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 2](#_Toc59011551)

[4. MODELS USED 3](#_Toc59011552)

[4.1 LINEAR REGRESSION 3](#_Toc59011553)

[4.2 SUPPORT VECTOR MACHINES 3](#_Toc59011554)

[4.3 RANDOM FOREST REGRESSION 3](#_Toc59011555)

[5. INPUT VARIABLES ENCODED 3](#_Toc59011556)

[6. ACCURACY EVALUATION CRITERIA 3](#_Toc59011557)

[7. RESULTS 3](#_Toc59011558)

[7.1 LINEAR REGRESSION 3](#_Toc59011559)

[7.2 SUPPORT VECTOR MACHINES 3](#_Toc59011560)

[7.3 RANDOM FOREST REGRESSION 3](#_Toc59011561)

[8. ENCOUNTERED DATASET PROBLEMS 3](#_Toc59011562)

[9. REFERENCES 3](#_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 sales demand. This impact also meant that digital sales has exponentially risen resulting in UK gaming outlets to struggle with endorsing sales from physical units. Severity of the impact can be assessed from one of UK’s leading video game outlets – ‘GAME’ store**3**.

## 1.1 PROBLEM DOMAIN

Our main focus is on the video game industry, especially video game retail stores. The problem  with the video game industry is that physical stores are dying due to events like coronavirus or just because digital copies of games are more convenient. However we know that there are a lot of people that love physical game copies as big console companies such as Microsoft and Sony still release consoles with a disc reader in them (PS5 and Xbox X series). Our aim is to help these stores predict which games will be the best sellers before they get released which will allow them to decide if they want to stock their store with a game that is sure to sell a lot compared to games that won’t.

The dataset we will be using identifies games based on Genres, Publisher, Platform, and many more which could give us multiple factors that can be useful for predicting a games success. This will help the video game market and video game companies as sales have been declining quite a lot in the past few years with sales falling by 4.7% in the first half of 2019, putting the popular known store called GAME in grave danger. Having this analysis allows Game stores to stock games that have a high chance of being successful to avoid losing money and going out of business.

https://www.thesun.co.uk/tech/9231044/game-stores-danger-sports-direct-takeover/

## 1.2 DATASET

The original dataset titled ‘Video Game Sales with Rating’ can be found on Kaggle (https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings).

The dataset itself contained 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 as within this period the dataset with sales from each region is consistent thus enabling better predictive analysis from which patterns can be identified.

# **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 the years 2010 to 2016, we found that a lot of games had their critic score and critic count missing. This would hinder our results as it won’t give an accurate representation between critic scores and sales. 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 website called Metacritic.com.

# **TECHNIQUES USED**

1. For our dataset, we decided to use Pearson's correlation coefficient in order to find out which 2 variables have the strongest relationship which will help us predict video game sales in the future.

# 

1. We used a heatmap to represent the coefficients for 2 variables. We can see that Sales for different regions have a strong relation to Global Sales which is to be expected. For our analysis we made Global\_Sales our dependent variable and as it counts for all regions we decided to drop the other sales features. So when disregarding 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

# **MODELS USED**

## 4.1 LINEAR REGRESSION

## 4.2 SUPPORT VECTOR MACHINES

## 4.3 RANDOM FOREST REGRESSION

# **INPUT VARIABLES ENCODED**

# **ACCURACY EVALUATION CRITERIA**

# **RESULTS**

## 7.1 LINEAR REGRESSION

## 7.2 SUPPORT VECTOR MACHINES

## 7.3 RANDOM FOREST REGRESSION

# **ENCOUNTERED DATASET PROBLEMS**

# **REFERENCES**

**1 Gamescrate. 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].