Contents

1 Introduction

The video gaming sector has grown significantly in the recent years. There are over two billion games across the world, which is around 26% of the population. The industry is predicted to grow even more in the next few years, making the industry one of the most revenue generating industries. The industry is involved in the development, marketing, and monetization of video games. As such, there are three popular consoles controlling the market of the industry, Microsoft(Xbox), Nintendo(Switch) and Sony(PlayStation). When it comes to the production of games, we have three categories of developers:

- **First Party:** When the console develop their own games, such as Nintendo developing Super Mario.
- **Second Party:** When a console collaborates with a game development company.
- **Third Party:** When game development companies develop their own games and sell them.

A major problem that third party developers face, is that they often have creative and original ideas that ideally should be popular, but the companies do not have the capabilities to make them popular. Such developers, also known as Indie-Developers, therefore end up with a lot of unrecognized games.

As the consoles dominating the market are no longer depending on hard copies of video games but rather are becoming more digitalized with their own websites, companies that previously depended solely on hard-copy video games are therefore encountering difficulty in transforming towards this new market trend. GameStop specifically, which was one of the largest video gaming retail seller, is facing major challenges to its business model from the internet. As more people buy video games through digital storefronts, fewer customers buy games on physical discs from GameStop. As such, while GameStop is working on this transformation to the digital world, our goal is to allow GameStop have a competitive advantage that will allow it to successfully compete in the market with a feature other competitors lack.

Finding the right video game is not that simple. There are many video games being released at the same time and it takes a lot of time to decide if this video game is worth spending money on. A lot of websites have tried to create recommender systems, however, these systems suffer from a lot of flaws such as obvious recommendations(such as FIFA 2021, if you've played FIFA 2020) and difficult user interface. This can leave the users frustrated as they suffer from the paradox of choice.

In this project, we will tackle this issue through creating a recommendation system for GameStop. The recommender system that we will be presenting as our solution generates recommendations on the basis of users' reviews and critics' reviews. Generally, business leverage their website data (such as rating from users, their browsing history etc.) and transaction data to recommend products to the consumer. The solution that the report will be presenting is different in a way that it will leverage users' and critics' reviews from a platform that is one stop for all these reviews for all the video games, Metacritic.com. The recommender system will be solely based on the reviews from critics and users. These reviews are usually in the form of free text, and can serve as a useful resource to derive information about the users' and critics' preferences. Among the information elements that can be extracted from reviews, opinions about

particular item aspects (i.e., characteristics, attributes or components) can be utilized to personalize recommendations for users. These recommender system will closely relate to the idea of content-based filtering since they will be using the various features mentioned in the reviews of different games to recommend games. One of the elements of the recommender system will also have an extra filter that will prioritize games that have low reviews but at the same time a high score. As a result of the four recommendation system variations that will be proposed (as shown in Figure 1), GameStop can have a competitive advantage in the market, and ultimately Indie-developers will have their games recognized.

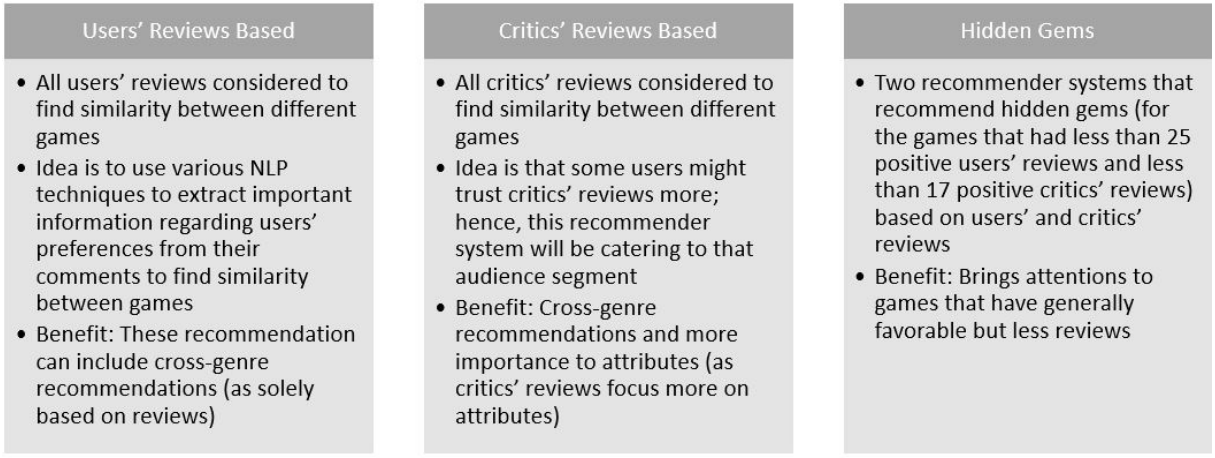

Figure 1: Different elements of recommendation system proposed as part of the solution

2 Data Description and Analysis

To successfully scrape the data from the Meta-critic website, the following steps were followed:

- **Step 1:** Extract a list of all game links for Nintendo Switch (for a manageable-sized dataset)
- **Step 2:** Extract each game information such as the name, genre and characteristics
- **Step 3:** Extract game critic and user reviews

2.1 Data Description

We ended up with 4 CSV files with the following data dictionary each:

Table 1: Switch Sales Data (sales-switch.csv)

Attribute Name	Meaning
title	Title of each Game
total-sales-usdmm	Total Sales Generated by each game in USD

Table 2: Switch Critic Reviews (switch-critic-review-filtered.csv)

Attribute Name	Meaning
index	Index of each game
title	Title of each Game
platform	Platform of each Game
critic	Name of the critic reviewer
date	Date the review was written
score	Score given to the game
text	Review

Table 3: Switch User Reviews (switch-user-review-filtered.csv)

Attribute Name	Meaning
index	Index of each game
title	Title of each Game
platform	Platform of each Game
critic	Name of the user reviewer
date	Date the review was written
score	Score given to the game
text	Review

Table 4: Switch Game Information (switch-game-info-filtered.csv)

Attribute Name	Meaning
title	The Name of the Game
platform	Platform of each Game
summary	The Summary description of the Game
release-date	The Release date of the Game
developer	The Developer of the Game
Genre	The Genre of the Game
Rating	The Rating of the Game
meta-overall,meta-positive,meta-negative,meta-mixed	Rating given by the critics
user-overall,user-positive,user-negative,user-mixed	Rating given by the users

2.2 Data Analysis

2.2.1 Data Pre-Processing:

First the comments were pre-processed so that they do not include nay unnecessary words like stop words or punctuation. Additionally, some of the words that appeared with very low frequency (less than 50) were removed as well. Some comments were not in English, hence some of those words from any other

language that showed up in high frequency were removed as well. The words were lemmatized as well. These steps ensured that that model only took relevant words, mostly attributes from comments. After the cleaning of the comments, they were used as an input for Tfidf vectorizer.

2.2.2 TF/IDF vectorizer:

This tokenized the aggregated comments, learnt the vocabulary and inverse document frequency weightings, which converted a collection of raw documents to a matrix of TF-IDF features.

2.2.3 Topic Modelling:

This unsupervised learning approach was used to cluster commments to discover the topics. Different dimension reduction techniques such as Non-negative Matrix Factorization(NMF), Latent Semantic Analysis (LSA), and Latent Dirichlet Allocation (LDA) were run. Each of these techniques follow a different approach to come up with topics. After experimenting with all these techniques and the number of topics, it was determined that best results were obtained using 25 number of topics with NMF.

2.2.4 Nearest Neighbors using Cosine Similarity:

After this the output after performing NMF was used an input for determining the similarity between games. The nearest neighbors method with cosine similarity was used to find similarity between games in order to generate recommendations.

The above mentioned approach was followed to generate recommendations for all the four variants. In case of hidden gems, a conditional statement was input into the original code which only gives the recommendations from hidden games list, which was generated on the basis of number of comments. For the users' reviews, the hidden gems were considered to be the games that had less than 25 reviews while for the critics', the reviews had to be less than 17.

3 Insights

As we performed data analysis using different techniques, the following insights were discovered through linear regression, as can be seen in Figure 2. The same results can be verified by merging sales data along with games data:

- Generally unfavorable games can still have high sales. This can be an indicator that popular games might have sales for franchise influence or developer popularity. This was inferred from the results when the overall sentiment of users' comments were compared to sales data. Some of the examples of such games included Pokemon Shield and Pokemon Sword, which had generally unfavorable user reviews but the sales were in top ten.
- Universally acclaim games do not necessarily have larger sales, emphasizing the fact that game quality is not a defining feature of commercial success. For example, one of the universally acclaimed games based on the users' review was Hollow Knight, which wasn't even in top ten in sales.

- Games developed by first party companies sell more, which may be related to marketing, popularity and customer top of mind.
- After trying a lot of games, it was observed that the recommendations generated from critics' comments had the high likelihood of being closer to the game entered by user in terms of genre. This could be due to the terminology/jargon that critics use in their comments, making the whole recommendation process relatively more accurate.

All these points were verified using regression as well, as shown in Figure 2.

OLS Regression Results

Dep. Variable:	total_sales_USDMM	R-squared:	0.810
Model:	OLS	Adj. R-squared:	0.788
Method:	Least Squares	F-statistic:	36.10
		Prob (F-statistic):	8.57e-29
		Log-Likelihood:	-643.53
No. Observations:	105	AIC:	1311.
Df Residuals:	93	BIC:	1343.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|
const	52.0706	23.259	2.239	0.028
genre_racing	257.9498	93.262	2.766	0.007
genre_shooter	151.7964	64.627	2.349	0.021
genre_simulation	-124.1876	60.091	-2.067	0.042
rating_E	69.3619	31.850	2.178	0.032
rating_M	-111.3395	40.928	-2.720	0.008
meta_overview_Generally favorable reviews	-219.5786	24.005	-9.147	0.000
meta_overview_Mixed or average reviews	-248.3889	27.331	-9.088	0.000
meta_overview_Universal acclaim	520.0381	39.984	13.006	0.000
user_overview_Generally favorable reviews	185.8357	35.362	5.255	0.000
user_overview_Generally unfavorable reviews	226.9570	65.181	3.482	0.001
user_overview_Mixed or average reviews	196.3866	37.782	5.198	0.000
user_overview_Universal acclaim	-557.1087	104.380	-5.337	0.000
Developer_type_First	197.9016	35.616	5.557	0.000

Figure 2: Regression results showing variable importance

*Note that these insights are based on the limited data set used for analysis. Hence, these insights may change if additional data is added. Another point to take in consideration is that the limited sales data might not be representative of the market.

4 Conclusion

The main objective of this project was to develop a recommendation system that will solely be based on users' and critics' review. This recommendation system generated four different types of recommendations with one focusing on users' reviews, second focusing on critics' reviews, third focusing on hidden gems obtained from users' reviews and the fourth focusing on hidden gems based on critics' review. The results of these recommendations varied a lot showing that it was a good idea to keep user and critics' reviews separately as users' opinions sometimes do not match with critics'. In fact, in most of the cases the recommendations generated using critics' reviews were more accurate as they greatly matched the characteristics and genre of the game the user input. This could be very well due to the fact that critics usually use proper terminology, making the recommendations more accurate. Another important point to consider is that the relevance of recommendations varied a lot depending on the game the user input. For example, sometimes the recommendations matched the genre of the user's preferred games and sometimes they did not. On one side, this is reasonable considering that the recommendations were solely based on reviews and were able to provide cross-genre recommendations, which can be sometimes difficult to obtain from traditional recommendation systems that generate recommendations based solely on games' genre and characteristics. This shows that there is room to improve the accuracy of the recommender system. The following steps can be taken to improve accuracy:

- Including other features such as games' ratings, genres, sentiment score of comment to generate recommendations.
- Additional data can be incorporated for different consoles and games. This will lead to a larger dataset, which can potentially yield better results.
- Data overlay sources should be explored to derive even deeper insights that could prove to be useful for generating recommendations. For example, for this report there was lack of sales data, which is why no significant insights were derived. However, with the help of proper sales data, the games recommended as hidden gems would have been more accurate.
- Gathering data from other similar websites. This will help in comparison and validity purposes.
- Using more sophisticated machine learning algorithms that can put more weight on the important attributes in the comments.
- Building an interactive recommendation engine using Dash to provide customers a better user experience.