IN3062 – INTRODUCTION TO ARTIFICIAL INTELLIGENCE

Coursework

Document Version 1.0

Authors: Kevin La (acvt726), Martin Doung (aczg157), Saffan Ahmed (acwf535)

TABLE OF CONTENTS

1. INTRODUCTION

Video game culture has exponentially grown to become a popular source of entertainment and consumed by modern pop culture, evolving from humble beginnings to a multibillion-dollar industry today[1]. Current predictions estimate the net global revenue of video games in 2020 to be over $160 billion dollars[2], an increase of over 7.6% from 2019. Throughout its history, the market for video games has only strengthened due to continuous advancements in computer technology, through an increase in processing power, graphical performance, and game design. With each successive year, more advanced consoles have spawned a plethora of video game titles and genres, with thousands of titles available to feed increased sale demands. This impact also meant that digital sales has exponentially risen resulting in UK gaming outlets to struggle selling physical units. Severity of the impact can be assessed from one of UK's leading commercial video game retailers – 'GAME' store[3].

1.1 PROBLEM DOMAIN

The distribution of video game sales in the UK reflects that 80% is sold digitally via micro-transactions, DLC and subscription services. Thus, digital revenue is up by 12.5% year-on-year whilst physical revenue goes down by 2.8%[4]. As a result, commercial retailers are threatened by losing profit from excess supply of physical video game units. This is reflected by the UK's prominent video games retailer 'GAME' store, case study below:

Fig 1: 'GAME' store administration.

Case Study: UK retailer GAME runs out of money, enters administration[5]

With supply of physical units increasing overtime, has meant that not enough revenue is generated from in-store transactions. To alleviate excess supply, unit pricing of popular games was reduced, resulting in profit-loss. Popularity of E-commerce overtime had led to excess overhead costs of maintaining store operations thus forcing 40 'GAME' stores to go into administration.

Digital formatted video games sold via E-commerce will continually rise, as consumers benefit with improved convenience, accessibility, and privacy. Developers of next-gen consoles will eventually want to move over to operating only on digital formatted video games in their bid to reduce carbon footprint[6] and maintain corporate social responsibility.

Our aim to help commercial video game retailers like 'GAME' to adapt to these changes, also make effective predictions on which game titles and platforms will be popular to maintain stock control i.e., best-selling. Goal is to prevent stock-overload and efficiently manage stock inventory with the purpose of maximising revenue and reducing overhead.

1.2 DATASET

The original dataset titled 'Video Game Sales with Rating' can be found on Kaggle, here is the link. The dataset itself contained more than 12,000 games including outdated gaming platforms and game titles thus opted to narrow down our dataset to the top 100 selling video games across 5 platforms that are released within the years 2010 and 2016 – PC, Xbox One, PS4, Nintendo Wii U and 3DS. To ensure accurate predictions are generated about future sales, we are focusing on video games released between 2010 & 2016, because within this date range the regional sales are consistent thus enabling better predictive analysis from which patterns can be identified. Moreover, the dataset highlights 5 individual classes: genre, publisher, platform amongst many other factors. From which correlations can be identified and could contribute to determine a video game's success rate.

2. REGRESSION

Our models will mainly be computed using supervised learning methods like Linear Regression, SVM and Random Forest Regression. The main goal is to allow us to predict new continuous data based on the trends and results of the datasets that have already been tested using the different regression algorithms and see which variables are most correlated to Video Games having high sales numbers. We will use Pearson's correlation coefficient to see which 2 variables have the best correlation. R value of 1 meaning perfect correlation, 0 meaning no correlation and -1 meaning perfect inverse correlation.

3. MISSING DATA

Having filtered the top 100 sales for each platform within time period 2010-2016, we found discrepancies within columns critic score and critic count missing. This would hinder our results as no correlation can be identified between user reviews and video games sold over time. We thought about using the mean or median of the critic score/count values to fill in the blanks, but this wouldn't be adequate. We concluded that since critic score and count is an opinion-based system, we decided to manually enter the missing data for each game which has been taken from a credible source such as 'Meta Critic' link.

4. TECHNIQUES USED

For our dataset, we decided to use **Pearson's Coefficient** in order to find out which 2 variables depict the strongest correlation that will help us predict video game sales in the future.

To better understand the Pearson Co-efficient correlation, we decided to use a coloured heatmap to represent the correlation. Greater coefficient value will correspond with a darker hue of Red and show a strong correlation and vice versa. We can see that 'sale figures' for different regions have a strong relation to Global Sales which is what we expected. For all our regression models we made Global Sales our dependent variable and as it counts for all regions, so we decided to drop the other sales features. So, the relationship between Global Sales and Region we can see that Critic Score and Critic Count have the next highest positive relationship with a value of **'0.28'** and **'0.36'**. Despite Critic Count having the higher value we decided to disregard this feature as yes it may have a relation with Global Sales however users buy games depending on the Critic Score and not Critic Count. Therefore, our main focus is Critic Score to Global Sales.

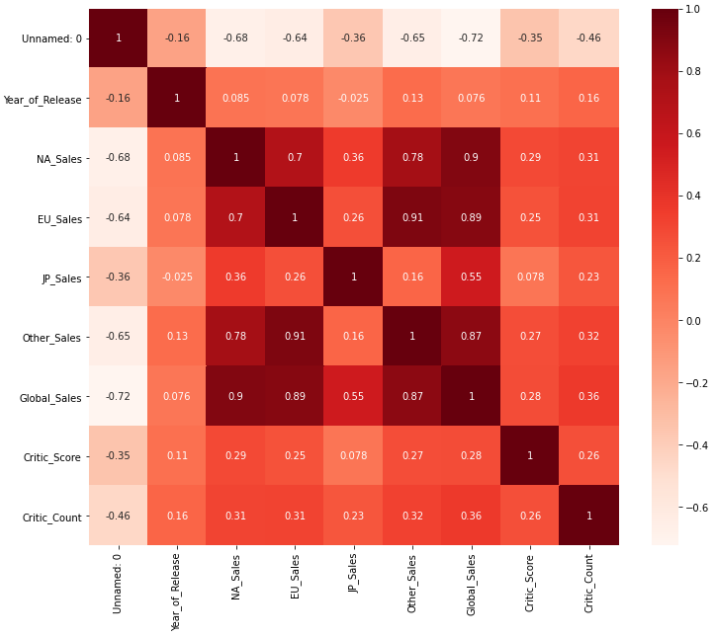


Fig 2: Heatmap for Pearson Co-efficient
- Highlights the correlation strengths between each Regions and consumers: Critic Score & Count.

5.

5. MODELS USED

5.1 LINEAR REGRESSION (LR)

LR is a statistical approach for modelling the relationship between a dependent variable with a given set of independent variables. It tries to fit data with the best hyper-plane which goes through the points. This model works well with our dataset as we look for the variation of critic scores that are dependent on global sales. The LR for our model depicts a positive correlation whereby a high critic score would equal a higher level of global sales. However, despite LR being simple to implement and interpret. This model is prone to underfitting; fails to capture data properly and often the line does not fit well due to the relationship between critic score and global sales being not entirely linear. Some data-points for global sales reflected irregular behaviour, whereby once thresholds of critic score reached (70-85%) global sales reflected 28.5% (14 million sales) greater sales than those critic scores that reached (90%) but global sales dropped (10 million sales) when it should reflect greater marginal sales. Further analysis within the 'Results' section is highlighted with RMSE.

5.2 SUPPORT VECTOR REGRESSION (SVR)

SVR uses the same basic idea as Support Vector Machine (SVM), a classification algorithm, but applies it to predict real values rather than a class. Our goal will be to try to fit the error in a certain threshold (unlike minimizing the error rate we were doing in the previous cases). The SVR model can easily be converted into a linear regression which we predicted will work well with our dataset given we had 8 classes that had to be separated from which we relied on 2 classes: 'Critic Score' and 'Global Sales'. This model works well in a low dimension space than other models thus should reflect a higher accuracy of results via RMSE. However, in comparison to the LR model the accuracy of results were lower. This poses a problem as it highlights a weak correlation between global sales and critic score. On the other hand, with the data being kernel comparisons can be made to identify similarity between data points. Thus, the majority of video game titles with a low critic score will reflect low volume of global sales but are set to gradually increase as critic score increases. Further analysis within the 'Results' section is highlighted with RMSE. Our aim is to help the commercial video games retailers be able to make effective predictions on certain aspects, by using SVR it will find an appropriate line of best fit and margin of error to help with this prediction.

5.3 RANDOM FOREST REGRESSION (RF)

A special form of decision tree algorithm which utilises multiple decision trees which are known as forest and glue them together to urge a more accurate and stable prediction. The RF model for our dataset was used to compare the volume of global sales with the baseline predicted value of sales. Further analysis of correlation is highlighted within the 'Results' section also compared with RMSE. Consequently, for a larger dataset it will reach a point whereby no matter how much this model is trained, the accuracy won't change thus resulting in a model becoming unreliable. In our case increasing the video games range would highlight some change for the predictive baseline of the model, however this change would not significantly reflect the specific output value. Thus, the only way around would be to model the dataset based on the top selling video games rather than the entire video games catalogue. In comparison.

Our aim was to investigate how these models reflect the volume of global sales of video games differ based upon the critic score. We expected both SVR and RF to perform much better than LR models however this was not the case as you can see from the 'Results' section. The RMSE coefficient that is used highlights the LR model reflects a stronger correlation between 'Global Sales' and 'Critic Score'.

6. INPUT VARIABLES ENCODED

The fields required for each AI algorithm model had been split up from the original dataset and placed into their individual CSV files i.e. for Random Forest Model will would encode Critic Score and Global Sales from 'RandomForestGlobalSales&CriticScore.csv' file. This was the case for all other models for our project. Using a simple function, we then extracted each column and placed into corresponding variables "x" and "y" in their respected numerical format. Utilising ***'sklearn. model_selection import train_test_split'*** the raw data from each class 'Critic Score' and 'Global Sales' had been split in terms of 2-dimensional array from which the components for 'x' and 'y' had been derived.
Utilising these classes formed the significant basis of comparison for our models, simply because they highlighted the most marginal difference and correlation compared to the other classes. For instance, our dataset featured regional sales of EU (Europe), JP (Japan) and NA (North America) however it was difficult to identify a linear correlation with critic score as a result had led to the cluster of data points to be irregular which in itself leads to inaccurate data reading for RMSE.

7. ACCURACY EVALUATION CRITERIA

Each model will be evaluated using the RMSE value. For a model to be accurate and strong at predicting future data, we say that the RMSE value should be roughly 10% or less than the MSE value. For us to calculate these values we import the sklearn metrics library to use algorithms that calculate the MSE and RMSE. 1 of the reasons why we use RMSE over MSE value is due to it being an absolute measure of fit. This then gives us an idea of errors in actual sales. This is backed up by a report about Assessing the fit of regression model which states that "*RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction." Taken from "theanalysisfactor.com/Assessing-the-fit-of-regression-models".*

8. RESULTS

8.1 LINEAR REGRESSION

For our dataset, we decided on the 2 variables based on the heat map in the above sections showing which 2 variables have the highest correlation coefficient. As a result, we had Global sales as our dependent variable which was paired with Critic Score as these 2 had the highest relationship after we disregarded other region sales and other variables that we felt weren't necessary for us to predict future sales.

After splitting the data into 80% training and 20% testing, we used Linear Regression function and fit the X_Train and y_train and plotted the data and included a red line to show the line of best fit. As you can see in 'Fig 3' the line is sloping upwards meaning that there is a positive correlation between the 2 variables as when Critic Score increases, Global Sales also increases.

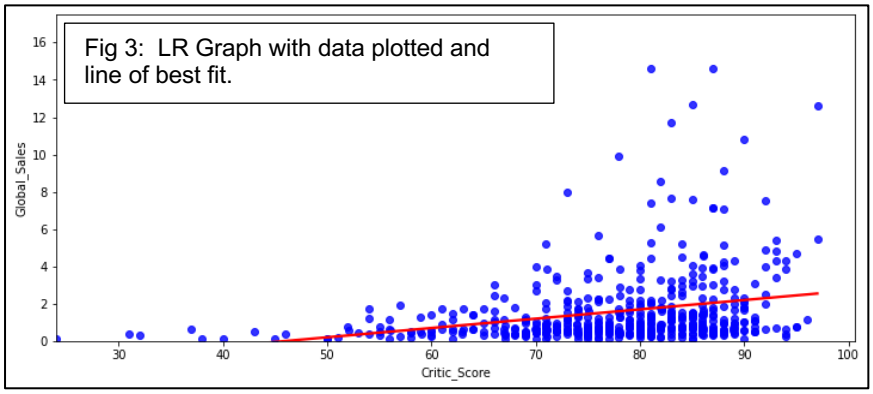


Fig 3: LR Graph with data plotted and line of best fit.

To see if the prediction made by the model is accurate, we did a side-by-side bar chart comparison of the Global Sales value near enough the same to the actual values which is a positive sign that the predictive strength of the model is a good reflection. (Seen in Fig '4').

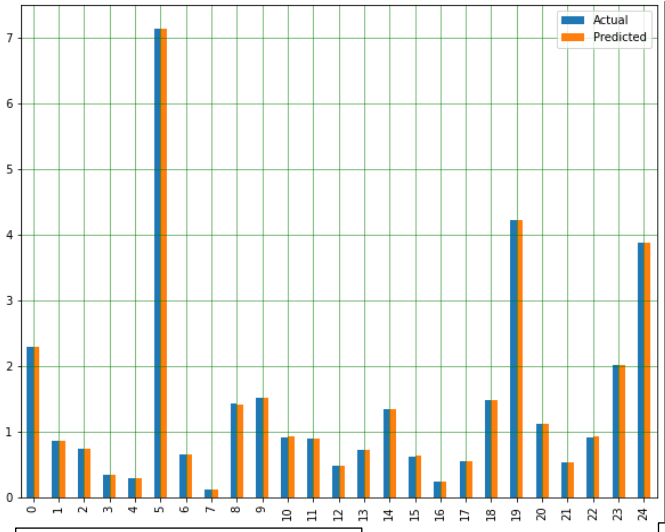


Fig 4: Bar chart showing predicted and actual values.

	Actual	Predicted
0	2.29	2.291213
1	0.87	0.870674
2	0.74	0.750295
3	0.35	0.349931
4	0.30	0.299976
5	7.14	7.139701
6	0.65	0.650515
7	0.13	0.120957
8	1.43	1.419800
9	1.51	1.509713
10	0.91	0.930717
11	0.89	0.890241
12	0.48	0.480162
13	0.73	0.730449
14	1.34	1.340680
15	0.63	0.630331
16	0.25	0.250052
17	0.56	0.560512
18	1.49	1.490390
19	4.22	4.219950
20	1.13	1.120311
21	0.53	0.529986
22	0.92	0.930389
23	2.01	2.020494
24	3.88	3.878821

Mean: 1.4878000000000002
Root Mean Squared Error: 0.0058489610963485174

Fig 5: Side by Side comparison of predicted and actual values along with MSE and RMSE value

To get a better sense of how good the model is we use the RMSE for scoring as RMSE is an absolute measure of fit. A reasonably good score is less than 10% of the MSE. We can depict that the RMSE is very low compared to the MSE As, seen in the Fig'5'. The RMSE value is around 0.4% of the MSE meaning that it is a very good score and the predictions made by the AI is extremely accurate.

8.2 SUPPORT VECTOR MACHINES

For SVR, we have used "Critic Score" to represent the x-axis and "Global Sale" for y. We chose to use SVR over SVM because we wanted to see if there is a correlation between critic score and global sales, and if that's the case then being able to make predictions based upon what the global sales are in relation with critic score for different games. Usually, a real-world dataset contains features that vary in magnitudes, units and range, so we normalized the dataset using feature scaling before fitting SVR to the dataset.

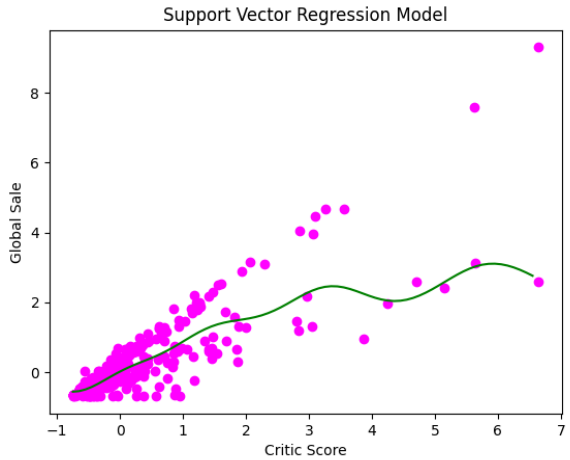


Fig 7: SVR Graph with data plotted and line of prediction (Global Sales)

From looking at this graph (Fig 7), you can see there is a cluster of points when critic score and global sales are low but as critic score increases, global sale increases showing a positive correlation. To create the hyperplane, first trained the dataset 80% training and 20% testing using a radial basis function kernel which will help to produce a smooth surface from the large data points therefore creating a more realistic and accurate prediction.

The RMSE value was roughly 40% which was not really expected as we assumed that it would be at least closer to the RMSE of Linear regression in %. To make the RMSE value lower, we tried to tune the SVR regression model by selecting the best parameters. The parameters chosen were C and epsilon because we can reduce the number of errors admitted to the graph to gain the desired accuracy of our model by reducing the value of epsilon. Then we set the values of C and epsilon to different numerical values to see if the RMSE value in % could be potential below what we have already which is 0.4 but unfortunately it couldn't so we left out the parameters.

8.3 RANDOM FOREST REGRESSION

With the random forest regression, we used the Critic Score and Global Sales for the X and Y axis again. We can see the line fluctuating as it goes higher up in the critic score and only spiking once it hits the peak Critic Score (see Fig 8). We can see that the prediction (Red line) does not follow the actual data that well. This may be due to the data not being linear for example, data with a Critic score of around 60 to 70 has higher sales than games that have above 90 critic score. As stated before, this may be due to games releasing in certain regions only but that may be the reason for the prediction line no following the data well.

In the code we calculated the RMSE value to see how accurate the prediction is and how well the line fits. As expected from looking at the graph, the RMSE value produced is quite bad. The RMSE value is around 50% of the MSE value which does not fit our criteria by a long way – 10% or lower is a reasonably good score. We then decided to change the number of trees created from the default value 100 to 2000 as that's said to increase accuracy and the values did not change much but the processing time increase tremendously.

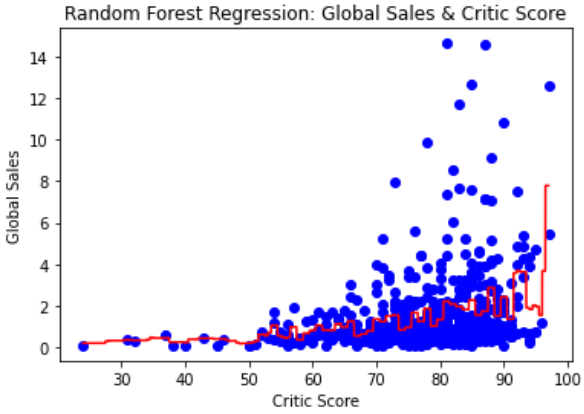


Fig 8: RF Graph with data plotted and line of prediction (Global Sales)

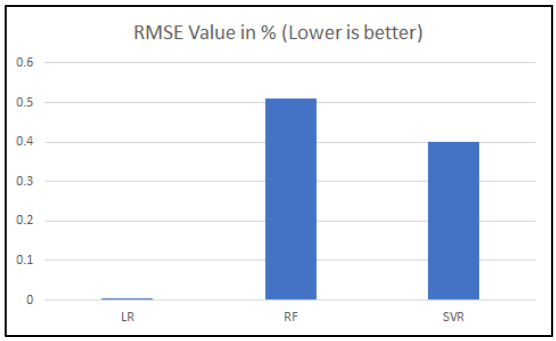


Fig 9: RMSE Comparison of All Models: LR, SVR and RF

Compiling all the RMSE scores for each model into 1 bar graph (Fig 9), we can see that LR has the lowest RMSE % value with the RMSE value being 0.04% of the MSE value meaning that the predictive accuracy and strength is the best out of the 3 models used. SVR is placed 2nd and RF placed last. Despite them placing 2nd and 3rd, those models are still not considered "accurate" as the RMSE value is more than 10% of the MSE value which is considered not a good score.

9. ENCOUNTERED DATASET PROBLEMS

Overall, it was very easy to train our models given filtered dataset of the top 100 for games released within the years 2010 to 2016 across 5 different platforms however we think that results could have been better if we had a larger dataset with variety of feature such as more platforms or a bigger range in years. We did encounter problems with missing data and

figured that filling those blanks with median or mean values was not the best solution due to critic score being opinion based, so we manually filled them in for a more accurate result. In terms of coding, there was no issue coding the models as there were many resources that taught us what some functions do and thus helping us go through each model.

10. CONCLUSION

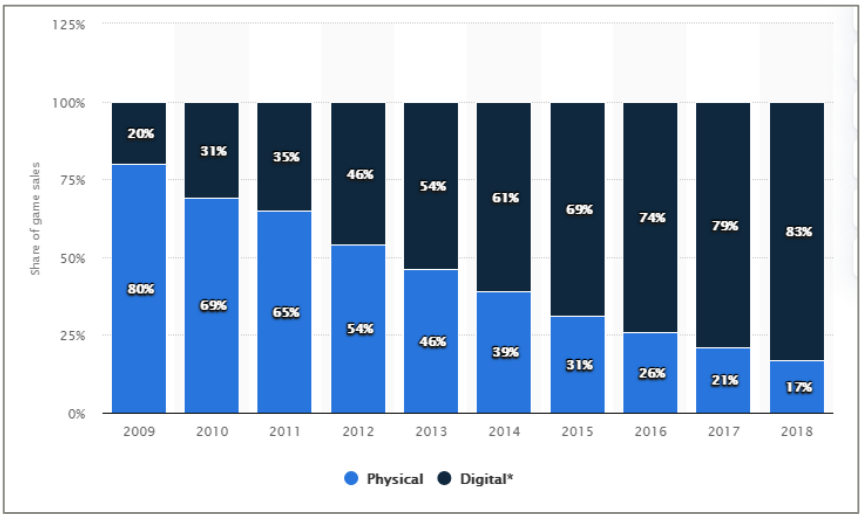
In conclusion our research suggests the regression model most significant to highlight the correlation of video games being sold globally is Linear Regression (LR) as it reflects the highest RMSE value; linear relationship between Critic Score and Global Sales being strongest for LR meant that those video game titles with highest Critic Score will most likely sell more physical units thus consumers reviewing video games will form a significant factor for selling video games. Moreover because of this it would suggest that popular media outlets like 'Metacritic' and content creators notably from Twitch a renowned platform for gaming community which also accounts for 3.8 video game streamers[8] and YouTube whereby creators are able to share reviews of different game titles and platforms, all of which will have a strong influence in helping consumers to decide whether to purchase a video game.

In alignment with these video game reviews from consumers that will be reflected within Critic Score, it enables video game developers and market leaders to identify those particular video game titles to specialise towards creating; those titles and genre with most popular reviews should highlight what consumers would want which in response will enable industry leaders to focus on creating games with these factors in mind.

Furthermore, due to the popular demand of next-gen consoles and added benefits of digital formatted video games of convenience, accessibility, and privacy. Developers will eventually want to move onto distributing digital formatted games as well as distributing physical copies of games. As a result, it will be in retailers and gaming outlet's like 'GAME' best interest to adapt to these changes by only stocking those video games that are most likely to sell.

Which our models prove, that top selling video games are those who have the highest rate of Critic Score in turn will result in greater distribution of video games sold globally including digital formats.

The rate of which digital sales are sold across can be analysed from the Fig 10 below:



Sales for digital formatted video games will gradually increase per year. In relation to our dataset those top selling video games can be broken down to an extent to show how much of those sales are Digital and Physical. E.g. Global Sales for 2016 reflects (74% Digital and 26% Physical).

Fig 10: Distribution of Video Games between Digital and Physical

11. REFERENCES

[1] **Games Crate. Angelo M. D'Argenio., 2020.** ***Statistically, Video Games Are Now The Most Popular and Profitable Form Of Entertainment*****. [online]**
Available at: <https://gamecrate.com/statistically-video-games-are-now-most-popular-and-profitable-form-entertainment/20087> [Accessed 16 December 2020].

[2] **Reuters. 2020.** ***Report: Gaming Revenue To Top $159B In 2020*****. [online]**
Available at: <https://uk.reuters.com/article/esports-business-gaming-revenues/report-gaming- revenue-to-top-159b-in-2020-idUSFLM8jkJMl> [Accessed 16 December 2020].

[3] **Eurogamer.net. 2020.** ***GAME Intends To Close 40 Stores In The UK*****. [online]**
Available at: <https://www.eurogamer.net/articles/2020-01-09-game-intends-to-close-40-stores-in-the-uk> [Accessed 15 December 2020].

[4] **Eurogamer.net. 2020.** ***UK Video Game Sales Now 80% Digital*****. [online]**
Available at: <https://www.eurogamer.net/articles/2019-01-03-uk-video-game-sales-now-80-percent-digital> [Accessed 15 December 2020].

[5] **BBC News. 2020.** ***Game Group Goes Into Administration, Closing 277 Stores*****. [online]**
Available at: <https://www.bbc.co.uk/news/business-17512143> [Accessed 15 December 2020].

[6] **Games Radar. 2020. The Recyclable PS5 Packaging Is The First Small Step For A Console** ***Generation That Must Reckon With Gaming's Carbon Footprint*****. [online]**
Available at: <https://www.gamesradar.com/the-recyclable-ps5-packaging-is-the-first-small-step-for-a-console-generation-that-must-reckon-with-gamings-carbon-footprint/> [Accessed 15 December 2020].

[7] **Grace-Martin, K., 2020.** ***Assessing The Fit Of Regression Models - The Analysis Factor*****. [online] The Analysis Factor.**
Available at: <https://www.theanalysisfactor.com/assessing-the-fit-of-regression models/#:~:text=Whereas%20R%2Dsquared%20is%20a,an%20absolute%20measure%20of%20fit.&text=Lower%20values%20of%20RMSE%20indicate,of%20the%20model%20is%20prediction.> [Accessed 22 December 2020].