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

IN3062 – INTRODUCTION TO ARTIFICIAL INTELLIGENCE

Coursework

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document.   
When you’re ready to add your content, just click here and start typing.]

Table of Contents

[INTRODUCTION 2](#_Toc59011069)

[PROBLEM DOMAIN 2](#_Toc59011070)

[DATASET 2](#_Toc59011071)

[REGRESSION 2](#_Toc59011072)

[MISSING DATA 2](#_Toc59011073)

[MODELS USED 3](#_Toc59011074)

[INPUT VARIABLES ENCODED 3](#_Toc59011075)

[ACCURACY EVALUATION CRITERIA 3](#_Toc59011076)

[RESULTS 3](#_Toc59011077)

[ENCOUNTERED DATASET PROBLEMS 3](#_Toc59011078)

[REFERENCES 3](#_Toc59011079)

# **INTRODUCTION**

Video games are becoming more and more popular around the world. This is evident by tech giants such as Google, Facebook, Amazon, and Apple all trying to have a piece in the gaming industry from Google stadia, Apple Arcade, Games made by Amazon to Facebook gaming streaming service. Because of these tech giants being involved, playing video games has never been easier as you do not even need a console anymore.

## 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/

## DATASET

For our dataset, we wanted to focus on variables that will help predict the next best games platform or genres. Our focus is critic score, Genres, Platform, and sales across continents. Due to there being many games with little to no video game sales and platforms that have been discontinued, we have narrowed down the data to the top 100 video game sales across 5 different platforms – 3DS, Xbox One, PC, Ps4 and WiiU. On top of that, we are focusing on games released from 2010 to 2016 so that is the time where gaming really starts to pick up and gain immense popularity. After filtering the data, we saw that a lot of rows had critic score and critic count missing. We initially tried using a method to fill in those values by using the mean or median values of all critic scores however, we concluded that this is not an accurate representation as critic scores is a system based on critics’ opinions and part of our goal is to see if critic scores affect video game sales. With this in mind, we filled in the critic scores ourselves using data from the website Metacritic which is the source for the rest of the data’s critic scores and counts.

# **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**

**\*Currently going through the spreadsheet filling in the missing critic scores\***

# **MODELS USED**

* Algorithms to be used:
  + Linear Regression - Supervised - Martin does exercise for Linear
  + \*SVM (Regression) - Supervised\*
  + Random Forest Regression

# **INPUT VARIABLES ENCODED**

# **ACCURACY EVALUATION CRITERIA**

# **RESULTS**

# **ENCOUNTERED DATASET PROBLEMS**

# **REFERENCES**