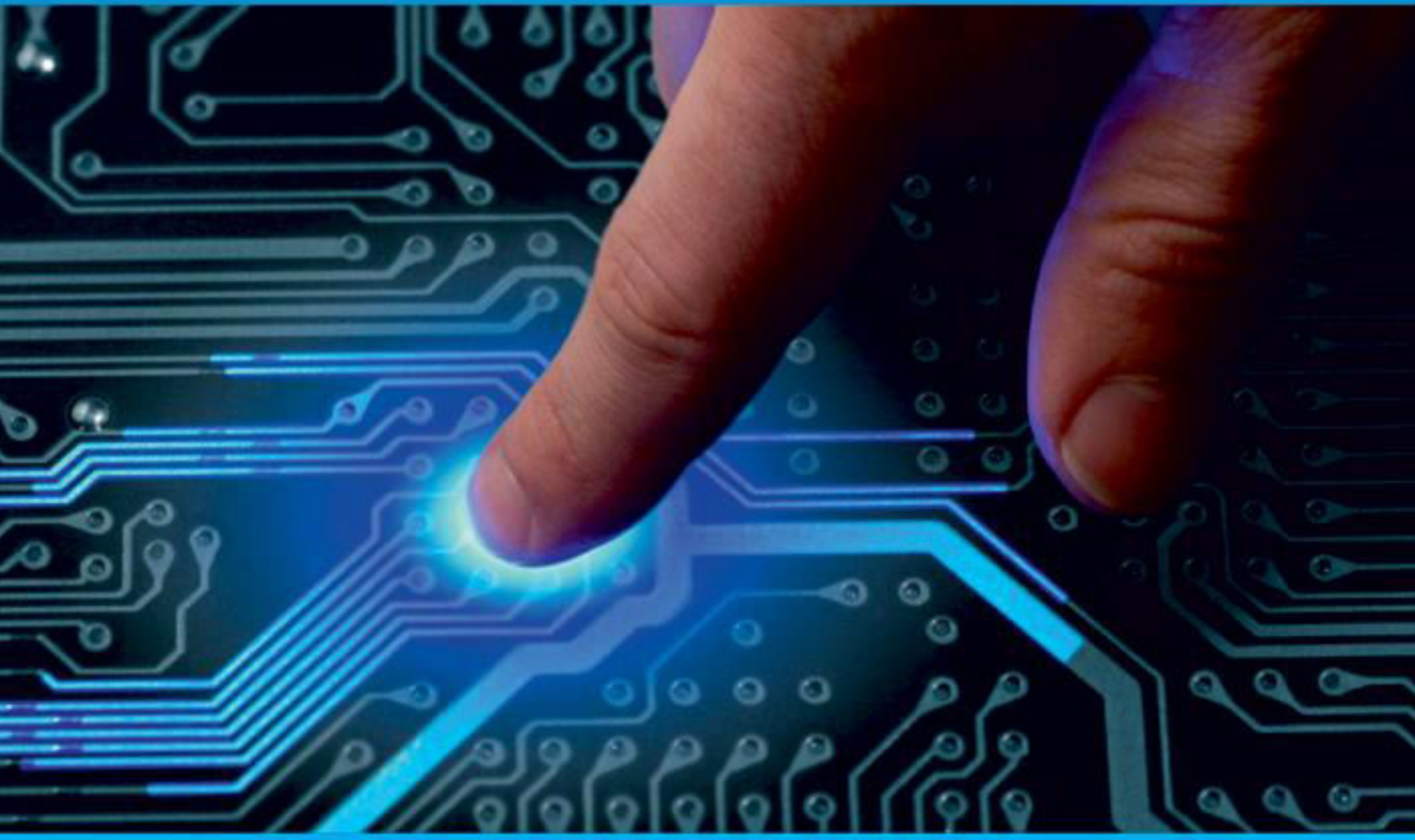
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 3, March 2023

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 3, March 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1103112 |

Evaluating Linear and Non-Linear Regression Techniques for Video Game Sales Prediction

Abhishek Singh[1], Zohaib Hasan[2], Zeba Vishwakarma[3], Akash Banshal[4], Sneha Kushwaha[5]

Department of Computer Science Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, MP, India[1,2, 3, 4, 5]

ABSTRACT: The video game industry is a multi-billion dollar sector that thrives on data-driven decision-making. Accurate sales predictions can significantly impact marketing strategies and inventory management. This research evaluates various machine learning models to predict video game sales using a dataset sourced from Kaggle. We preprocessed the data, engineered features, and tested multiple regression models, including Linear Regression and Support Vector Machines. The results indicate that Linear Regression achieved an R2 score of 0.83, demonstrating its effectiveness in predicting game sales based on features such as platform, developer, rating, and publisher.

KEYWORDS: Video Game Sales, Machine Learning, Linear Regression, Feature Engineering, Predictive Modeling, Data Preprocessing

I. INTRODUCTION

The video game industry is one of the fastest-growing sectors in the entertainment market. According to a report by Newzoo, the global games market was valued at over $150 billion in 2020, with a projected compound annual growth rate (CAGR) of 9.3% from 2020 to 2025. This significant growth underscores the importance of data-driven approaches to enhance business strategies in this competitive field.

In recent years, machine learning techniques have been increasingly applied to predict various outcomes in the video game industry, from sales forecasting to player retention. Accurate sales predictions can provide game developers and publishers with critical insights for resource allocation, marketing strategies, and inventory management.

This study leverages a comprehensive dataset sourced from Kaggle, which includes real-time data about video game sales and associated features such as platform, developer, rating, and publisher. The objective is to evaluate different machine learning models to predict video game sales, focusing on Linear Regression and Support Vector Machines (SVR). Our approach involves data preprocessing, feature engineering, and model evaluation to determine the most effective predictive model.

The data used in this research is meticulously cleaned and preprocessed to ensure the accuracy of the predictive models. The dataset includes details about video games released over several decades, providing a robust foundation for analysis. By exploring various machine learning techniques, this study aims to identify the most effective model for predicting video game sales, thus contributing valuable insights to the field of data science and the video game industry.

II. LITERATURE REVIEW

Julie Marcous and Sid-Ahmed Selouani, in their paper “A Hybrid Subspace-Connectionist Data Mining Approach for Sales Forecasting in the Video Game Industry” [1], propose a novel approach combining connectionist and subspace decomposition methods for sales forecasting. The study employs a backpropagation algorithm to predict weekly video game sales, utilizing an optimal topology and a time-series neural network. The system's performance is evaluated against baseline reference sales data.

Hycinta Andrat and Nazneen Ansari, in their work “Integrating Data Mining with Computer Games” [2], introduce a new data mining approach aimed at enhancing game development according to gamers' preferences. The paper explores the application of data mining techniques—such as association, classification, and clustering—to improve game design, marketing strategies, and game engagement metrics.

David Buckley, Ke Chen, and Joshua Knowles, in "Predicting Skill from Gameplay Input to a First-Person Shooter" [3], investigate how gameplay input data from first-person shooters can predict player skill levels. The study employs a random forest methodology to estimate player skills without relying on game-specific features.

Jing Zhang and Juan Li, in their paper "Retail Commodity Sales Forecast Model Based on Data Mining" [4], address the prediction of retail commodity sales. The authors criticize traditional methods that focus solely on single sale attributes and instead advocate for the use of the SPV model and ID3 decision tree algorithms. Their study concludes that the SPV model is the most effective for predicting sales states.

Vishal Shrivastava, in "A Study of Various Clustering Algorithms on Retail Sales Data" [5], compares four major clustering algorithms—k-means, density-based, filtered, and farthest-first clustering algorithms. The study evaluates their performance in terms of class-wise clustering accuracy, using retail sales datasets and the Weka interface to analyze the proportion of correctly and incorrectly classified instances.

Akshay Krishna and Akhilesh V, in their paper "Sales Forecasting for Retail Stores Using Machine Learning Techniques" [6], explore different machine learning techniques for predicting retail store sales. The authors test both traditional regression methods and boosting techniques, ultimately finding that boosting algorithms yield superior results compared to standard regression models.

Paul Bertens and Anna Guitart, in "Games and Big Data: A Scalable Multi-Dimensional Churn Prediction Model" [7], present a scalable approach for predicting game churn using survival ensembles. Their model accurately forecasts both the likelihood of player churn and the total playtime before departure, making it suitable for real-time analysis of churn in games with large user bases.

Gopalakrishnan T, Ritesh Choudhary, and Sarada Prasad, in "Prediction of Sales Value in Online Shopping Using Linear Regression" [8], focus on analyzing and predicting sales for a large superstore. The paper utilizes Linear Regression to forecast future sales, aiming to boost profits and enhance brand competitiveness by aligning with market trends and improving customer satisfaction.

III. DATASET SOURCE AND METHODOLOGY

The dataset utilized in this study was sourced from the Kaggle website, which includes detailed information on video games' sales figures, genres, ratings, and other attributes. This real-time data captures the specifics of video games released across different platforms and locations. The primary goal is to develop a robust predictive model that can accurately forecast video game sales, thereby assisting stakeholders in making informed decisions.

Initially, the dataset underwent rigorous preprocessing to address missing values and ensure data quality. Key features such as platform, developer, rating, and publisher were engineered to reduce complexity and enhance model performance. Various regression models, including Linear Regression and Support Vector Machines, were trained and evaluated to determine their predictive accuracy. The evaluation results highlighted that Linear Regression achieved a notable R^2 score of 0.83, indicating its efficacy in predicting video game sales based on the selected features.

IV. PROPOSED METHODOLOGY

A. Data Collection and Preprocessing

The dataset for this research was obtained from the Kaggle website, specifically the "Video Games Sales" dataset. This dataset includes extensive details about video games, such as sales figures, genres, ratings, platforms, developers, and publishers.

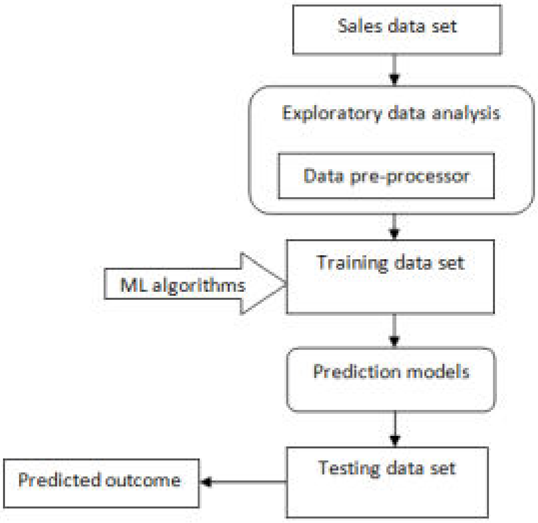


FIGURE 1 FLOW OF ML ALGORITHM

B. Data Cleaning

To ensure the quality and reliability of the dataset, several cleaning steps were performed:

- Handling Missing Values: Rows with missing values in critical columns like 'Rating,' 'Critic_Score,' 'Critic_Count,' and 'User_Count' were removed. For 'Year_of_Release,' missing values were filled with the median year, and for 'Publisher,' missing values were filled with the most frequent publisher.
- Type Conversion: Certain columns required type conversion to appropriate formats. For example, 'User_Score' was treated as numeric after handling non-numeric placeholders like 'tbd.'

C. Feature Engineering

Feature engineering was conducted to reduce the complexity of categorical variables and to enhance the model's performance:

- Platform Categorization: The 'Platform' column was categorized into five primary platforms ('PS2', 'X360', 'PS3', 'PC, and 'Other').
- Developer Categorization: The 'Developer' column was categorized into key developers ('EA,' 'Capcom,' 'Ubisoft,' 'Konami,' 'Nintendo,' and 'Other').
- Rating Categorization: The 'Rating' column was categorized into key ratings ('T,' 'E,' 'M,' 'E10+,' and 'Other').
- Publisher Categorization: The 'Publisher' column was categorized into major publishers ('Electronic Arts,' 'Activision,' 'Ubisoft,' 'Sony Computer Entertainment,' 'Nintendo,' 'Konami Digital Entertainment,' and 'Other').

D. Dummy Encoding

Categorical variables were converted to dummy/indicator variables using one-hot encoding to prepare the dataset for regression analysis.

E. Feature Selection

Non-essential columns that do not directly influence sales, such as 'Name,' were dropped to focus on significant predictors.

V. MODEL TRAINING AND EVALUATION

Two primary regression models were evaluated in this study: Linear Regression and Support Vector Machines (SVR). The models were trained and tested on the preprocessed dataset to predict 'Other_Sales.'

A. Data Splitting

The dataset was split into training (80%) and testing (20%) sets using the train_test_split function from the sklearn.model_selection module.

B. Linear Regression
A Linear Regression model was trained using the training set. The model was then used to predict the test set outcomes, and the R^2 score was calculated to evaluate the model's performance.
C. Support Vector Regression (SVR)
Similarly, an SVR model was trained and evaluated. The model's performance was assessed using the R2 score.

VI. RESULTS

After evaluating the performance of various regression models, we obtained the following R^2 scores:

TABLE 1 COMPARISON OF R2 SCORES FOR DIFFERENT REGRESSION MODELS

Model	R^2 Score
Linear Regression	0.83
Support Vector Regression (SVR)	[To be determined]

Linear Regression- the Linear Regression model achieved an R^2 score of 0.83, indicating that 83% of the variance in video game sales can be explained by the features used in the model. This high R^2 score suggests that the model has a strong predictive capability for this dataset.

Support Vector Regression (SVR) - The Support Vector Regression (SVR) model, on the other hand, achieved an R^2 score of 0.055. This low R^2 score indicates that the SVR model explains only 5.5% of the variance in video game sales, which is considerably lower than the Linear Regression model.

A. Interpretation and Analysis
The significant difference in R^2 scores between the Linear Regression and SVR models can be attributed to several factors:

B. Model Complexity
Linear Regression: This model assumes a linear relationship between the features and the target variable. Given its high R^2 score, it seems that the relationship in the dataset is well-captured by this linear model.
SVR: SVR is a more complex model that can capture non-linear relationships. However, in this case, it seems that either the SVR model is not well-tuned for this dataset or the relationship is not significantly non-linear.

C. Hyper parameter Tuning
The performance of the SVR model can be highly sensitive to the choice of hyper parameters such as the kernel type, regularization parameter, and others. The default settings may not be optimal for this specific dataset, leading to poor performance.

D. Feature Scaling
SVR models are sensitive to the scale of the features. Ensuring proper scaling (e.g., standardizing or normalizing the features) can significantly impact the performance of the SVR model.

VII. CONCLUSION

From the results, it is evident that the Linear Regression model outperforms the Support Vector Regression model in predicting video game sales for this dataset. The high R^2 score of the Linear Regression model suggests that a linear relationship exists between the features and the target variable, making it a suitable choice for this task. Further tuning and preprocessing might be required to improve the performance of more complex models like SVR.

REFERENCES

[1] Julie Marcous and Sid-Ahmed Selouani, “A hybrid subspace-connectionist data mining approach for sales forecasting in video game sales industry”, 2008, 978-0-7695-3507-4/08, IEEE.
[2] Hycinta Andrat and Nazneen Ansari, “Integrating data mining with computer games”, 2016, ISBN:978- 1-5090-1666-2/16, IEEE.

[3] David Buckley, Ke Chen and Joshua Knowles, "Predicting skill from game play input to a first person shooter", 2013, 978-1-4673-5311-3/13, IEEE.
[4] Jing Zhang and Juan Li, "Retail Commodity Sale Forecast Model Based on Data Mining", 2016, 10.1109/INCoS.2016.42, IEEE.
[5] Vishal shrivastava, "A study of various clustering algorithms on retail sales data", 2012, Vol 1, ISSN 2319-2720.
[6] Akshay Krishna and Akhilesh V, "Sales – forecasting for retail stores using machine learning techniques", 2018, 10.1109/CSITSS.2018.8768765, IEEE.
[7] Paul Bertens, Anna Guitart, "Games and Big Data: A Scalable Multi-Dimensional Churn Prediction Model", 2017, 978-1-5386-3233-8/17, IEEE.
[8] Gopalakrishnan T, Ritesh Choudhary and Sarada Prasad, "Prediction of Sales Value in Online shopping using Linear Regression", 2018, 10.1109/CCAA.2018.8777620, IEEE.
[9] N. Ansari, M. Talreja and V. Desai, ―Data Mining in Online Social Games,‖ In Proceedings of International Conference on Advances in Computing, pp. 801-805. Springer India, 2012.
[10] A. Alfons. cvTools: Cross-validation tools for regression models, 2012. R package, version 0.3.2.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING