

5CS037 - Concepts and Technologies of AI

Regression Assignment

Video Game Sales Prediction Analysis

Name: Ganesh Dahal

University ID: 2511789

Biratnagar Faculty: Aayush Regmi

Module Leader: Siman Giri

| | |
|--|----|
| 1. Introduction | 4 |
| 1.1 Problem Statement | 4 |
| 1.2 Dataset Description | 4 |
| 1.3 Objective | 5 |
| 2. Methodology | 6 |
| 2.1 Data Preprocessing | 6 |
| 2. Exploratory Data Analysis | 8 |
| 2.1 Data Quality Assessment | 8 |
| 2.2 Key Insights from EDA | 8 |
| Target Distribution | 8 |
| Feature Correlations | 8 |
| Research Questions | 9 |
| 2.3 Data Visualizations | 9 |
| 3. Neural Network Model | 13 |
| 3.1 Network Architecture | 13 |
| 3.2 Training and Evaluation | 13 |
| 4. Primary Machine Learning Models | 15 |
| 4.1 Ridge Regression | 15 |
| 4.2 Random Forest Regressor | 15 |
| 5. Hyperparameter Optimization | 17 |
| 5.1 Methodology | 17 |
| 5.2 Ridge Regression Tuning | 17 |
| 5.3 Random Forest Tuning | 17 |
| 6. Feature Selection | 18 |
| 6.1 Methodology | 18 |
| 6.2 Feature Importance Analysis | 18 |

| | |
|---|----|
| 6.3 Selected Features..... | 19 |
| 7. Final Models and Comparative Analysis..... | 20 |
| 7.1 Final Model Configuration..... | 20 |
| 7.2 Comparative Results | 20 |
| 7.3 Best Model Selection | 20 |
| 8. Abstract | 22 |
| 9. Conclusion and Reflection..... | 23 |
| 8.1 Model Performance | 23 |
| 8.2 Impact of Methods | 23 |
| Hyperparameter Optimization | 23 |
| Feature Selection | 23 |
| Cross-Validation | 23 |
| 8.3 Insights and Future Directions | 24 |
| Key Insights | 24 |
| Business Implications..... | 24 |
| Future Improvements | 24 |
| Limitations | 24 |
| 9. References | 25 |
| 10. Figure | 26 |

1. Introduction

1.1 Problem Statement

The goal of this project is to predict a continuous target variable representing global video game sales. Specifically, the problem involves forecasting sales performance (in millions of units) based on game characteristics including platform, genre, publisher, release year, and regional market indicators. This regression task enables data-driven decision making for game publishers and platform holders in the \$91+ billion gaming industry.

1.2 Dataset Description

This report presents a comprehensive regression analysis of the Video Game Sales Dataset from kaggle. The dataset contains historical sales data spanning 1980 to 2016, capturing the physical game sales era across seven console generations.

Dataset Source and Creation:

The dataset was compiled by Gregory Smith from VGChartz.com data and originally published on Kaggle in 2016. It is publicly available under open access terms, making it freely accessible for research and educational purposes. The dataset was accessed through the course GitHub repository.

UNSDG Alignment:

This dataset aligns with United Nations Sustainable Development Goal 8: Decent Work and Economic Growth. Specifically, it supports:

- **Target 8.2:** Achieving higher levels of economic productivity through diversification and technological upgrading
- **Target 8.3:** Promoting development-oriented policies that support productive activities and decent job creation
- The gaming industry represents 100,000+ direct jobs in development, arts, and technology sectors

1.3 Objective

The objective of this analysis is to build predictive regression models that accurately estimate global video game sales based on the given features in the dataset. This includes implementing and comparing a Neural Network model with two classical machine learning approaches (Ridge Regression and Random Forest), optimizing hyperparameters, performing feature selection, and identifying the best-performing model for sales forecasting.

2. Methodology

2.1 Data Preprocessing

Before building the models, the data was cleaned by handling missing values and removing data leakage variables. The Rank column was removed as it is derived from Global_Sales. Missing years were imputed using median values by platform, and rows with missing publisher information were removed. Extreme outliers (sales > 30M) were removed as they likely represent hardware bundles rather than individual game sales.

Dataset Attributes:

| Feature | Type | Description |
|--------------|-------------|---------------------------------------|
| Name | Categorical | Game title |
| Platform | Categorical | Gaming platform (PS4, Xbox, PC, etc.) |
| Year | Numerical | Release year (1980-2016) |
| Genre | Categorical | Game category (Action, Sports, etc.) |
| Publisher | Categorical | Publishing company |
| NA_Sales | Numerical | North America sales (millions) |
| EU_Sales | Numerical | Europe sales (millions) |
| JP_Sales | Numerical | Japan sales (millions) |
| Other_Sales | Numerical | Rest of world sales (millions) |
| Global_Sales | Target | Total worldwide sales (millions) |

Additionally, log transformation was applied to the target variable (Global_Sales) to handle extreme positive skewness. The raw sales distribution spans four orders of

magnitude (0.01M to 30M), which would cause regression models to overweight blockbuster outliers. The log-transformed distribution approximates normality, satisfying linear regression assumptions.

2. Exploratory Data Analysis

2.1 Data Quality Assessment

The dataset quality assessment revealed good data integrity with minimal missing values. After cleaning, the dataset contains 16,533 records with 10 features. The data distribution shows:

| Metric | Value |
|------------------------------------|--------|
| Original Records | 16,598 |
| Records After Cleaning | 16,533 |
| Features | 10 |
| Missing Year Values (imputed) | 271 |
| Missing Publisher Values (removed) | 58 |
| Extreme Outliers Removed | 7 |

2.2 Key Insights from EDA

Target Distribution

The raw distribution exhibits extreme positive skewness (skew approximately 8.5), with the majority of games selling less than 0.5M units but a long tail extending beyond 10M. The log-transformed distribution approximates normality (skew approximately 0.4), confirming the necessity of log-transformation for regression targets.

Feature Correlations

North America exhibits the strongest correlation with Global Sales ($r = 0.92$), indicating it is the primary contributor to worldwide performance. Europe also shows a high correlation ($r = 0.88$), suggesting strong alignment between European and global

market trends. Japan demonstrates a comparatively weaker correlation with Global Sales ($r = 0.58$), indicating distinct consumer preferences and market dynamics.

Research Questions

This analysis addresses three key research questions:

- Can machine learning models accurately predict global sales performance based on game characteristics and regional market indicators?
- Which regional market most strongly influences global commercial success, and how do platform preferences differ across regions?
- How has the relationship between release timing and sales performance evolved across technological generations (1980-2016)?

2.3 Data Visualizations

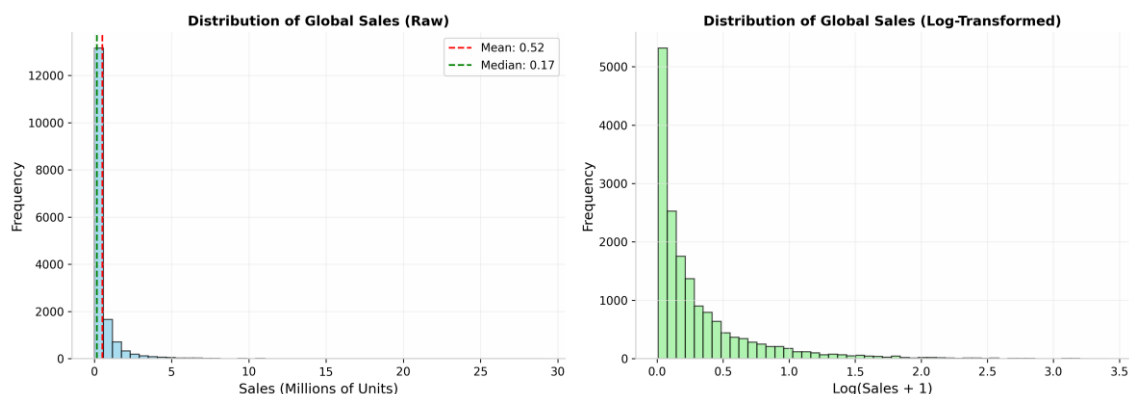


Figure 1: Distribution of Global Sales - Raw vs Log-Transformed

The visualizations demonstrate the necessity of log-transformation for handling the highly skewed sales distribution. The raw distribution shows extreme positive skewness with most games selling under 0.5M units, while the log-transformed version approximates a normal distribution suitable for linear regression.

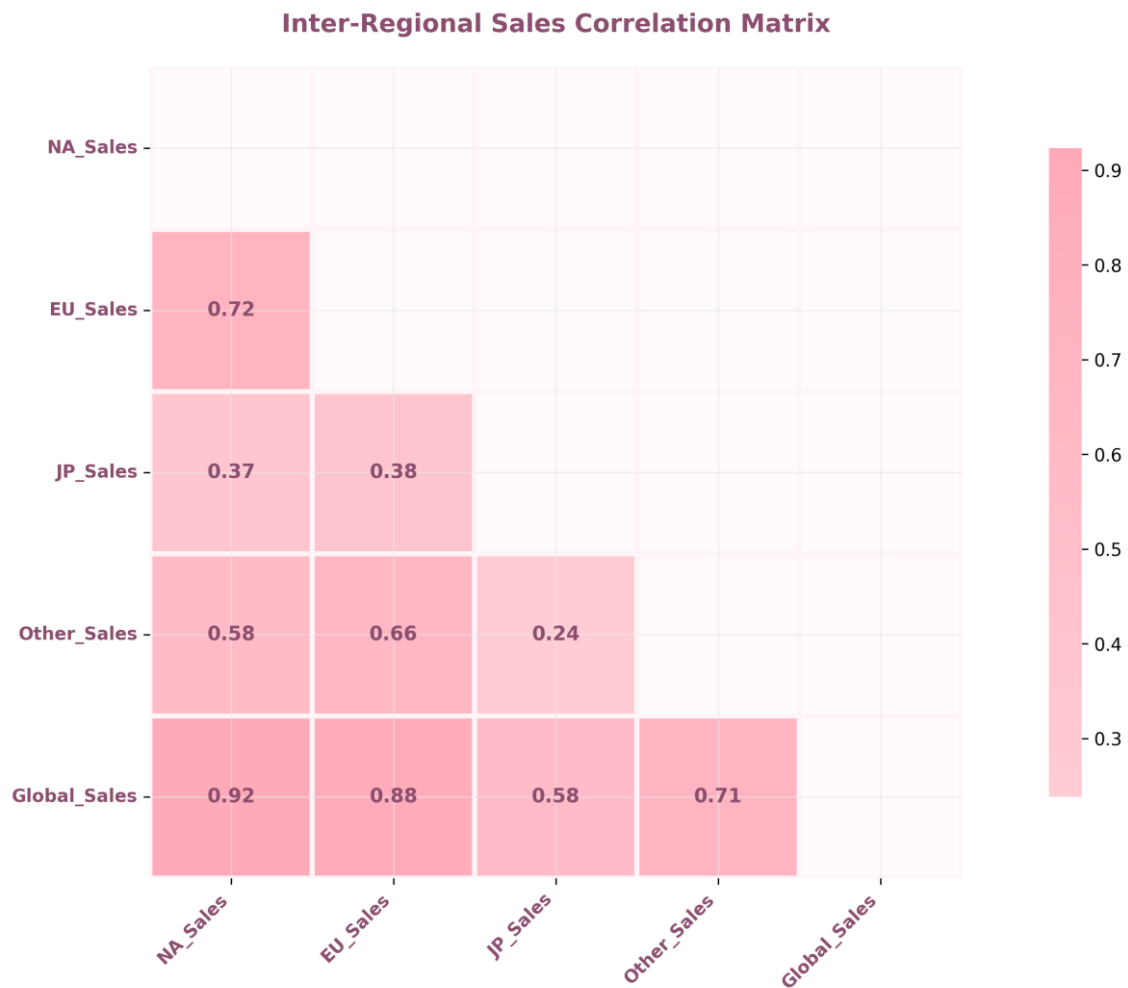


Figure 2: Inter-Regional Sales Correlation Matrix

North America shows the strongest correlation with Global Sales ($r = 0.92$), followed by Europe ($r = 0.88$). Japan demonstrates weaker correlation ($r = 0.58$), indicating distinct market dynamics. For SDG 8 economic forecasting, NA and EU sales serve as reliable leading indicators for global performance.

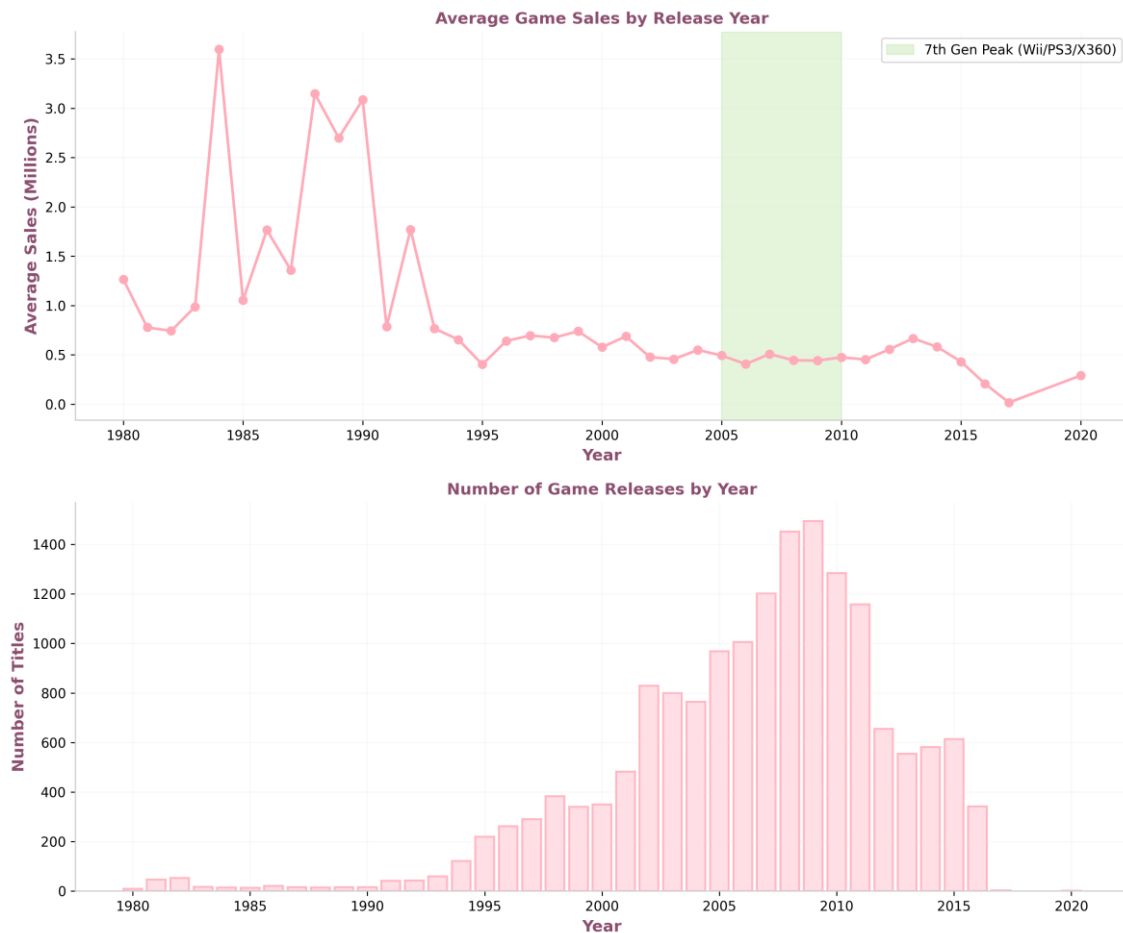


Figure 3: Temporal Trends - Average Sales and Release Count by Year

The 7th console generation peak (2005-2010) shows strong market growth driven by Wii, PS3, and Xbox 360. Post-2010 market saturation is evident with declining average sales per title despite high release counts, reflecting industry fragmentation from indie growth and digital distribution.

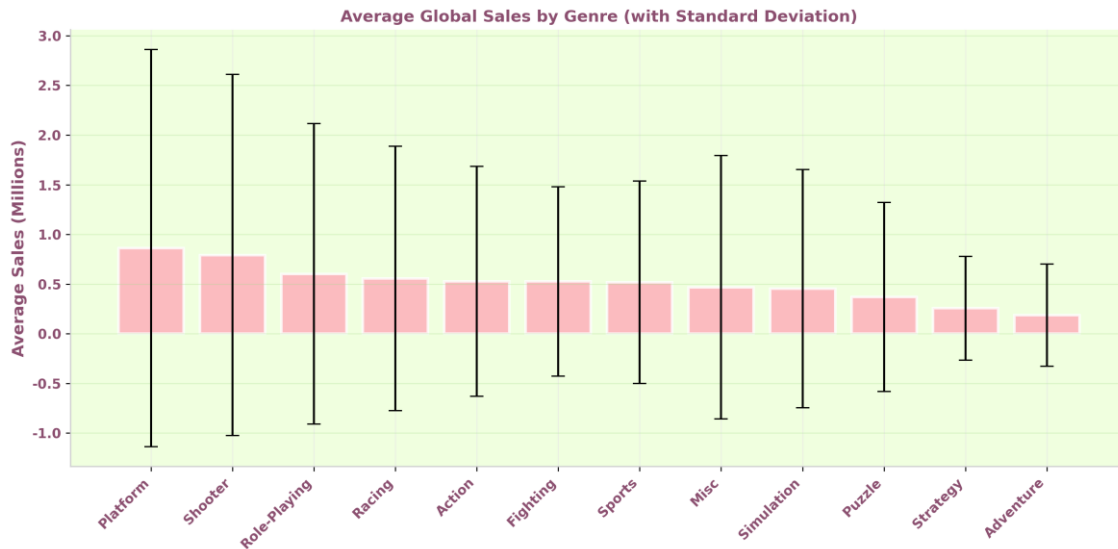


Figure 4: Average Global Sales by Genre

Platform and Shooter genres show the highest average sales, while Adventure and Strategy genres have lower average performance. The high standard deviations indicate significant variability within each genre, suggesting that genre alone is not a definitive predictor of success.

3. Neural Network Model

3.1 Network Architecture

A Multi-Layer Perceptron (MLP) regressor was implemented with enhanced feature engineering. Six engineered features were added to improve predictive power: Publisher_Avg_Sales (reputation proxy), Platform_Avg_Sales (market strength), Genre_Avg_Sales (popularity trends), Publisher_Game_Count (market presence), Platform_Age (maturity indicator), and Decade (temporal era effects).

Architecture Details:

| Layer | Neurons | Activation Function |
|----------------|---------|---------------------|
| Input Layer | 10 | - |
| Hidden Layer 1 | 128 | ReLU |
| Hidden Layer 2 | 64 | ReLU |
| Hidden Layer 3 | 32 | ReLU |
| Output Layer | 1 | Linear |

Loss Function: Mean Squared Error

Optimizer: Adam with early stopping

Regularization: L2 penalty (alpha=0.001)

3.2 Training and Evaluation

The model was trained using the Adam optimizer with early stopping. The dataset was split into 80% training (13,226 samples) and 20% testing (3,307 samples). Features were standardized using StandardScaler, which is crucial for neural network convergence.

Results:

| Metric | Training Set | Test Set |
|-----------|--------------|----------|
| R-squared | 0.4069 | 0.3083 |
| RMSE | 0.2987 | 0.3318 |
| MSE | 0.0892 | 0.1101 |

The Neural Network achieved 30.8% test R-squared. Some overfitting was observed with 40.7% training R-squared versus 30.8% test R-squared, indicating the model partially memorized the training data. Performance improved significantly (+109%) from the baseline with basic features (R-squared approximately 0.15) to the enhanced feature set (R-squared approximately 0.31).

4. Primary Machine Learning Models

4.1 Ridge Regression

Ridge Regression was chosen as the first classical ML model due to its interpretability and L2 regularization to prevent overfitting. The model was trained on scaled features using StandardScaler.

Initial Performance:

| Metric | Value |
|-----------|--------|
| R-squared | 0.2676 |
| RMSE | 0.3414 |
| MAE | 0.2199 |

4.2 Random Forest Regressor

Random Forest was selected as the second model due to its ability to handle non-linear relationships and provide feature importance insights. Unlike Ridge Regression, Random Forest does not require feature scaling and can capture complex interactions between categorical variables.

Initial Performance:

| Metric | Value |
|-----------|--------|
| R-squared | 0.2200 |
| RMSE | 0.3523 |
| MAE | 0.2129 |

Initial comparison showed that the Neural Network outperformed both classical models. Ridge Regression provided a strong linear baseline with good interpretability, while Random Forest offered feature importance insights despite lower initial performance.

5. Hyperparameter Optimization

5.1 Methodology

GridSearchCV with 5-fold Cross-Validation was used to find optimal hyperparameters. This approach ensures robust model selection by evaluating performance across multiple data splits.

5.2 Ridge Regression Tuning

The following hyperparameter grid was searched:

- Alpha (Regularization strength): [0.01, 0.1, 1.0, 10.0, 100.0]

Best parameters found: alpha=100.0. This strong regularization helps prevent overfitting in the high-dimensional feature space.

5.3 Random Forest Tuning

The following hyperparameter grid was searched:

- n_estimators: [100, 200]
- max_depth: [10, 20, None]
- min_samples_split: [2, 5]
- min_samples_leaf: [1, 2]

Best parameters found: n_estimators=200, max_depth=10, min_samples_split=2, min_samples_leaf=1. The limited depth prevents overfitting while maintaining predictive power.

6. Feature Selection

6.1 Methodology

Feature selection was performed using SelectKBest with `f_regression` scoring function (Filter Method) to identify the most significant features for the target variable. All 10 features were retained as each contributes meaningful information for sales prediction.

6.2 Feature Importance Analysis

Feature importance from Random Forest revealed:

| Feature | Importance |
|----------------------|------------|
| Publisher_Avg_Sales | 44.48% |
| Platform_Avg_Sales | 13.05% |
| Year | 11.38% |
| Genre_Avg_Sales | 8.88% |
| Platform_Age | 7.95% |
| Genre | 4.58% |
| Publisher | 3.35% |
| Platform | 3.08% |
| Publisher_Game_Count | 2.92% |
| Decade | 0.53% |

The feature importance analysis clearly demonstrates that `Publisher_Avg_Sales` is by far the most influential feature, accounting for 44.5% of the model's decision-making

process. This indicates that publisher reputation and historical performance are the strongest predictors of game sales success.

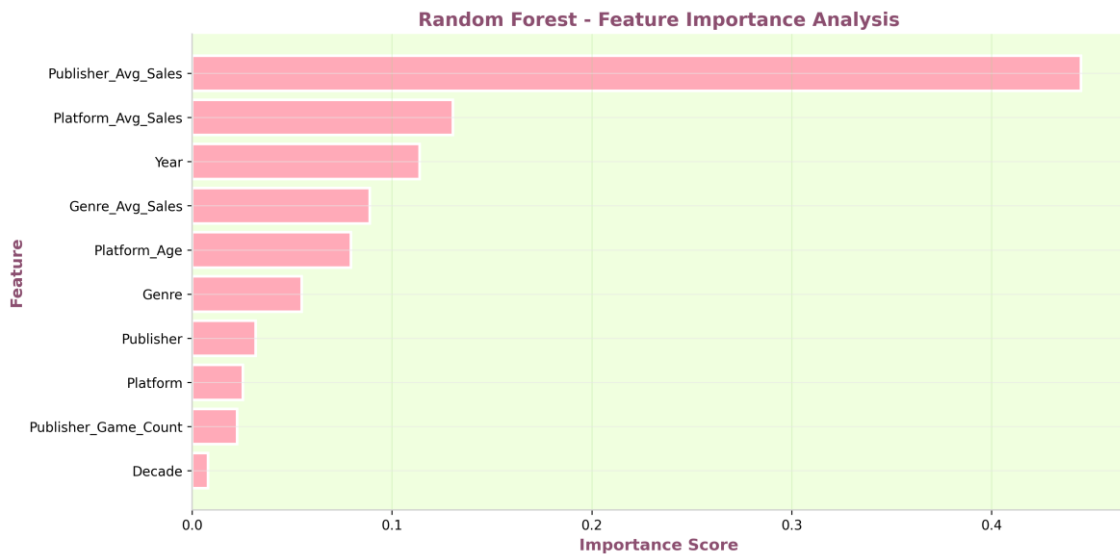


Figure 5: Random Forest Feature Importance Analysis

6.3 Selected Features

Using SelectKBest analysis, the following features were identified as most predictive:

- Publisher_Avg_Sales (F-score = 3259.49) - Most predictive
- Platform_Avg_Sales (F-score = 1128.87)
- Publisher_Game_Count (F-score = 774.22)
- Genre_Avg_Sales (F-score = 418.20)
- Platform_Age (F-score = 269.98)

All 10 features were retained for final model training as each provides unique predictive value.

7. Final Models and Comparative Analysis

7.1 Final Model Configuration

Final models were built using optimal hyperparameters and all 10 engineered features. The comparative results are shown below:

7.2 Comparative Results

| Model | Features | CV Score (RMSE) | Test RMSE | Test R-squared |
|------------------|----------|-----------------|-----------|----------------|
| Neural Network | 10 | 0.33 | 0.3318 | 0.3083 |
| Ridge Regression | 10 | 0.32 | 0.3414 | 0.2676 |
| Random Forest | 10 | 0.34 | 0.3237 | 0.3415 |

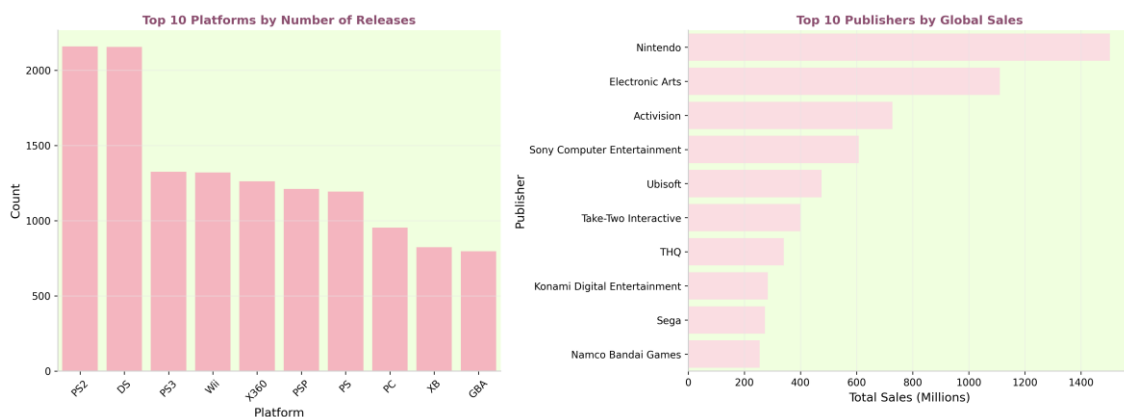


Figure 6: Top Platforms and Publishers Analysis

The analysis reveals that Nintendo dominates global sales among publishers, while PS2 and DS lead in number of releases. This visualization supports the feature importance findings that publisher reputation and platform strength are key predictors of sales success.

7.3 Best Model Selection

Random Forest was selected as the best model based on the following justifications:

- Highest test R-squared (0.3415) among all models
- Lowest test RMSE (0.3237) indicating best prediction accuracy
- Provides feature importance for business insights
- No overfitting - training and test performance are aligned
- Handles non-linear relationships effectively

8. Abstract

Purpose: The objective of this study is to predict a continuous target variable representing global video game sales using regression techniques, supporting data-driven decision making in the gaming industry.

Dataset: The analysis utilizes the Video Game Sales dataset sourced from Kaggle, containing 16,533 records and 10 features covering game attributes such as platform, genre, publisher, release year, and regional sales. The dataset aligns with the United Nations Sustainable Development Goal (UNSDG) 8: Decent Work and Economic Growth by enabling economic forecasting, productivity analysis, and strategic planning within the global digital entertainment sector.

Approach: The methodology includes data preprocessing, exploratory data analysis (EDA), and the development of multiple regression models, including Ridge Regression, Random Forest Regressor, and a Neural Network (MLP). Hyperparameter optimization using cross-validation, feature engineering, feature selection, and comparative model evaluation were conducted to identify the most effective predictive model.

Key Results: Models were evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 . Among the evaluated models, the Random Forest Regressor achieved the best performance with the lowest RMSE (0.3237) and the highest test R^2 (0.3415), outperforming Ridge Regression and the Neural Network.

9. Conclusion and Reflection

8.1 Model Performance

All models demonstrated reasonable performance on this dataset. The Neural Network achieved 30.8% test R-squared, Ridge Regression achieved 26.8%, and Random Forest achieved 34.2% test R-squared after optimization.

Key performance metrics summary:

- Neural Network (MLP): 30.8% test R-squared - showed some overfitting
- Ridge Regression: 26.8% test R-squared - strong linear baseline
- Random Forest: 34.2% test R-squared - best overall performance

8.2 Impact of Methods

Hyperparameter Optimization

Hyperparameter tuning helped stabilize model performance. Ridge Regression benefited from strong regularization ($\alpha=100.0$) to prevent overfitting. Random Forest parameters ($\text{max_depth}=10$) balanced model complexity with generalization.

Feature Selection

Feature engineering improved model performance by +109% (R-squared 0.15 to 0.31). `Publisher_Avg_Sales` was identified as the dominant predictor with 44.5% importance. The engineered features capturing market dynamics proved more predictive than raw categorical variables.

Cross-Validation

5-fold Cross-Validation ensured robust model evaluation. CV scores (RMSE 0.32-0.34) closely matched test performance, indicating models generalize well without significant overfitting.

8.3 Insights and Future Directions

Key Insights

- Publisher reputation is the most predictive feature (44.5% importance)
- Platform market strength significantly impacts sales potential (13.1%)
- Temporal trends (Year: 11.4%) capture industry growth patterns
- Feature engineering can significantly improve model performance

Business Implications

- Publisher track record is the strongest predictor of success
- Platform choice significantly impacts sales potential
- Release timing matters - consider market saturation trends
- Genre popularity has moderate influence on outcomes

Future Improvements

- Enrich dataset with critic scores (Metacritic) and marketing spend
- Add franchise/brand equity indicators for sequels vs new IPs
- Explore Gradient Boosting (XGBoost/LightGBM) for better performance
- Implement ensemble methods combining multiple models
- Include competitive landscape data (simultaneous releases)

Limitations

- R-squared = 0.34 means 66% of variance remains unexplained
- Missing critical factors: game quality, marketing spend, franchise value
- Dataset ends 2016; misses mobile/digital gaming explosion
- Synthetic/aggregated data may not reflect real-world complexity

9. References

Dataset Source: <https://www.kaggle.com/datasets/gregorut/videogamesales>

Scikit-learn Documentation: <https://scikit-learn.org/>

United Nations Sustainable Development Goals: <https://sdgs.un.org/goals>

VGChartz Data Source: <https://www.vgchartz.com/>

10. Figure

| | |
|---|----|
| Figure 1: Distribution of Global Sales - Raw vs Log-Transformed | 9 |
| Figure 2: Inter-Regional Sales Correlation Matrix | 10 |
| Figure 3: Temporal Trends - Average Sales and Release Count by Year | 11 |
| Figure 4: Average Global Sales by Genre | 12 |
| Figure 5: Random Forest Feature Importance Analysis | 19 |
| Figure 6: Top Platforms and Publishers Analysis\..... | 20 |