A Data Analysis of Steam's Game Catalog and Diverse Recommendation Strategies

Emina Salkanović
International Burch University
Francuske Revolucije BB., 71000
Sarajevo, Bosnia and Herzegovina

Nejla Zukorlić
International Burch University
Francuske Revolucije BB., 71000
Sarajevo, Bosnia and Herzegovina

Lamija Oković
International Burch University
Francuske Revolucije BB., 71000
Sarajevo, Bosnia and Herzegovina

Dino Kečo
International Burch University
Francuske Revolucije BB., 71000
Sarajevo, Bosnia and Herzegovina

ABSTRACT

This research paper deals with the video game recommendation system on the Steam platform. The goal is to recommend games to users that are similar in certain parameters but not focusing only on popular games. In addition, a detailed Exploratory Data Analysis (EDA) was conducted to find out which factors influence the popularity of the game and which games the players prefer the most. It is found that some of these factors are game genres, prices, ratings, and publisher reputation. Very important insights into the popularity of games and the connection between different features were gained through the testing of hypotheses. To provide better recommendations to users, advanced metrics and custom models of K-nearest neighbors (KNN) have been developed. The biggest advantage of this system is that it provides a balance between recommending games that are relevant based on features and at the same time provides a different variety of games. This approach can be very useful in the further creation of advanced recommendation systems and in addition offers very significant insights into the video game industry itself, which can be very effective in improving user engagement and satisfaction on a platform such as Steam.

General Terms

Recommendation Systems, Exploratory Data Analysis (EDA), Machine Learning, K-Nearest Neighbors (KNN), Video Game Analytics, User Engagement, Player Preferences, Gaming Industry Trends

Keywords

Steam, video game recommendation, Exploratory Data Analysis (EDA), content-based recommender system, K-Nearest Neighbors (KNN), game popularity, player preferences, user engagement, analytical framework

1. INTRODUCTION

Online platforms for video games are developing more and more every day, and they are becoming more and more popular. In addition, in today's world, video games change very quickly and players always want something new and different. For this reason, it has become increasingly important for the platform to recommend games to players that will satisfy their preferences. It has also become very important to recommend a diverse range of games to the player, not focusing only on the popular ones, because the player has certainly heard of them, but to recommend something different to improve his user experience. For this reason, it becomes very useful to analyze various databases of video game platforms in order to better understand the behavior of players and what they prefer. In this paper, Steam's game catalog has been analyzed, with various factors such as user interactions, game tags, and purchase history examined, as they all contribute to game recommendation algorithms.

This kind of analysis is very important, because it reveals insights into the behavior of players, the popularity of games, and what influences that popularity, and provides some important insights for the development of different recommendation systems. Content-based recommender systems are considered very useful for recommending games to players based on their attributes, such as game genres, game tags, game content, publisher, and others. Such systems are very important when the player wants to find a new game based solely on content. The problem arises if games that he often sees are always suggested to him. This happens when the model is biased towards popular games, so it always recommends the most popular games. In order to find a solution for such an issue, a detailed analysis of the Steam catalog of games was conducted, the weaknesses and strengths of different models were discussed, and an attempt was made to identify the best possible solution. The main goal is to provide as detailed an analysis as possible through testing different hypotheses, which could be useful further in the gaming industry. In addition, the goal is to offer the player as many different games as possible based on the content, while still respecting the various attributes that are important.

The structure of this paper is as follows: Section 2 provides a detailed review of the literature, where it summarizes the papers and their contribution to game recommendations. Section 3 describes the dataset and methodologies used in this analysis. Section 4 gives the results of hypothesis testing and the Diversity metric for the video game recommender. Section 5 discusses the implications of the findings, limitations of the study, and suggestions for future research. Finally, Section 6 concludes the paper with a summary of key contributions and their applicability to the gaming industry.

2. LITERATURE REVIEW

As part of this study, findings drawn from other related research works conducted with the intent of analyzing Steam's environment and the recommendation systems are compiled in this work. The papers cover text mining for understanding the data, finding patterns of data for making categories sharper, conducting EDA for encouraging discussions among the researchers and for backing up their discoveries. In a study by Windleharth et al., Steam tags were studied extensively, and 29 different categories were identified that exist in the platform and includes categories such as gameplay, genre, mechanics, progression, narrative genre/theme, setting, and feeling/mood [1]. What they already knew: Their research unveiled facets about the categorization of players and their understanding. By applying EDA, they checked their analysis, as well as discovered patterns and finetuning categories with the help of the Steam tags regarding the shortage of information essential for the development of video game metadata systems [1]. Another use of EDA was realized by Xiaozhou Li and Boyang Zhang, whereby utilization of user-generated tags from Steam unearthed prominent core tags like Indie, along with major game genres such as Strategy & Simulation [2]. They used the said methodology to define several tags and themes related to gameplay in order to analyze and identify the different genres of video games on Steam based on the players themselves [2]. In their article, G. Cheuque et al have brought out several points that revealed that game content and context should be taken into huge consideration while improving the game recommendations. In their study, 'Recommender Systems for Online Video Game Platforms: In the case ‘Gentrification and the Case of STEAM,’ they adapted several recommendation systems such as Alternating Least Squares, Factorization Machines, Deep Neural Network, Deep Factorization Machine (ALS, FM, DNN, DeepFM). This meant that although the ALS baseline model had been established when it was tested it proved ineffective and all the more highlighted the significance of content and contextual data [3]. The use of recommendation systems in the industry of video games involves the following strategies: implicit users’ data, collaborative filtering algorithms and creative models. Regarding collaborative filtering algorithms, seven different algorithms were also discussed and among them, SlopeOne and SVD++ algorithms work well [4]. Separately, a website recommendation system was designed by Man-Ching Yuen and Chi-Wai Yung using tools such as the Python programming language, the Tableau tool, and the Flask framework [5]. They have included an interface with other two-dimensional power point like buttons, an internal web hosted site, a typed box for the name of games, and a sun radial, a sunburst chart showing the ratio of games having positive results. The system employs a score that identifies the likeness between the input query and games [5]. The focus of the research work done by Yang highlights aspects such as personalization, game contextualization, and relations between players and performs better than the traditional GNN models [6]. Similarly, Balapriya & Srinivasan from Sathyabama University Another paper from Sathyabama University also tries to learn player interests and behaviors to improve the loot suggestion, in which classification and recommendation techniques, sentiment analysis, and Convolutional Neural Networks (CNN) have been employed [7]. Fasiha Ikram and Humera Farooq have jointly created a recommendation system to incorporate multimedia into the users’ profiles. More so, their deep visual semantic-based multimedia recommendation model identifies high accuracy especially compared to other current models [8]. Furthermore, a recommendation system with a recommending capability by paying attention to hours played, to optimize the already existent item-based CF algorithms was designed by the group led by Javier Pérez-Marcos [9]. Akash S. and his team introduced a recommendation system using collaborative filtering and compared three algorithms, the approach that requires prognosis of utilization values and the collaborative filtering with ALS, collaborative filtering with EM and SVD, and a content-based system [10]. Finally, it is found that the collaborative recommender by applying the ALS algorithm is the best performing system [10]. Finally, the paper by Markus Viljanen et al. concentrates on enhancing the multi-purpose models, applicable in any evaluation situation, where such aspects as the filter with the players’ cooperation and their interaction are taken into account [11]. They use cross-sectional survey data from different countries and discuss issues of practicality and understanding of player characteristics [11]. In conclusion, these studies play a valuable role in the evaluation of video game recommendation systems as they stress the importance of content, contexts, user behavior data, and creativity for timely and personalized game recommendations. Implementing these sorts of insights helps to advance recommendation systems within the video game trade last, permanently.

3. DATA AND METHODOLOGY

3.1 Data Source and Description

The primary dataset used for this investigation was collected by Nik Davis [12] and is based on games on the Steam Store. This data was collected up to May 2019 and covers all products that were released in the store before this date including most gaming products. The main dataset has 27,075 records and 18 features, including all the features to capture all the aspects of the full description of the wide range of games developed on Steam platforms. In the following Table 1. represents the attributes of video games, their description, and data type.

Table 1. Dataset Key Variables and Descriptions

Variable	Description	Data type
‘appid’	Unique identifier for each game	Numerical
‘name’	Name of the game	Categorical
‘release_date’	Release date of the game	Date
‘english’	Indicator if the game supports English	Categorical
‘developer’	Developer of the game	Categorical
‘publisher’	Publisher of the game	Categorical
‘platforms’	Platforms the game is available on	Categorical
‘required_age’	The age requirement to play the game	Numerical
‘categories’	Categories associated with the game	Categorical
‘genres’	Genres associated with the game	Categorical
‘steamspy_tags’	Tags assigned to the game on SteamSpy	Categorical
‘achievements’	Number of achievements available in the game	Numerical
‘positive_ratings’	Number of positive ratings	Numerical
‘negative_ratings’	Number of negative ratings	Numerical
‘average_playtime’	Average playtime in hours	Numerical

‘median_playtime’	Median playtime in hours	Numerical
‘owners’	Estimated number of owners of the game	Numerical
‘price’	The price of the game in USD	Numerical

3.2 Data Preprocessing, Normalization, and Scoring Logic for Recommender System Application

In this subsection, the data preprocessing steps are described in detail, and new attributes that will be needed later for the development of the recommender system are added. The first step is the normalization of game ratings, followed by the normalization of the average time spent by players playing the game, and the normalization of the number of owners of each game.

3.2.1 Normalized Rating

The normalized rating is obtained using a weighted average of the individual ratings and where α (alpha) is 10. The factor α (alpha) helps to smooth out the smooth out the rating, ensuring that unusual or extreme ratings don’t overly influence the overall score. The formula for the normalized rating is as follows:

$$Normalized\ Rating = \frac{Positive\ Ratings + \alpha}{Positive\ Ratings\ + Negative\ Ratings + 2 \cdot \alpha}$$

3.2.2 Owners Score

The owner score is calculated based on the number of owners each game has. The scores range from 0.1 to 1. A custom function, *getScoreByOwnersCount()*, which was written, is used for this normalization. It assigns specific values to different ranges of owners. For example, games with 0-20,000 owners are given a score of 0.1, while games with 50,000-100,000 owners are assigned a score of 0.3, and so on. This logic is applied to avoid bias toward only the most popular games.

3.2.3 Normalized Playtime

The game time was normalized based on the average game time spent by players and the total game time across all games. This type of normalization is used to ensure that recommendations are not biased toward only the most played games. The formula is as follows:

$$Normalized\ Playtime = \frac{log(Average\ Playtime + 1)}{log(Total\ Playtime + 1)}$$

3.2.4 Index Columns

In order to develop the recommender system, four new columns (indexes) have been generated, which will be key for the future. Those are: “time_engaged_index”, “player_acclaim_index”, “adopters_choice_index”, and “game_nexus_index”. Each of these indexes focuses on different aspects of the game, such as player time spent on each game, game rating, and number of owners. An overview of how each of these indices was generated is provided in Table 2., along with formulas that show how each index is calculated based on the specified weights for normalized rating, normalized playing time, and number of owners.

Table 2. Game Performance Index Formulas

Index Column	**Formula**
time_engaged_index	TEI = 0.05 * Ratings + 0.9 * Played Hours + 0.05 * Owners
player_acclaim_index	PAI = 0.9 * Ratings + 0.05 * Played Hours + 0.05 * Owners
adopters_choice_index	ACI = 0.05 * Ratings + 0.05 * Played Hours + 0.9 * Owners
game_nexus_index	GNI = 0.33 * Ratings + 0.33 * Played Hours + 0.33 * Owners

In the previous formulas in Table 2., it can be observed that each index has different weights multiplied by attributes, depending on the focus of a particular index. For example, in the formula for the “time_engaged_index”, the highest weight is multiplied by the time spent playing the game, thereby placing emphasis on game time. This means that when games are recommended by the system based on this index, the result will be games that are similar, with a focus on playing time.

3.2.5 Optimization and Categorization

After the previous four indices were generated, it became necessary to convert the numerical results into categorical string ratings. This step is important due to the subsequent application of the K-Nearest Neighbors (KNN) model. To assign categorical string ratings to the results of the indices, they were divided into ranges based on percentiles. In Table 3., the percentile ranges and their corresponding string ratings are shown.

Table 3. String Ratings Based on Percentile Ranges

Percentile Range	**String Rating**
Range 1	Unacceptable
Range 2	Subpar
Range 3	Mediocre
Range 4	Below Average
Range 5	Average
Range 6	Above Average
Range 7	Good
Range 8	Excellent
Range 9	Outstanding
Range 10	Masterpiece

3.3 Hypotheses Formulation

In this subsection, the hypotheses have been formulated based on the goals of the analysis, which aim to determine the factors influencing game popularity, game prices, and player preferences on the Steam Store. Table 4. presents the formulation of each hypothesis, along with the justification for each one.

Table 4. Hypotheses Formulation and Justification for Analysis

Hypothesis	Description	Justification
H1	Games in the “action” genre have a higher average number of owners compared to games in other genres.	Action games are often popular and widely marketed, attracting a larger player base.
H2	There is a negative correlation between game prices and the number of owners.	Lower-priced games are more accessible to a border audience, potentially leading to higher sales volumes.
H3	Games with higher ratings tend to have more owners.	Higher ratings reflect better game quality and user satisfaction, driving more purchases.
H4	Players may prefer games that blend multiple genres, leading to higher ratings and playtime compared to single-genre games.	Blending genres can provide a richer and more varied gameplay experience, appealing to a broader audience.
H5	The number of game releases on the Steam Store has surged since 2014, and these games may have higher number of owners.	The Steam Store has seen rapid growth in game availability, particularly in recent years.

3.4 Visualization Techniques for Hypotheses Testing

To test each of the aforementioned hypotheses, different visualization techniques were used. The methods were selected to best present whether a hypothesis is correct or not. In Table 5., it is clearly shown which visualization technique was used for each hypothesis.

Table 5. Summary of Visualization Techniques for Hypotheses

Hypothesis	Visualization Technique	Description
H1	Bar Chart	Visualizes the distribution of the number of owners across different genres, highlighting action games.
H2	Scatter Plot	Visualizes the relationship between game prices and the number of owners.
H3	Bar Chart	Visualizes the relationship between game ratings and the number of owners.
H4	Bar Chart	Compares the ratings and playtime distributions for multi-genre and single-genre games.
H5	Bar Chart	Visualizes the distribution of game releases by year and distribution of game owners by their release year.

3.5 Recommender System

The core of the recommender system is the K-nearest neighbors (KNN) model, which uses cosine distance to recommend games. Similar games are recommended based on attributes, which, in this case, are genres, categories, SteamSpy tags, along with the generated indexes. In this study, the process began with the basic KNN model, using cosine similarity as the baseline, to allow for later comparison with other models. Five custom KNN models, each with unique weight characteristics, were then introduced:

1. Basic KNN Model with Cosine Similarity
2. Custom KNN Model with Weighted Fundamental Features
3. Custom KNN Model with Weighted Fundamental Features and TEI Score
4. Custom KNN Model with Weighted Fundamental Features and PAI Score
5. Custom KNN Model with Weighted Fundamental Features and ACI Score
6. Custom KNN Model with Weighted Fundamental Features and GNI Score

Unlike collaborative filtering, games are recommended by content-based systems solely based on game content and certain specified attributes. Such an approach is very useful if, for example, a player who likes strategy games gets accurate recommendations for similar ones based on attributes. These models, which include Weighted Fundamental Features with Time Engaged Index (TEI), Player Acclaim Index (PAI), Adopters Choice Index (ACI), and Game Nexus Index (GNI) results, enable very diverse recommendations to players based on the content, while also allowing the player to discover something new. In this approach, the bias towards the most played and popular games, which the player frequently encounters, is bypassed. When the Diversity metric of these models was calculated, computational complexity was encountered, primarily due to difficulties with the calculation of feature weights when determining the cosine distance. In order to solve this issue, the original data set of 27,075 video games was reduced to 2,816, but while preserving the cumulative distribution curves of all observed features. It is important to emphasize that the reduced dataset was used only to calculate the Diversity metric and that the game recommendations were derived based on the original dataset. In Figure 1, it can be seen that the feature distributions of the reduced and original datasets are almost identical. This way of maintaining feature distributions includes cumulative distribution functions (CDF) for visual and quantitative analysis [13]. So, it can be observed that the reduced dataset is representative, and the results themselves are representative thanks to the almost equal distributions. Furthermore, it should be noted that in this paper, the recommendations were generated for the video game Dota 2. This game was selected because of its numerous characteristics and popularity, with the aim of assessing the diversity of the recommendations obtained.

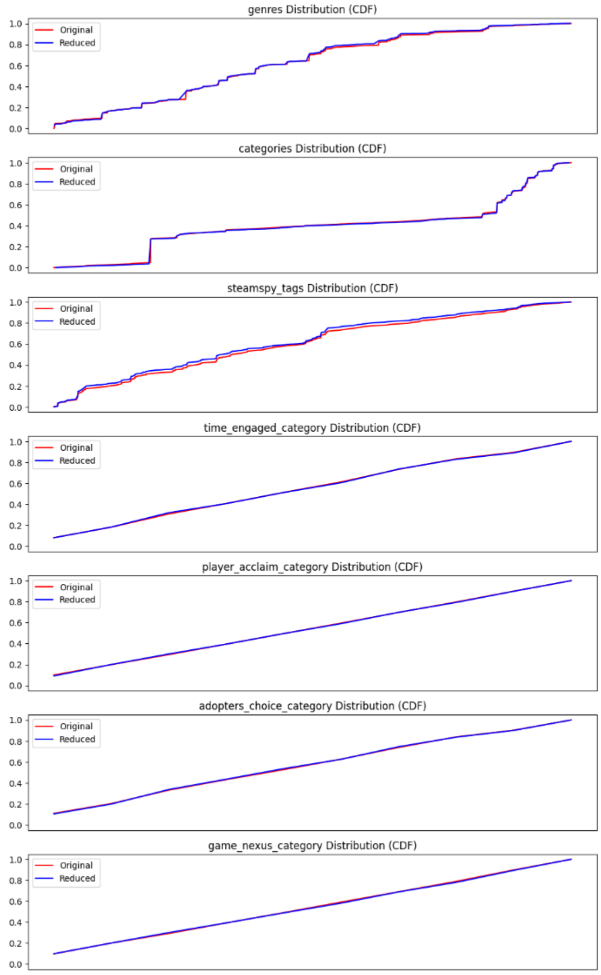


Figure 1: Feature-wise CDF Distributions: A Comparative Analysis of Original and Reduced Video Games Dataset

4. RESULTS AND DISCUSSION

In this section, some key findings of the research and analysis are presented, as well as the results of the Recommender System in terms of Diversity metrics. Each part of these findings will be discussed in detail, highlighting the significance of the results and their impact on the overall performance of the system.

4.1 Hypotheses Testing

The bar chart in Figure 2 clearly supports Hypothesis 1 by showing that the "Action" genre has the most owners out of the top 20 genres. This chart shows that "Action" games have over 500 million owners, which is much more than other genres. This big difference suggests that "Action" games are more popular and owned by more people than games from other genres. This figure strongly supports the idea that "Action" games have more owners on average.

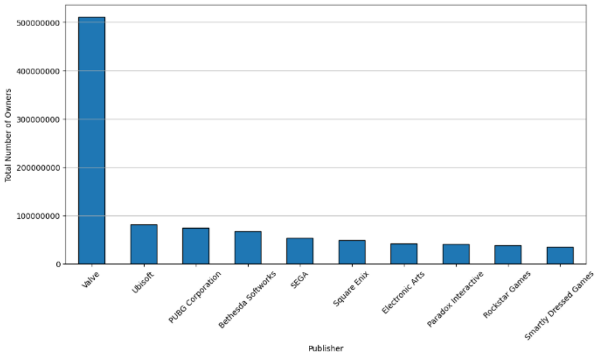


Figure 2: Total Number of Owners per Genre (Top 20 Genres)

The scatter plot in Figure 3 helps prove Hypothesis 2 by showing a link where game prices go down as the number of people who own them goes up. The chart shows that when game prices rise, fewer people own them. Most games that many people own are cheaper. On the other hand, games that cost more usually have fewer owners, suggesting that cheaper games are more liked by users.

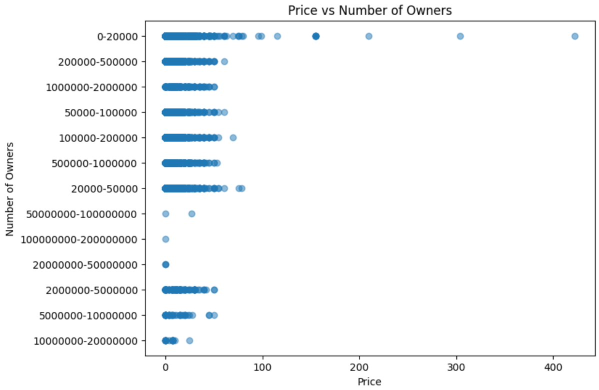


Figure 3: Relationship Between Game Price and Number of Owners

The bar chart shown in Figure 4 presents the distribution of positive and negative by the number of owners. It can be seen on the scatter plot that games with a larger number of owners mostly have a larger number of positive ratings. For example, games with a range of owners from 20,000,000 to 50,000,000 have the largest number of positive ratings, followed by other ranges such as 10,000,000-20,000,000 and 5,000,000-10,000,000. On the other hand, it is observed that games with a smaller number of owners, such as those in the ranges of 0-20,000 and 100,000-200,000, have relatively fewer positive ratings. Therefore, a positive correlation can be identified between the number of owners and the total number of positive ratings, indicating that games with a larger number of owners typically have a higher number of positive ratings. This supports the hypothesis that games with a higher number of positive ratings tend to have more owners, meaning that games with higher ratings attract more players.

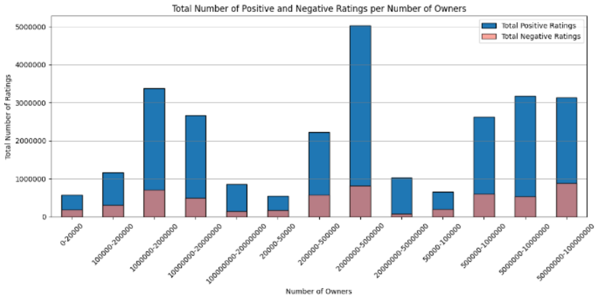
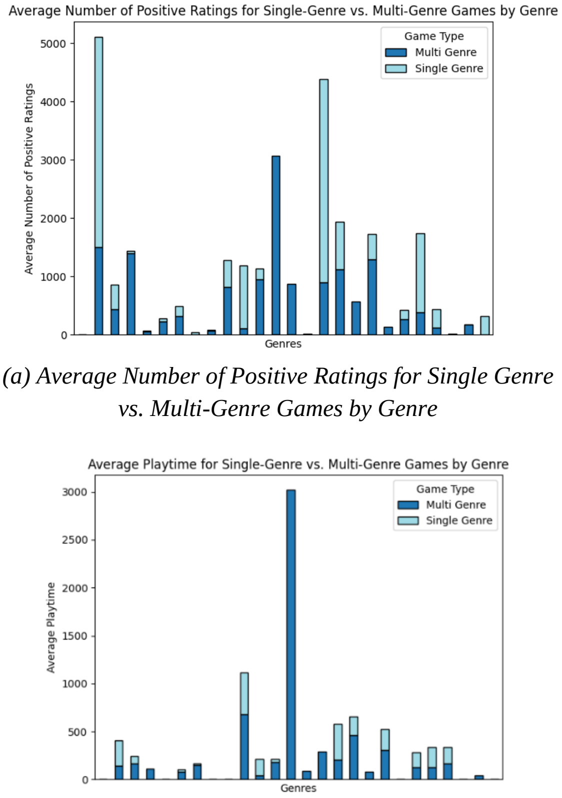


Figure 4: Total Number of Positive and Negative Ratings Across Different Ownership Ranges

The following diagrams shown in Figures 5 provide insight into player preferences for games that have multiple genres.



(a) Average Number of Positive Ratings for Single Genre vs. Multi-Genre Games by Genre

(b) Average playtime for Single-Genre vs. Multi-Genre Games by Genre

Figures 5: Comparison of Positive Ratings and Playtime for Single-Genre vs. Multi-Genre Games by Genre

The bar chart in Figure 5a. shows that multi-genre games generally receive a higher number of positive ratings compared to single-genre games across genres. This visualization supports the hypothesis that players prefer multi-genre games. This leads to a higher number of positive ratings for a certain game. Next, the bar chart in Figure 5b shows that multi-genre games also have higher average play times compared to single-genre games across many genres. For example, in genres such as "RPG", "Strategy" and "Adventure", multi-genre games have significantly longer playing times. Overall, Figure 5 supports Hypothesis 4 by showing that players generally prefer multi-genre games. The reason for this is probably that the presence of multiple genres in games probably provides a richer and more varied gaming experience, which appeals to a large number of gamers. This trend is evident in various genres, highlighting the advantage of multi-genre games in terms of player engagement and satisfaction. Regarding Hypothesis 5, first, Figure 6 shows a significant increase in the number of game releases starting around 2014, with releases peaking in the following years. This indicates a noticeable increase in the number of games featured in the Steam Store during this period.

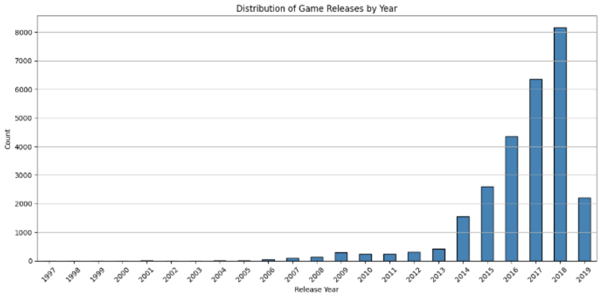


Figure 6: Distribution of Game Releases by Year

Figure 7 shows a noticeable leap in the number of owners for games released in 2013. The year 2013 was a very significant year with particularly popular game releases, probably due to the development of the gaming industry at that time. With this, the number of owners also increased. Despite this leap in 2013, the general trend observed from the figures is that the number of game releases increased significantly from 2014 onwards, and most of these games have a higher number of owners. The sharp leap from 2013 can be considered an exceptional case when looking at the broader trend. Therefore, while 2013 stands out, the overall increase supports the hypothesis that the increase in game releases since 2014 is correlated with higher ownership numbers for those games.

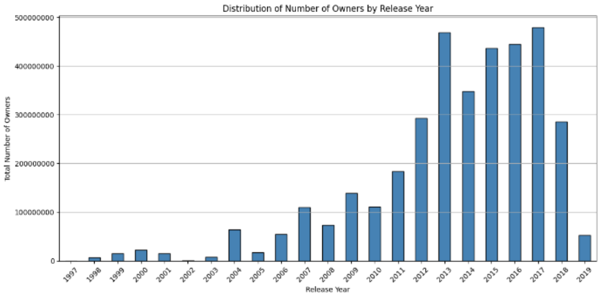


Figure 7: Distribution of Number of Owners by Release Year

Together, these figures illustrate a significant increase in game releases starting in 2014. This has been accompanied by an increase in the number of owners. This supports the hypothesis. This examination of the hypotheses provided a clear understanding of the factors influencing game popularity, pricing, and player preferences on the Steam platform. All hypotheses were proven, confirming that genre, pricing, publisher reputation, product quality, and platform availability are key factors affecting game success.

4.2 Diversity in Video Games Recommendations

In the comprehensive analysis of recommender system models, the focus was placed on evaluating the diversity and effectiveness of various K-nearest neighbors (KNN) models. The primary goal was to balance the diversity and relevance of in-game recommendations, ensuring a satisfying user experience. The models were assessed using different metrics:

Basic KNN with Cosine Similarity, Custom KNN with Weighted Fundamental Features, Custom KNN with Weighted Fundamental Features and Time Engaged Index (TEI) Score, Custom KNN with Weighted Fundamental Features and Player Acclaim Index (PAI) Score, Custom KNN with Weighted Fundamental Features and Adopters Choice Index (ACI) Score, and Custom KNN with Weighted Fundamental Features and Game Nexus Index (GNI) Score. Table 6. summarizes the diversity results of the Basic KNN Model and various Custom KNN Models, including their respective metrics, feature weights, and additional features such as TEI, PAI, ACI, and GNI. After thoroughly evaluating the models, it can be observed that across all Custom KNN models, Metric 4 consistently provided the best balance between diversity and relevance, ensuring that users receive a diverse set of game recommendations. Each Metric 4 effectively integrates additional metrics, TEI, PAI, and ACI with core features, offering a comprehensive recommendation system that enhances user satisfaction. This metric ensures a comprehensive and satisfying user experience by addressing varied preferences while maintaining a high level of diversity in the recommendations. The diversity results for Metric 4 under TEI, PAI, ACI, and GNI scores reveal significant insights into how weighting features and focusing on specific indices affect recommendations for Dota 2. As seen in Table 7., the recommendations based on the TEI score prioritize games likely to engage players for extended durations. The top recommendations, such as Dropzone (77%), Atlas Reactor (76.62%), and AirMech Strike (76.61%), share a gameplay style that demands strategic thinking and long-term player investment, similar to Dota 2. This metric successfully identifies games with mechanics designed for prolonged engagement, aligning closely with Dota 2's multiplayer online battle arena (MOBA) genre. However, it places less emphasis on broader categories, focusing instead on engagement patterns.

Table 7. Dota 2 Recommended Games using Custom KNN Model with Weighted Fundamental and TEI Score Metric

Game Name	Similarity (%)
Dropzone	77.00
Atlas Reactor	76.62
AirMech Strikes	76.61
SMITE®	76.26
Empires Mod	75.85
Hand Of the Gods	75.77
Duelyst	75.48
Realm Grinder	75.22
Emporea: Realms of War & Magic	75.10
Wuxing Master (CCG)	75.07

Table 8. represents the recommendations based on the PAI score, which highlights games with strong player reviews and reputation. Notably, AirMech Strike (76.61%), Empires Mod (75.85%), and Duelyst (75.48%) remain in the top positions, similar to TEI-based recommendations. However, unique games such as My Lands: Black Gem Hunting (74.47%) and Tactical Monsters Rumble Arena (74.47%) appear in this list. These results suggest that PAI diversifies recommendations by introducing games with a strong player reputation, even if they differ slightly from Dota 2's core gameplay style.

Table 8. Dota 2 Recommended Games using Custom KNN Model with Weighted Fundamental and PAI Score Metric

Game Name	Similarity (%)
AirMech Strike	76.61
Empires Mod	75.85
Duelyst	75.48
Realm Grinder	75.22
My Lands: Black Gem Hunting	74.47
Tactical Monsters Rumble Arena	74.47
Bloodline Champions	74.42
Zero-K	74.32
Istrolid	74.18
Battle for Wesnoth	74.05

Next, as shown in Table 9., the ACI score prioritizes games that are popular among a similar player base to Dota 2. This focus is evident in the inclusion of Prime World (95.72%), which has an exceptionally high similarity score compared to the other games. Other notable recommendations, such as Strife® (76.64%) and Dogs of War Online (76.76%), emphasize shared player demographics and community preferences. This model aligns more with the player type than gameplay mechanics, reflecting a tendency to recommend games that attract a similar audience to Dota 2.

Table 9. Dota 2 Recommended Games using Custom KNN Model with Weighted Fundamental and ACI Score Metric

Game Name	Similarity (%)
Prime World	95.72
Dropzone	77.00
Dogs of War Online	76.76
Strife®	76.64
Atlas Reactor	76.62
AirMech Strike	76.61
Games of Glory	76.54
SMITE®	76.26
Boid	76.26
Battle Islands: Commanders	75.94

Further, Table 10. illustrates the recommendations based on the GNI score. This metric combines gameplay relevance with broader trends, leading to a balance between diversity and similarity. While common recommendations like Dropzone (77%) and AirMech Strike (76.61%) remain, the inclusion of games like GUNS UP! (75.76%) and Forge of Gods (75.42%) highlights GNI's broader scope. GNI considers multiple dimensions, including category, genre, and popularity, providing a well-rounded recommendation set.

Table 10. Dota 2 Recommended Games using Custom KNN Model with Weighted Fundamental and GNI Score Metric

Game Name	Similarity (%)
Dropzone	77.00
Strife	76.64
Atlas Reactor	76.62
AirMech Strike	76.61
SMITE	76.26
Empires Mod	75.85
GUNS UP!	75.76
Duelyst	75.48
Forge of Gods (RPG)	75.42
Age of Conquest IV	75.22

The comparison across metrics highlights both significant overlaps and distinct variations in the recommendations, which depend heavily on the specific weighting strategies used for each model.

The similarities across the metrics reveal a consistent identification of certain games, such as Dropzone, Atlas Reactor, and AirMech Strike, which appear in almost every list. These games likely share core features with Dota 2, including multiplayer strategy gameplay, high player engagement, and real-time mechanics. This consistency suggests that the models effectively capture Dota 2's core characteristics, making these games universally relevant recommendations for its players. Furthermore, most of the recommended games align with the MOBA genre or adjacent strategy genres, emphasizing that all metrics prioritize gameplay experiences similar to Dota 2. Across all metrics, the focus on engagement-oriented titles stands out, as many recommended games are known for their ability to hold players' attention over long periods. This indicates that sustained player engagement is a critical factor in all the models.

Despite these commonalities, the metrics introduce unique differences that shape the diversity and relevance of the recommendations. Each metric brings its own perspective, prioritizing different features and player needs. For instance, PAI (Player Acclaim Index) emphasizes games with strong community reviews and acclaim, leading to the inclusion of unique titles like My Lands: Black Gem Hunting and Tactical Monsters Rumble Arena. These games reflect the metric's focus on highly-rated experiences, even if they differ from Dota 2's core gameplay. Similarly, ACI (Adopters Choice Index) focuses on popularity among similar player bases, leading to the recommendation of Prime World, which achieves an exceptionally high similarity score of 95.72%. This metric leans heavily toward demographic-driven recommendations, emphasizing relevance within the Dota 2 community but limiting diversity. On the other hand, GNI (Game Nexus Index) broadens the scope by incorporating games like Forge of Gods and GUNS UP!, which blend strategic depth with broader appeal. This metric attempts to balance relevance and diversity by considering a mix of factors like popularity, genre, and gameplay mechanics.

The weighting focus of each metric further explains these differences. TEI (Time Engaged Index) prioritizes games with long playtimes, aligning closely with Dota 2 players' engagement patterns. However, this focus can result in a narrower range of recommendations, often favoring games with similar engagement dynamics. In contrast, PAI shifts attention to highly-rated games, introducing diversity by including lesser-known but critically acclaimed titles. ACI, as seen in its emphasis on community overlap, results in more focused recommendations tailored to a specific demographic, while GNI seeks to balance these factors, offering a blend of diversity and relevance that caters to players with varied preferences.

Another notable difference lies in the similarity scores assigned to the games. For example, ACI produces the highest similarity score with Prime World (95.72%), reflecting its strong focus on community overlap. In contrast, TEI, PAI, and GNI show more balanced similarity percentages, reflecting their broader focus on diverse attributes. These variations indicate how the inclusion of weighted features significantly influences the output. TEI emphasizes in-game engagement, while ACI and PAI prioritize community-driven and critically acclaimed recommendations, respectively. GNI combines these factors into a more general approach.

The comparison also raises important questions about potential biases in the recommendation process. Metrics like ACI, which emphasize popularity, may overlook niche or unconventional games that could appeal to Dota 2 players. Similarly, PAI's reliance on high ratings might prioritize games that differ significantly from Dota 2's core gameplay style. TEI's engagement focus, while effective, might miss games with shorter but equally satisfying experiences. GNI attempts to address these biases by blending multiple factors, but its broader scope can sometimes lack the precision of the other metrics.

Overall, the similarities and differences across metrics highlight the strengths and weaknesses of each approach. While all metrics succeed in identifying games that align with Dota 2's core features, their individual weighting strategies significantly impact the diversity and relevance of the recommendations. Each metric caters to different aspects of player preferences, offering a nuanced understanding of how various factors influence recommendation quality. The results emphasize the importance of choosing the right metric based on the desired balance between diversity and relevance, ensuring that recommendations align with user goals while maintaining a satisfying and engaging user experience.

Table 6. Diversity Evaluation of KNN Models

<table>
<tr><th colspan="6">Basic KNN Model with Cosine Similarity</th></tr>
<tr><th></th><th>Steam Game Tags Weight</th><th>Genres Weight</th><th>Categories Weight</th><th>Additional Feature Weight</th><th>Diversity</th></tr>
<tr><td></td><td>-</td><td>-</td><td>-</td><td>-</td><td>0.79821</td></tr>
<tr><th colspan="6">Custom KNN Model With Weighted Fundamental Features</th></tr>
<tr><th></th><th>Steam Game Tags Weight</th><th>Genres Weight</th><th>Categories Weight</th><th>Additional Feature Weight</th><th>Diversity</th></tr>
<tr><td><b>Metric 1</b></td><td>0.1</td><td>0.1</td><td>0.8</td><td>-</td><td>0.54084</td></tr>
<tr><td><b>Metric 2</b></td><td>0.1</td><td>0.8</td><td>0.1</td><td>-</td><td>0.55787</td></tr>
<tr><td><b>Metric 3</b></td><td>0.8</td><td>0.1</td><td>0.1</td><td>-</td><td>0.73899</td></tr>
<tr><th colspan="6">Custom KNN Model With Weighted Fundamental Features and TEI Score</th></tr>
<tr><th></th><th>Steam Game Tags Weight</th><th>Genres Weight</th><th>Categories Weight</th><th>Additional Feature Weight</th><th>Diversity</th></tr>
</table>

Metric 1	0.1	0.05	0.05	0.8	0.15631
Metric 2	0.05	0.3	0.05	0.6	0.24409
Metric 3	0.3	0.05	0.05	0.6	0.30855
Metric 4	0.6	0.05	0.05	0.3	0.53692
Custom KNN Model With Weighted Fundamental Features and PAI Score					
	Steam Game Tags Weight	**Genres Weight**	**Categories Weight**	**Additional Feature Weight**	**Diversity**
Metric 1	0.1	0.05	0.05	0.8	0.15640
Metric 2	0.05	0.3	0.05	0.6	0.24414
Metric 3	0.3	0.05	0.05	0.6	0.30883
Metric 4	0.6	0.05	0.05	0.3	0.53748
Custom KNN Model With Weighted Fundamental Features and ACI Score					
	Steam Game Tags Weight	**Genres Weight**	**Categories Weight**	**Additional Feature Weight**	**Diversity**
Metric 1	0.1	0.05	0.05	0.8	0.15635
Metric 2	0.05	0.3	0.05	0.6	0.24411
Metric 3	0.3	0.05	0.05	0.6	0.30867
Metric 4	0.6	0.05	0.05	0.3	0.53715
Custom KNN Model With Weighted Fundamental Features and GNI Score					
	Steam Game Tags Weight	**Genres Weight**	**Categories Weight**	**Additional Feature Weight**	**Diversity**
Metric 1	0.1	0.05	0.05	0.8	0.15640
Metric 2	0.05	0.3	0.05	0.6	0.24413
Metric 3	0.3	0.05	0.05	0.6	0.53533
Metric 4	0.6	0.05	0.05	0.3	0.53745

4.3 Discussion of Diversity Results and Related Studies

Diversity in recommendations is very important in order to satisfy the player's preferences and improve his overall gaming experience. The study presented different KNN models with different metrics and feature weights, which actually affect the variety of recommended games. The basic KNN model with cosine similarity achieved a high diversity score of 0.79821, indicating a wide range of game proposals. However, in this case, the lack of weighted features can lead to the recommendations not always being closely aligned with the player's preferences when looking to play a game with similar content. On the other hand, the adjusted KNN models, which include weighted underlying features and different scores (TEI, PAI, ACI, and GNI), provided diverse recommendations but again retained some basic characteristics that make the games similar. These models showed diversity scores ranging from 0.15631 to 0.73899. Metric 4 consistently achieved the highest diversity scores across all fitted models. This emphasizes how efficient and good weighted features are in these cases for recommending similar games based on content. By including features such as TEI, PAI, ACI, and GNI in addition to the core features, Metric 4 provides a diverse and highly attractive set of recommendations, while maintaining a balance between popular and lesser-known games. Also, it can be noticed that Metric 4 in each fitted KNN model (weighted underlying features, TEI, PAI, ACI, and GNI scores) consistently yields the highest diversity scores. This balance between popular and diverse game recommendations increases user satisfaction. Therefore, the adoption of metric 4 in each custom KNN model is recommended to optimize the diversity and relevance of video game recommendations. Of course, this ensures a very engaging and varied experience for video game players. Further, this study differs from related works in several ways. Windleharth et al. and Li & Zhang [1] primarily focused on the Exploratory Data Analysis (EDA) of Steam and user-generated tags to enrich metadata systems and understand player perceptions. While their work provided valuable insights, this study goes further by using EDA not just for analysis but as a foundation for hypothesis testing and the development of a content-based recommender system, bridging analysis with practical application. Unlike their studies, this study's EDA insights formed a basis for hypothesis testing and also developed a content-based recommender system. Cheuque et al. [3] evaluated various recommendation models, highlighting the importance of content and context in recommendations. However, this study directly integrates these aspects into a more advanced KNN model with weighted metrics, improving the balance between diversity and relevance in-game recommendations. By incorporating more nuanced metrics such as TEI, PAI, and ACI, this approach offers recommendations that are both diverse and tailored. Ikram & Farooq's [8] multimedia recommendation model lacked the specific focus on video games, which are addressed with game-specific metrics, and Pérez-Marcos et al.'s playtime-based filtering approach is broadened in this work through additional features like owner scores and genre-blending [9]. Finally, Viljanen et al.'s emphasis on player interactions is extended in this study by incorporating both interaction data and content-based metrics, leading to a more comprehensive recommendation system [11]. The novelty of this approach lies in its unique combination of content-based features, weighted indices, and EDA-driven insights, offering a more holistic, tailored, and diverse recommendation system that better captures the complexities of game recommendations.

5. CONCLUSION

This study compares different video game recommendation systems, focusing on the Steam platform. The goal was to explore new methods and metrics to improve game recommendations based on game content. A review of past research helped us understand Steam's environment and the effectiveness of methods like Exploratory Data Analysis (EDA), collaborative filtering, and content-based systems. A number of hypotheses were implemented which helped us to get some ideas about what factors affect the popularity of games and players' choices on Steam. It was observed that action games had a more significant number of owners, lower-priced games drew more owners and high-rated games engaged the player community more. Moreover, multi-genres earned higher ratings as well as playtime people spent playing them. To improve the way games are recommended, this recommendation system used different scores as features for recommendations, such as Time Engaged Index (TEI), Player Acclaim Index (PAI), Adopters Choice Index (ACI), and Game Nexus Index (GNI) scores in K-Nearest Neighbor (KNN) models. A balance between relevance and diversity in game recommendations was successfully achieved by this approach, suggesting games from different genres while maintaining alignment with player preferences. As a result, player engagement and satisfaction were improved over time. The significance of this research is highly valuable for the gaming industry, especially for platforms like Steam, where players have access to a vast selection of games. With such a large variety, it becomes difficult for players to find games that truly match their interests and preferences. This research offers a solution by helping platforms improve the way they recommend games to users, making the process more personalized and effective. For example, the platform can gather data on what features are most important to each player, whether it’s the game genre, the game's popularity, the publisher, or other factors like user ratings and playtime. By understanding what drives a player’s interest in certain games, the platform can gain a clearer picture of each player's preferences. After the platform collects and analyzes this data, it can then integrate this recommendation system to provide more tailored suggestions. The system would not just recommend popular games but would offer a diverse selection that fits within the player’s preferences. So, this approach has the added benefit of helping players discover new and lesser-known games that they might not have considered otherwise. However, there are some limitations to this study. One issue is that a smaller dataset was used, which might limit how generalizable these findings are. Another limitation is that this study didn’t include direct user feedback in this system. In future research, a hybrid recommendation system could be developed. This hybrid system would combine both content - based methods (which focus on the game's characteristics) and collaborative filtering (which looks at player interactions and preferences). By combining these two approaches, an even more accurate and diverse recommendation system could be created, taking into account both game data and player behavior. In the broader field of recommendation systems, this study introduces a more nuanced approach to content-based recommendations. By incorporating advanced systems and tracking player behavior beyond just popular titles, it opens up new possibilities for personalized experiences in various domains, not just gaming. The methodology can be adapted for other fields where personalization and diversity are key, making it a valuable contribution to ongoing research and development in recommendation systems.

6. ACKNOWLEDGMENTS

Our thanks to the University and professors who helped us to complete this work.

7. REFERENCES

[1] T. W. Windleharth, J.J. S. and M. &. L. J. H., “Full Steam Ahead: A Conceptual Analysis of User-Supplied Tags on Steam,” *Conceptual Analysis of User-Supplied Tags on Steam. Cataloging & Classification Quarterly,* 2016.

[2] X. Li and B. Zhang, "A preliminary network analysis on steam game tags: another way of understanding game genres," in *AcademicMindtrek '20: Proceedings of the 23rd International Conference on Academic Mindtrek*, 2020.

[3] G. Cheuque, J. Guzmán and D. Parra, "Recommender Systems for Online Video Game Platforms: The Case of STEAM," in *WWW '19: Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

[4] R. Bunga, F. Batista and R. Ribeiro, "From implicit preferences to ratings: Video games recommendation based on collaborative filtering," in *13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR*, 2021.

[5] M.-C. Yuen, C.-W. Yung, W.-F. Cheng, H.-P. Tsang, C.-H. Kwan, C.-L. Chan and P.-Y. Li, "Game Recommendation System," *Frontiers in Artificial Intelligence and Applications*, 2023.

[6] L. Yang, Z. Liu, Y. Wang, C. Wang, Z. Fan and P. S. Yu, "Large-scale Personalized Video Game Recommendation via Social-aware Contextualized Graph Neural Network," in *WWW '22: Proceedings of the ACM Web Conference 2022*, 2022.

[7] S. Balapriya and D. N. Srinivasan, "A Multi-Level Adaptive Loot Box Recommendation System for Video Games," *International Journal of Aquatic Science*, 2021.

[8] F. Ikram and H. Farooq, "Multimedia Recommendation System for Video Game Based on High-Level Visual Semantic Features," *Scientific Programming*, 2022.

[9] J. Pérez-Marcos, D. Sánchez-Moreno, V. López Batista and M. Dolores Muñoz, "Estimated Rating Based on Hours Played for Video Game Recommendation," in *Distributed Computing and Artificial Intelligence, Special Sessions, 15th International Conference*, 2019.

[10] C. BharathiPriya, A. Sreenivasu and S. Kumar, "Online Video Game Recommendation System Using Content and

Collaborative Filtering Techniques," in *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 2021.

[11] M. Viljanen, J. Vahlo, A. Koponen and T. Pahikkala, "Content Based Player and Game Interaction Model for Game Recommendation in the Cold Start setting," *arXiv*, 2020.

[12] N. Davis, "Kaggle," 2019. [Online].

[13] O. Shukurovich Sharipov, "Glivenko-Cantelli Theorems," in *International Encyclopedia of Statistical Science*, Heidelberg, Springer, 2014, pp. 612-614.