

NAME: EDWARD SARAH MFON

MATRIC NUMBER; VUG/CSC/22/7054

COURSE CODE; CSC 308

CONTENT

1. INTRODUCTION
2. DATASET SUMMARY
3. CLEANING AND PREPROCESSING
4. EXPLORATORY DATA ANALYSIS
5. VISUALIZATION
6. MACHINE LEARNING MODEL
7. CONCLUSION

**Introduction**

The video game industry has evolved into one of the largest and most dynamic sectors in the entertainment market, generating billions of dollars in revenue annually across various platforms and genres. With the increasing importance of data-driven decision-making in modern businesses, analyzing video game sales can provide valuable insights into consumer behavior, market trends, and the factors that contribute to a game's commercial success. Understanding these patterns not only benefits game developers and publishers but also helps investors, marketers, and analysts make informed decisions. This project leverages machine learning techniques to predict global sales of video games based on several key attributes, including platform, genre, publisher, release year, and regional sales performance.

The primary motivation behind this project is to demonstrate the practical application of machine learning algorithms to real-world problems involving structured data. By building a predictive model, the goal is to identify the most influential factors affecting video game sales and accurately estimate a game's potential performance in the global market. Moreover, the project aims to practice essential data science skills such as data cleaning, exploratory data analysis (EDA), visualization, model building, evaluation, and effective communication of results.

The dataset used for this project, commonly known as vgsales.csv, was obtained from Kaggle, a popular online platform for data science projects. It contains detailed sales data for over 16,000 video games released across different platforms, genres, and years. The dataset includes features such as the name of the game, the platform it was released on, the publisher, the genre, sales figures in different regions (North America, Europe, Japan, and Others), and the total global sales.

The research question guiding this project is:

"Can we accurately predict the global sales of a video game based on its platform, genre, publisher, release year, and regional sales data?"

Answering this question will help uncover the relationships between game attributes and their market success, providing actionable insights for stakeholders in the video game industry. It also serves as a comprehensive exercise in applying machine learning methods to solve regression problems on real-world datasets.

Dataset Summary

The dataset used in this project is the Video Game Sales dataset (vgsales.csv) from Kaggle. It provides detailed information on video game sales across regions and platforms, making it ideal for a regression-based machine learning project. Initially containing 16,598 records and 11 features, the dataset was cleaned by dropping entries with missing values, resulting in 16,269 records.

Key features include Platform, Year, Genre, Publisher, and regional sales (NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales), with Global\_Sales as the target variable. The Name feature was dropped during cleaning. Regional sales and categorical features serve as important predictors.

The dataset’s combination of categorical and numerical variables allows for in-depth exploratory analysis and robust machine learning modeling, offering insights into market trends, consumer behavior, and global game sales dynamics.

Data Cleaning and Preprocessing

Data cleaning and preprocessing are critical steps in any machine learning project, as they ensure the quality and usability of the dataset for modeling. In this project, I worked with the "vgsales.csv" dataset, which initially contained 16,598 records and 11 columns. Before proceeding with analysis and modeling, it was essential to inspect the dataset for inconsistencies, missing values, irrelevant columns, and improper data types.

The first step was to check for missing data using .isnull().sum(). It was observed that two columns, Year and Publisher, had missing values. Specifically, there were 271 missing values in the Year column and 58 missing values in the Publisher column. Since these features are crucial for understanding the release timing and the responsible companies for the games, simply imputing them with an average or a placeholder could have introduced bias. Therefore, I made the decision to drop any rows where Year or Publisher were missing. This approach preserved the overall integrity of the dataset, with only a minimal loss of data (~2% of the entire dataset).

Next, I examined the data types using the .info() method. It was identified that the Year column was stored as a float. Since the year of release should logically be an integer (e.g., 2006, 2010), I converted the Year column to integer type using .astype(int). This conversion ensures that the model treats Year as a discrete value instead of a continuous one, which is more appropriate for categorical time-related data.

Further, I noticed two columns that were not necessary for prediction tasks: Rank and Name. The Rank field simply repeated the sorting order of global sales, and Name is a text field representing the game title, which would not contribute meaningfully to the prediction model without additional feature engineering. Therefore, both columns were dropped to streamline the dataset.

The dataset also included categorical features such as Platform, Genre, and Publisher. Since machine learning algorithms typically require numerical input, these categorical variables were encoded numerically using Label Encoding. For each column, a separate LabelEncoder was fitted, transforming each unique category into a unique integer. This method is appropriate because there is no ordinal relationship between the categories.

**Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) is essential for uncovering patterns and guiding the modeling process. After data cleaning, I performed a comprehensive EDA on the video game sales dataset, which consisted of 16,269 rows and 9 columns, including key categorical features (Platform, Genre, Publisher) and numerical features (Year, regional sales, Global\_Sales).

Using .describe(), I found that sales figures varied widely across regions, with North America leading. Global Sales were right-skewed, indicating that most games sold relatively few copies, while a few became major hits. .value\_counts() showed that platforms like PlayStation 2 and Nintendo DS had the highest number of releases, and Action, Sports, and Shooter were the most common genres.

Correlation analysis revealed strong positive relationships between Global Sales and regional sales (especially NA\_Sales and EU\_Sales), but weaker ties with categorical features. Temporal analysis showed a peak in sales between 2006 and 2008, coinciding with a surge in popular console and game releases.

Through this EDA process, several important insights were gathered:

* Sales are heavily influenced by region, especially North America.
* A few publishers and platforms dominate the market.
* Sales figures are highly skewed, with only a small percentage of games achieving blockbuster sales.
* The gaming industry experienced significant growth during the mid-2000s.

**DATA VISUALIZATION**

Data visualization plays a crucial role in any data science project by simplifying complex information for easier understanding. In this project, Matplotlib and Seaborn were used to create six visualizations that provided deeper insights into the video game sales dataset.

First, a histogram of Global Sales showed a right-skewed distribution, revealing that most games sold under 1 million copies, with only a few blockbuster hits. A count plot of Genres indicated that Action games were the most common, followed by Sports and Shooter genres, reflecting market trends and player preferences. Another count plot displayed the distribution of games across Platforms, with PlayStation 2 and Nintendo DS leading, likely due to their popularity and longevity.

A scatter plot of Year versus Global Sales revealed that the mid-2000s were peak years for sales, with a decline post-2010, possibly linked to shifts toward digital and mobile gaming. A heatmap of feature correlations showed strong positive correlations between regional sales and Global Sales, while Platform and Genre had weaker links. Finally, a boxplot of Global Sales by Genre highlighted how sales varied across genres, with Shooter and Platform games generally achieving higher median sales.

Clear titles, labels, and formatting enhanced readability, making these visualizations vital for interpreting data and guiding predictive modeling.

**MACHINE LEARNING MODEL**

The goal of the machine learning phase was to predict Global Sales of video games using features like Platform, Year, Genre, Publisher, and regional sales. Framed as a regression task, the dataset was split 80/20 into training and testing sets to ensure reliable evaluation. Two models were trained and compared: Linear Regression, a simple baseline assuming linear relationships, and Random Forest Regressor, an ensemble method better at handling non-linear patterns.

After training, model performance was assessed using R² Score and RMSE. Random Forest outperformed Linear Regression, achieving a higher R² and lower RMSE, indicating more accurate predictions. This showed that complex models better capture real-world sales patterns.

**CONCLUSION**

In this project, the Video Game Sales dataset from Kaggle was used to explore and predict global sales outcomes. After careful data cleaning and exploratory analysis, key insights into market trends and sales distributions were uncovered. Machine learning models, including Linear Regression and Random Forest Regressor, were trained and evaluated, with Random Forest delivering superior performance. The combination of categorical and numerical features provided a strong foundation for building accurate models. Overall, this project demonstrated the importance of thorough data preparation, insightful visualization, and careful model selection in creating reliable predictive systems within the video game industry.