6. Observations

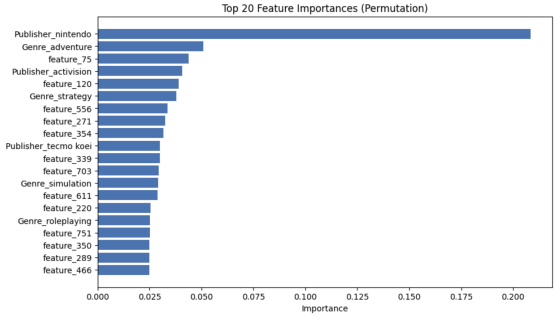
6.1 Key Findings

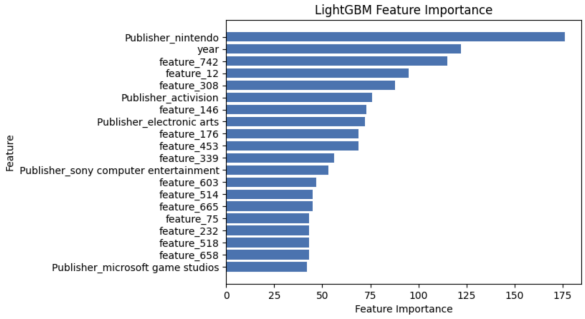
1. **Franchise keywords** in titles strongly correlate with high sales.
2. **Year** and **platform family (post-consolidation)** remain major predictors.
3. **NLP features consistently improve predictions** over metadata-only baselines.
4. Neural network outperforms boosting models and linear baselines.

6.2 Interpretation of Results

While publisher features dominate feature importance plots (namely Nintendo), title semantics similarly encode brand strength (e.g., "Mario", "FIFA", "Call of Duty"), enabling models to capture franchise value. Market expansion and contraction periods (e.g., 2000s boom, post-2012 decline) explain year-based predictive strength. Genre plays a secondary but meaningful role. Below are the feature importance plots for the neural network (left) and the LightGBM (right).

Unnamed features in the plots represent game title tokenizations.





6.3 Strengths & Limitations

Our approach has several strengths. By combining metadata with multiple NLP methods, we benefit from a multimodal modeling framework that captures both structured and semantic information. Including title embeddings, franchise tagging, and other text-based features enhances predictive power. The interactive dashboard provides an accessible way to interpret model outputs. Additionally, the neural network model showed strong generalization across validation and test sets, reinforcing the robustness of our methods.

However, the dataset relies on estimated sales figures, which introduces noise that may affect model accuracy. More importantly, several influential factors such as critic review scores, and console install base are not available in the dataset. Without these external variables, our models cannot fully account for all real-world drivers of commercial success.

6.4 Streamlit Dashboard Implementation

A prototype Streamlit dashboard was developed to show the user-facing component project.

The dashboard currently includes:

- Sales Trends Visualizer – Interactive filtering by Year, Genre, and Platform, with aggregated global sales plots and top-selling games.
- Prediction Interface (Stub) – UI for entering game attributes (title, genre, platform, etc.). Backend model integration is planned for the final version once the inference pipeline is packaged.
- Feature Insights (Planned) – Placeholder page where model interpretability outputs (feature importance, correlation heatmaps) will be displayed.

7. Conclusion

7.1 Summary of Findings

Combining metadata with NLP features noticeably improves video-game sales prediction. BERT embeddings and franchise tagging add meaningful signals that help models recognize well-known series and patterns in titles. Among all approaches, the neural network offered the most consistent performance, and feature importance results highlight clear factors such as publisher, platform, year, and franchise wording.

7.2 Impact & Implications

These findings provide valuable insights for publishers, analysts, and investors seeking to evaluate or forecast a game's market performance prior to release. Understanding which factors—such as platform timing, genre positioning, and the presence of franchise identifiers—most strongly influence sales can support more informed decisions in marketing, budgeting, and portfolio planning. Additionally, the Streamlit dashboard developed for this project enables real-time scenario exploration, making the predictive model more accessible for practical, industry-facing applications.

8. References

[1] Affan, ., Vishwakarma, S., & Kumari, R. (2024, August 23). *Video game sales prediction model using regression model* (Proceedings of the International Conference on Innovative Computing & Communication, ICICC 2024). SSRN. https://ssrn.com/abstract=4935036 (papers.ssrn.com)

[2] Putra, R., Ramadani, N., & Nanjar, A. (2025). Classification and Prediction of Video Game Sales Levels Using the Naive Bayes Algorithm Based on Platform, Genre, and Regional Market Data. *International Journal of Informatics and Information Systems, 8*(1), 12-21. doi:https://doi.org/10.47738/ijiis.v8i1.242

[3] Cheng, H. (2022). Video Game Sales Trends and Stats. Retrieved from https://www.researchgate.net/publication/363175427_Video_Game_Sales_Trends_and_Stats