Analysis of the Video Game Sales Dataset Using Linear Regression and Random Forest Regression

Justin Childs

Professor Hongfei Xue

ITCS 3156-091

5 May 2025

Introduction

Problem Statement

For the final project for my ITSC-3156 class, I have chosen to track and analyze the trend of sales data regarding sales of video games. For this I've used two machine learning algorithms to do this: Linear Regression and Random Forest Regression. The goal of this project is to see if the sales data can be properly analyzed and predicted for future trends in sales, especially considering the growth of the industry in the last two decades.

Motivation and Challenges

The video games industry is one of the most profitable industries in the world (Statista's *Video Games Industry*). My motivation here stems from my interests in video games as well as being a developer myself. The information and, more importantly, success of sales is very important to me as it could affect my very livelihood one day. The sales figures quite literally dictate whether a game is a success or not.

Approach:

This program utilizes both Linear Regression and Random Forest Regression.

Linear Regression is a simple approach that will model the relationship between a video game's release year and its global sales using a straight-line fit. It serves as a baseline to understand if a linear trend explains market behavior. The model is trained on an 80/20 split of the data, and its performance is evaluated using RMSE and MAE metrics.

Random Forest Regression is an ensemble method that aggregates multiple decision trees to capture non-linear relationships between video game release years and global sales. This model is chosen to address the limitations of linear regression by capturing complex interactions and improving predictive accuracy. Like the linear model, the random forest is trained on the

same data split. Its performance is also evaluated using RMSE and MAE, supplemented by residual analysis to assess model error distribution.

Data

Data Introduction

The data set used for this project is the Video Game Sales dataset by Gregory Smith. This dataset contains a list of over 16,500 games that each sold more than 100,000 copies. It has been updated until the year 2016. Each game entry has the following features:

- Rank - Ranking of overall sales
- Name - The games name
- Platform - Platform of the games' release (i.e. PC, PS4, etc.)
- Year - Year of the game's release
- Genre - Genre of the game
- Publisher - Publisher of the game
- NA_Sales - Sales in North America (in millions)
- EU_Sales - Sales in Europe (in millions)
- JP_Sales - Sales in Japan (in millions)
- Other_Sales - Sales in the rest of the world (in millions)
- Global_Sales - Total worldwide sales.

The goal of this project is to properly analyze the sales trend to reliably predict future trends for the industry.

Data Visualization and Analysis

The data set has some important things to note. There are two large spikes between the years 2005 and 2009 in both global sales and distribution numbers.

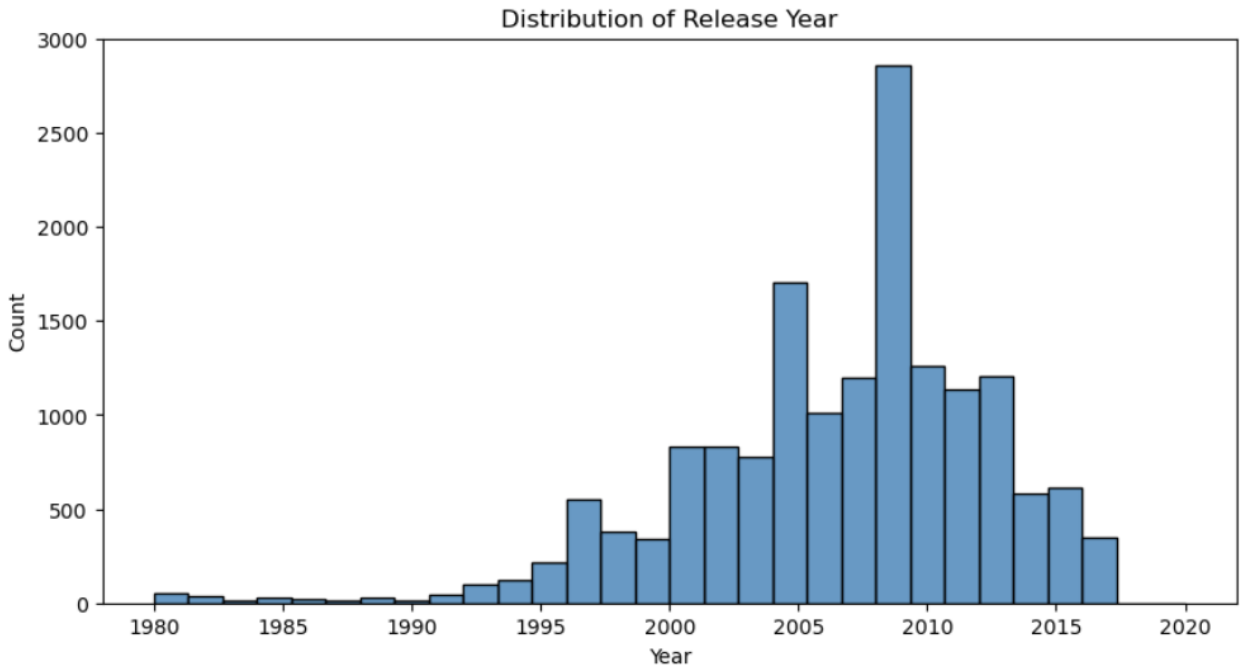

Figure 1: Distribution of Video Games by Release Year

Figure 1 shows a histogram of the number of releases in the data set per year. There is a large spike for both 2004-2005 and 2008-2009, with a sharp decrease after 2013-2014.

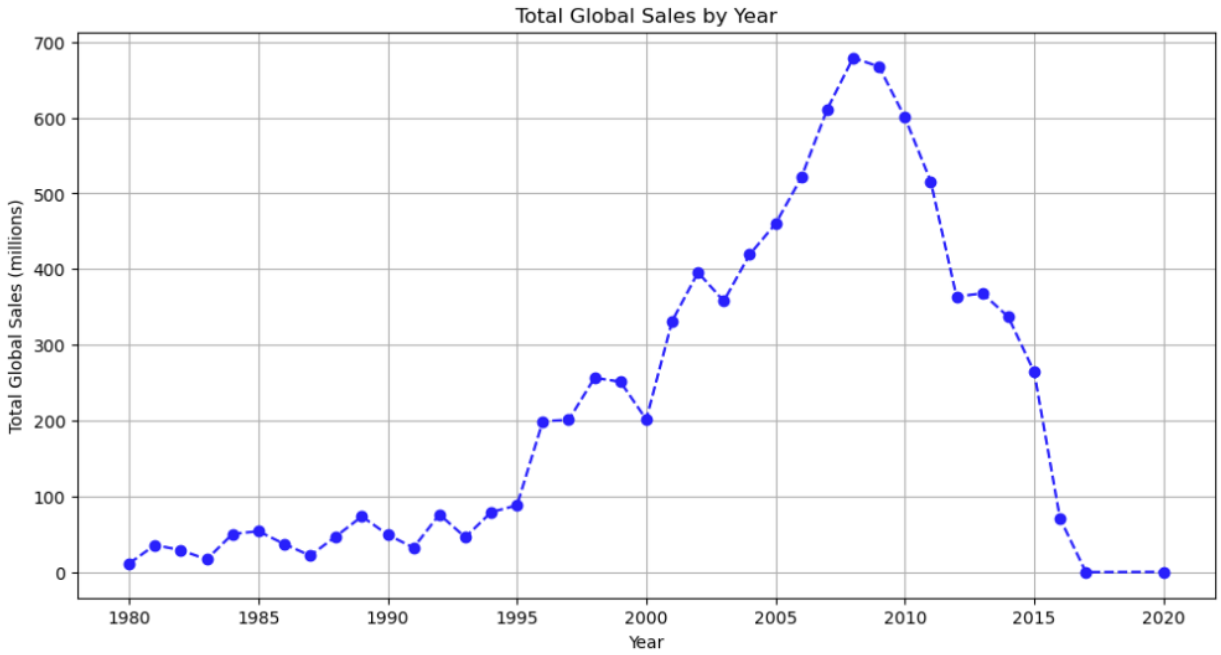

Figure 2: Total Sales Each Year

Figure 2 is a graph of total sales each year and demonstrates the same pattern in *Figure 1,* though it is important to note that the decrease from the two spikes mentioned before is not as prevalent in *Figure 2.*

Data Preprocessing*:*

Multiple steps were taken to prepare the Linear Regression and Random Forest Regression models. The data was cleaned by removing records without essential values (i.e., missing Year or Global_Sales) and converting the Year to an integer. The aggregated data was then used to develop visualizations and support the temporal analysis of sales trends.

Methods

As mentioned above, the project uses the two algorithms: Linear Regression and Random Forest Regression.

Linear Regression:

The Linear Regression method models the relationship between the dependent variable global video game sales and the independent variable the release year. The model fits a straight line through the data, minimizing the sum of squared errors between the actual and predicted values.

Reasons for Linear Regression:

- The simplicity of Linear Regression allows for easy interpretation of the data trends
- The performance of the model can serve well as a baseline for more complex models

The dataset was split into an 80/20 training/testing set. The Linear Regression model was imported from scikit-learn and was trained using the release year as the sole predictor of global sales, and its performance was evaluated using error metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

Random Forest Regression:

The Random Forest Regression method constructs multiple decision trees and combines their predictions for a more accurate output (*"Random Forest Regression in Python."*). Each decision tree is built on a bootstrapped sample of the data, and the final prediction is obtained as the average output of all the trees.

Reasons for Random Forest Regression:

- The ability for the model to handle more nonlinear data such as the sudden spikes displayed in *Figure 1* and *Figure 2.*
- The ability for the model to handle outliers more effectively than Linear Regression

The Random Forest model was also imported from scikit-learn and trained on the same 80/20 split of the dataset. The hyperparameter settings, such as the number of trees (n_estimators=100), were chosen based on preliminary tests. The model was evaluated using the same set of performance metrics (RMSE and MAE) to allow for a direct comparison with the linear model. Additionally, residual analysis was conducted to assess the dispersion of errors and affirm the model's robustness.

Results

Experimental Setup

The dataset is partitioned into training and testing sets using an 80/20 split. Both models are trained with the same data partitioning. The setup ensures consistent performance evaluation through metrics such as RMSE and MAE.

Test Results:

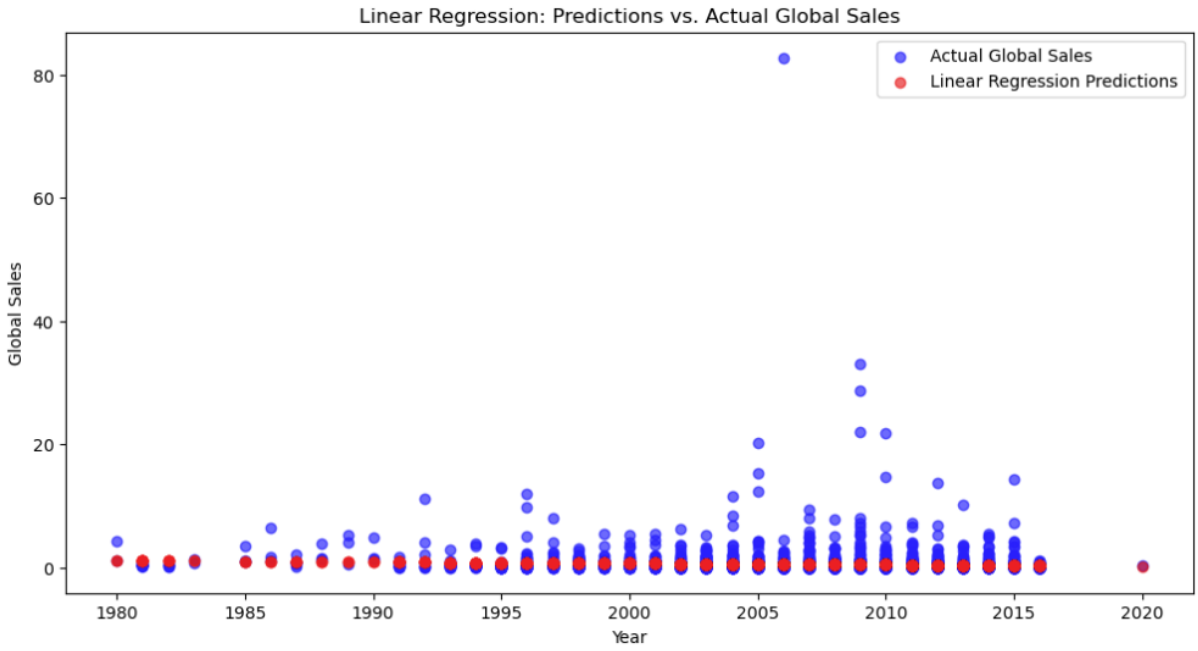

Figure 3: Linear Regression Prediction vs Actual

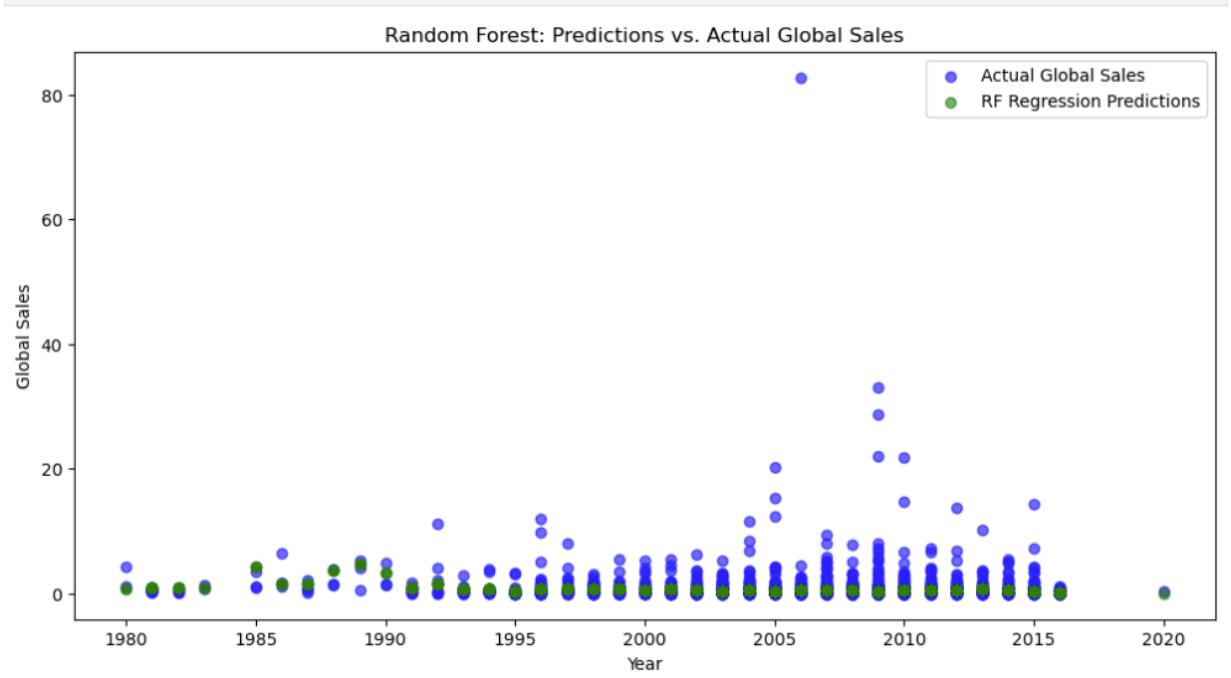

Figure 4: Random Forest Regression Prediction vs Actual

Metric	Linear Regression	Random Forest Regression
RMSE	2.0669	2.0665
MAE	0.6160	0.6022

Table 1: Performance Evaluation

Observations/Analysis:

The analysis of the results reveals both models performed very similarly. The Linear Regression model, with an RMSE of 2.07 and an MAE of 0.62, indicates that while most predictions are relatively accurate, but a few significant outliers substantially increased the overall error. Residual analysis further supports this observation by exposing non-random error patterns, suggesting that the simple linear model struggles to capture the inherent non-linear complexities of the market. The outliers themselves are likely to be very popular games that sold more than average.

In contrast, the Random Forest model demonstrated improved accuracy by effectively handling the non-linear relationships within the data, as evidenced by lower error metrics and a more uniform residual distribution. However, the two performed similarly enough to determine that the Random Forest Regression model was not able to properly account for the outliers very well, though it did notably better than the Linear Regression model which seems to be more of a baseline.

Supporting Experiments:

To support the analysis, multiple experiments were conducted in a systematic manner. First, the dataset was divided using an 80/20 train-test split, ensuring that both models were evaluated on the same data partitions using RMSE and MAE as performance metrics. Scatter plots comparing actual versus predicted global sales were then generated to assess the accuracy of each model, particularly noting how extreme values were handled. In addition, histograms of residuals were created to analyze the distribution of prediction errors and to uncover any systematic biases. In addition, key hyperparameters for the Random Forest model, such as varying the number of trees, were tuned to validate that the improved performance was robust across different configurations.

Conclusion

Overall, the models performed well I believe, even with the outliers taken into account. The Random Forest Regression model performing better than the Linear Regression model, although not by much, should have been expected. Linear Regression, simple as it is, was way less likely to consider and properly account for the outliers present in the data set. I would be very eager to see how they'd perform in regards to video game sales from 2016-present.

This project was not without its challenges. The tools and techniques learned from the class slides as well as the homework assignments over the course of the semester really helped to prepare me for this project. It was also very interesting to look at the video games industry through a different lens. I'm normally looking at the development side of things, not sales.

References & Citations

Clement, J. "Topic: Video Game Industry." *Statista*, 6 Nov. 2024, https://www.statista.com/topics/868/video-games/.

GregorySmith. "Video Game Sales." *Kaggle*, 26 Oct. 2016, https://www.kaggle.com/datasets/gregorut/videogamesales/data.

"Random Forest Regression in Python." *GeeksforGeeks*, 7 Apr. 2025, https://www.geeksforgeeks.org/random-forest-regression-in-python/.

Acknowledgements

- Information from the slides provided by Professor Xue was used in the preparation of this report.
- Microsoft's Copilot was used to help create an outline and format this project
- Various online sources were used in regard to information gathering for this project which are referenced in the References/Citations section

Source Code

The link to the GitHub Repository is here.