Shape

Description automatically generated with low confidence Logo, company name

Description automatically generated

**CAPSTONE PROJECT DRAFT**

**TITLE - AMAZON PRODUCT RECOMMENDATION SYSTEM**

**SUBMITTED TO – PROF. DAYA R.**

**BY-ANANYA SHARMA**

**PRANJALI PATIL**

**DATE- 21th MAY 2022**

**Introduction**

The purpose of the recommendation system is to understand the various products that are being used by customers and then recommend along similar lines products used by other people, to increase revenue generation. This is quite like the visual merchandising approach in brick-and-mortar stores where the store people organize sweeteners next to coffee and juice condiments so that people feel the need to buy both the products together.

Providing a recommendation system helps to streamline the products categorically to make it easier for the customers to purchase them. Netflix gave an option to pay one million dollars to anyone who could make a good, robust system for a recommendation of movies. Spotify, Amazon, Flipkart, Good Reads, and majorly all brands use this system to enhance the shopping experience for their clients and create the nearest experience for them in lieu of in-person visits.

In some businesses, recommender systems are crucial since they may produce a large amount of revenue or serve as a means to differentiate yourself from the competition.

**Dimensions of the Data Set**

We have tried to follow the metrics given in the project. The data set has been picked up from the following link:

Jianmo Ni, Jiacheng Li, Julian McAuley Empirical Methods in Natural Language Processing (EMNLP), 2019 [pdf](http://cseweb.ucsd.edu/~jmcauley/pdfs/emnlp19a.pdf)

The first dataset is for the category amazon fashion and the second is for amazon appliances. The dataset consists of a total of 20,000 records and 20 variables (vote, summary, Description). The dataset consists of 233.1 million reviews and data on ratings

**The rationale for Dataset**

The Product recommendation system is useful for both the user and the customer. We are flooded with data in today's world, and this data offers us vital information. However, the user is unable to extract the information that is of relevance to them from these data. Recommendation systems were created to assist users in obtaining information about a product. The Recommendation system establishes a resemblance between the user and the products and uses that similarity to provide recommendations.

**Questions to Investigate**

1. How to recommend the products based on the reviews and the items of interest?

2. Help websites to improve user engagement

3. Identity products that are most relevant to users

4. How to recommend the products based on the similarity between different users?

5. Exploring the best algorithms that enhance the recommendation system.

**Exploratory Data Analytics**

We must perform the basic cleaning in the dataset so that we don’t have null values in the project.

We are using the columns like reviewer ID, ASIN (Amazon standard identification number), Reviewer Name, Reviewer Time, Verified, Overall Rating, and Review Text to analyze the dataset taken from Amazon Website.

Graphical user interface, text, application

Description automatically generated with medium confidence

**Figure 1:** Columns in a Data frame

Chart, bar chart

Description automatically generated

**Figure 2:** Distribution of Overall Rating

The above figure shows the number of rating products received on varied products.

Chart

Description automatically generated with medium confidence

**Figure 3:** Quantiles and Ratings

The above figure depicts the relationship between quantiles and the ratings of the customers for fashion-based products on Amazon.

Table

Description automatically generated

**Figure 4:** Omitting missing Values

In the figure above, we have tried to explore the null values to understand which data we can use to derive good exploratory data analysis. We have refrained from dropping the values unnecessarily.

Chart, bar chart

Description automatically generated

**Figure 5:** Top 5 products purchased as per ReviewerID

**Predictive Models**

**1. Popularity-based suggestion:**

A popularity-based recommendation system is non-personalized and works on the premise of popularity or anything that is in style. It suggests things that are in flair. The difficulty with a popularity-based recommendation system is that personalisation is not possible, thus even if you know the user's preferences, you can't propose goods based on those preferences.

**2. Collaborative Filtering-**

Based on the similarity between various customers and items offered on an eCommerce website, this is considered one of the most sophisticated suggestion systems. It bases its suggestions on the preferences of other users. Collaborative filtering can be performed using the matrix approach as well. This approach faces a prominent issue, where this method can’t be recommended to new users because they don’t have any preferences set. In this case, the popular items checked by the most active person can be thrown towards as recommendations for the user. This filtering is major of two types: memory-based and model-based.

Square

Description automatically generated

**Figure 6:** Corrplot

Chart

Description automatically generated

**Figure 7:** ALJ66O1Y65LHA has the highest count of recommendations amongst products

3. **Content-based: -** A content-based recommendation uses information provided by the user, either by ratings or search terms. A user profile is created based on this information, and it is then utilized to provide recommendations to the user. The engine becomes increasingly accurate as the user offers more inputs or acts on the recommendations. Machine learning and data mining techniques are used in these procedures. The objective is to develop algorithms that can make judgments. We could, for example, leverage current user-item interactions to train a model that predicts a user's top-5 favourite things. When compared to alternative methods like memory-based approaches, these methods have the benefit of being able to propose a bigger number of things to a larger number of people. Even when working with big sparse matrices, they have a lot of coverage.

**Chart, scatter chart, box and whisker chart

Description automatically generated**

**Figure 8:** Joint Plot for Overall Ratings

**Interpretation and Conclusion**

**Text

Description automatically generated with low confidence**

**Figure 9:** Types of Recommender Systems

One of the major challenges I faced while running the recommender system was setting up the KNN design it took a lot of time while trying its best to suggest the nearest neighbour. The entire process took more than one hour, and I had to change the approach to the algorithm completely. The complexity of the KNN algorithm is O(ndk), where n is the number of users, d is the number of items, and k is the number of considered neighbours. We may make use of the sparsity of the interaction matrix when developing our algorithm or utilize approximate closest neighbours approaches to make calculations more tractable for large systems. Furthermore, another issue which we faced was that the model started recommending only the most popular items, thereby the sale of the other products which may have been the nearest match was not picked up. I am yet to fix this issue and I have not been able to generate a more sophisticated model yet. Model-based collaborative techniques only employ information about user-item interactions and presume a predictive model to explain them. Matrix factorisation techniques, for example, decompose the large and sparse user-item interaction matrix into two smaller and dense matrices.

Furthermore, the file was available to us in JSON format, however, we have converted it into CSV format for ease of usage. Due to the conversion, the live or the URL based links have become suppressed, so we could not delve further into analysis with the KNN based approach. I failed to load the KNN based approach with live images since I had 80,000 recommendations. The time for recommendation analysis increased beyond one hour, leading to the termination of the codes. We have not discussed hybrid recommendation systems here as well.

Last but not the least although some traditional metrics such as MSE, accuracy, recall, or precision can be used, some desirable properties such as uniqueness and modifiability cannot be assessed this way; real-world evaluation such as A/B testing or sample testing is the only true way to evaluate a new recommender system, but it requires a certain sense of credibility in the model. The user-user technique is based on finding users who have had comparable experiences with goods. Because most users only interact with a few things, the approach is extremely sensitive to any recorded interactions (high variance). However, because the final suggestion is based solely on interactions recorded for users who are like our target user, we get more tailored results (low bias).

The item-item strategy, on the other hand, is based on a search for related objects in terms of user-item interactions. Because many people have engaged with an item in general, the neighbourhood search is less sensitive to single interactions (lower variance). In contrast, interactions from all types of users (even extremely diverse users)

**References**

Rocca, B. (2021, December 10). *Introduction to recommender systems - Towards Data Science*. Medium. <https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada>

*ImportError in importing from sklearn: cannot import name check\_build*. (2013, March 7). Stack Overflow. <https://stackoverflow.com/questions/15274696/importerror-in-importing-from-sklearn-cannot-import-name-check-build>