DESIGNING A RECOMMENDATION SYSTEM USING COLLABORATIVE FILTERING

MAHENDRA CHAGAM

PRASAD DUDDUMPUDI

TUSHAR KOTANGLE

SWATHI BHAT

VIDISHA VIJAY

Date: 22-11-2021

GROUP E-SECTION 1

Table of Contents

Executive Summary:

The report deals with the implementation of collaborative and content-based filtering approaches to predict movie ratings of existing and new users. The survey data was split into training and test data for each model. In Part 1, prediction of ratings for the 3 movies that haven't been watched by any group members, user-user Non-Normalized Euclidean distance model turned out to be the best model with the lowest *RMSE*[1]. In Part 2, prediction of the ratings for the 3 specific new movies yields the Item-Item Non-Normalized cosine as the best model based on RMSE. For Part 3 to predict the ratings for new customers is carried out with certain criteria like gender and demographics, wherein User-User mean centered Cosine model outdid the other models based on the RMSE. User-user Non-Normalized Cosine was used in Part 4 with the new information available about the customers. Scope for enhancing the prediction is also discussed in the report. The scarcity and scalability issues limit the accuracy of the models, which can be resolved by Single Value Decomposition and Principal Component Analysis. Content based filtering can be leveraged to improve the prediction. Recommendation systems improve discovery experience for users and have immense potential for revenue generation.

Introduction:

In this report, to predict the movie ratings we used the collaborative and content-based filtering method which are recommendation system implementations. A recommender system anticipates consumers' product ratings and preferences. The major function of recommender systems is to establish a link between users and goods to maximize user-product interaction. The collaborative filtering strategy is based on the notion that if two people have the same opinion on a collection of things, A is more likely to share B's view for a specific item than a randomly picked individual. The analysis is multifold wherein firstly the group member's ratings are predicted followed by prediction of ratings for three specific movies and eventually prediction of ratings for new users.

[1] The root mean square error (RMSE) is the standard deviation of the residuals. The RMSE tells you how clustered the data is around the best-fit line.

A comparative analysis is conducted to determine which predicts the best. An evaluation of whether incorporating new information modifies the priority is carried out.

Problem Formulation:

Using collaborative filtering, we will be using the Cosine similarity to calculate similarity between the preferences of two users for a movie. To predict group members' ratings user-user mean centered and item-item mean centered methods are utilized. Estimation is carried out using the Non-Normal Euclidean and the mean centered Euclidean parameters for predicting group member's ratings for movies that have not been rated previously and for the three specific movies respectively. Collaborative and content-based filtering is also briefly assessed for enhancing prediction accuracy and predicting the rating for the new users with no previous data.

Data Description:

There are two datasets present. The first dataset contains 50 movies out of which 47 movies have been rated by 96 students of UC Davis MSBA gathered through Google forms survey. The ratings are in the integer format on the scale ranging from 1-5, while null values are assigned to rate movies that has not been watched by the student. 3 new customers are also present in the dataset and will be considered to predict their movie ratings. Another DBMI dataset consists of ratings by 184 anonymous users for the same 50 movies as in first dataset.

Model Development:

For predicting the group member's rating, we chose 3 movies not watched by anyone in the group. Those are *Call Me by Your Name*, *The Secret in Their Eyes* and *Three Billboards Outside Ebbing, Missouri*. To start with we took the first dataset and split to 80% into training data and 20% into test data. We identified the best model among various User-User and Item-Item methods using Cosine similarity and Euclidean distance techniques. For part 1, the best model turned out to be the user-user Non-Normalized Euclidean Distance model with the least RMSE.

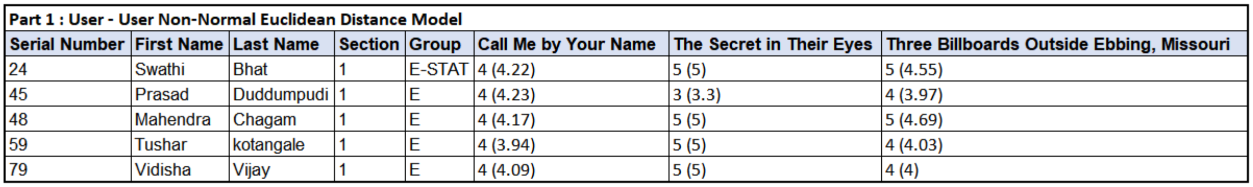
For part 2, to predict the ratings of new movies *Winter's Bone*, *A Serious Man* and *Son of Saul* we leveraged the DBMI dataset in which we have few ratings for the above movies. We selected a limited set of users who rated either one of the above movies from DBMI dataset along with the group's data. Again, we split the data into training and test dataset and the best model turned out to be the item-item Non-Normalized Cosine model based on RMSE. Other ways to predict the ratings was to use content-based filtering using movie characteristics or to rate by associating the popularity of movies and finding the nearest match.

For part 3, to predict the ratings for new users *Camille*, *Shachi* and *Amy* we included new characteristics of gender and demographics into the data. Since, we don't have any information about their past ratings, we added these new characteristics to closely predict their ratings based on similarities with other users in the data. Similarly, we can also leverage their interests and daily activities data in real life to predict the ratings. However, for the scope of the report we took two characteristics which are gender and demographics (Domestic or International) available at our disposal for all the people in the data. Moreover, we can perform content-based filtering based on movie characteristics and link them with the similarities from other user characteristics. But in this part, based on the gender and demographics similarities between new customers and old customers, we extrapolated the movie ratings from the old customers to the new customers. In this part we used user-user mean centered Cosine with least RMSE to predict their rating.

For part 4, since we have new information available about the new users *Camille*, *Shachi* and *Amy.* We used a similar method used in part 2 to come up with the best model of user-user Non-Normalized Cosine with least RMSE. We used only the new ratings available and ignored the movie characteristics. However, other ways to improve the rating prediction is to leverage content-based filtering based on movie and user characteristics. We will discuss all models' accuracy outputs in the appendix.

<u>Results:</u>

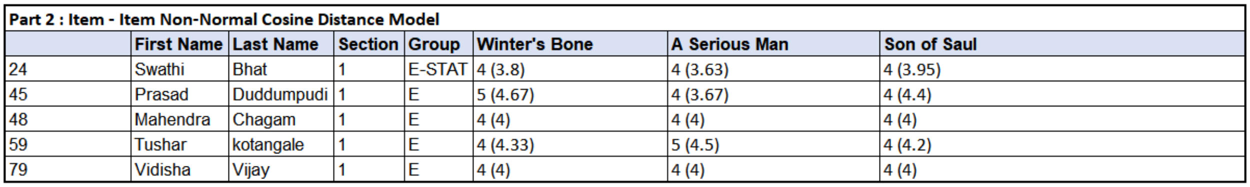
For part 1, using user-user Non-Normalized Euclidean distance model below are the ratings for the 3 movies not watched by anyone in the group. The kindest member turned out to be *Mahendra* and the harshest was *Prasad* in terms of reviewing the movies.



Part 1 : User - User Non-Normal Euclidean Distance Model

Serial Number	First Name	Last Name	Section	Group	Call Me by Your Name	The Secret in Their Eyes	Three Billboards Outside Ebbing, Missouri
24	Swathi	Bhat	1	E-STAT	4 (4.22)	5 (5)	5 (4.55)
45	Prasad	Duddumpudi	1	E	4 (4.23)	3 (3.3)	4 (3.97)
48	Mahendra	Chagam	1	E	4 (4.17)	5 (5)	5 (4.69)
59	Tushar	kotangale	1	E	4 (3.94)	5 (5)	4 (4.03)
79	Vidisha	Vijay	1	E	4 (4.09)	5 (5)	4 (4)

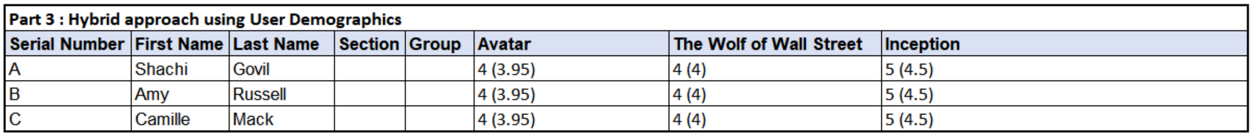
For part 2, using the item-item Non-Normalized Cosine model below are the ratings for the new movies *Winter's Bone, A Serious Man* and *Son of Saul* for the group members:



Part 2 : Item - Item Non-Normal Cosine Distance Model

	First Name	Last Name	Section	Group	Winter's Bone	A Serious Man	Son of Saul
24	Swathi	Bhat	1	E-STAT	4 (3.8)	4 (3.63)	4 (3.95)
45	Prasad	Duddumpudi	1	E	5 (4.67)	4 (3.67)	4 (4.4)
48	Mahendra	Chagam	1	E	4 (4)	4 (4)	4 (4)
59	Tushar	kotangale	1	E	4 (4.33)	5 (4.5)	4 (4.2)
79	Vidisha	Vijay	1	E	4 (4)	4 (4)	4 (4)

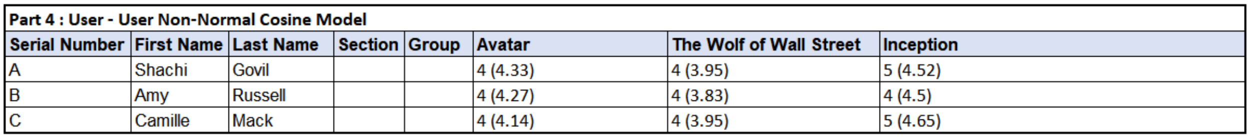
For part 3, using the user-user mean centered Cosine model the ratings for the for new users are:



Part 3 : Hybrid approach using User Demographics

Serial Number	First Name	Last Name	Section	Group	Avatar	The Wolf of Wall Street	Inception
A	Shachi	Govil			4 (3.95)	4 (4)	5 (4.5)
B	Amy	Russell			4 (3.95)	4 (4)	5 (4.5)
C	Camille	Mack			4 (3.95)	4 (4)	5 (4.5)

For part 4, using the user-user Non-Normalized Cosine model below are the ratings for the new users. The ratings are different from what we predicted in part 3 and we believe that these are better in terms of prediction. This is due to new information available leading to better predictions.



Part 4 : User - User Non-Normal Cosine Model

Serial Number	First Name	Last Name	Section	Group	Avatar	The Wolf of Wall Street	Inception
A	Shachi	Govil			4 (4.33)	4 (3.95)	5 (4.52)
B	Amy	Russell			4 (4.27)	4 (3.83)	4 (4.5)
C	Camille	Mack			4 (4.14)	4 (3.95)	5 (4.65)

Recommendations and Managerial Implications:

Recommender systems have wide ranging applications across many domains. Web based companies utilize collaborative filtering to enhance user and discovery experience to increase interaction while also increasing the business potential. Maximization ROI and sales growth are other benefits. It can also shape strategy for companies as the reports generated can help make key decisions to drive revenue growth. Today, consumers are inclined towards more personalized experiences. Building recommender systems using collaborative filtering and content-based filtering can help connect consumers with the right product. A hybrid approach with content and collaborative filtering is more common especially in cases where you don't have enough data to train the recommender system.

The current methods can be improved by solving the scarcity and scalability issue. Scarcity of data points could be a potential problem in effectively rating the product by the customers. Even the active customers rate very few products and the recommender system based on those data points has reduced coverage of recommendations. In such cases, the accuracy of the recommendation system would be poor. Scalability is another problem with the present approach as it is based on distance, which is calculated between each of the customers and products. Singular value decomposition and Principal component analysis are methods which help in reducing the dimensionality of the data. For e-commerce retailers, where we have millions of products listed and thousands of options to choose from, a recommendation system is something which is helps customers finalize on purchase by offering a subset of listings. This improves customer experience and could possibly reduce the bounce rate, exit rate, and improve conversions. The objective function used to refine the algorithm is reducing the options and predicting the scores for unrated/unpurchased items for the users. In customer analytics, recommender systems could be used to upsell and cross sell products to the customers with similar behavior.

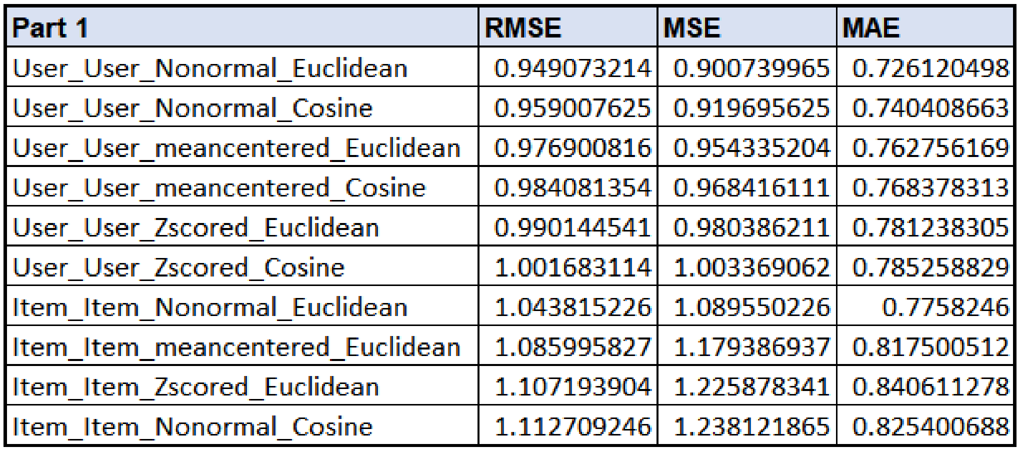
Conclusion:

To summarize, the better we understand the patterns of customers, the better our product or service will be, which can ultimately lead to self-branding and increased ROI. Developing a recommender system based on collaborative filtering by addressing data scarcity and scalability challenges can assist any organization in scaling up through upselling and cross-selling.

Appendix:

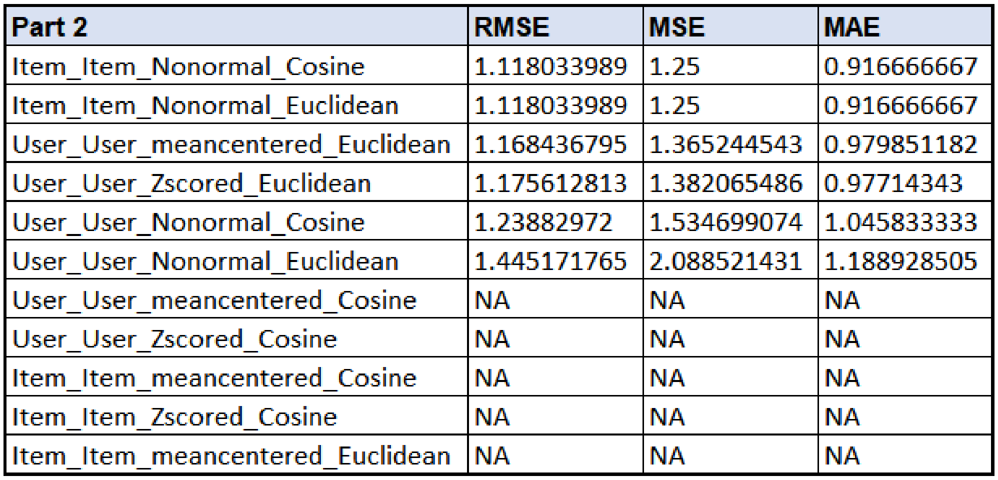
Reducing the options and predicting the scores for unrated/unpurchased items for the users are few of the objective 'value' functions that could be used to refine the recommendation system process. Recommender systems could be used to upsell and cross sell products to the customers with similar behavior which could be used to improve areas in customer analytics.

Part 1: RMSE Scores of all the models



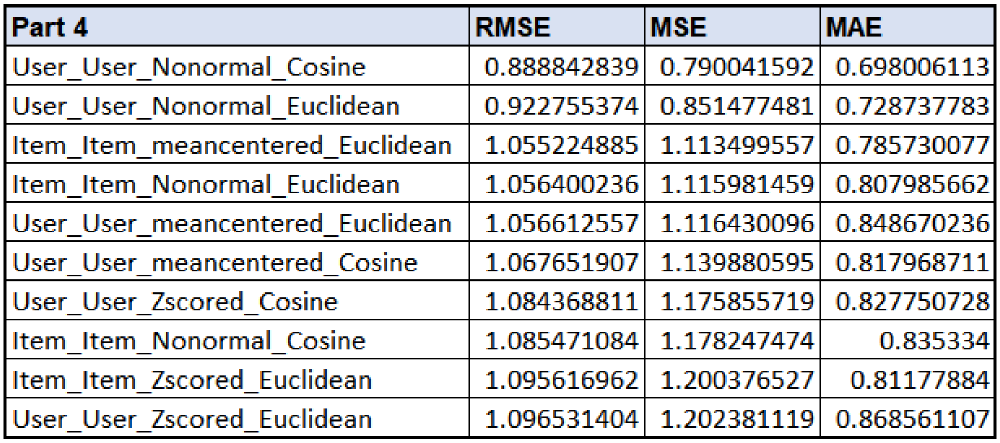
Part 1	RMSE	MSE	MAE
User_User_Nonormal_Euclidean	0.949073214	0.900739965	0.726120498
User_User_Nonormal_Cosine	0.959007625	0.919695625	0.740408663
User_User_meancentered_Euclidean	0.976900816	0.954335204	0.762756169
User_User_meancentered_Cosine	0.984081354	0.968416111	0.768378313
User_User_Zscored_Euclidean	0.990144541	0.980386211	0.781238305
User_User_Zscored_Cosine	1.001683114	1.003369062	0.785258829
Item_Item_Nonormal_Euclidean	1.043815226	1.089550226	0.7758246
Item_Item_meancentered_Euclidean	1.085995827	1.179386937	0.817500512
Item_Item_Zscored_Euclidean	1.107193904	1.225878341	0.840611278
Item_Item_Nonormal_Cosine	1.112709246	1.238121865	0.825400688

Part 2: RMSE Scores of all the models



Part 2	RMSE	MSE	MAE
Item_Item_Nonormal_Cosine	1.118033989	1.25	0.916666667
Item_Item_Nonormal_Euclidean	1.118033989	1.25	0.916666667
User_User_meancentered_Euclidean	1.168436795	1.365244543	0.979851182
User_User_Zscored_Euclidean	1.175612813	1.382065486	0.97714343
User_User_Nonormal_Cosine	1.23882972	1.534699074	1.045833333
User_User_Nonormal_Euclidean	1.445171765	2.088521431	1.188928505
User_User_meancentered_Cosine	NA	NA	NA
User_User_Zscored_Cosine	NA	NA	NA
Item_Item_meancentered_Cosine	NA	NA	NA
Item_Item_Zscored_Cosine	NA	NA	NA
Item_Item_meancentered_Euclidean	NA	NA	NA

Part 4: RMSE Scores of all the models



Part 4	RMSE	MSE	MAE
User_User_Nonormal_Cosine	0.888842839	0.790041592	0.698006113
User_User_Nonormal_Euclidean	0.922755374	0.851477481	0.728737783
Item_Item_meancentered_Euclidean	1.055224885	1.113499557	0.785730077
Item_Item_Nonormal_Euclidean	1.056400236	1.115981459	0.807985662
User_User_meancentered_Euclidean	1.056612557	1.116430096	0.848670236
User_User_meancentered_Cosine	1.067651907	1.139880595	0.817968711
User_User_Zscored_Cosine	1.084368811	1.175855719	0.827750728
Item_Item_Nonormal_Cosine	1.085471084	1.178247474	0.835334
Item_Item_Zscored_Euclidean	1.095616962	1.200376527	0.81177884
User_User_Zscored_Euclidean	1.096531404	1.202381119	0.868561107