**NJOKU CHDIDIEBERE CHIDUBEM**

**Dataset Link**: https://www.kaggle.com/datasets/ibriiee/video-games-sales-dataset-2022-updated-extra-feat

**THE TASK**

The taskis to predict the sales performance of several video games worldwide; to classify and cluster the video games based on some categorical variables. Attempt the following questions: a. Which of the variables in the video game dataset or a combination of them best predicts “global sales” of video games and why? Provide quantitative justifications for your answers. b. What effect will the number of critics and users as well as their review scores have on the sales of Video games in North America, EU and Japan? c. What propelled the choice of your regressor for this task? Aptly discuss with quantitative reasons! 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.

**1.0 ABSTRACT**

This report is about a video game dataset that offers current data on the performance of sales and global popularity of numerous video games. This report addresses the cleaning of the dataset, justification of why multiple linear regression variables best predicted global sales on video games by using cross-validation evaluation metrics and providing effects of critics and users on sales of the game on NA Sales, EU Sales and Japan Sales, justification of why the rating was the best categorical variable in data classification and why my classification model can be deployed in practice based on their performance.

**2.0 INTRODUCTION**

The essence of data cleaning can not be overemphasized as duplicates must be eliminated, blanks must be filled in, and errors must be fixed during the cleaning process. Using machine learning methods, a predictive model is created after the dataset has been cleaned, which allows it to analyze the data and produce forecasts.

**3.0 METHODOLOGY**

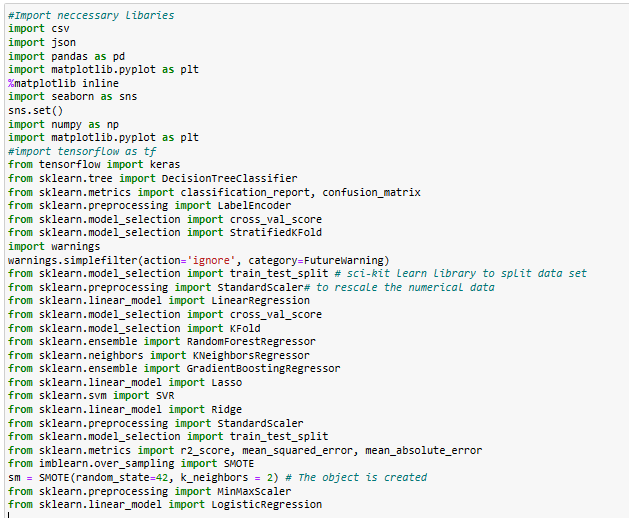
**3.1 DATA COLLECTION, TOOLS AND CLEANING**

**3.1.1 DATA COLLECTION**

I obtained the dataset from Kaggle on the performance of sales and global popularity of video games. The dataset contained 16719 rows and 16 columns which include Name, platform, year of release, genre, publisher, and sales in North America, Europe, Japan, and other territories are all included in the data. The critic score, the critic count, the user score, the user count, the developer, and the rating.

**3.1.2 IMPORTATION OF NEEDED LIBRARIES**

The needed libraries were imported to enable us to access varieties of arrays,functions and cleaning



**Fig1**-imported libraries

**3.1.3 DATA CLEANING**

Qualitative care was taken in the data cleaning process, which entailed the meticulous inspection of each data column. Series of checks were conducted, which includes:

**DATA TYPES**: After careful consideration of data type of each column, it was observed that user score was needed to be converted to float from an object.

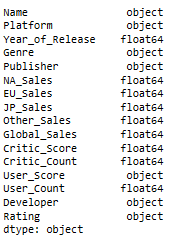
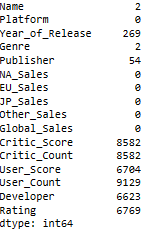
 

Fig2-Information about the dataframe fig3-Dataframe before cleaning with null values

**Dropping entries:** Two rows with missing name were taken out of the dataset under the ‘Name column’, as we cannot logically establish a relationship as to what name to fill in there.

**Null and blank values:** The year of release column, was cleaned up by applying linear interpolation and filling up the null values because it is a numeric data and to avoid skewing our dataset.

The Publisher null values were cleaned up and replaced by the mode because it is categorical variable and the most common value in the column.

The critic score was cleaned by applying linear interpolation and filling up the null values because it is a numeric data and to avoid skewing our dataset.

The critic count was cleaned by applying linear interpolation and filling up the null values because it is a numeric data and to avoid skewing our dataset.

The user score was cleaned by first converting the TBD values to nan, converted to float and finally applied linear interpolation because we don’t need the column to have an object for easier data manipulation

The user count was cleaned by applying linear interpolation and filling up the null values because it is a numeric data and to avoid skewing our dataset.

The developer column was cleaned by applying mode because it is categorical and the most common value in the column.

The rating column was cleaned by applying mode because it was categorical and the most common value in the column.

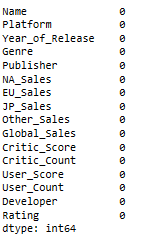
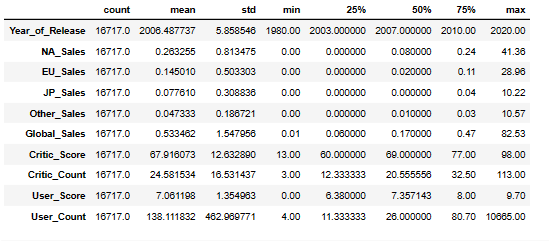


Fig4-Dataframe after cleaning

Our cleaned dataset we will now be using for training , testing , predicting and evaluation becomes :

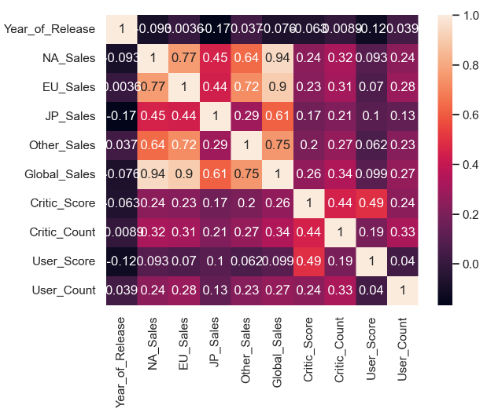


**Fig5**- Statistical overview of the dataset

**3.2 FEATURE SELECTION**

I selected important features based on the domain knowledge and most importantly based on the correlation analysis as shown in fig5 below

**DATA EXPLORATORY ANALYSIS**



**Fig 5-**Heatmap correlation showing the level of impact of several sales, critic and user counts have on each other.

**3.3 MODEL SELECTION AND EVALUATIONS**

I evaluated several machine learning algorithms but Multiple linear regressors was chosen as the best model to predict Global\_Sales using cross-validation evaluation metric because of its high accuracy and least minimal errors.

In addition, the critic, users and their review scores had less impact on our dataset.

Furthermore, the best regressor used was the gradient boosting regressor because it had the best performance metric (highest coefficient of determination) than others.

The rating was the best categorical variable in the data classification because it had the best performance metric measures using random forest.

The rating was use the best categorical variable that described the groups formed because it had the best internal and external evaluation metric

**3.4 PERFORMANCE METRIC**

I evaluated the performance of the respective models using several metrics like accuracy , coefficient of determination ,precision , recall , f1score , root mean square, root absolute error, v-measure score ,silhouette coefficient, Davies-Bouldin index.

**4.0 RESULT:**

1. Multiple linear regression(NA\_Sales, EU\_Sales,JP\_Sales and Other\_Sales) best predicted the global sales because it had the least minimal errors showing that the predicted values are close to the actual values and and highest coefficient of determination which is 1 showing a perfect fit for the model.

|  |  |  |
| --- | --- | --- |
| Performance evaluation metric | Reference Model(MultipleLinear regression) | Second model (Simple linear regression) |
| Coefficient of determination(R^2) | 1 | 0.86 |
| Root mean square error(RMSE) | 0.01 | 0.55 |
| Mean absolute error(MAE) | 0.00 | 0.2 |
| Mean square error(MSE) | 0.00 | 0.3 |

Fig6-performance measuring metrics table

Ref: Reference model (multiple linear regression here is our reference model)

Justification:

Percentage change in Coefficient of determination (%) = Ref (R^2) - Second model(R^2) × 100

REF R2

Percentage change in R^2= 1 - 0.86 × 100

1. =14% .

From the above table fig5, we picked the best model and the second best model for our justification:

There was an increase of 14% in percentage of coefficient of determination showing that multiple linear regression has the highest Coefficient of determination(R^2) which is 1 and had least number of errors.

1. The effect that the number of critic scores and users have on the sales of video games in North America, EU and Japan is actually less impactful/effective as shown by the heatmap correlation in fig 6. This is also buttressed by the low performance evaluation metric values seen in our respective models, where the coefficient of determination ranges from (0 to 0.21).
2. The regressor i chose was gradient boosting regressor because it gave me the best coefficient of determination with respect to NA\_Sales, EU\_Sales and JP\_Sales. it also gave me the least number of errors.

|  |  |  |  |
| --- | --- | --- | --- |
| Performance evaluation metrics | NA\_Sales Values | EU\_Sales values | JP\_Sales Values |
| R^2 | 0.21 | 0.15 | 0.008 |
| RMSE | 0.74 | 0.47 | 0.30 |
| MAE | 0.24 | 0.15 | 0.11 |
| MSE | 0.54 | 0.22 | 0.09 |

Fig7 Performance evaluation metric using gradient boosting regressor

|  |  |  |  |
| --- | --- | --- | --- |
| Performance evaluation metrics | NA\_Sales Values | EU\_Sales values | JP\_Sales Values |
| R^2 | 0.15 | 0.16 | 0.06 |
| RMSE | 0.76 | 0.47 | 0.30 |
| MAE | 0.28 | 0.17 | 0.12 |
| MSE | 0.15 | 0.22 | 0.09 |

Fig8 Performance evaluation metric using multi linear regressor

|  |  |  |  |
| --- | --- | --- | --- |
| Performance evaluation metrics | NA\_Sales Values | EU\_Sales values | JP\_Sales Values |
| R^2 | 0.18 | 0.16 | 0.01 |
| RMSE | 0.75 | 0.47 | 0.31 |
| MAE | 0.22 | 0.15 | 0.13 |
| MSE | 0.56 | 0.22 | 0.31 |

Fig9 Performance evaluation metric using support vector regressor

From the above tables we can see that gradient boosting regressor performed the best with the coefficient of determination of 0.21 for NA\_Sales, 0.15 for EU\_Sales and 0.008 for JP\_Sales, which are averagely high(closer to 1) compared to other values given by other type of regressors. In addition, gradient boosting regressor had the least number of errors which are closer to zero compared to the other type of regressors.

1. The variable that performed best in classifying the dataset in Rating is NA\_Sales','EU\_Sales','JP\_Sales','Other\_Sales' using random forest classifier with an accuracy of 63% which is more than others.

E. I checked if the accuracy of the models is consistent across all Kfolds

F. Yes , My classification can be deployed in practice based on their performances because they performed better with higher accuracy and other measuring metrics.

G. The best categorical variable that best describes the groups formed is rating using kmeans, as it has high v-measure of 0.055, silhouette coefficient of 0.192 and Davies Bouldin(DB) index of 1.751

**5.0 CONCLUSION**

In conclusion the best models were chosen because of their optimal evaluation performance metrics.