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

IN3062 – INTRODUCTION TO ARTIFICIAL INTELLIGENCE

Coursework

Document Version 1.0

TABLE OF CONTENTS

[1. INTRODUCTION 2](#_Toc59011547)

[1.1 PROBLEM DOMAIN 2](#_Toc59011548)

[1.2 DATASET 2](#_Toc59011549)

[2. REGRESSION 2](#_Toc59011550)

[3. MISSING DATA 3](#_Toc59011551)

[4. MODELS USED 3](#_Toc59011552)

[4.1 LINEAR REGRESSION 3](#_Toc59011553)

[4.2 SUPPORT VECTOR MACHINES 3](#_Toc59011554)

[4.3 RANDOM FOREST REGRESSION 4](#_Toc59011555)

[5. INPUT VARIABLES ENCODED 4](#_Toc59011556)

[6. ACCURACY EVALUATION CRITERIA 4](#_Toc59011557)

[7. RESULTS 4](#_Toc59011558)

[7.1 LINEAR REGRESSION 4](#_Toc59011559)

[7.2 SUPPORT VECTOR MACHINES 4](#_Toc59011560)

[7.3 RANDOM FOREST REGRESSION 4](#_Toc59011561)

[8. ENCOUNTERED DATASET PROBLEMS 4](#_Toc59011562)

[9. REFERENCES 4](#_Toc59011563)

# **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:

**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.



Fig 1: ‘GAME’ store 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](https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings). 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.

# **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.

# **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’](https://www.metacritic.com/) link.

# **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.

# **MODELS USED**

## 4.1 LINEAR REGRESSION (LR)

LR is a statistical approach for modelling 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 work wells with our dataset as we look for the variation of critic score that is dependent on global sales. LR model depicts a positive correlation whereby a high critic score would equal higher level of global sales. Despite LR being simple to implement and easier to interpret the output coefficients. This model is prone to underfitting; fails to capture data properly and often line does not fit well due to the relationship between independent and dependent variable being not entirely linear. This was the case for our dataset as when the critic score peaked past 70+ and global sales being dispersed above hyperplane.

## 4.2 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. SVR acknowledges the presence of non-linearity in the data and provides a proficient prediction model.

SVR is a regression model in which we try to fit the error in a certain threshold (unlike minimizing the error rate we were doing in the previous cases). SVR can work for linear as well as non-linear problems depending on the kernel we choose. There is an implicit relationship between the variables, unlike the previous models, where the relationship was defined explicitly by an equation (coefficients are sufficient to balance the scale of variables). Therefore, feature scaling is required here.

## 4.3 RANDOM FOREST REGRESSION (RF)

For one o

# **INPUT VARIABLES ENCODED**

The data has been put in a specific folder structure such as, “images/class\_name/image.jpg” using this we have made a simple function that extracts each image and its corresponding label into the variables “X” and “y”. the images are kept in their RGB format and converted into image arrays, while the string labels are one- hot codded using sklearn’s “LabelBinarizer”.

The fields required for each model had been split up from the original dataset and placed into their individual CSV files i.e. for Random Forest

Feature engineering is the process of extracting new features from the feature space using domain knowledge to increase the amount of data we have (Wikipedia, 2019). We wanted to test out different types of data with our models, so we made 3 different data types. “just pixels” refers to the flattened image arrays, “extracted features” are just the extracted features, and “all features” is the mix of the previous two. The extracted features consist of performing the following statistical operations on all the pixels and on each RGB channel, which are finding the mean, median, and standard deviation. Thus, we have 12 new features we were able to extract using our knowledge of RGB channels and our previous analysis of grey levels. We hope this slight increase in feature space helps the models perform better.

# **ACCURACY EVALUATION CRITERIA**

# **RESULTS**

## 7.1 LINEAR REGRESSION

## 7.2 SUPPORT VECTOR MACHINES

## 7.3 RANDOM FOREST REGRESSION

# **ENCOUNTERED DATASET PROBLEMS**

# **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].