Group B Assignment

RECOMMENDATION ENGINES

Daniel R., Tomas F., Albin S., Jorge R., Nicolas G., & Ckalib N
IE UNIVERSITY SCHOOL OF HUMAN SCIENCES AND TECHNOLOGY

Table of Contents

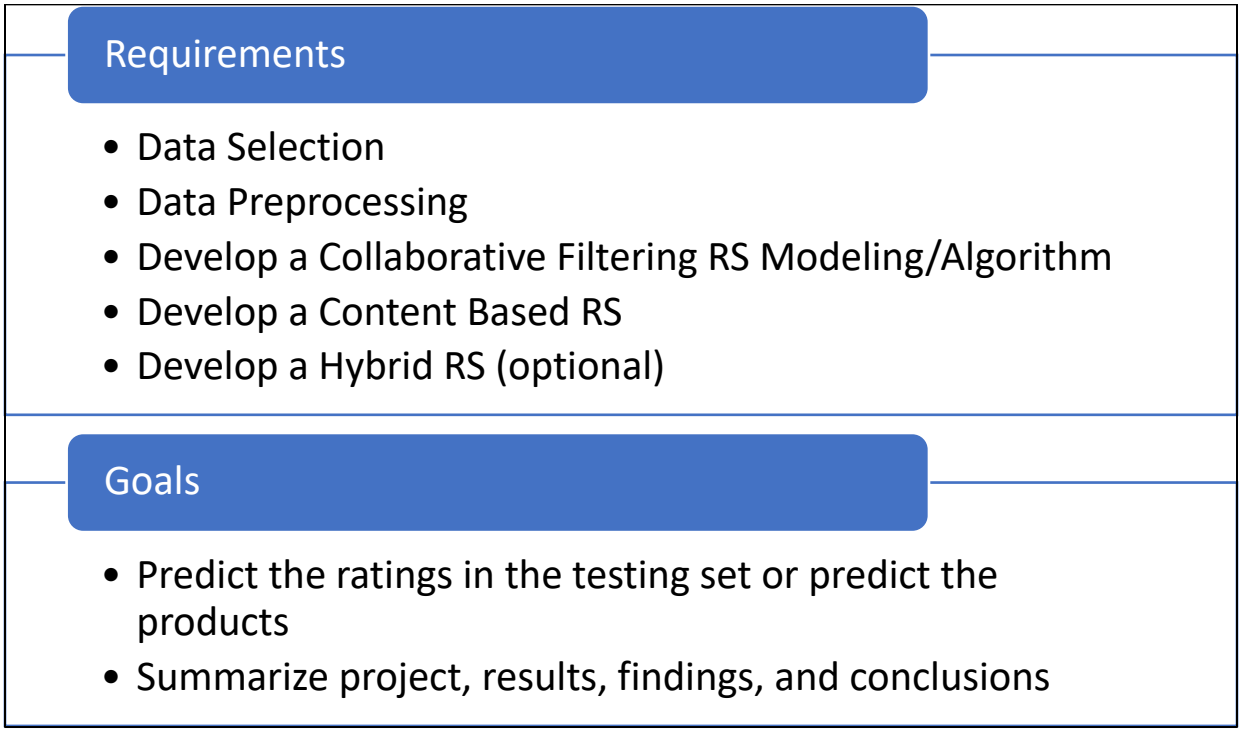
Introduction

For this assignment Group B will build a recommendation system with business impact. From the available 24 datasets, each representing a product domain, we will focus on Luxury & Beauty domain.

Group B used their technical knowledge to train and test the dataset before building a **Collaborative Filtering Recommender System (CFRS)** and a **Content-based Recommender System (CBRS)**. Finally, we also briefly elaborate, how we would approach building a **Hybrid Recommender System (HRS)** combining both CFRS and CBRS. Whereas **CFRS** generally recommends items that similar users like, if and only if other users have similar consumer activity to you, **CBRS** recommends similar items that you've consumed based on their features. **An HRS** uses combination of the recommendation types. This report, built with the *Cross Industry Standard Process for Data Mining (CRISP-DM)* in mind, serves as a summary of our project, results, findings, and conclusions.

Business Understanding

During the Business Understanding Phase, Group B thoroughly identified, analyzed, and understood the business problem to translate it into an analytical problem. Accordingly, requirements and goals of the project were identified and listed below:

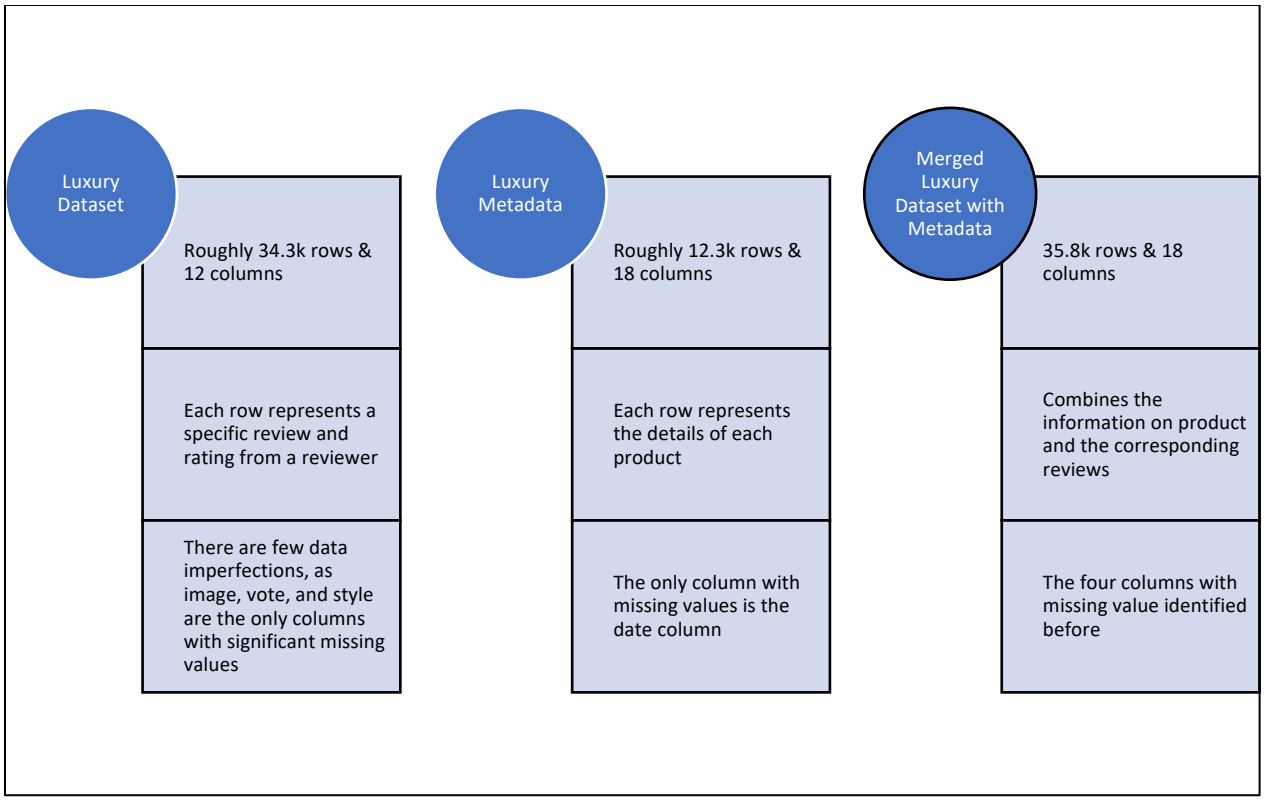


Data Understanding

During the Data Understanding Phase, Group B collected the data to understand and explore it to extract information and insights. After collecting the primary dataset and the metadata, an exploratory data analysis (EDA) was conducted for a thorough understanding of all the features. During the EDA, Group B was able to understand which features could be used to develop all the recommender systems.

Exploratory Data Analysis

After completing the EDA, Group B made the following observations about the dataset:



Based on the summary statistics, we can verify that the main category is indeed "Luxury Beauty" as you expect from the initially chosen dataset. Most reviews (195) were made on the 19th June 2017.

As part of our EDA we also looked at the rating distribution. We can observe that over half of the ratings have a score of 5, which corresponds to the maximum score. As a result, we will have to evaluate recommendations on a higher average score.

Similarly, skewed data can be observed when looking at the distribution number of ratings per product ID (variable *Asin*). From the more than 1,500 unique products, less than 100 received more than 50 ratings. This means there are a small number of products that receive most of the reviews and a large portion of products which have only a few or no reviews at all.

Also, the distribution of number of ratings per user shows a similar picture as a few of the users rated many products while the majority of the users have only rated few or no products at all.

Data Preparation

During the Data Preparation Phase, Group B analyzed the data to decide if any features needed to be removed, created, or amended.

Collaborative Filtering Recommender System (CFRS)

For the Collaborative Filtering Recommender System (CFRS), the variables were not amended in any way. However, only the variables *reviewerID, title, asin,* and *overall* were used.

Content Based Recommender System (CBRS)

For the Content-based Recommender System (CBRS), the price was used to create a new categorization with four levels (*low, lower_medium, upper_medium* and *high*). The reason why we decided to use the price feature instead of another text feature is because there are limited insightful features in this dataset for the purposes of building a CBRS.

	overall	reviewerID	asin	title	price	price_category_low	price_category_lower_medium	price_category_upper_medium	price_category_high
0	5	A2HOI48JK8838M	B00004U9V2	Crabtree & Evelyn - Gardener's Ultra-Moist...	30.0	0	0	1	0
1	5	A2HOI48JK8838M	B00004U9V2	Crabtree & Evelyn - Gardener's Ultra-Moist...	30.0	0	0	1	0
2	5	A1YIPEY7HX73S7	B00004U9V2	Crabtree & Evelyn - Gardener's Ultra-Moist...	30.0	0	0	1	0
3	5	A1YIPEY7HX73S7	B00004U9V2	Crabtree & Evelyn - Gardener's Ultra-Moist...	30.0	0	0	1	0
4	5	A2QCGHIJ2TCLVP	B00004U9V2	Crabtree & Evelyn - Gardener's Ultra-Moist...	30.0	0	0	1	0

The price category is then used to generate the user profiles based on the price category they rated. In a next step, the products most related to the user preference are identified based on the price categories.

Modeling

During the modeling phase, Group B built analytical models to extract knowledge from the data and solve the business problem.

Collaborative Filtering Recommender System (CFRS)

For the CFRS we tried several approaches. The first one was Neighborhood-based Collaborative Filtering. By using the KNN Model following the baseline approach, we got an RMSE of 0.87. This is a good starting point to be used as a baseline.

Then we tried using a random rating based on the training set with the Normal Predictor to compare the results of KNN, and we got an RMSE of 1.3, meaning the KNN Baseline model results are much better and the algorithm is learning from the dataset.

Finally, we tried tunning the parameters with Grid search which gave us a slightly better result of 0.81. The best configuration corresponds to an item-based configuration using Mean Square distance with a min support equals to 3, which can slightly reduce the RMSE to 0.81.

Content Based Recommender System (CBRS)

Regarding the CBRS, by using the overall rating and the price categorization, we can determine a score per product. Essentially, the higher the score, the more likely the user will desire the product. For example, products that a user has not purchased yet shares a similar price categorization as those products the user has purchased, might be part of the recommendation list.

Evaluation

During the Evaluation Phase, Group B evaluated the results to determine if the results addressed the business problem.

Collaborative Filtering Recommender System (CFRS)

Although the baseline approach yielded impressive results, we decided to compare it to several algorithms to determine which one yields the best results:

Algorithm	test_rmse	fit_time	test_time
SVD	0.814982	1.498966	0.089609
SVDpp	0.820181	5.924189	0.332951
KNNBaseline	0.859128	0.725680	3.125812
CoClustering	0.875552	0.970630	0.065424
SlopeOne	0.899861	0.123832	0.285145
NMF	0.911949	1.997482	0.077594
BaselineOnly	0.917363	0.072764	0.061672
NormalPredictor	1.334133	0.043138	0.089276

As you can see, SVD and SVD++ are the most accurate algorithms, however they take longer to train than KNNBaseline. Given how much more familiar Group B is with the KNNBaseline algorithm than the remaining algorithms and the impressive performance of the algorithm, we decided to base the CFRS from the KNNBaseline algorithm.

Content Based Recommender System (CBRS)

The final recommendations per the CBRS do seem to align with our approach of generating user profiles based on price categorization. As we'll see in the conclusion, all of the recommendations in the list have high scores and these scores have a price categorization that the user prefers and a favorable overall rating. Indeed, we should be worried if the highest scores also had a price categorization that was low or lower medium!

Conclusion

In regard to the CFRS, when testing the KNN Baseline algorithm, we returned the 10 nearest neighbors for the product **“Crabtree & Evelyn - Gardener's Ultra-Moisturizing Hand Therapy Pump - 250g/8.8 OZ”** and got the following results:

asin	title
B0007WX0EY	Mustela Gentle Shampoo, Tear Free Baby Shampoo with Natural Avocado Perseose, Gently Cleanses and Detangles Kids' Hair, Available in 6.76 and 16.9 fl. oz
B000GAO1PE	MenScience Androceuticals Advanced Deodorant, 2.6 oz.
B000MEKG30	Marvis Whitening Mint Toothpaste
B000PHTC20	EltaMD UV Sport Sunscreen Broad-Spectrum SPF 50, 3.0 oz
B000Q39MBY	Crabtree & Evelyn Nail and Cuticle Therapy Gardeners, 0.52 Fl Oz
B00172XBOC	LAFCO New York House & Home Candle
B00FASVFI8	Crabtree & Evelyn Ultra-Moisturising Hand Cream Therapy, Tarocco Orange, Eucalyptus & Sage, 3.5 oz
B00FRERO7G	Crabtree & Evelyn Gardeners Ultra-Moisturising Hand Cream Therapy - 3.5 oz
B00TBJWP86	La Roche-Posay Effaclar BB Blur with SPF 20, 1.01 Fl. Oz.
B01B3QIEPC	SkinMedica HA5 Rejuvenating Hydrator

While some products are similar, others are not; however, we could group most of them in a relaxing therapeutical product category which serves the purpose of the nearest neighbor's approach.

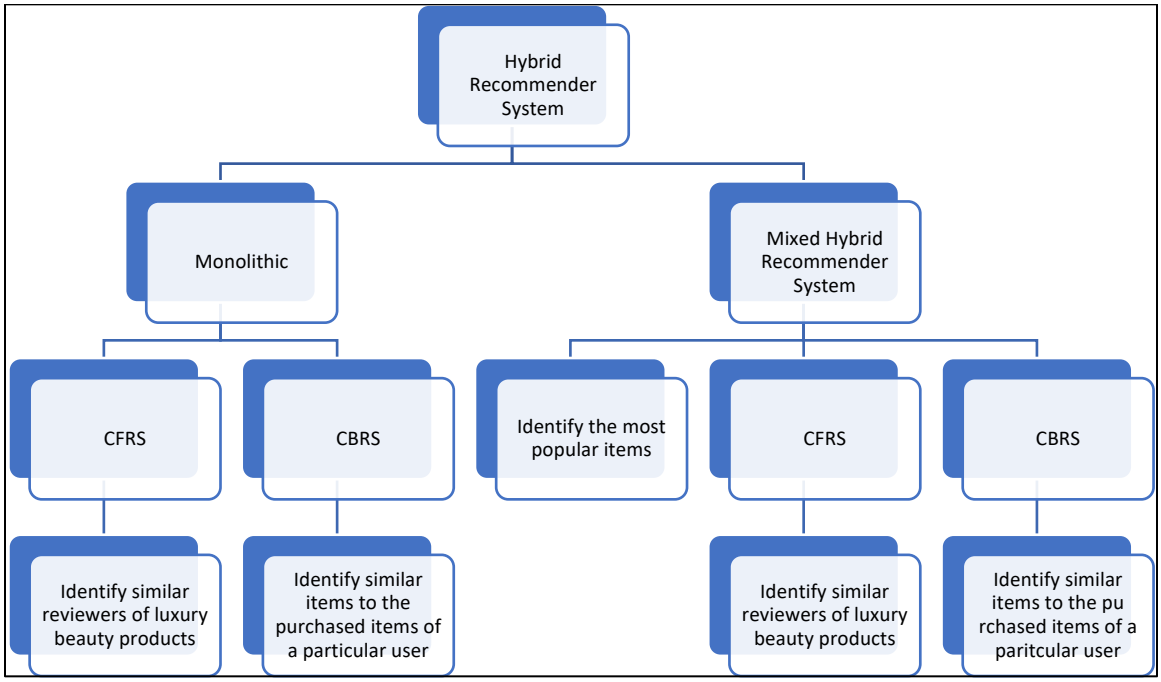
Regarding the CBRS, we returned top ten products with the scores based from their overall rating and price categorization:

	overall	asin	title	price	score
49	5	B00004U9V2	Crabtree & Evelyn - Gardener's Ultra-Moist...	30.000000	0.410256
65	5	B006XCUR5C	Vichy LiftActiv Supreme Anti-Wrinkle Eye Cream...	33.500000	0.410256
438	4	B00G6H88FK	POMMISST Hydration Spray	33.000000	0.410256
432	4	B00FW6ZTW6	Vichy Dermafinish Corrective Full Coverage Con...	28.000000	0.410256
389	5	B00F6XZNLM	Obagi Nu-Derm Gentle Cleanser, 6.7 fl. oz.	35.700001	0.410256
355	5	B00F6XZNGM	Obagi Nu-Derm Toner, 6.7 fl. oz.	35.700001	0.410256
327	5	B00B1IFMH6	Juice Beauty Stem Cellular CC Cream, 1.7 Fl Oz	39.000000	0.410256
320	5	B000IIA5UO	HOT TOOLS Professional 24k Gold Extra-Long Bar...	40.180000	0.410256
216	4	B00AKOU452	LORAC POREfection Mattifying Face Primer, 1.7 ...	33.000000	0.410256
208	5	B00AAR9I60	Obagi Hydrate Facial Moisturizer, 1.7 oz.	42.500000	0.410256
169	5	B00A8JWJRA	Oxygenetix Oxygenating Foundation	39.599998	0.410256
164	5	B00A87CW0G	SkinMedica Facial Cleanser, 6 oz.	38.000000	0.410256
160	4	B009JCVSV0	Juice Beauty Blemish Clearing Peel, 2 fl. oz.	42.000000	0.410256
137	5	B0089AABLQ	WEN by Chaz Dean Sweet Almond Mint Replenishin...	33.000000	0.410256
132	4	B00846IUA4	Agave Healing Oil - Oil Treatment. Hydrating L...	40.000000	0.410256
118	4	B007Y550PO	Vichy Capital Idéal Soleil Sunscreen SP...	28.500000	0.410256
102	5	B007PCB5OQ	Eau Thermale Avène Skin Recovery Cream,...	35.000000	0.410256
85	4	B007EUSL5U	La Roche-Posay Anthelios Ultra-Light Sunscreen...	35.990002	0.410256
77	3	B0073FET84	Davines Vegetarian Miracle Conditioner, 8.77 oz	35.000000	0.410256
496	4	B00G930HW8	Vichy Idéalia Radiance Boosting Antioxi...	39.000000	0.410256

The value of this approach is that an e-commerce site could tailor a users' preferences based off their willingness to pay. Given how expensive luxury beauty products can be, and the various socioeconomic levels of reviewers, the most optimal strategy might be deploying a recommended list to users that are within the price range of the particular users.

Hybrid RS

Although we did not include a hybrid recommender system, here is an idea of the two Hybrid RSs we could build given the knowledge we have about the data and the recommendation systems already built:



It might also be the case that the scores from the CBRS could be leveraged in a CFRS approach. We can imagine that the products with the most favorable scores from the CBRS align with the nearest neighbors suggested by the CFRS.