Proposal: Predicting Video Game Success using Logistic Regression and Binary Classification

Puttisan “Ninja” Mukneam

Feb 13, 2023

Introduction

The video game industry is experiencing tremendous growth, and accurate sales predictions are crucial for game developers and publishers in making informed decisions. In this project, we aim to leverage Logistic Regression to build a model that predicts the success of a video game based on Steam ratings, critic scores, and sales.

Motivation

This project serves as an alternative frameworks to our previous work and aims to provide deeper technical and practical insights into Logistic Regression and efficient modeling. The original project can be found at https://github.com/pmukneam/CSCI036-video_games_analysis

Methods

1. Collect datasets of video game sales, Meta critics score, players’ reviews score from online sources.
2. Convert datasets into binary classification of “hit” or “miss” (hit := $\geq$ 85% postive players’ reviews).
3. Train a Logistic Regression model to predict whether a video game will be a “hit” or “miss”.
4. Evaluate the model’s performance of the model using various tools.

Data sets

1) 80000 Steam Games DataSet

- Public URL: https://www.kaggle.com/datasets/deepann/80000-steam-games-dataset?select=steam_data.csv
- contains various information of games including players’ reviews that was scrapable through Steam API link using (https://store.steampowered.com/appreviewhistogram/945360) and (https://store.steampowered.com/appreviews/945360?json=1)

2) Video Games Dataset

- Public URL: https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-07-30/video_games.csv
- contains various information of games including metascore (critic score). The data was obtained from Steam Spy scraped and provided by Liza Wood and modified by jthomasmock in his tidytuesday repository.

3) Video Game Sales

- Public URL: https://data.world/sumitrock/video-games-sales

- contains various information of games with sales more than 100,000 copies including global sales and critic score. The data was obtained from https://www.vgchartz.com/ by scraping using Python library "BeautifulSoup."

Expected Outcomes

1) A trained model that can accurately predict whether a video game will be a "hit" or "miss".
2) A deeper understanding of Logistic Regression and how it can be used for binary classification problems.