Amazon recommendation based on rating and reviews using bert analysis and collaborative filtering

BY - ARPEET KUMAR

Agenda

- ❑ Introduction
- ❑ Background
- ❑ Literature Review
- ❑ Problem Statement
- ❑ Methodology
  - ❑ Data Selection
  - ❑ Data Cleaning
  - ❑ Data Exploration
  - ❑ Bert Sentiment Analysis
  - ❑ Collaborative Filtering
  - ❑ Hybridizing the above Two Models
- ❑ Results and Discussion
- ❑ Conclusion and Future works

Introduction

Welcome to an exciting journey into the heart of Amazon's recommendation engine. In a world flooded with choices, the need for precise and personalized suggestions is more crucial than ever. Today, we delve into a groundbreaking approach — combining the analytical prowess of collaborative filtering with the linguistic finesse of BERT analysis. Our mission is clear: revolutionize Amazon's recommendation system by decoding user sentiments from reviews and translating them into tailored suggestions. Join us as we explore the fusion of data-driven insights and advanced natural language processing, aiming to elevate the Amazon shopping experience to new heights

Background

- ❑ **Recommender System Shift:** E-commerce, especially on Amazon, has transformed with recommender systems. These systems are vital for user experience, adapting to dynamic preferences and a vast product array.
- ❑ **Synergy Approach:** Elevating Amazon's recommendation system through a combined approach, leveraging collaborative filtering and advanced BERT analysis.
- ❑ **Challenges and Solutions:** Collaborative filtering struggles with sparse data and capturing sentiments. BERT analysis is introduced to address these challenges by excelling in understanding contextual nuances and sentiments in textual data.
- ❑ **Evolution of Recommender Systems:** A brief overview of the historical evolution, emphasizing the transition from basic collaborative filtering to intricate hybrid models, showcasing the industry's commitment to refining personalized recommendations.
- ❑ **Amazon Case Study:** Amazon's extensive product catalog and diverse user base make it an ideal case study. The research aims to create a nuanced, accurate, and context-aware recommendation system.
- ❑ **Addressing Challenges:** Tackling challenges like the cold start problem and scalability issues in collaborative filtering through the incorporation of BERT analysis. The model aims to understand the underlying reasons behind user preferences, paving the way for an emotionally intelligent recommendation engine on the Amazon platform.

Literature Review

Historical Progression:

- Foundational Stages: From early rule-based algorithms in the 1960s to sophisticated machine learning models, recommender systems have undergone transformative journey.
- Dominance of Collaborative Filtering: The 1990s marked the dominance of collaborative filtering, influenced by projects like GroupLens and Netflix's recommendation challenge.

Rise of Hybrid Models:

- Integration of Content-Based Filtering: The 2000s witnessed the integration of content-based filtering to address collaborative filtering limitations.
- Matrix Factorization and Deep Learning: The 2010s brought advancements like matrix factorization and deep learning, enhancing recommendation accuracy for platforms like Amazon and Netflix.

Applications and Challenges:

- Diverse Applications: Recommender systems find applications in e-commerce, streaming, healthcare, and personalized marketing.
- Challenges: Persistent challenges include the cold start problem, scalability issues, and the delicate balance between personalized recommendations and diversity.

Future of Personalized Experiences:

- Continuous Advancements: As technology evolves, recommender systems continue shaping the future of personalized digital experiences,promising improved understanding and prediction of user preferences.

Problem Statement

While the current system thrives on quantitative metrics like purchase history, a pivotal shift awaits—a seamless integration of qualitative insights from customer reviews and ratings. Let's explore how this evolution will bridge the gap, offering users a richer, more nuanced recommendation experience.

- **Quantitative Emphasis:** Amazon's current recommendation system excels in leveraging quantitative data from user behavior, focusing on purchase history and browsing patterns for effective suggestions.

- **Untapped Richness in Reviews:** The existing system falls short in fully utilizing the qualitative wealth embedded in customer reviews and ratings, missing out on nuanced insights into product strengths, weaknesses, and user satisfaction.

- **Balancing Act with Ratings:** While the platform effectively uses quantitative ratings, it lacks the ability to comprehensively integrate both quantitative and qualitative aspects, hindering a holistic understanding of user preferences and experiences.

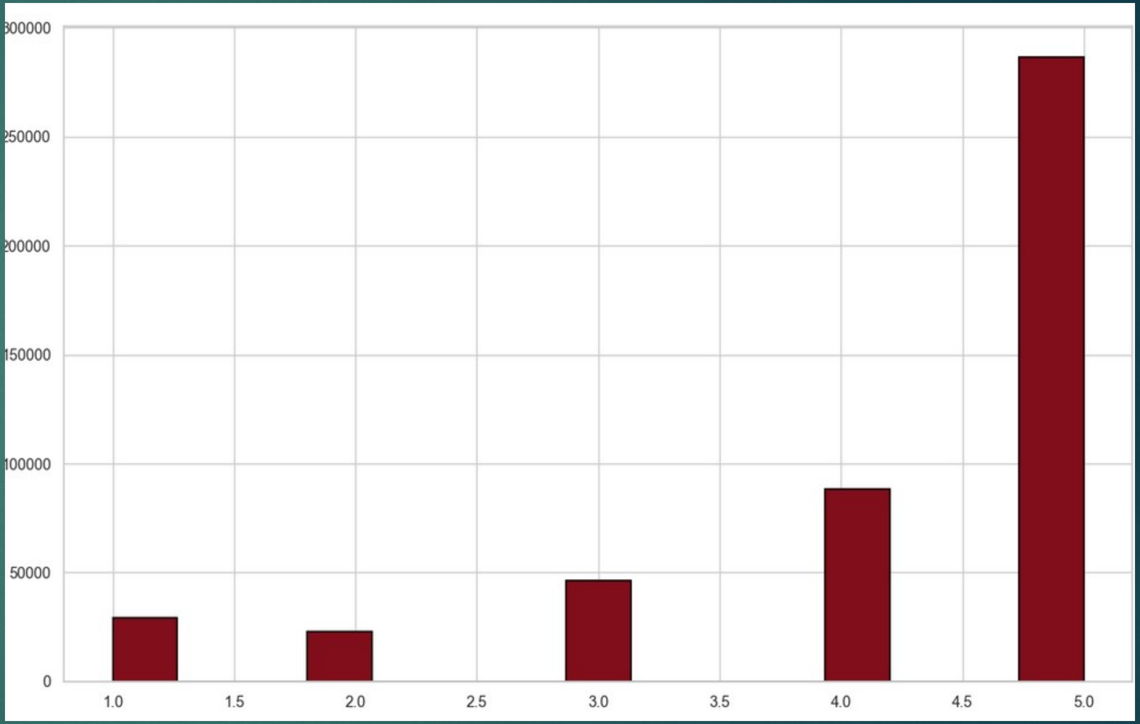
Methodology

- Data Selection
  - Comprehensive Dataset: Unleash the potential of a dataset encompassing 497,577 product reviews and 84,819 detailed product entries, each uniquely identified by ASIN.
  - User Sentiments Insights: Explore diverse user opinions through key columns like 'overall,' 'verified,' 'reviewerID,' and 'reviewText,' gaining valuable insights into sentiments.
  - Reliability and Temporal Aspects: Ensure reliability with the 'verified' column for authenticated reviews and address temporal considerations with 'reviewTime' and 'unixReviewTime.'
  - Contextual Product Information: Leverage the product dataset's 84,819 entries, featuring 'category,' 'brand,' and 'main_cat,' for content-based filtering, enriched by quantitative attributes like 'rank' and 'price.'
  - Hybrid Approach and Challenges: Overcome challenges such as occasional missing values and sparse data, with strategic data preprocessing. Embrace a hybrid approach, integrating dimensions like 'category,' 'description,' and 'brand' for a sophisticated recommender system.

- Data Cleaning
    - For the first dataset we observe that there are a lot columns that are not required because they have null values or they are not required for traing our model so we exclude them and only keep the ratings, productID,Reviews,Summary,Verified and Reviewername
    - For the Second dataset we had 19 columns but not all were required.It had to cloumns for category,alsobuy,brand to mention some we filtered it out according to our needs and kept what was required.
    - We merged the two dataset into one to make it better for our use.We also went ahead and renamed most of the coulmns so that it gets easier for us.
    - Since the dataset was huge and computing such huge data would take a lot of time so we used random sampling of 20% to make it easier for our use.

Exploratory Analysis

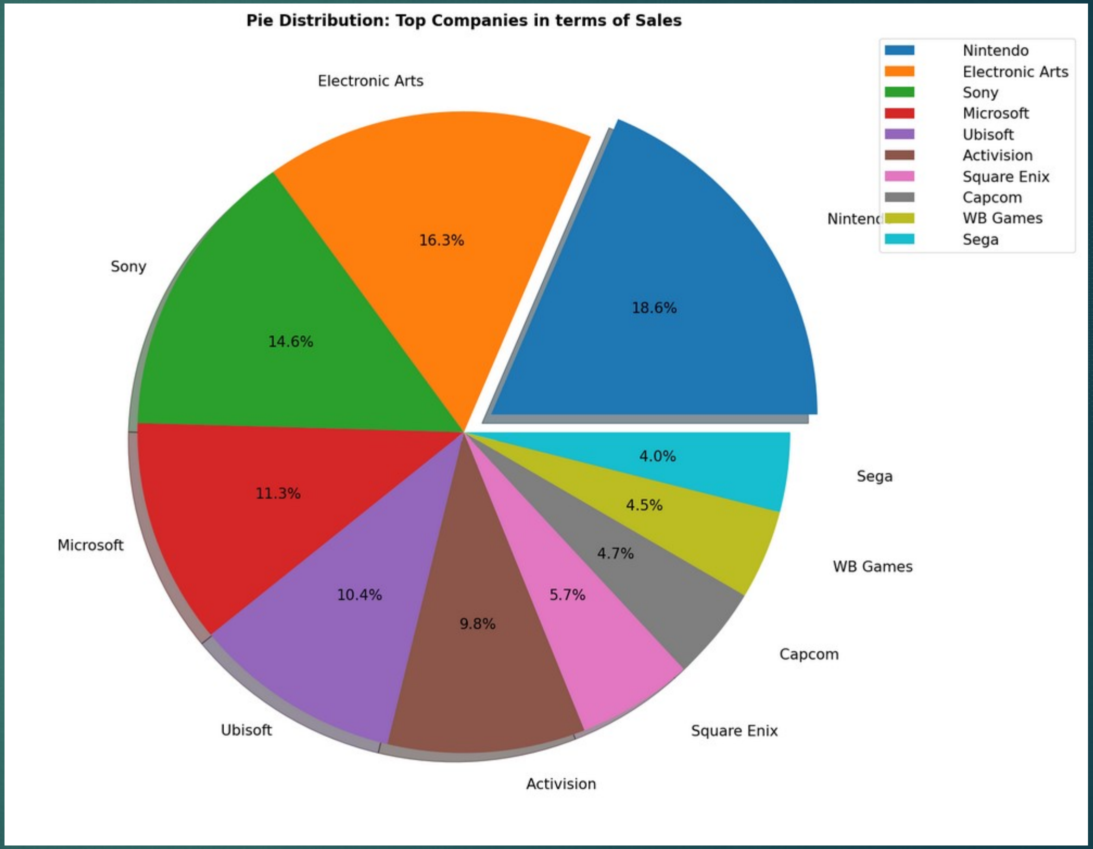
Ratings based on the number of Reviews



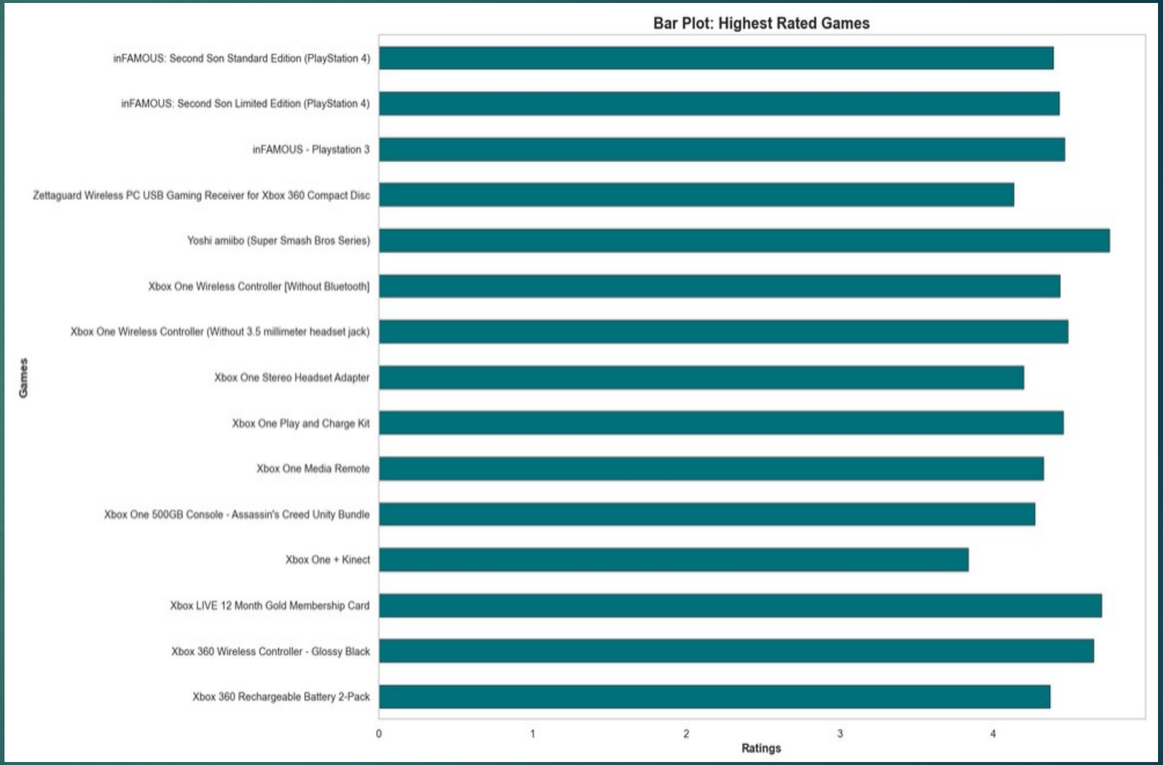
- ❑ It can be observed that there are a lot 5 stars ratings with the dataset.
- ❑ Coparitvely the other ratings are very low we could go ahead and filter the dataset to include only the 5 stars rated Products

Top Companies in terms of Sales.

- We can see that Nintendo has the most number of Products in terms of sales.
- Other companies also have a decent amount of sales but they are not very far behind.
- The top companies are close in terms of sales but in cannot filter based on these because so customers might prefer a specifc company.



Games with the highest amount of Rating



❑ Here we can see the top 15 games in terms of Rating.

❑ We can observe thatthese games are very cloesly rated.

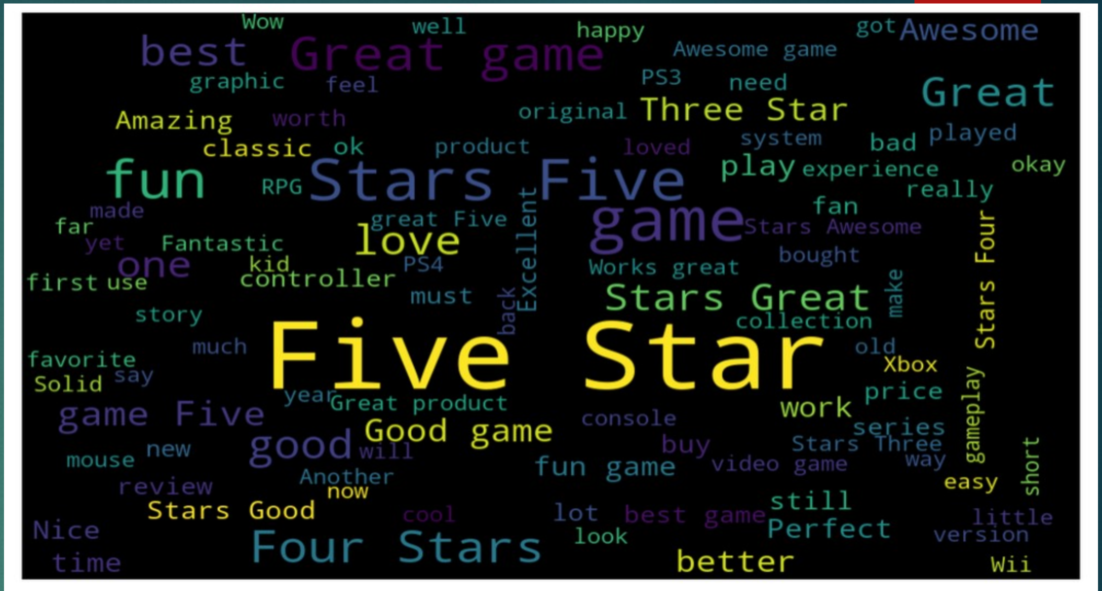
❑ We can go ahead and filter out based on the top 25 or top 30 games for our recommendations to the customers.

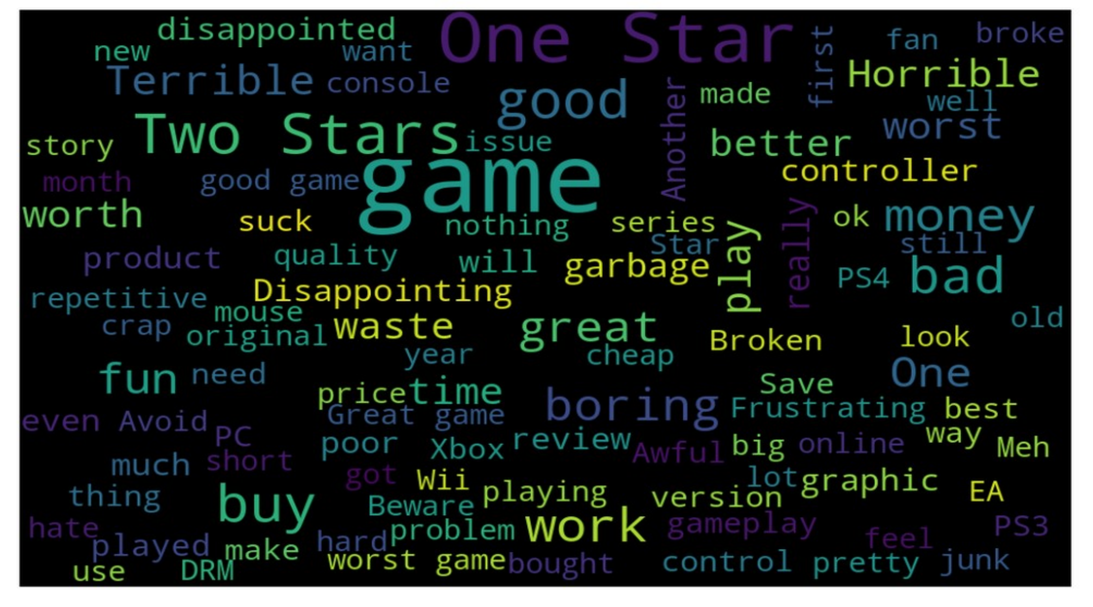
Collaborative Filtering

- Data Preparation:
  - Loading and Splitting: Utilizing the Surprise library, the code loads a dataset into a Surprise Dataset object, splitting it into training and testing sets (30% test size) for collaborative filtering.
- Algorithm Configuration:
  - k-NN with Means: The chosen collaborative filtering algorithm is k-NN with means, configured with a neighborhood size of 5 and Pearson baseline similarity. Item-based collaborative filtering is employed for recommendations.
- Model Training and Evaluation:
  - Training and Testing: The model is trained on the training set and evaluated on the test set. Root Mean Square Error (RMSE) is calculated to assess the accuracy of the collaborative filtering model's predictions.
- Matrix Decomposition with SVD:
  - Latent Factor Extraction: The code performs matrix decomposition using truncated Singular Value Decomposition (SVD) on a user-item ratings matrix. It creates latent factors capturing underlying patterns, essential for collaborative filtering-based recommendation systems.

Sentiment Analysis using BERT

- The first WordCloud shows the Positve reviews based on the summary column.
    - We can observe that there are alot of Five stars,Good Game,great and fantastic to mention some of them that have been added as reviews.
- The second WordCloud shows the Positve reviews based on the summary column.
    - We can observe that there are alot of Dissapointment,garbage,broken,waste and repitive to mention some of them that have been added as reviews.
- Observing both these WordCloud we conclude that it would be a better to filter out the Negative reviews and only keep the Positve reviews for our analysis.





Results and Discussion

- We calculated the Root Mean Squared error for the collaborative filtering and go the values as such:

  Item-based Model : Test Set

  Root Mean Squared Error : 1.1212

  1.1211888393138132

- After hybridizing the model by including BERT sentiment Analysis we are able to bring down the RMSE to:

  Item-based Model : Test Set

  RMSE: 0.7358

  0.7358377523181904

Conclusion

- Item-Based Collaborative Filtering Model:
  - Performance: RMSE of 1.1212 on the test set.
  - Insight: The model, while effective, demonstrates room for refinement in making recommendations based on item similarities.

- Hybrid Model (BERT + Collaborative Filtering):
  - Performance Leap: Significant improvement with an RMSE of 0.7358 on the test set.
  - Key Advantage: Integration of BERT sentiment analysis enhances personalized recommendations, showcasing a more nuanced understanding of user preferences.

- System Efficacy:
  - Affirmation: Results validate the hybrid model's efficacy in navigating diverse user preferences and product interactions on Amazon.
  - Enriched Understanding: The collaboration of advanced language understanding and collaborative filtering leads to a more accurate and context-aware recommendation system.

Future works

- ❑ **Sentiment Analysis Refinement:** Continuous improvement of sentiment analysis using advanced NLP techniques to discern nuanced emotions and context from user reviews, contributing to more precise sentiment-aware recommendations.

- ❑ **Deep Learning for Collaborative Filtering:** Exploration of deep learning architectures, such as neural collaborative filtering and attention mechanisms, to enhance recommendation accuracy, particularly in scenarios with sparse data.

- ❑ **Incorporation of Contextual Features:** Integration of additional contextual features, including temporal dynamics and user behavior patterns, for a more holistic understanding of user preferences, leading to accurate and adaptive suggestions.

- ❑ **Scalability and Performance Optimization:** Research and development efforts focused on scalability, employing strategies like distributed computing or parallel processing to ensure efficiency and effectiveness in handling large-scale datasets.

THANK YOU