Prediction of Video Game Sales Using Machine Learning Models

Ramakrishna Mission Vivekananda Educational & Research Institute

Belur Math, Howrah, West Bengal

Department of Computer Science

Machine Learning – Course Project Report

Student Name: Ananyo Sen Student Id: B2530066

1. Problem Statement

The video game industry continues to expand rapidly, with thousands of new titles released each year across multiple platforms. For developers, publishers, and analysts, predicting a game's potential sales performance is essential for strategic decision-making. Accurately *estimating global sales* helps *optimize marketing efforts*, manage *production budgets*, and *forecast revenue*, ultimately supporting data-driven planning in an increasingly competitive market.

This project aims to leverage historical sales data of *video games* to develop a **machine learning regression model** that predicts the **global sales** of a game (in millions of units). Using features such as **genre**, **year of release**, **platform**, **publisher**, and **regional sales distributions**, the model learns continuous relationships instead of discrete classes. The regression system estimates the *expected sales value* for a new or existing title, providing deeper numerical insights into **game performance trends**, **market potential**, and **feature influence**.

The objectives for this report are:

1. To analyze historical video game sales data and identify patterns related to game success.
2. To build a **regression model** that predicts the global sales of a game.
3. To compare and evaluate different machine learning algorithms for regression to determine the most effective methods.

This approach demonstrates the application of **data analytics and machine learning** in the entertainment industry, providing actionable insights for data-driven decision-making in game development and publishing.

2. Proposed Methodology

The objective of this methodology is to **predict the global sales of video games** using regression techniques and to **justify the choices** made for data preprocessing, feature engineering, model selection, and evaluation. The methodology follows a **step-by-step**, **data-driven approach** to ensure clarity, reproducibility, and accuracy in forecasting sales performance across various genres, platforms, and publishers.

I. Data Preprocessing & Feature Engineering

To prepare the dataset for machine learning, several preprocessing steps were undertaken:

- **Handling Missing Values**: Missing entries in the **Year** column were filled with the *median year*, and missing **Publisher** entries were replaced with "*Unknown*" to maintain data consistency.
- **Rare Category Grouping**: Publishers with fewer than five occurrences were grouped under "Other" to reduce encoding noise.
- **Regional Sales Features**: New proportion-based features (*NA_Share*, *EU_Share*, *JP_Share*, *Other_Share*) were created to represent each region's contribution to *Global_Sales*.
- **Feature Interactions**: Combined categorical features such as *Genre_Platform*, *Publisher_Platform*, and G*enre_Pub_Platform* were generated to capture cross-relationships.
- **Franchise Indicator**: A binary feature (*Is_Franchise*) was added to denote whether a game title appeared multiple times.
- **Dominance and Diversity**: *Max_Region_Share* identified the region with the highest sales share, while *Sales_Diversity* measured the spread of sales across regions.
- **Temporal Feature**: *Years_Since_Release* was calculated to capture the time difference from the most recent release year.

- **Feature Scaling**: The *Year* column was explicitly normalized using **StandardScaler** to ensure proper scaling for regression models.
- **Categorical Encoding**: Categorical variables (**Genre**, **Platform**, **Publisher**, and **other engineered**) were numerically encoded using Target Encoding with 0.3 smoothing based on their relationship with global sales.

II. Feature Selection

To enhance model performance and capture meaningful relationships:

- A comprehensive set of features was selected, including *Year*, *Years_Since_Release*, *Genre*, *Platform*, and *Publisher*.
- Additional engineered features such as *NA_Share*, *EU_Share*, *JP_Share*, *Other_Share*, *Max_Region_Share*, and *Sales_Diversity* were included to represent regional sales behavior.
- Interaction features like *Genre_Platform*, *Publisher_Platform*, and *Genre_Pub_Platform* were added to model complex relationships among categorical variables.
- The binary feature *Is_Franchise* was incorporated to account for repeated game titles, enhancing the model's predictive depth.

III. Model Selection

- **Linear and Regularized Regression Models: Linear Regression, Ridge Regression, and Lasso Regression**
  - Baseline models used to assess linear relationships. Ridge (L2) and Lasso (L1) regularization manage correlated and expanded predictors but cannot capture non-linear effects.
- **Tree-Based Ensemble Models: Random Forest, Gradient Boosting, and XGBoost Regressors**
  - Capture non-linear interactions among numerical and encoded categorical features, effectively modeling engineered relationships such as *Genre_Platform* or respective *Region Shares* and more.

IV. Model Training

- The dataset was divided into **training (70%)**, **validation (20%)**, and **testing (10%)** subsets to ensure reliable performance evaluation on unseen data.
- All regression models — including *Linear*, *Ridge*, *Lasso*, *Random Forest*, *Gradient Boosting*, and *XGBoost* — were trained on the encoded training set and evaluated using the validation and test subsets.
- Each model was initialized with optimized hyperparameters and a fixed random state (**42**) to maintain reproducibility.

V. Evaluation

- Regression performance was evaluated using **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R^2 Score** to assess prediction accuracy and model fit.
- These metrics were computed on both **validation** and **test** datasets to assess model generalization and prevent overfitting.
- Results from multiple regression models — Linear, Ridge, Lasso, Random Forest, Gradient Boosting, and XGBoost — were compared in a table.

VI. Visualization

- **R^2 Comparison Plot**: A bar chart compared validation and test R^2 scores across models, illustrating relative performance and consistency.
- **Residual Distribution**: Plots of residuals (actual − predicted values) of log-scaled values were generated for each model to analyze error patterns and model bias.
- **Regression Fit Plot**: Scatter plots of predicted versus actual global sales were used to visualize the goodness of fit and identify deviations from perfect predictions.
- **Feature Importance**: For both linear and tree-based models, feature importance charts were created to highlight the most influential variables driving global sales predictions.

3. Dataset Details

The dataset used in this study, **vgsales.csv** (sourced from **Kaggle**), contains over **16,000 video game titles** described by attributes such as *Name*, *Year*, *Platform*, *Genre*, *Publisher*, and regional sales (*NA_Sales*, *EU_Sales*, *JP_Sales*, *Other_Sales*), along with *Global_Sales*.

For regression modeling, **Global_Sales**, since was highly skewed and so was transformed on a log scale, served as the target variable, while features such as *Year*, *Genre*, *Platform*, and *Publisher* were used as key predictors and for other compound features engineered.

Missing Year values were imputed using the **median**, rare publishers were grouped under "Other," and categorical variables were numerically encoded using **Target Encoding**. The dataset was divided into **70% training**, **20% validation**, and **10% testing** subsets to enable robust model evaluation.

Attribute	Type	Description
Game	Categorical	Title of the video game
Year	Numerical	Year of release
Platform	Categorical	Gaming console or system on which the game was released
Genre	Categorical	Type or category of the game
Publisher	Categorical	Company publishing the game
NA_Sales, EU_Sales, JP_Sales, Other_Sales	Numerical	Regional sales (in millions)
Global_Sales	Numerical	Total worldwide sales (in millions)

Table 1: Dataset attributes from vgsales.csv

4. Comparative Analysis & Results

This section presents the comparative evaluation of various regression algorithms used to predict the global sales of video games. The goal is to analyze which model performs in which way on the given dataset based on key attributes such as Year, Genre, Platform, and Publisher, and other engineered features. Also, we would discuss the reason for the respective model performance.

A. Model Comparison

Six regression models — **Linear**, **Ridge**, **Lasso**, **Random Forest**, **Gradient Boosting**, and **XGBoost** — were trained on the same preprocessed dataset, split into 70% training, 20% validation, and 10% testing. Model performance was evaluated using MAE, RMSE, and R^2 Score on log scale value of Global_Sales.

Model	Val MAE	Val RMSE	Val R^2	Test MAE	Test RMSE	Test R^2	Remarks
Linear Regression	0.2188	0.3374	0.2879	0.2166	0.3260	0.3475	Captures linear trends but misses complex patterns.
Ridge Regression	0.2187	0.3375	0.2874	0.2167	0.3262	0.3466	Performs moderately well but limited by linear assumptions.
Lasso Regression	0.2195	0.3404	0.2753	0.2190	0.3298	0.3320	Slightly weaker due to coefficient shrinkage reducing model flexibility.
Random Forest	0.1261	0.2364	0.6505	0.1247	0.2281	0.6804	Captures non-linear patterns effectively with ensemble averaging.
Gradient Boosting	0.1180	0.2252	0.6826	0.1175	0.2151	0.7159	Learns complex relationships through sequential model improvement.
XGBoost	0.1222	0.2344	0.6563	0.1244	0.2286	0.6740	Achieves strong generalization via optimized boosting and regularization.

Table 2: Comparative Performance of Machine Learning Models

B. Discussion

The comparative results reflect how each model's inherent traits interact with the structure of the engineered dataset. **Linear Regression** provided a stable but limited baseline, as its assumption of linearity prevented it from capturing the complex, non-linear relationships introduced through engineered features such as *regional sales proportions*, *feature interactions*, and *sales diversity*.

Ridge Regression, through L2 regularization, slightly improved stability by shrinking coefficients and reducing overfitting, while **Lasso Regression** (L1 regularization) promoted sparsity by eliminating weak predictors — useful for high-dimensional data but at the cost of underfitting subtle patterns. These models could approximate general trends but failed to capture localized variations across genres or platforms.

Among tree-based models, **Random Forest Regressor** handled non-linear and categorical interactions effectively through ensemble averaging of deep decision trees. Its ability to model threshold-based splits across encoded categorical features led to improved accuracy, though the averaging process tended to smooth out extreme predictions.

Gradient Boosting Regressor excelled by sequentially correcting residual errors, combining weak learners to minimize bias and capture intricate interactions, particularly those derived from compound categorical features.

XGBoost achieved similar accuracy, aided by optimized gradient computation, built-in regularization, and parallel processing. However, its higher sensitivity to hyperparameters like learning rate and tree depth introduced minor fluctuations across validation and test sets.

Overall, the results indicate that boosting-based ensemble methods are best suited for modeling the complex, high-interaction, and non-linear relationships in the dataset, achieving a balanced trade-off between bias and variance.

C. Visualization and Interpretation

To visually support the quantitative findings, several plots were created. Results from each models are presented in this section.

(a) **R^2 Comparison Plot**

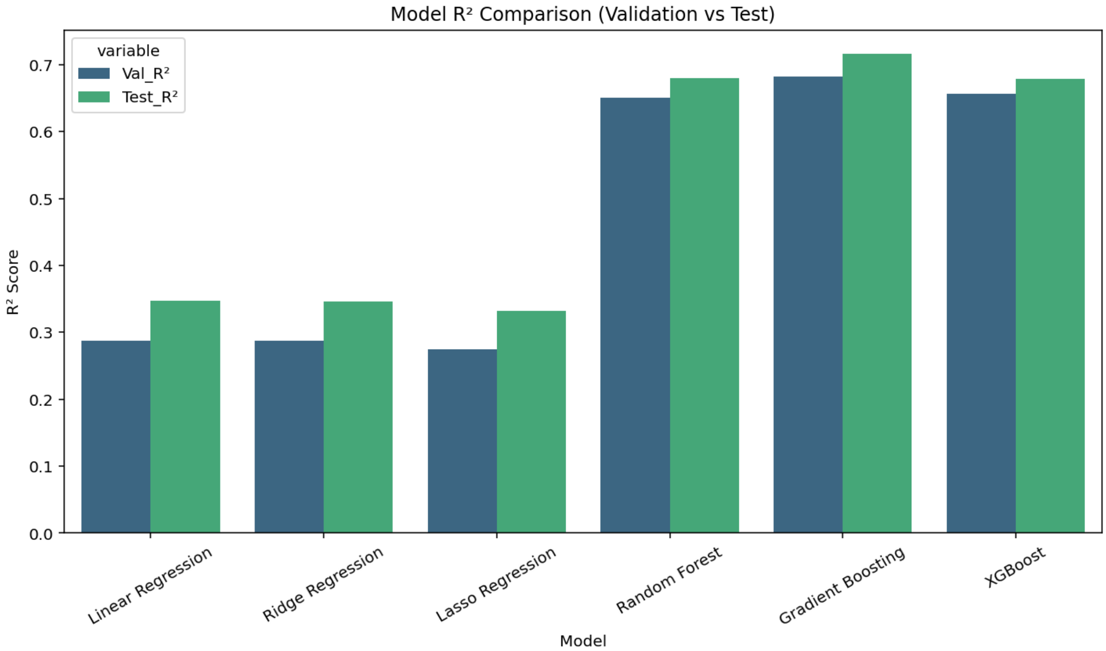
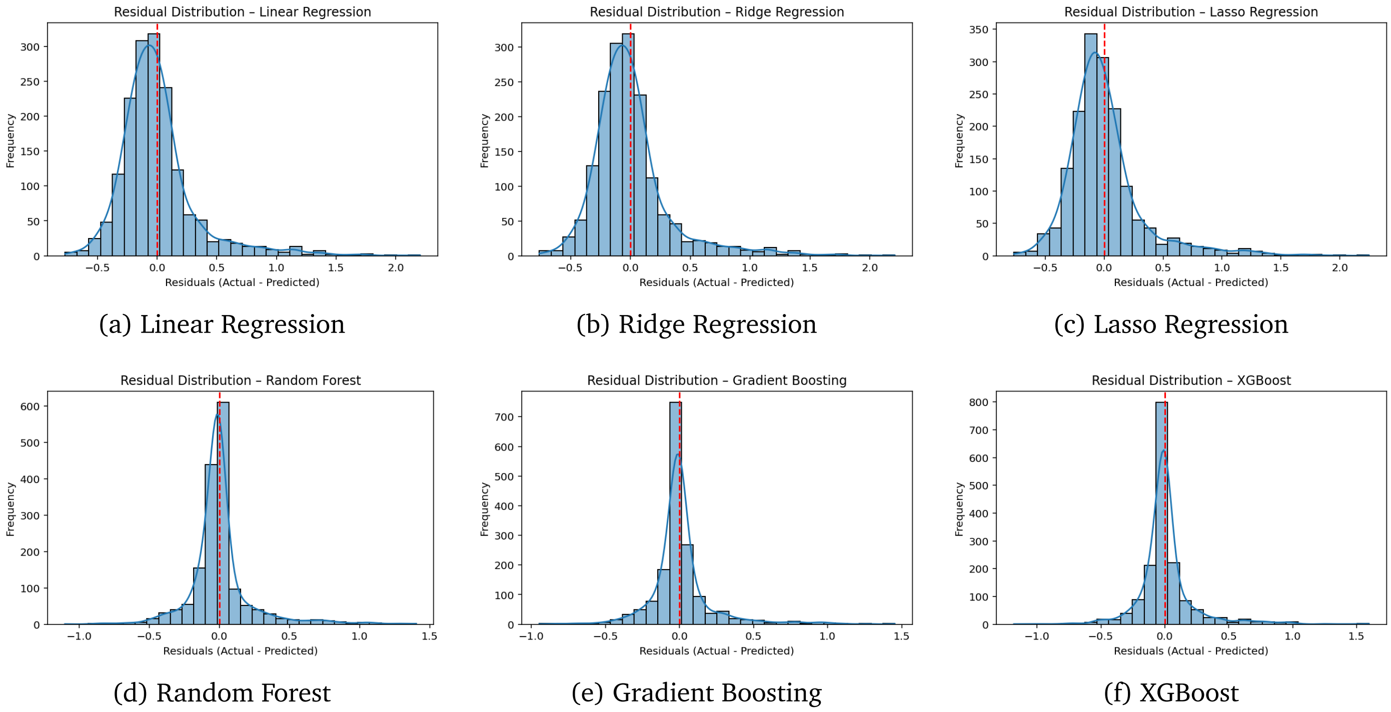


Figure 1: Model R^2 Comparison (Validation vs. Test).

Purpose: Compares model performance based on validation and test R^2 scores.
Highlights the superior generalization of ensemble models, particularly Gradient Boosting and XGBoost.

(b) Residual Error Distribution



(a) Linear Regression

(b) Ridge Regression

(c) Lasso Regression

(d) Random Forest

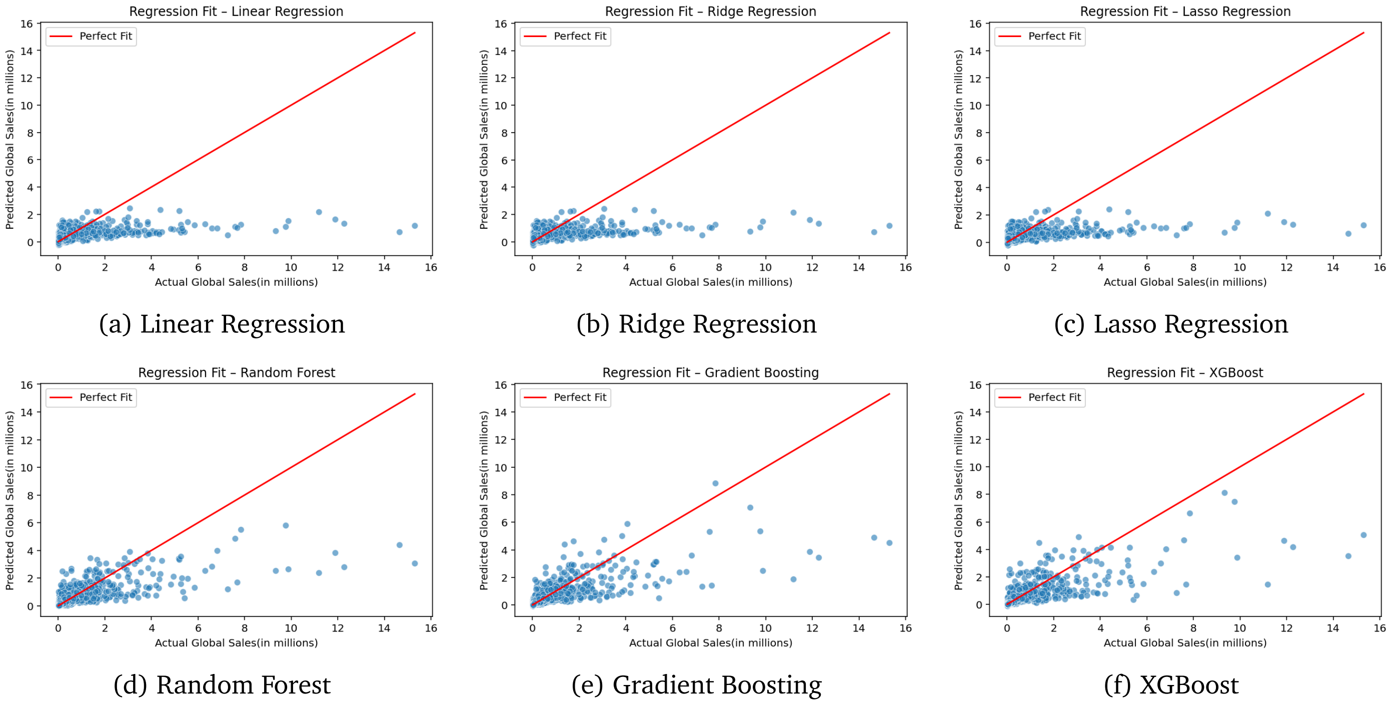
(e) Gradient Boosting

(f) XGBoost

Figure 2: Residual Distribution Plot for all six regressor.

Purpose: Visualizes the distribution of residual errors of log scale value for each regression model. A narrower and more centered spread around zero indicates higher predictive accuracy and reduced bias.

(c) Regression Fit



(a) Linear Regression

(b) Ridge Regression

(c) Lasso Regression

(d) Random Forest

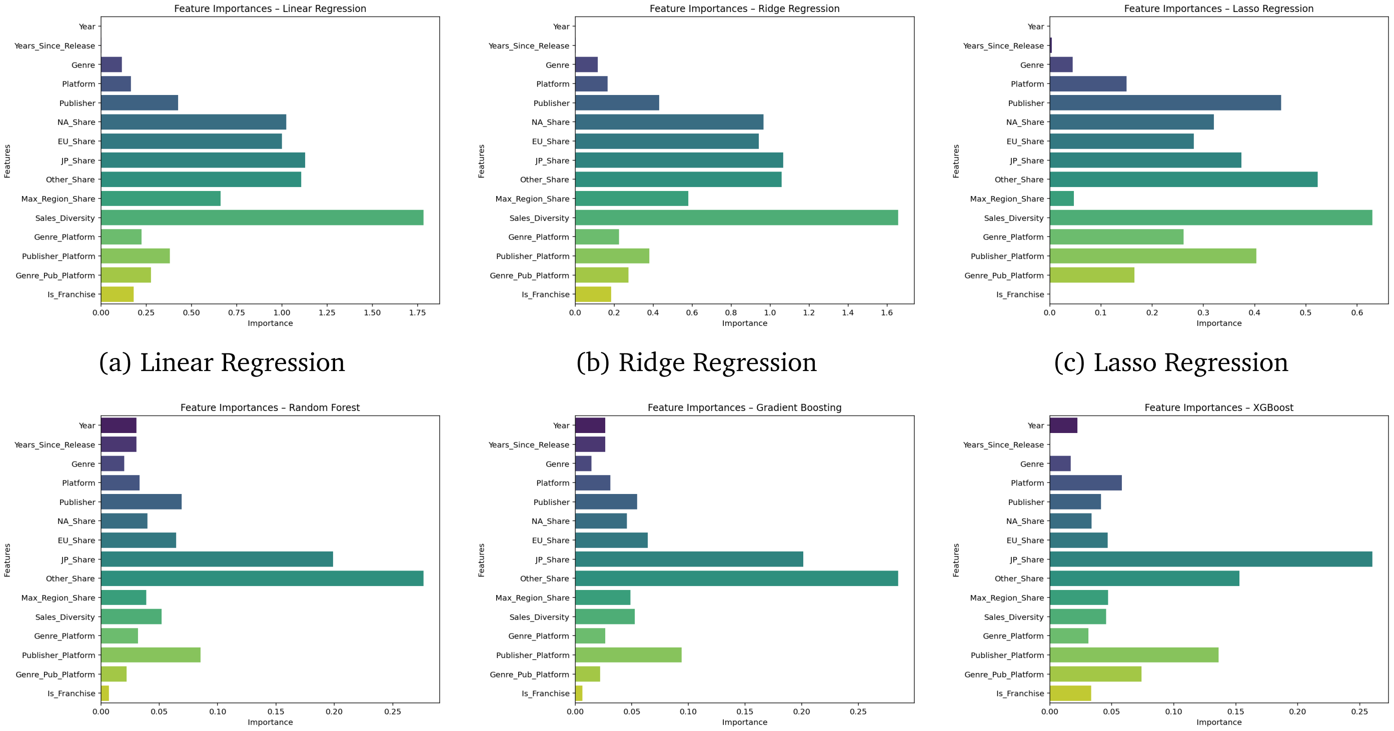
(e) Gradient Boosting

(f) XGBoost

Figure 3: Regression Fit Curve for all six regressors.

Purpose: Illustrates the relationship between actual and predicted sales values for each regression model. Points closely aligned along the diagonal line indicate strong model fit and accurate prediction performance.

(d) Feature Importance



(a) Linear Regression

(b) Ridge Regression

(c) Lasso Regression

(d) Random Forest

(e) Gradient Boosting

(f) XGBoost

Figure 4: Feature Importance Plot for all six regressors.

Purpose: Identifies the key features influencing predicted global sales across regression models. Regional sales shares, especially *Other_Share* and *JP_Share*, were the most influential predictors, with moderate effects from engineered features like *Publisher_Platform*, and *Sales_Diversity*.

5. Conclusion

This study applied multiple regression models to predict global video game sales using original and engineered features such as Year, Genre, Platform, Publisher, and regional sales distributions and other compound features. Linear, Ridge, and Lasso Regression provided interpretable baselines but were limited by their linear assumptions, failing to capture complex feature interactions.

Tree-based ensemble models—Random Forest, Gradient Boosting, and XGBoost—handled these non-linear dependencies more effectively through hierarchical splits and sequential learning. Gradient Boosting offered consistent performance, while XGBoost achieved comparable results with greater sensitivity to tuning.

Overall, the results indicate that ensemble tree-based methods are more suited for modeling the multi-factor relationships influencing video game sales and form a foundation for future work incorporating temporal or regional trends.

References

[1] Kaggle, "Video Game Sales Dataset," Available: `https://www.kaggle.com/gregorut/videogamesales`

[2] Scikit-learn Developers, "Scikit-learn Documentation," Available: `https://scikit-learn.org/stable/`

[3] XGBoost Developers, "XGBoost Documentation," Available: `https://xgboost.readthedocs.io/`

[4] GeeksforGeeks, "Introduction to Gradient Boosting," Available: `https://www.geeksforgeeks.org/introduction-to-gradient-boosting/`

[5] Towards Data Science, "Understanding Random Forests and Ensemble Learning," Available: `https://towardsdatascience.com/`