

Project Proposal:

Predicting the Videogame Average Monthly Sales

Darragh Hurley (16319250), Joespeh Carnec (17341978), William O’Sullivan (16321101)

Due 2020-11-13

Abstract

The goal of this project will be to create a Machine Learning model which is able to estimate the average monthly sales of videogames as described on Steam, a videogame marketplace application. The input features for each game comprises of 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 and controller support, among others.

1 Project Motivation

The videogame industry is projected to be worth over \$90 billion USD by 2020. As such, streamlining the production of profitable games is in the interest of this growing market.

One of the motivations for this project is to explore the ability for a ML system to predict the sales of a without taking into consideration its critical reception. 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 may 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 may want to determine if adding certain features like multiplayer or OSX support is in the financial interests of the studio.

2 Dataset

To create our dataset we will extract product information for a randomly selected subset of Steam store pages by scraping the website. Steam has over 10000 games on sale which would be a restrictively large number of pages to scrape, hence the smaller sample size. This will provide all of the inputs for the model. Information, like 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.

3 Methods

As will be regression problem and we are planning to try a variety of different models, including include Ridge, Lasso and K-nearest neighbors methods. Naturally, these models will all require tuning and adjustment of penalties.

4 Intended Experiments

The main point of investigation in this project will be in classification of features in successful games - do certain features dominate over others in how good a game will be, does our model require additional, more specific data to make good predictions etc. The primary goal of our experiments will be to construct a model that is good at predicting sales, but secondary results would be obtained by discovering which features have the highest impact on sales.

5 Evaluation

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, it will be possible to see how well the model is able to predict sales for games. The model will be compared to a reasonable baseline to see if the model is improving on a simple heuristic approach such as predicting the average sale of games on Steam.