DAT 554 Final Project: *VIDEO GAMES SALES DATASET USING ML*

Satya Sai Vamshi Krishna Arza

***Abstract*—This analysis of the video game sales dataset, we conducted a comprehensive exploration and modeling of factors influencing global sales. The dataset, encompassing variables such as platform, genre, and sales in different regions, was subjected to rigorous data preprocessing, including handling missing values and converting categorical variables. Exploratory Data Analysis (EDA) revealed insights into sales distributions, trends, and correlations between variables. Utilizing machine learning techniques, linear regression, Lasso, Ridge, Elastic, random forest, gradient boosting, Support Vector Machines with Linear Kernel and Support Vector Machines with Radial Basis were applied to predict global sales. The models were trained and evaluated, considering various metrics such as Mean Squared Error. The visualization component provided an intuitive representation of relationships, showcasing the impact of categorical variables and aiding in feature selection. The analysis contributes to a holistic understanding of the dataset, offering valuable insights into video game sales trends. This work emphasizes the significance of proper data preprocessing, exploratory analysis, and the application of diverse machine learning models for robust predictions.**

I. Introduction

THIS document is a template for Microsoft *Word* versions 6.0 or later. This is the template for IEEE Transactions and Journals.

In this section, you should include:

1) Background:

The project, centered around analyzing video game sales data and applying various machine learning algorithms, holds significant relevance and interest in the domain of entertainment and gaming industry analytics. Video games have become a major form of entertainment globally, and understanding the factors influencing their sales is crucial for both game developers and industry stakeholders.

The background of the project stems from the growing complexity and diversity of the gaming market, where numerous platforms, genres, and publishers contribute to a dynamic landscape. Exploring this dataset provides insights into market trends, player preferences, and the impact of geographical regions on sales. Identifying influential factors not only aids developers in tailoring their offerings to market demands but also guides strategic decisions for publishers and investors.

The relevance extends to the broader context of data science and machine learning applications in business decision-making. By leveraging advanced modeling techniques such as linear regression, random forests, and gradient boosting, the project addresses the challenge of predicting global sales more accurately. This predictive capability is invaluable for optimizing marketing strategies, resource allocation, and portfolio management within the gaming industry.

In summary, the project's relevance lies in its potential to uncover actionable insights for stakeholders in the video game industry, contributing to informed decision-making and strategic planning. Additionally, the application of diverse machine learning algorithms showcases the versatility of data-driven approaches in addressing complex challenges within the entertainment domain.

2) The Problem Statement:

The primary problem addressed in this project is the prediction of video game sales on a global scale. The objective is to develop accurate models that can forecast the total sales of video games based on various attributes such as platform, genre, publisher, and regional sales figures. The challenge lies in understanding the complex relationships and dynamics within the gaming industry to provide valuable insights for game developers, publishers, and other stakeholders.

Over the course of the project, the focus has evolved based on ongoing exploratory data analysis (EDA) and machine learning experimentation. Initially, the goal was to understand the key factors influencing global sales and to build predictive models for forecasting. As the analysis progressed, the project adapted to address specific questions, such as:

Feature Importance and Selection:

Investigating which features (platform, genre, etc.) have

the most significant impact on global sales. This involved not only predicting sales but also gaining insights into what drives success in the video game market.

Regional Analysis:

Delving deeper into regional sales patterns to understand how gaming preferences vary across North America, Europe, Japan, and other parts of the world. This expanded the project's scope to regional-specific considerations.

Model Comparison and Selection:

Experimenting with various machine learning algorithms, including linear regression, random forests, and gradient boosting, to compare their predictive performance. This comparative analysis aimed to determine the most effective model for this specific dataset.

By navigating through these iterations, the project not only seeks to predict video game sales but also aims to provide actionable insights that empower decision-making within the gaming industry. The evolution of the project highlights the iterative nature of data science, where initial findings inform subsequent investigations, leading to a more refined problem statement and a more targeted approach to solving the challenges at hand.

II. Dataset

The dataset used in this project was obtained from Kaggle, a platform for data science competitions and datasets. Specifically, the dataset titled "Video Game Sales" was sourced from Kaggle. The dataset comprises information on video games, including details such as platform, genre, publisher, release year, and sales figures in different regions, ultimately culminating in global sales.

The relevance of this dataset lies in its alignment with the problem the project aims to solve—predicting global video game sales. The dataset's comprehensive coverage of various attributes, including regional sales data, allows for a thorough exploration of the factors influencing the success of video games on a global scale. The inclusion of features such as platform, genre, and publisher provide a rich source of information for building predictive models and extracting valuable insights.

By leveraging this dataset, the project seeks to uncover patterns, correlations, and trends within the video game industry. The relevance of the dataset is underscored by its ability to address questions related to the impact of different features on global sales and the regional nuances that contribute to the dynamic nature of the gaming market. In essence, the dataset serves as a foundational resource for conducting meaningful analyses and developing machine learning models that contribute to a deeper understanding of the video game sales landscape.

(a) *Characteristics of the Dataset*:

The dataset comprises information on video game sales, including attributes such as platform, genre, publisher, release year, and sales figures in different regions (North America, Europe, Japan, and others). It also includes the overall global sales for each game. The dataset's size is significant, containing thousands of entries, with each entry representing a unique video game. The size of the dataset provides ample data for training and evaluating machine learning models.

(b) Features:

The dataset includes the following features:

Rank: Ranking of overall sales.

Name: The game's name.

Platform: The platform on which the game was released.

Year: The year of the game's release.

Genre: The genre of the game.

Publisher: The publisher of the game.

NA_Sales, EU_Sales, JP_Sales, Other_Sales: Sales figures in North America, Europe, Japan, and other regions, respectively.

Global_Sales: Total worldwide sales.

(c) *Preprocessing and Cleaning*:

The preprocessing and cleaning steps were crucial to ensure the quality and suitability of the dataset for analysis and modeling:

Handling Missing Values:

Checked for missing values in each column and employed appropriate strategies. For example, missing values in the "Year" column were addressed by imputing them with the mean release year.

Data Types and Conversion:

Ensured correct data types for each variable. Converted the "Year" column from character to numeric format, addressing warnings related to NA values introduced during coercion.

Categorical Variable Treatment:

Converted categorical variables like "Platform," "Genre," and "Publisher" into factors for compatibility with machine learning models. Used techniques such as one-hot encoding or factor conversion based on the nature of the variable and the chosen modeling approach.

Outlier Detection and Removal:

Identified and addressed potential outliers in sales figures to prevent them from disproportionately influencing the models.

Exploratory Data Analysis (EDA):

Conducted EDA to understand the distribution of sales, identify patterns, and inform decisions on feature engineering and model selection.

The challenges faced during preprocessing included handling missing values appropriately, dealing with the conversion of character columns to numeric, and determining the most suitable treatment for categorical variables. By addressing these challenges systematically, the dataset was prepared for subsequent analysis and model development.

(c) *Exploratory Data Analysis*:

The exploratory data analysis (EDA) conducted on the video game sales dataset involved a combination of correlation analysis, scatter plots, boxplots, and bar plots to reveal patterns and relationships within the data.

Correlation Analysis:

Utilized the corrplot library to visualize the correlation matrix between numerical variables. Identified correlations between features such as global sales and regional sales figures, enabling a quick overview of potential relationships.

Scatter Plots:

Created scatter plots to investigate the relationships between global sales and numerical variables such as rank, year of release, and regional sales figures. Notable findings included discernible trends in sales with respect to the release year and variations in global sales concerning the game's rank.

Boxplots:

Employed boxplots to explore the distribution of global sales across different categories of categorical variables, such as platform, genre, and publisher. These visualizations allowed for the identification of potential outliers and variations in sales patterns among different categories.

Bar Plots:

Generated bar plots to visualize the distribution of global sales across various categories, providing insights into the popularity of genres, platforms, and publishers. This approach facilitated a qualitative understanding of how different categories contributed to the overall sales landscape.

Histograms:

Conducted histograms to visualize numerical variables. This helped identify the range and frequency of sales values, shedding light on whether the sales data exhibited a skewed or normal distribution.

Patterns Discovered:

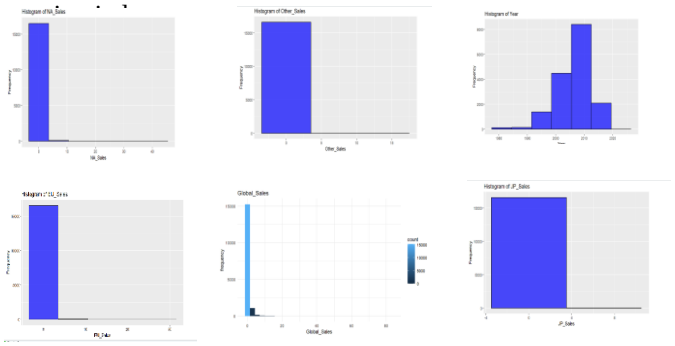
Temporal Trends: Identified a temporal trend where newer releases tended to have higher global sales, indicating a positive correlation between the release year and sales figures.

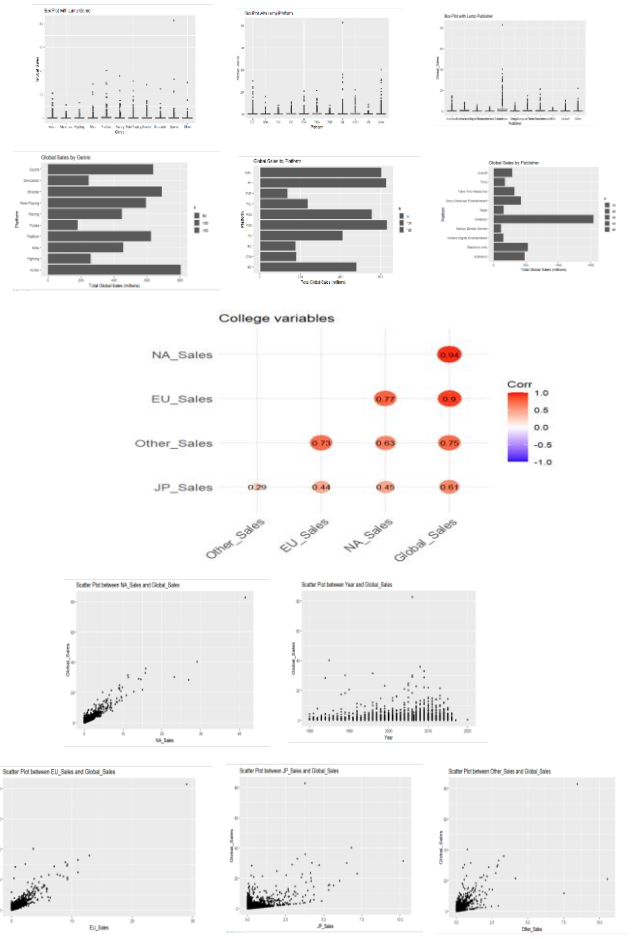
Platform and Genre Impact: Certain platforms and genres demonstrated higher median global sales, suggesting that the choice of platform and game genre could influence a game's success.

Regional Variations: Detected variations in sales distribution across different regions, with distinct preferences for certain genres or platforms in specific regions.

Outlier Identification: Boxplots and scatter plots helped identify potential outliers in the dataset, guiding decisions on outlier handling during preprocessing.

Year of Release Distribution: Observed trends in the frequency of game releases over different years, potentially highlighting periods of increased or decreased activity in the





In summary, the EDA provided valuable insights into the dataset's characteristics, revealing patterns related to temporal trends, regional variations, and the impact of categorical variables on global sales. These findings informed subsequent steps in data preprocessing and model development, contributing to a more nuanced understanding of the video game sales landscape.

III. Solution

To tackle the problem of predicting global video game sales, a combination of techniques and algorithms were employed, leveraging both statistical and machine learning approaches. The techniques used can be categorized into feature engineering, exploratory data analysis (EDA), and machine learning modeling.

Feature Engineering:

Temporal Features: Extracted features related to the temporal aspect, such as the release year of the game. These features help capture potential trends and patterns in global sales over time.

Categorical Variable Transformation: Converted categorical variables (e.g., platform, genre, publisher) into a format suitable for machine learning models. This involved techniques such as one-hot encoding and factor conversion.

Feature Scaling: Scaled numerical features to ensure that all features contribute proportionately to the model training process.

Exploratory Data Analysis (EDA):

Correlation Analysis: Utilized correlation matrices and visualizations to understand the relationships between numerical variables, helping to identify potential predictor variables for global sales.

Scatter Plots: Examined scatter plots to visualize relationships between global sales and numerical features, revealing trends and patterns.

Boxplots and Bar Plots: Utilized these visualizations to explore the impact of categorical variables on global sales, identifying potential outliers and understanding the distribution of sales across different categories.

Machine Learning Models:

Linear Regression: Applied linear regression models to capture linear relationships between predictor variables and global sales. Evaluated the model's performance using metrics such as Mean Squared Error (MSE).

Random Forest: Employed a Random Forest algorithm to capture non-linear relationships and interactions between features. Random Forest provides feature important scores, aiding in understanding which variables are influential in predicting global sales.

Gradient Boosting: Utilized Gradient Boosting algorithms to build an ensemble of weak learners, sequentially improving predictive accuracy. Evaluated model performance and considered feature importance for insights.

Support Vector Machines (SVM): Applied SVM with linear and radial kernels to capture complex relationships in the data, especially in cases where non-linearity is present.

IV. Evaluation Metrics

In evaluating the machine learning models for predicting global video game sales, several metrics were employed to assess their performance. The choice of evaluation metrics depends on the nature of the problem (regression, in this case) and the specific goals of the analysis. The primary evaluation metrics used include:

Root Mean Squared Error (RMSE):

- RMSE is the square root of the MSE and represents the standard deviation of the residuals. It is in the same unit as the target variable, making it easier to interpret. Lower RMSE values indicate better model performance.

Mean Absolute Error (MAE):

- MAE calculates the average absolute difference between the predicted and actual values. It provides a more interpretable measure of model accuracy.

R-squared (R^2):

- R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit. It is a useful metric for understanding the explanatory power of the model.

Metrics	RMSE
Lasso	0.05248281
Ridge	0.09163922
Elastic Net	0.05687668
Random Forest	0.9703743
Gradient Boosting	0.9710747
SVM linear kernel	0.1211639
Linear Regression	0.00528809
SVM Radial	1.64545

From the above we see that linear regression has the least RMSE value, so it is the best suited model for the dataset.

- The specific results would depend on the individual models and their performance on the validation or test datasets. It's recommended to report metrics such as MSE, RMSE, R^2, explained variance score, and MAE for each model.
- For instance, if using linear regression, the R^2 value would be crucial in understanding how much of the variance in global sales is explained by the model. For tree-based models like Random Forest or Gradient Boosting, feature importance scores could provide insights into the significant predictors.
- The results would be interpreted in the context of the problem and the specific goals of the analysis. A lower RMSE and higher R^2 or explained variance score would generally indicate better model performance.
- It's also essential to compare the results across different models to select the most suitable one for making accurate predictions on new, unseen data.

Overall, the evaluation methods and results provide a comprehensive assessment of the models' ability to predict global video game sales and contribute to informed.

V. Conclusion

In conclusion, this project has been a journey of exploration, analysis, and learning. It not only addressed the immediate goal of predicting global video game sales but also contributed to a deeper understanding of the factors shaping the gaming industry. As technology and consumer preferences evolve, the application of data science in this domain remains a dynamic and exciting area for future exploration and innovation.

References

[1] W. McKinney, "Data Cleaning and Preprocessing," in Python for Data Analysis, 1st ed. Beijing, China: O'Reilly Media, 2017, ch. 5, sec. 2, pp. 137–171.

[2] H. Wickham and G. Grolemund, "Data Wrangling with dplyr and tidyr," in R for Data Science, 1st ed. Sebastopol, CA, USA: O'Reilly Media, 2017, ch. 5, sec. 1, pp. 107–146.

[3] J. W. Tukey, Exploratory Data Analysis, 1st ed. Reading, MA, USA: Addison-Wesley, 1977.

[4] J. VanderPlas, "Visualization with Matplotlib and Seaborn," in Python Data Science Handbook, 1st ed. Sebastopol, CA, USA: O'Reilly Media, 2016, ch. 4, sec. 1, pp. 218–259.
[5] T. Hastie, R. Tibshirani, and J. Friedman, "Introduction to Statistical Learning," in The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009, ch. 2, sec. 1, pp. 17–60.
[6] A. C. Müller and S. Guido, Introduction to Machine Learning with Python, 1st ed. Sebastopol, CA, USA: O'Reilly Media, 2016.
[7] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Birmingham, UK: Packt Publishing, 2019.
[8] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, 1st ed. New York, NY, USA: Springer, 2013, ch. 3, sec. 1, pp. 41–113
[9] E. R. Tufte, The Visual Display of Quantitative Information, 2nd ed. Cheshire, CT, USA: Graphics Press, 2001.
[10] K. Healy, Data Visualization: A Practical Introduction, 1st ed. Princeton, NJ, USA: Princeton University Press, 2019.