VIDEO GAME RECOMMENDATION SYSTEM

BrainStation Capstone Project

Abstract

Recommendation system for video games based on a dataset of Amazon product reivews from 1999 to 2014, and their associated product metadata.

Joshua Sunga

Introduction

This project aims to create a recommendation system for video games based on a dataset of Amazon product reviews from 1999 to 2014, and their associated product metadata.

Business Question

What video games are customers likely to buy based on reviews provided by likeminded customers?

Business Value

The value of this project is two-fold:

- It allows customers to see products that they would like based on previously expressed preferences by likeminded customers.
- It allows ecommerce merchandisers and category managers to determine which products to continue featuring on their online stores based not only on revenue considerations but based on customer preferences as well.

Source of Data and Citation

I would like to acknowledge that I obtained the product review dataset from Professor Julian McAuley, who kindly provided them here: http://jmcauley.ucsd.edu/data/amazon/links.html.

The review dataset is part of the following research papers:

Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering
R. He, J. McAuley
WWW, 2016
pdf

Image-based recommendations on styles and substitutes
J. McAuley, C. Targett, J. Shi, A. van den Hengel
SIGIR, 2015
pdf

I would also like to acknowledge that the metadata dataset was obtained from Professor Jianmo Ni, who kindly provided them here:
http://deepyeti.ucsd.edu/jianmo/amazon/index.html

The metadata dataset is part of the following paper:

Justifying recommendations using distantly-labeled reviews and fined-grained aspects
Jianmo Ni, Jiacheng Li, Julian McAuley
Empirical Methods in Natural Language Processing (EMNLP), 2019
pdf

The product review dataset had 231,780 reviews; while the metadata dataset had 84,893 products.

The product review dataset consisted of the following columns:

1. reviewerID
2. asin (product number)
3. reviewerName
4. helpful (number of votes that considered a review helpful / number of total votes)
5. reviewText
6. overall (product rating)
7. summary (review title)
8. unixReviewTime
9. reviewTime (date of review)

The metadata dataset consisted of the following columns:

1. category
2. tech1 (technical feature column - empty)
3. description
4. fit (size information not pertinent to video games - empty)
5. title
6. also_buy (list of products that were bought with the product in question)
7. tech2 (technical feature column – empty)
8. brand
9. feature (special features associated with a video game)
10. rank (ranking based on sales performance)
11. also_view (list of products that were viewed with the product in question)
12. main_cat (main_category – largely redundant)
13. similar_item (list of products that were similar to the product in question)
14. date (launch date of product – empty)
15. price (empty)
16. asin (product number)
17. imageURL
18. imageURLHighRes

Summary of Data Processing

Given that the project relied on two datasets (the product reviews dataset and the metadata dataset), significant data cleaning as well as data reduction was required.

For the product reviews dataset, the only columns that remained were reviewerID, asin, helpful, overall, and reviewTime. The helpful column was later expanded into two new columns: helpfulness_votes, total_votes. Out of the helpfulness_votes and total_votes column, a third column was created: helpfulness_rating, which was the percentage of helpful votes cast in favor of reviews out of a total number of votes cast.

The product review dataset did not have any duplicates.

For the metadata dataset, the only columns that remained were the category, title, and asin columns.

Significant cleaning was required for the category columns as it contained the overall category (i.e. 'Video Games') and the console (i.e. 'PlayStation 4'). Unfortunately, the game genres were not part of the column and were not included in the dataset at all, so I had to settle with using the console type.

Consoles themselves as well as related accessories were included in the metadata dataset. These had to be dropped.

There were also products that had no category information at all, as well as duplicated rows. These were also dropped.

The two datasets were merged into one dataframe, main_df.

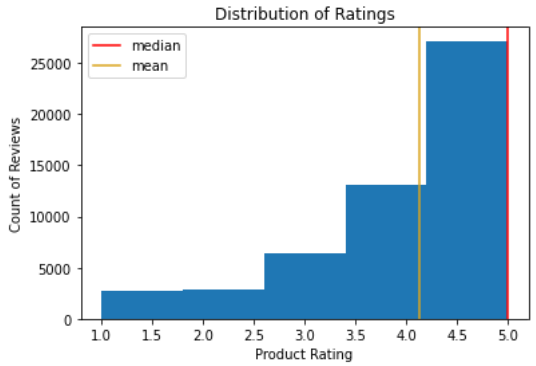
After the two datasets were merged, further processing was needed to reduce the number of rows, so I calculated the 80% quantile for the number of reviews for products (26) and the number of reviews given by reviewers (8). Both products and reviewerse that fell below these newly established 'minimums' were dropped from the main dataset.

Exploratory Data Analysis

As part of the EDA section, I examined the distribution of ratings, number of reviews by category, average rating by category, ratings over time, helpfulness rating, total votes received, unique reviewers, and unique products reviewed. *I will provide some insights in this report. The insights from this data can be read in full in the EDA section of Part 1.*

Distribution of Ratings

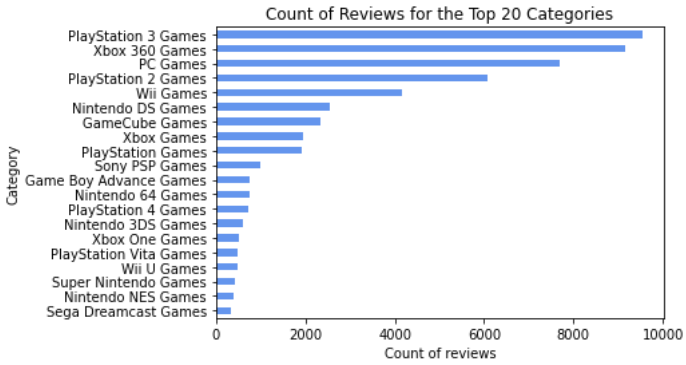
The bulk of ratings provided were at the 5-star level, followed by 4 stars. The ratings distribution suggests that the bulk of reviewers tend to give five-star reviews. Reviewers who are dissatisfied with their purchase tend to give 4-star reviews, which reflects why the mean rating is only 4.12, which is still high.



Number of Reviews by Category

The PlayStation 3 Games category has the highest number of reviews, followed by Xbox 360 Games and PC Games.

PlayStation (1-4) products dominate the product reviews. Interesting to note that older consoles like GameCube still made it to the top 20 list for number of reviews.



Average Rating by Category

For categories with more than 1,000 reviews, the PlayStation Games category is the highest performing, with 1,922 reviews and an average rating of 4.43.

The category with the most reviews is PlayStation 3, with an average rating of 4.17.

Xbox 360 Games and PC Games are the categories with. A high quantity of reviews but a lower average rating, at 3.95 and 3.81, respectively.

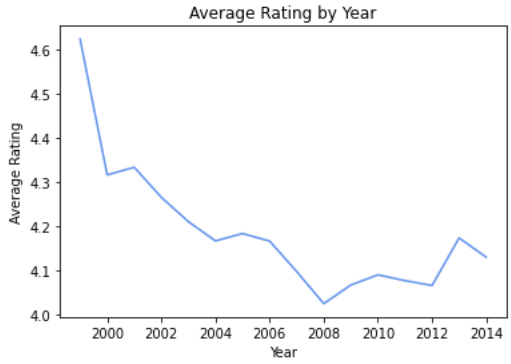
PlayStation Vita, despite having only 13 reviews, has the lowest average rating at 2.76.

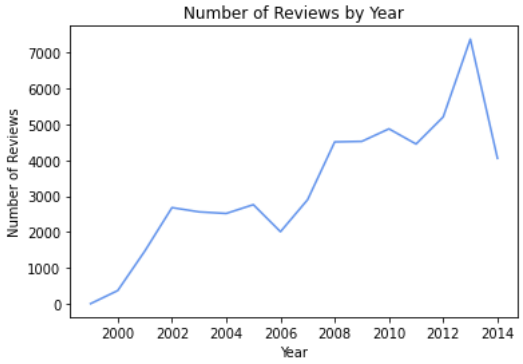
Also, it is interesting to note that the highest performing categories in terms of rating are Super Genesis Games and Super Nintendo Games, both considered ‘vintage’ for the purposes of this project given that Sega Genesis was first launched in 1988 and Super Nintendo was first launched in 1990. Given this, there seems to be a nostalgia factor when it comes to older games and their average product rating.

category	mean	count
Sega Genesis Games	4.811321	53
Super Nintendo Games	4.691566	415
Game Boy Games	4.675676	37
Nintendo NES Games	4.569975	393
Sega Saturn Games	4.518519	27
Sega Dreamcast Games	4.506173	324
Nintendo 64 Games	4.504673	749
Nintendo 3DS & 2DS	4.468750	64
PlayStation Games	4.429761	1922
Game Boy Advance Games	4.398950	762
PlayStation Vita Games	4.383333	480
Nintendo DS Games	4.362633	2537
GameCube Games	4.355241	2328
Game Boy Advance	4.327273	55
Wii Games	4.288948	4153
Nintendo 3DS & 2DS Games	4.260870	230
Wii U Games	4.255230	478
Sony PSP Games	4.252280	987
Xbox Games	4.207305	1944
Nintendo 3DS Games	4.174837	612
PlayStation 3 Games	4.174683	9543
PlayStation 2 Games	4.146570	6079
Xbox 360 Games	3.951009	9165
PlayStation 4 Games	3.848315	712
PC Games	3.812216	7695
Xbox One Games	3.736434	516
PlayStation Vita	2.769231	13

Ratings over Time

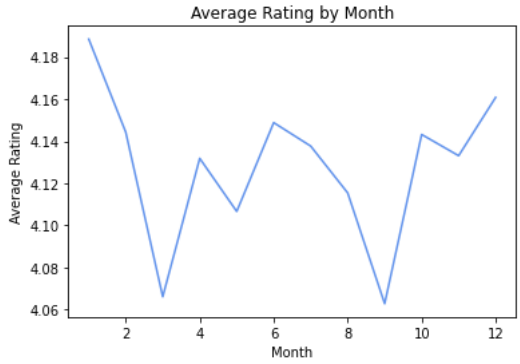
By calendar year

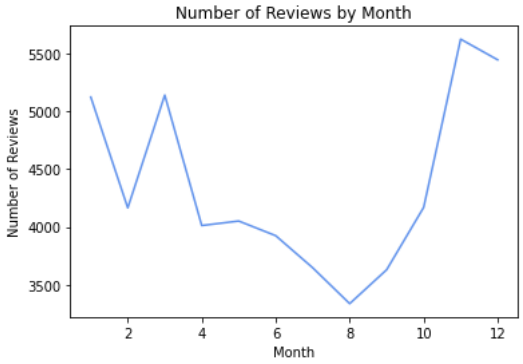




We can see in the two graphs above that as the number of reviews increases by each passing year, the average rating decreases, with the share of negative ratings increasing year over year.

By calendar month





The average rating tends to be high in January, June, October, and December.

The number of reviews submitted tends to be high in January, March, November, and December.

Despite having many reviews in January, November, and December, the average rating for these months is high. In contrast, the number of reviews submitted during the summer months is low, yet the average rating is high for this period. This suggests that there is an inverse relationship between the average rating and number of reviews submitted for most months, but during January, November, and December, ratings tend to be high despite the large quantity of reviews submitted.

Recommendation System

We used the Funk Singular Value Decomposition (Funk SVD) Algorithm to build the recommendation system. Based on the algorithm tuning using GridSearchCV, I used the following parameters to obtain the best possible FCP score:

- n_factors: 100
- n_epochs: 10
- lr_all: 0.1
- biased: False

At the train test split, using a test size of 25%, I found that 61% of the predictions were almost accurate.

At the full trainset, I found that the FCP score was 60%, which was the ideal score as I wanted to ensure that novelty would be introduced in the recommendations provided.

I found that there were 150 latent factors that determined how products were recommended.

Conclusion

I found that the results fulfilled the aim of the project. A final FCP score of 60% was desirable, given that it would provide accurate enough recommendations to customers while at the same time introducing novelty to customers.