Product recommendation based on regular cart/basket purchases in a product line category.

Deepanshu uppal

Business Analytics and Big Data

Thapar University

***Abstract —* Recommendation systems are an essential component of machine learning algorithms that make relevant recommendations to users based on their requests. Machine learning techniques have sparked a surge of interest in recommender systems. Because there are so many implicit and explicit features that may be utilised to estimate user preference, scalable and accurate algorithms are required, as well as a system that is highly available and scalable. As the demand for e-commerce websites grows, so does the amount of information accessible, making it harder for consumers to discover relevant information that matches their interests. This challenge is solved by recommender systems, which search through a massive amount of dynamically created data to present users with tailored content and services. To see this solution and to compare between different machine learning models we have implemented the recommendation system, through a variety of machine learning approaches. Content-Based Filtering, Hybrid Content-Collaborative Based Filtering and Collaborative Based Filtering are among these strategies.**

***Index Terms—*CF, SPA, CBF, Classification, Machine Learning, Product recommendation model**

1. **INTRODUCTION**

Consumer purchasing habits alter on a daily basis. In this age of information overload, personalised recommendation systems for online products and services are becoming increasingly important. Such systems allow consumers to locate precisely what they want without wasting time, and sellers to present purchasers with the things they are most likely to buy, benefiting both sides. Many online stores now provide recommendations as a standard feature. Amazon has built it to recommend products, YouTube utilizes it to figure out which video to show next in the autoplay, and Facebook leverages it to recommend sites to like and people to connect. There are other significant recommender systems for domains such as restaurants, movies, and online dating. Recommender systems have been used to investigate research articles, experts, collaborators, and financial services.

Shopping is a human culture's way of life. It is something that is consistent in different parts of the world. Consumer-generated content is becoming increasingly crucial in supporting user purchasing decisions as online businesses embrace the realm of the social web.

***A. About Problem***

Product recommendation, as we all know, has been there for a very long time, and as time has gone, it has gotten increasingly difficult to produce new and unique pieces of the model since someone else may have previously done something that resembles or is similar to that in the past. Due to an inefficient filtering method, product mismatch is a common occurrence in several categories. The main issue in the market is time to investigate accessories and complementary products/sub-products for the bought or searched goods, the issue of compatibility with this product, the need for a low or best price for the same, and many times the lack of knowledge, information, and awareness about such items. This product recommendation model solves all of these problems. A well-designed recommendation system will aid businesses in improving their website's shopper experience, resulting in increased client acquisition and retention. This paper also focuses on the question of what if there were different datasets? i,e categorical and numerical) what different models can be deployed then and which one then in turn would yield the best result.

1. ***Motivation***

Product purchase is subjective; if some individuals buy a product with another, it is more probable that others will buy it based on reviews and ratings. Each individual does not necessarily have his or her own preferences and likes; instead, they rely on suggestions and feedback, which is akin to word-of-mouth marketing; hence, the key reason we picked this idea was for the potential to build and deliver a fantastic product recommendation experience for different users based on their preferences and views. Ecommerce services such as Amazon, Flipkart, and eBay have recommendation systems that recommend new products based on the products you see or like, whereas here we are taking a similar approach where instead of recommending just the complementary product, we are generating a basket through our model inspired by the purchase history and low-price strategy, resulting in an entirely new experience for the user. This also serves as a check for the consumer; if there are a good number and variety of accessories and attachments available for the sought product, he or she will buy it; otherwise, he or she will go on to the alternative/substitute.

1. **RELATED WORK**

There are a variety of models proposed about product recommendation through different approaches and using different techniques of machine learning.

**Zeinab et al. [1],** proposes a technique of content-based filtering which requires users to click information and user profile to recommend items to the user. XGBoost is used as a classifier and for the prediction process.

**Keerthana et al. [2]** suggest that we first divide the user transaction data into 2 parts: part one is considered as purchase history and is used by the conditional probability model to make recommendations, part two is considered as recommendation ground truth. Different algorithms are used so that the second part only has 1 transaction. This aims to predict category of the users next purchase or recommend deals that are most likely to be accepted.

**Kim et al. [3]** focus on recent purchase patterns of customers by applying sliding – window (PRMSW) scheme and integrating – window (PRMIW) scheme. It also elaborates on different data mining methods out of which decision tree using bagging technique shows the most promise. It was also concluded that PMRIW achieves better results than PMRSW. **Choi et al. [4]** proposes a new collaborative filtering approach which uses implicit rating information which can be computed by making use of transactions rather than using explicit rating information. Also, CF and SPA (Sequential Pattern Analysis) were integrated to make HOPE (Hybrid Online Product rEcommendation) system to achieve better recommendation models.

**Chen et al. [5]** talks about a product recommender system using Apache Spark and its ALS algorithm. The program is developed over Amazon Web Service (AWS) to enhance the scalability of the algorithm. Results were calculated by the root mean square error (RMSE) method and gave a good output.

**Soma et al. [6]**, in this proposed work an e-commerce based online recommendation system has been implemented for customers by clustering different apparel products based on their brands, size, price and income of customers.

**Michael et al. [7],** a new Classification-based approach to the recommendation of helpful product reviews. Various classifiers has been considered and their performance was examined in terms of accuracy, test and training size, feature selection etc. A.B.M.

**Fahim et al [8]**, studied the effect of recommender system when integrated with eye tracking. This method can significantly help users in decision making while shopping online.

**Sanket et al. [9]**, this investigation proposes a novel pam identification scheme in particular SpamDup based on techniques of ranking. SpamDup can sort out the importance of each feature without a prepared range

Table 1 Summary of related work

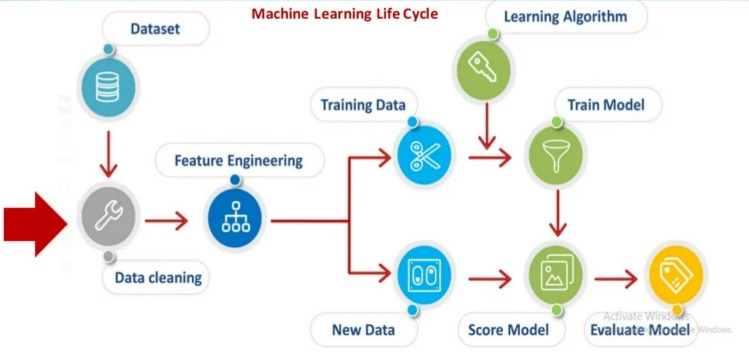
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| References | Technique Used | Dataset Used | Details | Performance measure used |
| Zeinab et al. [1] | Hybrid of the contest - based filtering and collaborative filtering | Jeju online shopping mall records | Uses Content – based filtering and XGBoost for their recommendation model. | Mean average Error  Mean zero – one test error |
| Keerthana et al. [2] | Memory based filtering and ‘model-based filtering |  | Using recommendation techniques, we can find the frequent products purchased by the customers and suggest products to customers or users | Condition probability method |
| Kim et al. [3] | User – based collaborative filtering  Item – Based filtering | Datasets are created by using multiple company data which includes (customer info, purchase, product, competitor, and channel) | Dataset is divided into 2 parts and then the filtering is done. | Average Success rate  Deviation from success rate using bagging techniques |
| K. Choi et al. [4] | Collaborative Filtering and Sequential Pattern Analysis | Online  Shopping mall dataset Korea | They used CF and SPA techniques for hybrid recommendation and tested them on the shopping mall dataset. | Precision, recall and F1 |
| L. Chen et al. [5] | Altering least squares factorization method | Spark core’s basic spark components including (RDDs) | Uses Apache Spark developed over AWS to propose a product recommender model | RMSE (root mean squared error) |
| Soma et al.[6] | K-means and PCA | Dataset Self-Made through google forma | Use of One Hot Encoding in case of PCA to make clusters from categorical data along with the use of k-means clustering. | Elbow method. |
| Michael et al. [7] | Naïve Bayes and Decision Tree algorithm | Dataset comprises of reviews of Chicago hotels gathered from TripAdvisor Service. | Use of Classification as an approach to achieve Recommendation. | -- |
| A.B.M. Fahim et al [8], | Content Based and collaborative based | Data is taken from the user itself through eye tracker | Use of eye gazer module along with collaborative filtering to assist the user in their decision-making process | Unsupervised k-means clustering |
| Sanket et al. [9] | Fake information Network(FIN) classification and weighted algorithm | Amazon review database | Use FIN to achieve classification and remove the least like products from the model | Weighted calculation |

1. **PROPOSED METHODOLOGY**

In the current context, it has been noticed that e-commerce product segmentation and product recommendation to buyers are difficult and complex jobs, and its dominance in the researching space is expanding rapidly. The database of various products/items purchased by clients is taken into account in this job.

1. **Procedures and Methods**

We first started with searching the dataset or data gathering for the problem identified. This step included: Identifying various data sources like Kaggle, UCI, UCSD repository; then Collecting data and lastly, integrating the data obtained from different sources. Next, we performed Data Pre-processing to convert the raw data into a clean data set and performed operations to fix the issue of missing values, duplicity or redundancy, invalid data and noise in the data set by first importing all the required libraries from loading the dataset to performing manipulation and finding and handling the missing values, encoding the categorical data, also performing feature scaling when and where needed to do the Explanatory Data Analysis with visualization to get meaning full conclusions and insights from the data. In EDA, we checked the head, tail, shape, info, describe, null values, unique ones and so on, followed by use of graphs and charts like bar, histogram, scatterplot, heatmap, pair plot, line plot, and others.



*Figure: Steps in Machine Learning*

For the Online Retail dataset, we followed the below steps:

1. Loading the Libraries/modules
2. Importing Data
3. Finding and handling the missing data
4. Extraction of data
5. Handling the categorical data
6. Split the dataset
7. Feature scaling
8. Visualization
9. Creating the Recommender/Model
10. Run Recommender systems/Models
11. Analyze Recommendations/Prediction
12. End

For the Electronics (Amazon Rating) dataset, we followed the below steps:

1. Loading the Libraries/modules
2. Importing Data
3. Finding and handling the missing data
4. Visualization
5. Creating the Recommender/Model
6. Run Recommender systems/Models
7. Analyze Recommendations/Prediction
8. End
9. **Proposed model and implementation**

As our both the datasets are having independent variables, we had to perform the unsupervised learning techniques in which in which a machine learns without any supervision or when the leaning involves training by using unlabeled data and allowing the model to act on that information without guidance. So, we need to allow model to work on its own to discover information and patterns. Here are the models we used for the analysis of the data and to make a recommender system:

* 1. **KNN (K Nearest Neighbor)**

For classification and regression, the K Nearest Neighbor approach is a form of supervised learning methodology and a versatile method that may be used to fill in blanks and resample datasets. As the name suggests, KNN evaluates the neighbors to estimate the class or continuous value for a new point.

To explain how K-NN works, use the following method:

Step 1: Determine the number of neighbours (K).

Step 2: Calculate the Euclidean distance between K neighbouring points.

Step 3: Determine the K nearest neighbours using the calculated Euclidean distance.

Step 4: Among these k neighbours, count the number of data points in each category.

Step 5: Assign the newly acquired data points to the category with the most neighbours.

Step 6: We've finished our model.

* 1. **Logistic regression**

Another strong supervised machine learning approach for binary classification issues is logistic regression (when target is categorical). The easiest way to think of logistic regression is as a type of linear regression that is used to solve classification difficulties. The logistic function defined below is used to represent a binary output variable in logistic regression.



* 1. **Correlation Matrix**

The correlation matrix is a matrix that displays how variables are related. In a matrix format, it shows the correlation between all possible pairs of values.

A correlation matrix is used to summarise a huge dataset, find trends, and make decisions based on them. We can also determine which variables are more associated with each other and visualise our findings.

Correlation Matrix with Truncated SVD was used: Singular Value Decomposition (SVD), a linear algebra approach, is becoming increasingly prominent in data science and machine learning. Its prominence arises from its usage in the development of recommender systems. There are several online user-centric programmes, such as video players, music players, e-commerce applications, and so on, that provide users with additional stuff to engage with.

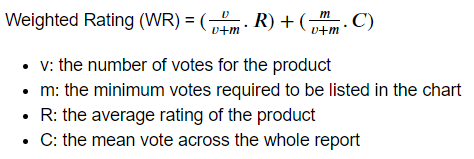
It's a type of matrix factorization that's comparable to PCA (principal component analysis). On the data matrix, we utilise Truncated SVD or any SVD, and on the covariance matrix, we use PCA.

The number of columns in a truncated SVD factorised data matrix is equal to the truncation. It mathematically shortens the value of float digits by dropping the digits following the decimal place.

* 1. **Popularity-Based Recommender (Weighted)**

It provides users with broad product suggestions based on product popularity. This approach gives all users the same product recommendations and does not provide individualised recommendations.

We used Popularity-Based Recommender to recommend products to the new customers as per the maximum number of best ratings received by a product.



* 1. **Content-Based Recommender**

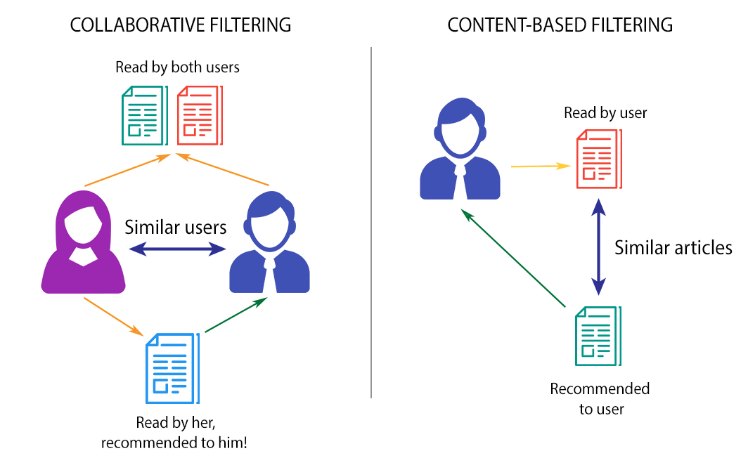
It is a different form of recommendation system that is based on related material. If a user is viewing a movie, the system will look for additional films with comparable material or that are in the same genre as the one they are watching. When screening for comparable material, a variety of essential criteria are employed to compute similarity.

It creates an engine that calculates product similarity based on criteria like description and proposes items that are most similar to a product that a consumer likes. The basic premise underlying these recommender systems is that if a person likes one item, he or she will also enjoy another similar.

* 1. **Collaborative Filtering Recommender**

It is regarded as one of the most intelligent recommender systems, based on the similarities between various users and things such as e-commerce websites and online movie websites. It makes suggestions based on the preferences of comparable users.

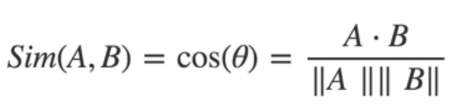
Similarity is not limited to the user's preference; it may also be considered between distinct objects. If we have a big amount of data on people and things, the algorithm will make more efficient recommendations.

**

*Figure: Collaborative and Content-Based Filtering*

This method connects people who share similar interests and makes suggestions based on that connection. Collaborative filtering is based on the premise that users who are similar to a certain user may be used to estimate how much that user would appreciate a product or service that those users have used/experienced but that user has not.

1. **Item-based nearest-neighbor collaborative filtering:**

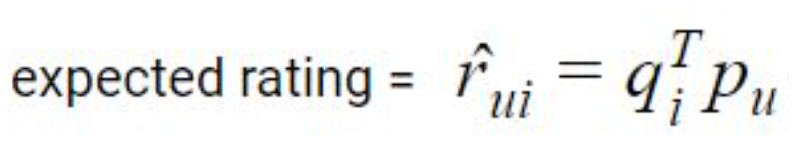
In the Electronic Rating dataset, we employed KNN with Cosine Similarity. We may establish how similar or distinct two sets of objects are using the Cosine function and the K-Nearest Neighbor method, and then use that information to classify them. In high-dimensional space, the Cosine function is used to calculate the similarity or distance between data.

The algorithm looks for things that are comparable to those purchased by the consumer. For the prediction, the similarity between distinct things is determined based on the items rather than the users. Users X and Y both bought goods A and B, indicating that their likes are comparable.

1. **SVD: Matrix Factorization Based Algorithm**

In recommendation systems, singular value decomposition, often known as the SVD algorithm, is employed as a collaborative filtering approach. SVD is a matrix factorization method that reduces the number of features in data by lowering the dimensions from N to K, where (KN) is the number of dimensions.

For the part of the recommendation, the only part which is taken care of is matrix factorization that is done the user-item rating matrix. Matrix-factorization is all about taking 2 matrices whose product is the original matrix. Vectors are used to represent item ‘qi’ and user ‘pu’ such that their dot product is the expected rating.

**

1. **EXPERIMENTAL RESULT AND ANALYSIS**

The experimentation is conducted on Windows 11 OS with 16 GB RAM and 512 GB SSD Hard disk. We used libraries such as Scikit-learn, Surprise, Matplotlib, Seaborn, Numpy, and Pandas.

We worked on the Online Retail Dataset having data of one year from customers of an online retail store obtained from the University of California Irwin (UCI) repository and Electronics (Ratings) data set of Amazon obtained from the University of California San Diego (UCSD) repository to evaluate the performance of the proposed methodology. In this part, the dataset description and step-by-step analysis and results are explained:

1. **Datasets Description**
2. **Online Retail Dataset**

This is a transnational data collection [11] that covers all transactions for a UK-based and registered non-store internet retailer between December 1, 2010 and December 9, 2011. The firm specializes on selling one-of-a-kind presents for any occasion. Wholesalers make up a large portion of the company's clientele.

The dataset is multivariate, sequential, and time-series in nature, with eight columns or characteristics.

Invoice No is the number of the invoice. Each transaction is given a nominal number, which is a six-digit integral number. This code denotes a cancellation if it begins with the letter 'c.'

Stock Code is the code for the product (item). Each different product is allocated a 5-digit integral number known as a nominal.

Description: The name of the product (item) (Nominal).

Quantity: The number of units of each product (item) sold in a single transaction. (Numeric).

Invoice Date: The date and time of the invoice. Each transaction's numeric value, as well as the date and time it was created.Unit Price: Price per unit. Product price per unit in sterling, numeric.

CustomerID is a unique identifier for each customer. Each client is given a 5-digit integral number that is unique to them.

Country. The name of the nation in which each consumer resides is nominal.

1. **Electronics Dataset (Amazon Electronics products Ratings Data Set)**

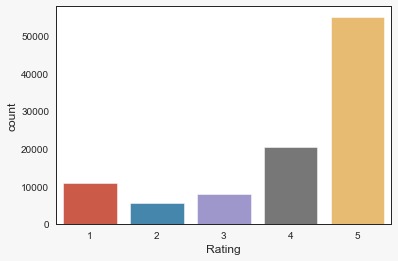
There are various datasets in the Amazon Reviews data source. We are utilising the Electronics dataset [10] for our case study, which includes:

userId: Each user is assigned a unique id.

productId: Each product is assigned a unique id.

Rating: The associated product's rating as determined by the timestamp of the relevant user:

Timestamp: The time of the evaluation (we ignore this column in for the recommender system)

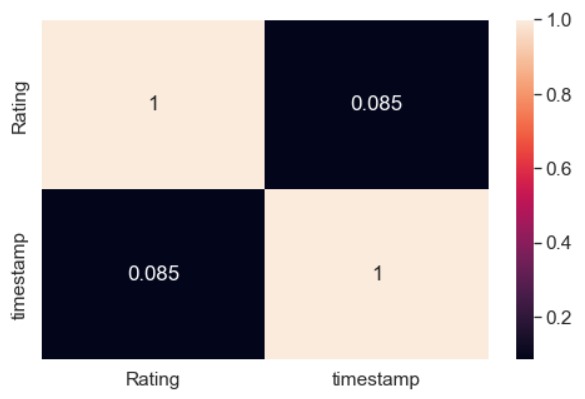


*Figure: Plot between Rating and Votes*

1. **Performance Metrics and Evaluation**



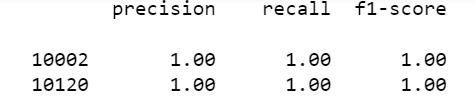
*Figure dataset 1 heatmap*



*Figure dataset 2 heatmap*

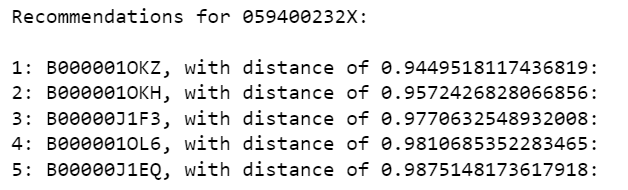
|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| KNN | 93.18% |
| Logistic Regression | 97.96% |

Table 2: Model and Accuracy

****

*Figure* *Logistic Regression Classification Report*

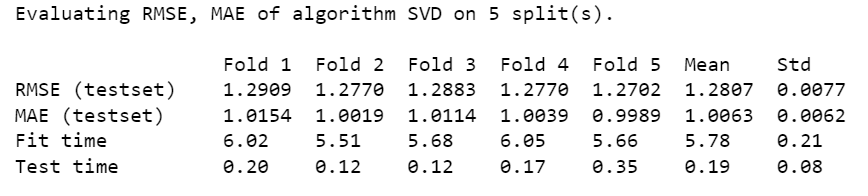
As it is evident from the above data that the logistic regression is more accurate for product recommendation.



*Figure Collaborative Filtering using Cosine function*

The distance is the measure of how accurate the given collaborative filtering model with cosine similarity function is, the closer it is to 1 the better the model is.

The popularity-based model utilizes the voting system i.e., if the a certain product receives highest number of votes based upon the rating it is recommended to the user, which means that the product is currently in high demand amongst the users.



Collaborative filtering using svd model uses the same algorithm as traditional collaborative filtering method using cosine similarity but it also uses matrix factorization to reduce the model’s computational time.

1. **Discussion**

Data preprocessing is done before and model is deployed on the dataset so that it doesn’t face any issues while running the model.

For the Online Retail dataset, KNN model, Logistic Regression and correlation model were used.

KNN model, it is a model that classifies data points based on the points that are most similar to it. Dataset was divided in 80:20 ratio of training and test dataset respectively and the KNN model was deployed on test dataset to predict the stock code of the product.

Logistic Regression is a classification model used when the data is not linear in nature, it uses a sigmoid function to predict the required data.

Correlation Model uses preliminary the group-by function to turn the datasets into subsets and correlates with each value and give recommendations on the basis of strongest correlational values. As a generic explanation correlation means the relationship between two variables and the distance between them.

For the Electronics dataset,

Popularity based recommender model is used where new vote count column is added in the dataset on the basis of products rating and the most popular products are displayed. For we should:

- Create a metric to score or rate the products.

- Calculate the score for every product.

- Sort the scores and recommend the best rated product to the users.

we used IMDB's weighted rating formula to score the products, as follows:

Weighted Rating (WR) = (𝑣.*R*/(𝑣+𝑚))+(𝑚.𝐶 /(𝑣+𝑚))

Collaborative Filtering Method, here our approach is to find similar products by using the ratings provided by the users. It uses cosine metrics with brute algorithm to give the required recommendations, cosine similarity uses distance between two vectors of an inner product space.

Collaborative Filtering Method (using Surprise library or SVD Method), it reduces the number of features the model has to generally go through in order to recommend a product and thus reducing the computational time, it uses matrix factorization method to achieve this

1. **CONCLUSION AND FUTURE WORK**

We utilized and constructed a variety of algorithm-based recommendation systems. The following are the details:

First, we do recommendation of product as per description for a new customer visiting the ecommerce.

We tried – 1. KNN and 2. Logistic Regression.

Our conclusion, LR is better with 97% accuracy.

Second, we did Content based recommendation system using Correlation Model from input of Product description to help the visiting customer with some suggestions.

Correlation: This model is used to find the most suitable product by using the correlation matrix which shows correlation between variables and whichever variables are more closely corelated helps us find the desired result.

Third, we used Item-based recommendation from 2 Popularity based recommenders –

1. A weighted rating method using no. of votes and mean ratings

2. Popularity method of showing highest count rating in desc order.

Recommender Based on Popularity: To determine the top 10 productIds, this algorithm employed the overall Number of Ratings and the VMean Rating. The IMDB Weighted Rating System was utilised to produce the scores that were then used to order the movies.

Our conclusion, Weighted is better.

Lastly, we used Collaborative filtering approach -

1. KNN with Cosine metric using Brute algorithm

2. Correlation Matrix model with Truncated SVD

The above explained model can is a bit faster in terms of total computational time than the traditional collaborative filtering technique. The system has to go through less amount of key features in order to recommend a single or multiple products to the user.

3. Surprise module SVD (stands for Simple Python Recommendation System Engine.)

We built a collaborative filter based on single value decomposition using the powerful Surprise Library. The system projected anticipated ratings for a particular user and product with an RMSE of around 1.28.

Our conclusion, Surprise SVD is better.

**REFERENCES**

1. Zeinab Shahbazi, Yung-Cheol Byun\* (2020). Product Recommendation Based on Content-based Filtering Using XGBoost Classifier. International Journal of Advanced Science and Technology. 6979 – 6988
2. T. Keerthana1 ,T. Bhavani2 , N. Suma Priya3 , V. Sai Prathyusha4 , K.Santhi Sri5 (2020). FLIPKART PRODUCT RECOMMENDATION SYSTEM. ISSN NO:0377-9254
3. Haein Kim, Geunho Yang, Hosang Jung, Sang Ho Lee, Jae Joon Ahn(2018). An Intelligent Product Recommendation Model to Reflect the Recent Purchasing Patterns of Customers 163–170
4. Keunho Choi, Donghee Yoo, Gunwoo Kim, Yongmoo Suh. (2012). A hybrid online-product recommendation system: Combining implicitrating-based collaborative filtering and sequential pattern analysis. ECRA, 11, 309-317.
5. Lin Chen, Rui Li, Yige Liu, Ruixuan Zhang, Diane Myung-kyung Woodbridge. Machine Learning-based Product Recommendation using Apache Spark. Manning publications. 2016
6. Soma Bandyopadhyay, S. S. Thakur1, J. K. Mandal, “Product recommendation for e‐commerce business by applying principal component analysis (PCA) and K‐means clustering: benefit for the society”, Springer-Verlag London Ltd., part of Springer Nature 2020.
7. Michael P. O’Mahony, P ́adraig Cunningham and Barry Smyth,” An Assessment of Machine Learning Techniques for Review Recommendation Marlin, B. Modeling user rating profiles for collaborative filtering. Advances in Neural Information Processing Systems, 16, 2003, 627–634.
8. A.B.M.Fahim Shahriar, Mahedee Zaman Moon, Hasan Mahmud, Kamrul Hasan, “Online Product Recommendation System by Using Eye Gaze Data”
9. Sanket Hanmant Pawar, Avadhut Prashant Lohar, Omkar Nitin Thopate, Rushikesh Sukhdev Auti, Dr.Y.B.Gurav. “PRODUCT RECOMMENDATION SYSTEM USING MACHINE LEARNING”, || Volume 6 || Issue 4 || April 2021 || ISSN (Online) 2456-0774
10. Image-based recommendations on styles and substitutes J. McAuley, C. Targett, J. Shi, A. van den Hengel SIGIR, 2015
11. Daqing Chen, Sai Liang Sain, and Kun Guo, Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197â€“208, 2012 (Published online before print: 27 August 2012. doi: 10.1057/dbm.2012.17).s