Video Game Sales Analysis

1. Introduction

This dataset offers information on the sales performance and popularity of different video games worldwide. It includes the game's Name, release year, platform, Genre, publisher, and sales in different regions. The dataset also features ratings from critics and users, such as average critic score, number of critics reviewed, average user score, number of users reviewed, developer, and rating. This dataset provides valuable insights into the international video game market and is a valuable resource for market researchers, industry professionals, and gamers. The objective is to forecast how well a number of video games will sell globally and categorize and group the games according to certain descriptive factors. This task will be addressed as follows.

2 Data Exploratory data analysis and Modelling method

This section introduces the data pre-processing steps (**Exploratory data analysis)** and the different models used to predict the global sales performance of multiple video games and also to classify and group the games based on certain criteria.

2.1 Exploratory data analysis

2.1.1 Data Understanding

There are missing values in the year of release, Genre, Publisher, Critic score, Critic count, User score, User count column, Developer and Rating, and Name.

2.1.2 Data cleaning

a) **Critic score, Critic count and User count:** The missing values were replaced with zero.
b) **User score:** The missing values were replaced by the median value.
c) Rating: The missing values were replaced with Unknown.
d) Genre and Name: The two missing values were dropped.
e) **Year of release, Publisher:** These columns were not treated because they were deemed insignificant to our model.

2.1.3 Other consideration

a) **Platform:** Platform type may contribute to global sales, and there are many types. One-hot encoding too many platforms can lead to too many columns and overfitting. So, a "new_platform" column was created, grouping platforms with less than 1000 occurrences as "others."
b) **Outlier treatment:** The outliers in our data set were treated so that the model would satisfy the treat assumption of linear regression (assumption of linearity, Normality, and Independence). When outliers are present in the data, the model may be skewed towards these outliers, leading to inaccurate and unreliable predictions.
c) **Rating**: For the rating column entry “unknown” (corresponding to the missing values in the rating column),” EC”, “K-A”, “RP” and AO were not considered in our modelling.

2.2 Modelling methods

The methodology involved importing relevant libraries to build the model. For the regression analysis, numerical variables were selected as predictors, and global sales were the target variable. The data was split into training and testing sets, and standard scaling was used to normalize the data. Various regression algorithms were used, including Lasso, Ridge, Gradient Boosting, Random Forest, Support Vector, K-Nearest Neighbors, and Linear Regression. For classification, rating, platform, and Genre were chosen as the target variables; the numerical variables were the independent variables. Categorical variables were converted to numeric using label encoding, and the data was balanced using SMOTE. The data was then split into training and testing sets and normalized using min-max scaling. Different classifiers, including Support Vector, Decision Tree, K-Nearest Neighbors, Random Forest, and Logistic Regression, were used to build the classification model. For clustering, the columns "Name," "Global_Sales_log," "Platform," "Year_of_Release," "Publisher," and "Developer" were dropped, and numerical columns were used for clustering. The data were normalized using standard scaling, and the number of clusters was determined using the elbow method. The resulting clusters were described using rating and Genre.

2.2.1 Model evaluation

The performance of the regression models was evaluated using various metrics such as mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), R2, adjusted R2, and mean absolute percentage error MAPE. The performance of the classification model was evaluated using metrics like confusion matrix, accuracy, recall, precision, and f1 score. On the other hand, clustering analysis was assessed by utilizing external metrics, such

as V-measure Score, Rand Index Score, and Mutual Information Score. Internal evaluation measures, such as Davies-Bouldin Index, Silhouette Coefficient, and Calinski Harabasz Score, were also used to evaluate the clustering model's performance.

3 Result and discussion:

3.1 Best feature in predicting global sales

To determine the variables in the video game dataset that best predict global sales, simple linear regression was first carried out using the numeric variable whose correlation with global sales is greater than 0.5. Based on the evaluation metrics in Fig.1, the feature that best predicts global sales is NA_Sales. This is because it has the lowest MSE, RMSE, MAE and MAPE values. Additionally, it has the highest value of R2 and Adj.R2. Overall, the model's performance is relatively better using sales in North America and other regions compared to Europe and Japan

Model performance of NA_sales (correlation to golbal sale =0.94)

	MSE	RMSE	MAE	R2	Adj. R2	MAPE
0	0.727802	0.853113	0.641517	0.658432	0.65833	4.858293e+08

Model performance of EU_sales (correlation to golbal sale =0.9)

	MSE	RMSE	MAE	R2	Adj. R2	MAPE
0	0.966538	0.983127	0.760014	0.54639	0.546254	2.869063e+08

Model performance of JP_sales (correlation to golbal sale =0.61)

	MSE	RMSE	MAE	R2	Adj. R2	MAPE
0	1.981565	1.407681	1.154542	0.070023	0.069745	8.856782e+08

Model performance of others_sales (correlation to golbal sale =0.75)

	MSE	RMSE	MAE	R2	Adj. R2	MAPE
0	0.814165	0.902311	0.695988	0.617901	0.617786	1.983194e+08

Fig.1 Model performance using simple linear regression

Furthermore, NA Sales, EU Sales, JP Sales, Others- Sales, Critic score, Critic count, User score, and User count were used to perform multiple linear regression to see if they better predict global sales. As shown in Table 1 below, the multiple linear model performed better at predicting global sales

Table 1. Model performance using multiple linear regression.

Features	MSE	RMSE	MAE	R2	Adj.R2	MAPE
NA_Sales, EU_Sales, JP_Sales, Others_Sales,Critic_score,Critic_count, User_score, User_count	0.419	0.647	0.517	0.803	0.803	3.360e+08

The multiple linear model has lower MSE, RMSE, and MAE values, indicating that it has the smallest errors in predicting global sales values. Additionally, it has the highest R-squared value and Adj. R-squared value, indicating a better fit. Finally, the lower value of MAPE for multiple linear regression indicates that it has a smaller percentage error in predicting global sales. Figure 2 below shows each feature's contribution in predicting global sales.

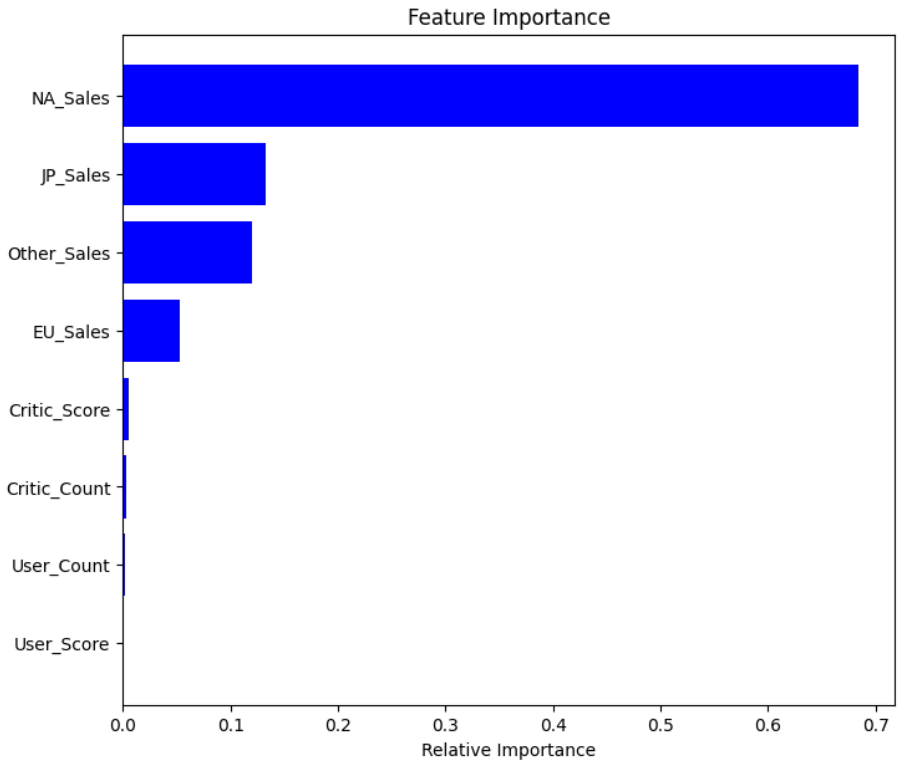

Fig.2 Feature importance to global sales prediction

3.2 Effect of the number of critics and users, as well as their review scores on the sales of Video games in North America, EU and Japan

Multiple linear regression was used to examine the relationship between the predictor variables (critic score, critic count, user score, and user count) and target variables (sales in North America, Japan and Europe). The coefficients for each variable were used to determine the strength and direction of their respective relationships.

a) Relationship to sales in North America:

$$NA\_sales\_y = 0.161 + 0.003 * Critic\_score + 0.039 * Critic\_Count + 0.000 * User\_Score + 0.040 * User\_count \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots eq1$$

As shown in equation 1, NA_sales have a positive relationship with Critic Count and User Count and a very weak positive relationship with Critic Score. The equation suggests that higher ratings by critics may also have a small positive impact on sales in North America, but this effect is not as strong as the influence of Critic Count or User Count. However, user ratings do not significantly affect sales in North America.

b) Relationship to sales in Europe:

$$EU\_sales\_y = 0.071 - 0.004 * Critic\_score + 0.016 * Critic\_Count + 0.000 * User\_Score + 0.031 * User\_count \ldots\ldots\ldots\ldots\ldots\ldots eq2$$

Equation 2 shows a weak negative relationship with Critic Score, meaning that higher ratings by critics may negatively impact sales in Europe. However, the positive relationship with Critic Count and User Count suggests that sales in Europe may increase with more ratings from critics and users. Again, there is no significant relationship with User Score.

a) Relationship to sales in Japan:

$$JP\_sales\_y = 0.024 - 0.015 * Critic\_score + 0.005 * Critic\_Count + 0.000 * User\_Score + 0.007 * User\_count \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots eq3$$

For Japan sales, equation 3 shows a weak negative relationship with Critic Score, suggesting that higher ratings by critics may have a small negative impact on sales in Japan. However, as with the other regions, the positive relationship with Critic Count and User Count indicates that sales in Japan may increase with more ratings from critics and users. There is also no significant relationship with User Score in Japan.

3.3 Performance of different regressors

Based on the evaluation metrics in Table 2, the Gradient Boosting Regressor (GB), Random Forest Regressor (RF) and outperform the other regressors in terms of MSE, RMSE, MAE, R2, and Adj. R2. Linear Regression (LR), Ridge Regression (RG), and Lasso Regression (LS) perform similarly and have higher errors and lower R2 and Adj. R2 values compared to GBR and RF. Support Vector Regression (SVR) and Kneigbor Regression (KB) have better performance than LS, LR, and RG

Table.2: Evaluation metrics of the different regressors

Regressor	MSE	RMSE	MAE	R2	Adj. R2	MAPE
LS	0.419	0.647	0.517	0.803	0.803	3.339e+08
RG	0.419	0.647	0.517	0.803	0.803	3.360e+08
GB	0.089	0.296	0.165	0.959	0.959	2.273e+07
RF	0.086	0.294	0.148	0.960	0.960	4.609e+07
SVR	0.439	0.663	0.502	0.794	0.793	2.747e+08
LR	0.419	0.647	0.517	0.803	0.803	3.360e+08
KB	0.116	0.340	0.192	0.946	0.945	7.446e+07

Overall, the RF and GB Regression performed better because they have lower MSE, RMSE, and MAE, and the highest R2, as shown in Fig.3. and deemed the most effective regression technique based on this result.

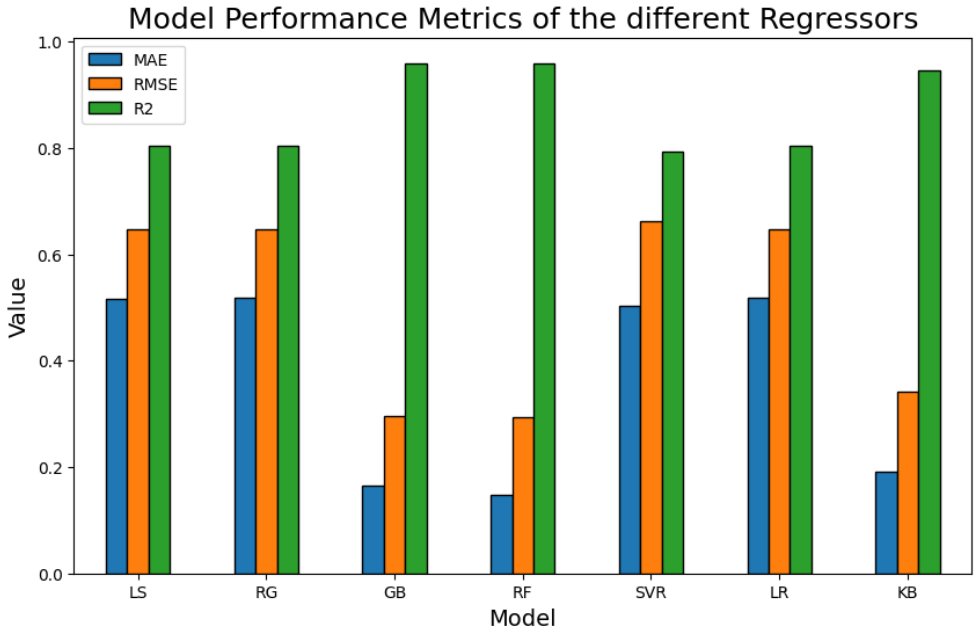

Fig. 3 Model performance of the different regressors

Based on their performance on the training and testing in Fig.4, the RF regression model performed better on both training and test data than the GB regression model. The RF model had a lower MSE, RMSE, and MAPE on the test data, indicating better performance in making accurate predictions on new data. Additionally, the RF model had a higher R2 and Adj. R2 on the test data, indicating that it explains more variance in the target variable. Therefore, the RF regression model is the better-performing model.

Model performance of RandomForest regression on train data

	MSE	RMSE	MAE	R2	Adj. R2	MAPE
0	0.049282	0.221995	0.092818	0.977049	0.977035	2.434744e+08

Model performance of RandomForest regression on test data

	MSE	RMSE	MAE	R2	Adj. R2	MAPE
0	0.086145	0.293505	0.147731	0.959571	0.959474	4.609006e+07

Model performance of Gradient Boosting regression on train data

	MSE	RMSE	MAE	R2	Adj. R2	MAPE
0	0.080265	0.283312	0.159722	0.962619	0.962597	4.150045e+08

Model performance of Gradient Boosting regression on test data

	MSE	RMSE	MAE	R2	Adj. R2	MAPE
0	0.087831	0.296362	0.164699	0.95878	0.958681	2.272803e+07

Fig.4 Model performance of RF and GB on test and training data set

Furthermore, using K-fold cross-validation, the performance of random forest regression was slightly improved. Increasing the value of k from 5 to 15 did not improve the model's performance.

Table. 3 Performance of RF using cross-validation

K = values	MSE	RMSE	MAE	R2	Adj. R2	MAPE
5	0.084	0.290	0.145	0.960	0.960	4.391e+08
10	0.083	0.287	0.145	0.961	0.961	3.386e+08
10	0.083	0.288	0.144	0.961	0.961	3.495e+08

3.4 Classification

The performance of different classifiers while setting the rating as the target variable is shown in Fig 5. Notably, the random forest classifier exhibited superior accuracy and F-1 score when compared to other classifiers. Consequently, random forest classifier was applied in the classification of the dataset with Genre and Platform as the target variables. The subsequent performance is presented in Table 4.

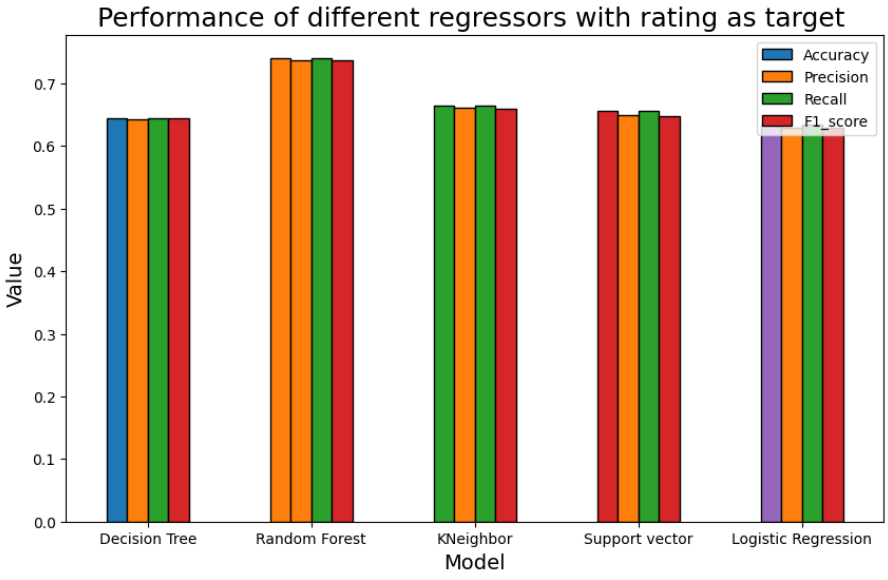

Fig.5 Model performance of different regression using rating as the target

Table 4: Performance of different categorical variables

Feature	Accuracy	Precision	Recall	F1-score
Rating	0.741	0.736	0.741	0.736
Genre	0.550	0.547	0.550	0.547
Platform	0.709	0.713	0.709	0.709

Based on the results presented in Fig 6, it can be concluded that the model achieved the best performance when the target variable was set as the rating. This is evident from the highest values of accuracy, precision, recall, and F1-score obtained for this model, compared to the other two models. Hence, setting the rating as the target variable is the optimal choice for accurately classifying the data set. From Table 5, the model performance was not improved by cross-validation; even with the increase in the k value from 5 to 15, the model still did not perform better.

Table 5 Model performance using cross-validation with rating set as target

K values	Accuracy	Precision	Recall	F1-score
5	0.726	0.719	0.727	0.723
10	0.735	0.729	0.735	0.733
15	0.740	0.736	0.739	0.738

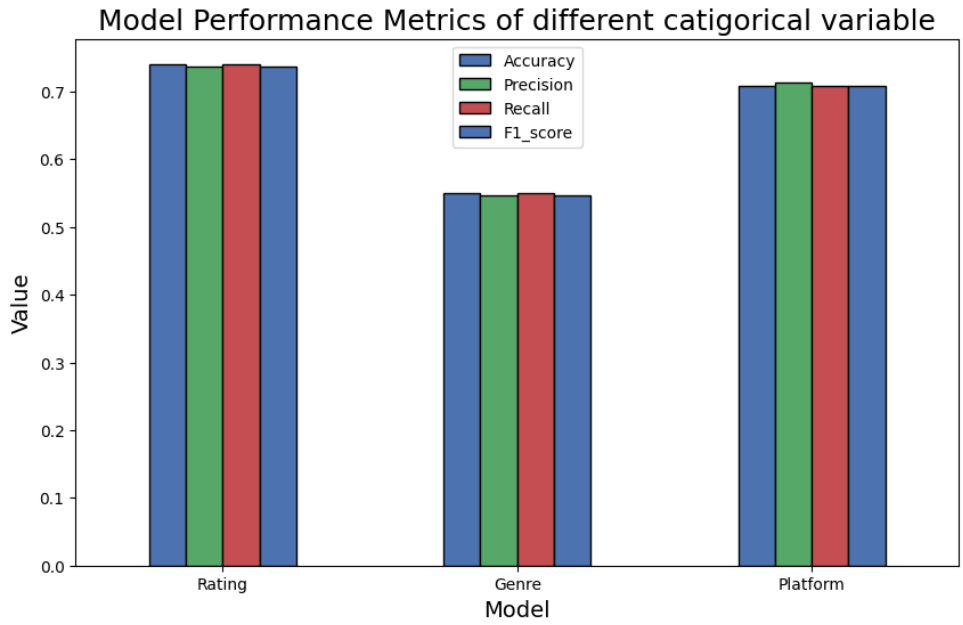

Fig. 6 Performance of different categorical variable in classifying our data set

3.5 Checking for Overfitting

From Table 6, the training set has very high accuracy, precision, recall, and F1-score, suggesting the model overfitting the training data. Furthermore, the Fig 7 learning curve illustrates that the training accuracy is exceptionally high. In contrast, the testing accuracy is comparatively lower and not exhibiting significant improvements with the inclusion of additional training samples. This suggests that the model is simply memorizing the training data and not generalizing well on new data.

Table 6: Model performance on the training and testing data set

Data set	Accuracy	Precision	Recall	F1-score
Training	0.996	0.996	0.996	0.996
Testing	0.741	0.736	0.741	0.736

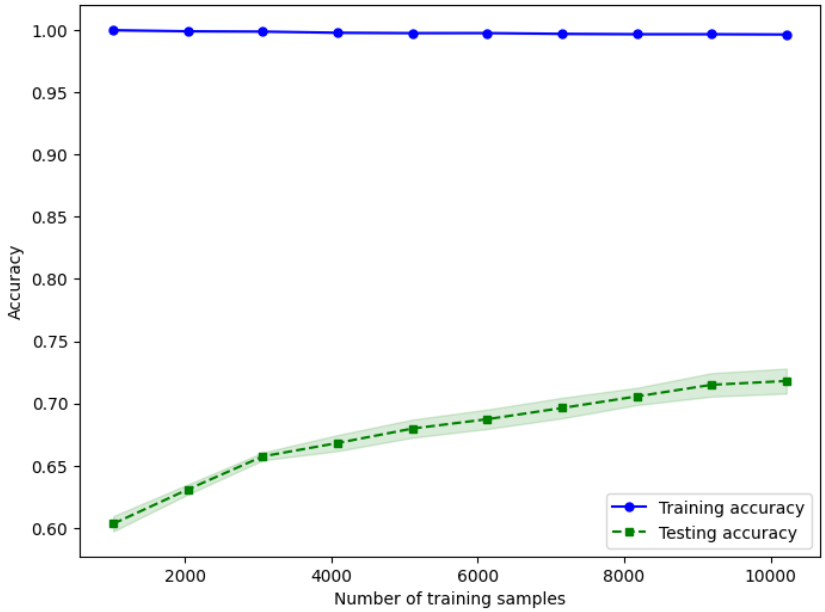

Fig. 7 Performance of the model on the training and testing data set

3.6 Can the model be deployed in practice

Based on the results presented in Table 6 and Fig.7, it can be concluded that the model did not perform well on new data. Although the accuracy, precision, recall, and F1-score on the testing set are moderately satisfactory, they are not high enough. This suggests the model may require additional optimization to perform well on new data. Therefore, deploying the model in practice is not recommended until further improvements are made.