Ecommerce Product Recommendation System Project Update

Arpita Ghosh,Reg No: 2017331008
Farjana Sultana,Reg No: 2017331030

October 29, 2022

For the eCommerce product recommendation system, we have found mainly two types of data: rating-based and review-based. The paper by Rezaei[1], used Amazon review dataset where customers who made purchases on Amazon provide reviews by rating the product from 1 to 5 stars and sharing a text summary of their experience and opinion of the product. It conducted traditional models along with proposed deep neural network (DNN) architecture to predict the reviews' rating score. And It's implemented recommender system with DNN provides the lowest MSE score(0.51-0.53). Also This paper by Dwivedi et al. [2] worked on the ratings based data from Amazon for electronic products and run user based nearest-neighbour collaborative filtering approach to recommend users 5 top products. The performance evaluation provide RMSE score 0.0027 and MAE score(0.0019) which are significantly low enough. In this paper by Balush et al. [3], it has used different approach for product filtering like Collaborative filtering, Content-based filtering, Hybrid filtration to be able to optimize the search engine with the possibilities of recommendations. We are more interested for the literature review for the betterment of our solution. We have already collected those datasets assosiated with the mentioned papers. We will try to implement the hybrid of those mentioned models for improving the accuracy or apart from those approaches, we can come up with some novel approaches.

To facilitate user with more relevent recommendations, we proceed with a particular type of dataset from amazon review data. We choose 5 core Kindle Store data where each of the users and items have at least 5 reviews. The dataset contains 2,222,983 samples and 12 features. To find the corresponding titles of the books, we use the metadata along with it.

Methods: To add the title of the book in the main data, we have merged the metadata with it based on book id. From the rating distribution, we noticed that the dataset is biased with good reviews. The timeframe analysis shows that the dataset contains mostly recent reviews. It helps us to build more accurate model for future. Sorting the reviewers and books according to the average rating doesn't provide true insights. Because there may be many users who rated 5 books as 5 resulting in average of 5 and many books with only 5 reviews getting 5 ratings resulting in the same. So we sort the data according to the highest reviewers and highly reviewed books.
After this, we perform the sentiment analysis on highly reviewed books. We classify the reviews according to the ratings and classify them as positive, negative, and neutral sentiments. In the preprocessing stage, we convert the reviews in lower case and remove punctuation and stop words. Then, we generate the word clouds for the three sentiments. For positive reviews, 'Good', 'book', 'love', 'western', 'good' words are highlighted. 'Read', 'time', 'kindle', 'supposed' are highlighted words for negative. For neutral reviews, 'events', 'book', 'western', 'best' look like more characteristic than others. To get more insights about a particular sentiment, we proceed with frequency analysis of the reviews. The analysis shows that 'book' and 'story' are most common words in the three classes of sentiments. 'read' appears more frequently for the positive and neutral sentiments and 'like' for the negatives. Most probably, the reason is that computer counts every word and removes stop words such as 'did not'. If the most common words are 'did not like' for negative, it is counted as 'like'.

Results: We vectorize the review text and apply different models like logistic regression, decision tree, and random forest classifier to better classify the reviews into positive and negative sentiments. The random forest classifier performs best with 99% train accuracy and 85% test accuracy.
The recommendation system is built using collaborative filtering. To perform the filtering, we take samples with reviewers and books having greater than 150 reviews. The Singular Value Decomposition (SVD) is applied to factorize the reviewer-book matrix. The dot product of the decomposed matrices gives the predicted rating of each book by each reviewer. Then, we recommend user with top ten books based on recommendation strength.

Future Scope: In this project, we have built a recommendation system for a particular genre. But, in future, we wish to expand this for different types of ecommerce products. For the further improvement of the sentiment analysis, the models applied can be tuned to get better results. And, we would also like to apply deep learning models with Neural Networks.

References

[1] Mohammad R Rezaei. Amazon product recommender system. *arXiv preprint arXiv:2102.04238*, 2021.

[2] Rohit Dwivedi, Abhineet Anand, Prashant Johri, Arpit Banerji, and NK Gaur. Product based recommendation system on amazon data.

[3] Illia Balush, Victoria Vysotska, and Solomiia Albota. Recommendation system development based on intelligent search nlp and machine learning methods. In *CEUR Workshop Proceedings*, volume 2917, pages 584–617, 2021.