Analysis on the Sales Performance of Video Games

1 Introduction

Companies around the world use a variety of techniques to understand the relationship between different variables and their products and to assess how these variables affect financial performance. In addition, these techniques are also used to predict financial performance in other to evaluate if the business is on course to meet set financial goals. The outcome of several analysis conducted a video games dataset extracted from Kaggle in presented in this report. The sections below provide an overview of the methodologies adopted, the experiments conducted and the discussions about these results.

2 Methodology

2.1 Data collection and modelling

The dataset used for the analysis was extracted from Kaggle (Naeem, 2023) and includes information relating to some of the factors that could possibly affect the sales of video games. This information include:

- **Name**: The name of the video game.
- **Platform**: platform on which the game was released, such as PlayStation, Xbox, Nintendo, etc.
- **Year of Release**: The year in which the game was released.
- **Genre**: The genre of the video game, such as action, adventure, sports, etc.
- **Publisher**: The company responsible for publishing the game.
- **NA Sales**: The sales of the game in North America.
- **EU Sales**: The sales of the game in Europe.
- **JP Sale**: The sales of the game in Japan.
- **Other Sales**: The sales of the game in other regions.
- **Global Sales**: The total sales of the game across the world.
- **Critic Score**: The average score given to the game by professional critics.
- **Critic Count**: The number of critics who reviewed the game.
- **User Score**: The average score given to the game by users.
- **User Count**: The number of users who reviewed the game.

- **Developer**: The company responsible for developing the game.
- **Rating**: The rating assigned to the game by organizations such as the ESRB or PEGI.

The data was cleaned using a variety of techniques for imputing missing and null values. Also, models such as Random Forest, Logistic Regression, Linear Regression, K-Means were employed in the process of answering various questions about the data. The experiments were carried out in a Python environment and in Jupyter notebook. The models implemented during this study were carried out using the Scikit-Learn Library.

3 Results and Discussion

3.1 Predicting Global Sales

This section of the report includes an analysis of the variables in the dataset or a combination of them that best predicts "global sales" of video games. To do this, Linear Regression, Random Forest, and Gradient Boosting Regressors were implemented. From the analysis of the model coefficients and feature importance observed, NA, EU, JP, and other sales individually and collectively have a significant important effect on global sales. For instance, in the case of NA sales the coefficient for the Logistic Regression model is 0.863644. The simple interpretation of this is that for every 1-unit growth in NA sales, global sales increases by 0.863644 unit assuming all other variables are constant. This show a significant correlation of NA sales with global sales and is a valid predictor of global sales in the absence of data about global sales. The same logic applies in terms of the quantitative effect of variables such as EU sales, JP sales, other sales, all of which have positive and relatively high coefficients with global sales. Table 1 below presents relevant metrics on the performance of the model and the coefficients of the variables in the dataset.

Also, the same logic holds true for the feature importance of the Random Forest and Gradient Boosting Regressors. However, other than these variables, the results of the models show that the other variables have little to no effect on the outcome of global sales with some of them showing negative impact on global sales.

Nonetheless, one important point to investigate is the effect of these other variables on the best predictors of global sales – for example, if NA, EU, JP, and other sales have positive and strong correlation with global sales, what variables in the dataset influence the sales in the regions and would be seen as an indirect effect on global sales? This question will be answered in the next section.

In summary, NA, EU, JP, and other sales are the best predictors of global sales. Across all 3 models of Logistic Regression, Random Forest, and Gradient Boosting regressors observed, NA sales best predicts the video games sales globally with coefficients of 0.863644, 0.76464, and 0.690188 respectively

Table 1: Results of the regression analysis

Variable	Linear Regressor Coefficients	Random Forest Feature Importance	Gradient Boosting Feature Importance
NA Sales	**0.863644**	**0.76464**	**0.690188**
EU Sales	**0.528437**	**0.14977**	**0.204638**
JP Sales	**0.313251**	**0.02393**	**0.041656**
Other Sales	**0.193753**	**0.02132**	**0.025684**
Rating	-0.008221	0.00170	0.00331
Platform	0.007836	0.00153	0.000117
Publisher	-0.006563	0.00096	0.000059
Developer	-0.006374	0.00072	0.000003
Genre	-0.006193	0.00058	0.000003
Year of Release	0.000580	0.00000	0
Name	0.000333	0.00000	0
User Count	0.000244	0.00000	0
User Score	-0.000188	-	0
Critic Count	-0.000138	-	0
Critic Score	-0.000076	-	0

3.2 What is the effect of critics and users scores and reviews on regional sales?

The analysis conducted aimed to understand the relationship between sales in the different regions and critic and user counts and scores. Two analyses were conducted which included:

1. A correlation analysis
2. Multiple linear regression analysis

Table 2: Results of the correlation analysis

Variable	NA Sales	EU Sales	JP Sales
User Score	7.6%	5.9%	10.5%
User Count	18.1%	23.9%	5.9%
Critic Score	16.2%	16.6%	12.5%
Critic Count	20.4%	22.6%	8.8%

Table 2 above presents the correlation analysis user score, user count, critic score and critic count with sales in North America, EU, and Japan. From the above, the highest correlation observed is between

User count and sales in EU at 23.9%. Across all the variables, the correlation scores are below 50% and the analysis does not suggest significant direct impact on sales across all the regions.

By analysing the results of the multiple linear regression, the following conclusions were reached:

- For NA sales, there are positive relationships with critic scores, count, and user count with coefficients scores of 0.007, 0.007 and 0.0002 respectively unlike user score with negative relationship of -0.0109. The coefficients indicate that for example in the case of a one-unit increase in critic score and count, there is an increase of 0.0077 increase in NA sales with all other variables constant.
- The same trend as above holds true for EU sales with positive relationships with critic score, count and user count, and user score having a negative relationship.
- However, in the case of Japan, all the variables have positive relationships with sales.

Looking at the results of the regression analysis in conjunction with the correlation analysis, a deduction can be made that for NA and EU sales, other than user score, all other variables have a positive relationship with sales and an increase or decrease potentially have an impact, albeit minimal. On the other hand, for Japan, all the variables have a positive relationship with sales.

Table 3: Coefficients for multiple linear regression

Variable	NA Sales	EU Sales	JP Sales
User Score	-0.0109	-0.0163	0.0130
User Count	0.002	0.002	0.00001469
Critic Score	0.007	0.0050	0.0019
Critic Count	0.007	0.0048	0.0012

The analysis above show that these variables do not have a significant impact on sales in these regions. Also, this analysis validates the results of the analysis carried out earlier on the effect of all the variables in the dataset in predicting global sale.

3.3 What propelled the choice of your regressor for this task? Aptly discuss with quantitative reasons.

In predicting global sales, as noted earlier, Linear Regression, Random Forest, and Gradient Boosting Regressors were implemented. The performance of the models was evaluated using 4 common regression metrics of R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Of all the models, across the selected models the Linear regressor performed the best achieving the highest R-squared of 0.99 for the training and validation data and lowest MAE (0.05, 0.05), MSE (0.000006, 0.000006)and RMSE (0.0024, 0.0077) scores across the training and testing data respectively.

Therefore, regressor of choice based on the analysis conducted so far is the Linear Regression model propelled by the fact that it achieved the best R-squared score of 0.999997368 for training data and 0.999970234 for validation data indicating that the model was able to explain a high proportion of the variance in the data, along with the other results explained above.

Table 4: Average evaluation metrics of the selected models

Model Evaluation Metric	Dataset	Linear Regressor	Random Forest Regressor	Gradient Boosting Regressor
R-Squared	Training	0.999997368	0.988734222	0.999017625
	Validation	0.999970234	0.933050823	0.949622953
MAE	Training	0.005606880	0.032799486	0.035639721
	Validation	0.005606880	0.032799486	0.035639721
MSE	Training	0.000006005	0.029234006	0.002304043
	Validation	0.000060545	0.227603509	0.181742514
RMSE	Training	0.002450509	0.164266988	0.047922088
	Validation	0.007779946	0.383913822	0.322659690

3.4 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.

Although the dataset contained six categorical variables, the relevant categorical variables that were used for the analysis include Genre and Rating. These target variables were selected after deciding against using any variable with more than 20 classes. Random Forest and Logistic Regression classifiers were the models used for the analysis.

Table 5: Performance metrics of classifiers

Target variable	Variable	Random Forest	Logistic Regression
Genre	Accuracy	70%	63%
	Precision	74%	62%
	Recall	63%	57%
	F1-Score	67%	59%
Rating	Accuracy	79%	77%
	Precision	80%	75%
	Recall	72%	72%
	F1-Score	75%	73%

Of the two models that were used to classify the data, Random Forest performed best across both target variables with 70% accuracy versus 63% for Logistic Regression for the genre target variable. In the case of the rating target variable, again Random Forest outperformed the Logistic Regression model with its 79% accuracy score and 77% respectively. Overall, the models performed better on the rating target variable versus the genre target variable, and this can be attributed to the fact that the genre column had 12 unique labels versus 4 unique labels for rating. Therefore, based on the results of the classifiers and an assumption that the number of labels potentially influences the performance of classifiers, the rating label performs best in classifying the data.

3.5 How did you check whether your models did not overfit?

Generally, models are considered to overfit if they perform well on the training data but perform poorly on the test data. In the case of the classification models, the training and test scores of the models were assessed in determining if the models overfit. All 4 models that were developed across the 2 target variables were observed to overfit as showed in the table below, achieving high scores on the training data but performing poorly on the test data. The table below presents the training and testing scores of the classifiers.

Table 6: Accuracy score of classifiers

Target variable	Metric	Random Forest	Logistic Regression
Genre	Training Accuracy	100%	88%
	Test Accuracy	70%	63%
Rating	Training Accuracy	100%	92%
	Test Accuracy	79%	77%

3.6 Can your classification models be deployed in practice based on their performances? Explain

The rating variable was performed best in classifying the data and the Random Forest and Logistic Regression models has accuracy scores of 79% and 77% respectively. In this context of this study, these scores are relatively good, especially when compared to the performance of the models when the genre variable was used as the target. However, to deploy the model in practice, more investigations will need to be carried out including optimizing for parameters. In addition, while cleaning the dataset, some of the labels originally in the rating column were discarded because the value counts for these labels were low at counts of 1 and therefore irrelevant for the analysis. In that case, this will need to be investigated as oversampling with SMOTE will not have worked due to a lack of neighbours for these observations. Therefore, the classification models developed using the rating column will create more problems than solutions and therefore cannot be deployed without more investigation carried out on the dataset.

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

Using the K-Means model, groups were formed with the non-categorical variables in the dataset with user and critic score being the variable of comparison for all the groups formed. Also, the categorical variables of genre and rating were used as the variables to perform the external evaluation. The internal evaluation metrics considered included the Davies-Bouldin Index, the Silhouette Coefficient, and the Calinski Harabasz Score. The external evaluation metrics considered included the v measure, Rand Index, and Mutual Information score. The results of the internal and external evaluation metrics are presented in the table below.

The results of the external evaluation metrics across all the clusters developed show that genre is better than rating in describing the groups as it is consistently higher than the metrics for rating.

Table 7: Internal and evaluation performance results for K-Means clusters

Cluster	External Metric	Genre	Rating
User Score and NA Sales	V Measure	0.016	0.011
	Rand Index	0.007	0.011
	Mutual Information Score	0.015	0.010
User Score and EU Sales	V Measure	0.011	0.005
	Rand Index	-0.001	0.005
	Mutual Information Score	0.011	0.005
User Score and JP Sales	V Measure	0.013	0.004
	Rand Index	0.000	0.003
	Mutual Information Score	0.013	0.004
User Score and Global Sales	V Measure	0.017	0.011
	Rand Index	0.007	0.013
	Mutual Information Score	0.016	0.010
User Score and Other Sales	V Measure	0.011	0.004
	Rand Index	-0.001	0.005
	Mutual Information Score	0.010	0.004
Critic Score and NA Sales	V Measure	0.011	0.004
	Rand Index	0.001	0.011
	Mutual Information Score	0.010	0.004
Critic Score and EU Sales	V Measure	0.010	0.004
	Rand Index	-0.000	0.011
	Mutual Information Score	0.010	0.004
Critic Score and JP Sales	V Measure	0.012	0.003

	Rand Index	0.001	0.011
	Mutual Information Score	0.012	0.003
Critic Score and Global Sales	V Measure	0.011	0.004
	Rand Index	0.001	0.011
	Mutual Information Score	0.010	0.004
Critic Score and Other Sales	V Measure	0.010	0.004
	Rand Index	-0.000	0.011
	Mutual Information Score	0.010	0.003

4 Conclusion

The analysis conducted aimed to understand the relationship between the variables in the video games sales dataset from Naeem (2023). The findings show that in predicting global sales, NA, EU, JP, and other sales were the best variables to do this. Also, while there does appear to be some relationship between regional sales and user count, critic score and count, it is not significant. In addition, the report presents the findings of classification and clustering experiments that were conducted.

5 References

Naeem, I. M. (2023) *Video Game Sales Dataset Updated -Extra Feat.* Available online: https://www.kaggle.com/datasets/ibriiee/video-games-sales-dataset-2022-updated-extra-feat [Accessed 11/04/2023].