ML Group Project Proposal

Joseph Carnec (17341978) XXXX XXXX XXXX XXXX

November 12, 2020

1 Introduction

Are group project will have the goal of creating an ML model which is able to estimate the average monthly sales of video games on the website store.steampowered.com. The input features for each game comprises information which is available on the website's store page. This will involve scraping the product page for a random sample games on the website in order to extract information including: genre, access to multiplayer, publisher, age rating, controller support among others.

2 Motivation

One of the motivations for this project is to explore the ability for a ML systems to predict the sales of a without taking into consideration the *gameplay quality* of the game. A similar example in another form of media would be predicting box office revenues based on the genre and cast of a film. This would give insights into how decisions made outside of game play design might influence game sales. Another motivation for this project is exploring how ML could potentially help publishers and developers make decisions on which type of games they should prioritise. For example a developer is interested in how the inclusion of local multiplayer might influence the sales of a racing game or how adding OSX support for a PC game might boost sales.

3 Dataset

To create our dataset We will extract product information for a randomly selected subset of the steam store page by scraping the website. The reason for this is that steam has over 10000 games on sale which would be a restrictively large number of pages to scrape. This will provide all of the inputs to our model. Information such as publisher or genre will be represented by a set of indicator variables. The target variable of sales is not publicly available but is estimated to 1% accuracy by the Steamspy API.

4 Methods

This will be regression problem and we are planning to try a variety of different models. We will have a large number of features to input in the model. We will try a number of methods and penalties to find the best performing model. These models include Ridge, Lasso and K-nearest neighbors.

5 Evalution

The model will be evaluated by splitting our dataset into a training, validation and test. Once the model trained and hyper-parameters are chosen, with the use of the test set, we will be able to see how well the model is able to predict sales for games. We will compare the model to a reasonable baseline to see if the model is improving on a simple heuristic such as predicting the average sale of games on steam.