COMPONENT TWO

Video Game Sales Dataset

Abstract

The aim of this report is to predict the sales performance of several video games, classify and cluster the video games based on some categorical variables. The analysis was carried out on the video games sales dataset obtained from Kaggle. The dataset contains information on video games including the platform, critic and user score, genre, year of release, platform, and global sales. The report discusses the methods used to achieve this goal, including regression analysis, classification, and clustering. The performance of the models is evaluated using quantitative metrics and findings are presented in this report.

Introduction

Video games have become a popular form of entertainment worldwide. The video game sales dataset used in this report contains information on various variables that are potentially useful in predicting video game sales. The primary objective of this report is to investigate which of the variables best predicts global sales of video games.

Methodology

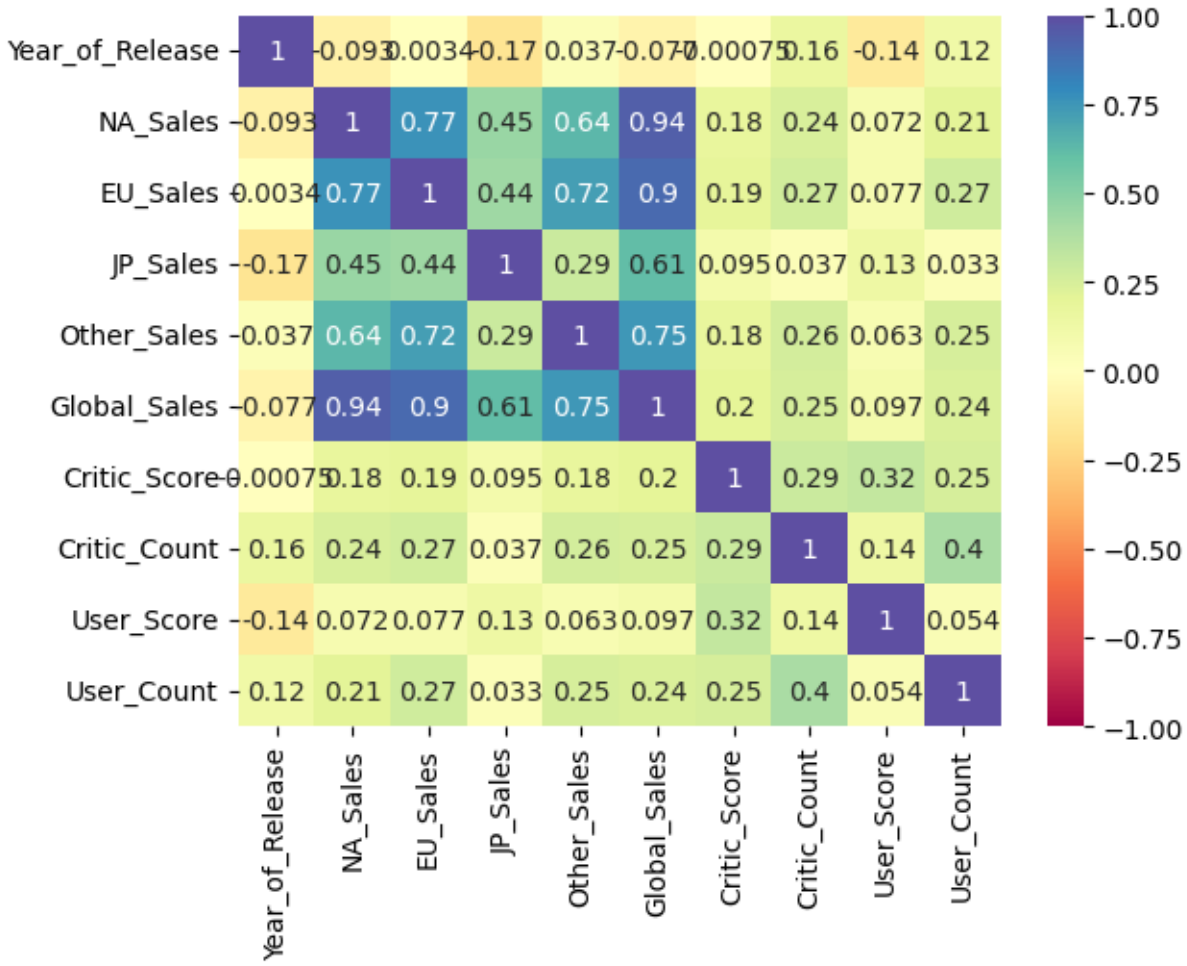
We perform regression analysis to predict video games sales based on the available variables. The effect of the number of critics and user score is analyzed. The performance of different classification models is evaluated, and clustering analysis and the performance of the models is evaluated using internal and external evaluation metrics.

Addressing the questions asked

a. Which of the variables in the video game data set or a combination of them best predicts global sales of video game and why

Answer

I plotted a heat map showing the correlation of each variable with Global Sales.



From the correlation matrix, the strongest correlation with Global Sales is the NA Sales, which is 0.94, followed by EU Sales which is 0.9 and Other Sales which is 0.75. Critic _Score and critic count also have a moderate correlation with global sales. User score and User count do not show a clear correlation with global sales.

To verify this, I performed different regression models to determine which variables best predict the global sales of video games.

1. **LINEAR REGRESSION**

```
Coefficients: [ 7.54513063e-01  4.41863469e-01  3.08071596e-01  1.80867589e-01
 -9.89409289e-06 -6.79430732e-05 -5.50430044e-05 -1.05053474e-04]
Intercept: 0.5255543925233646
The result of the mean squared error for the linear regression is : 2.783484696132191e-05
The coefficient of determination for the linear regression is: 0.9999932569888903
```

From these results, we can see that the Linear Regression model performed very well with a low MSE of 2.78e-05 and a R-squared value of 0.999993.

Coefficient for each variable

	Variable	Coefficient
0	NA_Sales	0.754513
1	EU_Sales	0.441863
2	JP_Sales	0.308072
3	Other_Sales	0.180868
4	Critic_Score	-0.000010
6	Critic_Count	-0.000055
5	User_Score	-0.000068
7	User_Count	-0.000105

The coefficients of the predictor variables indicate their impact on the target variable in the multiple linear regression model for predicting global sales of video games. We can see that NA Sales has the highest coefficient, followed by EU Sales, JP Sales, and Other Sales. NA Sales has the strongest impact on the target variable, followed by the other sales regions.

On the other hand, the variables Critic Score, Critic Count, User Score, and User Count have negative coefficients, indicating that they have a negative relationship with global sales.

2. RANDOM FOREST REGRESSION

```
The result of the mean squared error for Random Forest is : 0.7303005363192598
The coefficient of determination for Random Forest Regression is: 0.8230841852069443
```

From these results we can see that the model performed well with a low MSE of 0.73 and a R squared value of 0.8230.

	Feature	Importance
0	NA_Sales	0.847200
1	EU_Sales	0.100299
2	JP_Sales	0.032692
3	Other_Sales	0.015752
4	User_Count	0.001377
5	User_Score	0.001278
6	Critic_Score	0.000703
7	Critic_Count	0.000699

The most important feature for predicting global sales is NA Sales. This is followed by EU Sales, JP Sales, Other sales. Random Forest Regressor model shows that the most important variables for predicting global sales of video games are the sales performance in North America, Europe and Japan.

3. RIDGE REGRESSION

```
Coefficients: [ 7.54477136e-01  4.41860571e-01  3.08074266e-01  1.80878657e-01
 -7.05437682e-06 -6.77296659e-05 -5.07533973e-05 -1.01165761e-04]
Intercept: 0.5255543925233646
The result of the mean squared error for ridge regression is : 2.7819147401899513e-05
The coefficient of determination for ridge regression is: 0.9999932607921196
```

	Variable	Coefficient
0	NA_Sales	0.754477
1	EU_Sales	0.441861
2	JP_Sales	0.308074
3	Other_Sales	0.180879
4	Critic_Score	-0.000007
6	Critic_Count	-0.000051
5	User_Score	-0.000068
7	User_Count	-0.000101

From these results, we can see that the Ridge Regression model performs very similarly to the Linear Regression with a high R squared value of 0.9999 and a low mean squared error of 2.78e-05 in predicting the global sales of video games. The model indicates that NA Sales has the highest coefficient, followed by EU Sales, JP Sales, and Other Sales. This suggests that sales in North America have the strongest impact on global sales, followed by sales in Europe and Japan. The impact of critic and user scores and counts is very small and has a negative impact on the global sales of video games.

The mean squared error and R-squared values indicate that the model fits the data very well, with a very high degree of accuracy in predicting the global sales of video games. The Ridge Regression model is a good choice for this dataset, providing accurate predictions and insight into the impact of different independent variables on the sales performance of video games.

4. SUPPORT VECTOR REGRESSION

```
Support Vector Regression MSE: 0.005977788737775397
Coefficients: [[ 0.75047819  0.43755661  0.30844844  0.18189074 -0.00222653  0.00464231
   0.00192165  0.0007604 ]]
Intercept: [0.60202856]
The result of the mean squared error is : 0.005977788737775397
The coefficient of determination is: 0.9985518765047956
```

The SVR model with a linear kernel achieved a very high coefficient of determination of 0.9985 and a low mean squared error of 0.00597, indicating that it fits the data very well. The model indicates that NA Sales has the highest coefficient, followed by EU Sales, JP Sales, and Other Sales. This suggests that sales in North America have the strongest impact on global sales, followed by sales in Europe and Japan. The impact of critic and user scores and counts are very small and does not really have any significant impact on the global sales of video games.

5. DECISION TREE REGRESSION

```
The result of the mean squared error for decision tree regressor is : 0.6068986036952043
The coefficient of determination for decision tree regressor is: 0.8529783895399375
```

	Feature	Importance
0	NA_Sales	0.850569
1	EU_Sales	0.108908
2	JP_Sales	0.027093
3	Other_Sales	0.010557
4	Critic_Count	0.001037
5	Critic_Score	0.000807
6	User_Count	0.000746
7	User_Score	0.000281

The output shows that the mean squared error is 0.60 and the R-squared value is 0.8505. This indicates that the model has a good fit and accurately predicts the global sales of video games. This suggests that sales in North America have the strongest impact on global sales above other features.

Based on the R-squared values and MSEs, Linear Regression and Ridge Regression both performed very well in predicting global sales, with R-squared values close to 1 and very low MSEs. The other regression models also have relatively good performance, but not as good as these two. Therefore, Linear Regression and Ridge Regression are the best choices for predicting global sales based on these results.

From all the regressors, we can see that the variables that best predict global sales are the NA Sales, EU Sales, JP Sales and other sales The sales performance in North America has the highest positive impact on the global sales, followed by the sales performance in Europe and Japan, and finally the sales performance in other regions.

b. What effect will the number of critics and users as well as their review scores have?

```
North America Sales:
Critic_Count: 0.007747731649555218
User_Count: 0.00025171333387730187
Critic_Score: 0.008118862694802938
User_Score: 0.003435461054734925
------------------------
EU Sales:
Critic_Count: 0.00475919836231607
User_Count: 0.00024135085589987678
Critic_Score: 0.0045632150248557345
User_Score: 0.0028062101301833562
------------------------
Japan Sales:
Critic_Count: -1.9421629402386274e-05
User_Count: 1.064812882232974e-05
Critic_Score: 0.0017683749319312462
User_Score: 0.012348620289477568
```

Based on the coefficients provided, we can see that for North America and EU sales, none of the variables seem to have a significant impact on sales, as all the coefficients are very small. However, for Japan sales, we can see that the User Count has the highest coefficient of 1.06 million units, which suggests that it has the strongest effect on sales in Japan and the critic count also have a negative coefficient of -1.94 on Japan sales which means that for the critic count, sales in Japan decrease by 1.94 million units.

c. What propelled the choice of your regressor for this task?

Answer

Linear Regression: Linear regression is commonly used for predictive modelling techniques. It can handle numerical variables and can interpret and understand the effects of the predictor variables on global sales. The linear regression performed very well in predicting the global sales with a low mean squared error of 2.78e-05 and a high R-squared value of 0.9999.

Random Forest Regression: Random Forest Regression is a powerful algorithm that can handle numerical variables and complex datasets. It can help identify which variables are most important in predicting global sales. The model has a good fit and accurately predicts global sales with a low mean squared error of 0.73 and high R-squared value of 0.8230.

Ridge Regression: Ridge regression is a regularized linear regression model that can handle many features relative to the number of samples in the dataset. Since the video game sales dataset likely contains many features. The ridge regression performed very well in predicting the global sales with a high R squared error of 0.9999 and a low mean squared error of 2.78e-05.

Support Vector Regression: SVR is a robust and versatile regression technique that can handle a wide variety of regression tasks, including video game sales prediction. SVR is good for predicting the sales performance of video games in the dataset due to its ability to handle non-linear relationships, outliers, and its flexibility in kernel functions. The SVR model with a linear kernel achieved a very high coefficient of determination of 0.9985 and a low mean squared error of 0.00597, indicating that it fits the data very well.

Decision Tree Regression: Decision tree regression can help to identify the features that are most important in predicting video game sales. The decision tree regression accurately predicts the global sales with a high R squared error of 0.8505 and a low mean squared error of 0.60.

d. Use all the relevant categorical variables in the video game dataset as the target variable at each instance and determine which of the variables performed best in classifying the dataset. Explain your findings.

Answer

This is the accuracy scores data frame that contains the accuracy scores for each variable and classifier.

Variables	Random Forest	Naïve Bayes	Logistic Regression	Decision Tree
Rating	0.80	0.42	0.69	0.75
Platform	0.58	0.21	0.19	0.51
Genre	0.41	0.18	0.22	0.36
Publisher	0.38	0.06	0.12	0.34
Developer	0.60	0.18	0.42	0.55

From these results, The Random Forest classifier performs the best for all target variables while Naïve Bayes performed poorly. However, the Rating variable performs the best overall, with the highest accuracy scores for all classifiers. This suggests that the Rating variable is the most informative in predicting the sales of video games. On the other hand, the Genre and publisher variable has the lowest accuracy scores for all classifiers, indicating that it is the least informative in predicting the sales of video games. This could be because the Genre and publisher of a video game does not necessarily correlate with its sales. Overall, the Rating variable is the most important variable in predicting the sales of video games in this dataset. The Genre and publisher variable is the least important in predicting sales.

e. How did you check whether your models did not overfit?

Answer

Overfitting was checked by comparing the performance of the model on the training and testing set.

f. Can your classification models be deployed in practice based on their performance?

Answer

Based on the poor performance of the classification models in classifying the video game dataset using categorical variables, it is not advisable to deploy these models in practice.

g. In the video game dataset, use a relevant categorical variable and other relevant non-categorical variables to form groups at each instance. By employing internal and external evaluation metrics, determine which categorical variable best describes the groups formed.

Answer

Rating

```
External Evaluation Measures
******************************
V-measure Score: 0.081
Rand Index Score: 0.186
Mutual Information Score: 0.081

Internal Evaluation Measures
******************************
Davies-Bouldin Index: 1.241
Silhouette Coefficient: 0.462
Calinski Harabasz Score: 4362.943
```

The external evaluation metrics V-measure score, Rand Index score, and Mutual Information score evaluate how well the clustering algorithm grouped the data based on the true categories. In this case, the V-measure score is 0.081, which is a moderate score and indicates that the algorithm was slightly effective at grouping the data. The Rand Index score is 0.186, which is better than random grouping but still relatively low. The Mutual Information score is 0.081, which indicates a weak relationship between the true categories and the resulting clusters.

The internal evaluation metrics evaluate the quality of the clustering algorithm. The Davies-Bouldin Index is 1.241, which is relatively high. The Silhouette Coefficient is 0.462, which the is fair. The Calinski Harabasz score is 4362.943, which is relatively high and indicates good clustering quality.

Overall, the results suggest that the algorithm is slightly effective at clustering the data based on the categorical variable (Rating) and other non-categorical variables, as shown by the moderate external evaluation metrics scores. Therefore, the Rating Variable best describes the group formed compared to other categorical variables.