Università della Svizzera italiana | Institute of Computational Science ICS

Data Analytics **Academic Year 2019/2020**

Assignment 2 - Report
Due date: 20th May 2020, 00:30 PM

Sanchit Chiplunkar

IMPORTANT NOTE: Please keep the .ipynb file in the same folder where the .csv file is kept to succcefully run the code.

Introduction

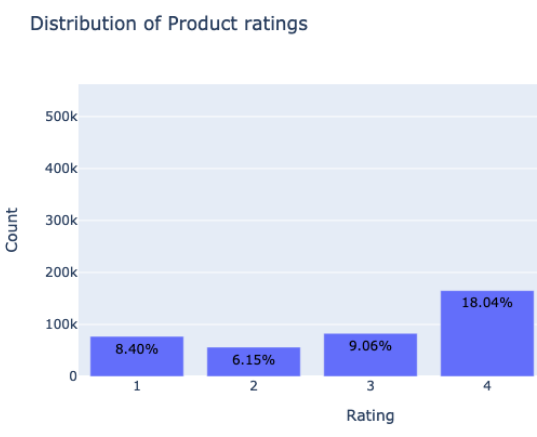
The task in this given assignment is to build a Recommendation System on a user preference data on Baby products.The data consists of 915446 entries and 3 columns which are UserID,ProductID and rating given by the users on that particular product. In the analysis that i have done i have used packages such as **surprise**,**LightFM**, pandas,sklearn,numpy seaborn matplotlib,plotly. The analysis consists of Data Exploration followed by some data removal and finally training of model using Collaborative and Hybrid algorithms and choosing the best one on the basis of RMSE or AUC metric in the test case.

In the next sections we will see some of the plot and figures which gives us more insight about the dataset.

Data Exploration

As can be seen from figure 1 60 % of our users have given 5 ratings to product while not a lot of products are given lower ratings such as 1 or 2.

Figure 2 shows us distribution of number of ratings per baby product. While Figure 3 shows us that most of the products are not rated by users and product with product id B000IDSLOG is the most rated product with 3648 ratings.



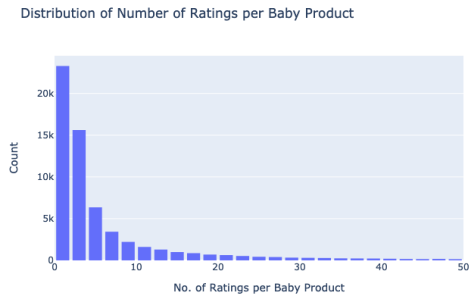


Figure 2. Distribution of Number of Ratings per baby product

Figure 4 and 5 tells us that most of the users don't give rating to the product and user with user id ARIFCL50JD5SK has rated 155 products the most among any user

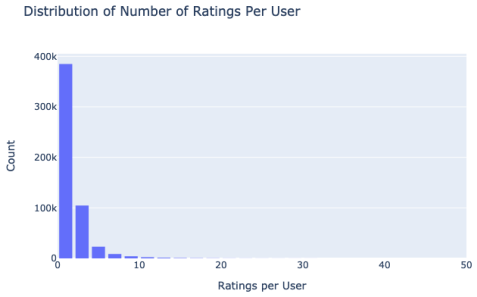


Figure 4. Distribution of Number of Ratings per User

	productId	ratings
5641	B000IDSLOG	3648
2996	B000BNQC58	2923
18267	B00295MQLU	2832
35870	B0052QYLUM	2830
9873	B000YDDF6O	2682
1025	B0000DEW8N	2458
6687	B000LXQVA4	2211
10534	B0011URFRE	2185
618	B0000635WI	2085
15919	B001OC5UMQ	1928

Figure 3. Table of product rated by most users

Preprocessing Steps

As can be seen from the plots above in data exploration we have a lot of products which aren't rated by users and a lot of users which haven't given any rating therefore for the scalability and ease of training of model i decided to remove those productId which have been rated by less than 2 users and those Users who have rated less than 2 products, doing this i reduced my overall dataset from 915446 to 363430.

	userId	ratings
498578	ARIFCL50JD5SK	155
466858	AJGU56YG8G1DQ	140
450551	AF8SREA2XE7BJ	122
466373	AJC88791BZEW7	105
169306	A276OI0NHBYORX	95
241126	A2PNW6QDW8OPY0	93
86966	A1M5ZT35YX6TIN	82
486301	AOEUN9718KVRD	81
526651	AYNNJ0DBGL5H7	81
562	A100L918633LUO	81

Figure 5. Table of users who have given most ratings to product

Model training

I built a recommender system using 2 methods: 1.) The first one is Collaborative Model which was trained using the Surprise package 2.) Secondly i trained a Hybrid model using the LightFM package. A Hybrid recommender system is a special type of recommender system that combines both content and collaborative filtering method and doing so it hides the disadvantages of each individual model and is thus more likely to perform better than either pure collaborative model or content based model.

Collaborative model using Surprise package:
In the first method I trained my model on 5 different algorithm: **SVD,SVDpp, SlopeOne, NMF and NormalPredictor** Using the RMSE as a criteria to judge my model i decided to pick SVD as my final algorithm as can be seen from the table 1.

Following this i split my dataset into trainset and testset using 75% data in the trainset and made prediction on testset. The RMSE on the testset was found to be 1.112. Figure 6 shows some of the predictions made by our model.

	uid	iid	rui	est	details	Iu	Ui	err
0	A2JWGBR1F593II	B003939VGE	5.0	4.824629	{'was_impossible': False}	2	178	0.175371
1	A2TQ9HHZJWLMWF	B00BZOF6T0	5.0	4.462259	{'was_impossible': False}	2	13	0.537741
2	A2KZUHTU589XVH	B004G8QSYO	5.0	4.571455	{'was_impossible': False}	2	208	0.428545
3	A2MIZ6QH53W0D7	B002Y5XJKW	5.0	4.340871	{'was_impossible': False}	12	35	0.659129
4	ALNFHVS3SC4FV	B0002CRS9M	5.0	3.906646	{'was_impossible': False}	4	22	1.093354

Figure 6. Predictions

	uid	iid	rui	est	details	Iu	Ui	err
77086	A2LZL2MTKUFRWF	B004NONY8E	5.0	5.0	{'was_impossible': False}	3	158	0.0
65898	A3BJG56NIQ4Q7U	B001AG0YL8	5.0	5.0	{'was_impossible': False}	19	386	0.0
16043	A2WTP6MMH0OGK5	B001I481LM	5.0	5.0	{'was_impossible': False}	10	227	0.0
79175	A3ESQAU2PEKN8Z	B0001CTZ8K	5.0	5.0	{'was_impossible': False}	6	53	0.0
89151	A1GVPO7N5CNC0I	B006Z6E8AG	5.0	5.0	{'was_impossible': False}	16	258	0.0

Figure 7. Top 5 Predictions

Hybrid model using LightFM package

In the second method as mentioned earlier i trained a Hybrid method using the LightFM package. The metadata json file of the dataset has data about every products such as : 1.) ProductID 2.)Product Name, 3.) Brand Name, 4.) Description of the product,5.) Related Product.
The brand name basically tells about the name of the company/brand which makes the product.

	uid	iid	rui	est	details	Iu	Ui	err
77086	A2LZL2MTKUFRWF	B004NONY8E	5.0	5.0	{'was_impossible': False}	3	158	0.0
65898	A3BJG56NIQ4Q7U	B001AG0YL8	5.0	5.0	{'was_impossible': False}	19	386	0.0
16043	A2WTP6MMH0OGK5	B001I481LM	5.0	5.0	{'was_impossible': False}	10	227	0.0
79175	A3ESQAU2PEKN8Z	B0001CTZ8K	5.0	5.0	{'was_impossible': False}	6	53	0.0
89151	A1GVPO7N5CNC0I	B006Z6E8AG	5.0	5.0	{'was_impossible': False}	16	258	0.0

Figure 8. Worst 5 prediction

The related product metric tells us about what other product were viewed or bought after buying this product. It has multiple other subcategories in it such as 1.) Items viewed after buying this product 2.) Items bought after buying this product 3.) Items bought together, I trained three Hybrid model using two different loss function(so in total 6).
A.) In the first model i trained a pure collaborative model using the LightFM package this model will serve as a baseline as to how much better we can do with a Hybrid model.

B.) In the Second model i made a hybrid model using Brand as an item features, in total there were 1741 unique brand such as Time Too, SoftPlay, Ftbstyle Baby, Naugthy Baby, My Baby Essentials etc.. i trained the mode using WARP and BPR loss function and computed the AUC and top 10 precision values which are mentioned in the Table 2 and Table 3 respectively.

c.) In the third model i made a hybrid model using related metric as an item features trained on WARP and BPR loss function. The AUC and the top 10 precision values for which are shown in the Table 2 and Table 3 respectively.

Results and Conclusion

Table 1. RMSE Error of Collaborative model

Model	RMSE
SVD	1.11
SVDpp	1.12
SlopeOne	1.35
NMF	1.37
NormalPredictod	1.52

As can be seen from Table 1 SVD algorithm give use the lowest RMSE loss on Collaborative model. Figure 7 and 8 shows us the Best and the worst predictions of this model

The reason why the product got such rating is that that majority of the user rated it 5 while in the test case user gave it a rating of 1 which was an outlier which can be confirmed using the distribution plot in Figure 9

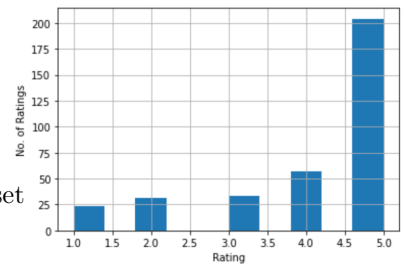


Figure 9. Distribution of the product with the worst prediction

Table 2. AUC values in train and test set

Model/Loss	BPR	Warp
—-	Train|Test	Train|Test
Pure Collaborative	97% | 54%	99% | 75%
Brand	98% | 97%	99% | 99%
Product Bought	95% |94%	95%| 95%

For our Hybrid models trained using the LightFM package we can see from Table 2 and Table 3 that Hybrid models have performed better than the Pure Collaborative model. We get the

Table 3. Precision at 10 values in train and test set

Model/Loss	BPR	Warp
—-	Train|Test	Train|Test
Pure Collaborative	.05 | 0.0	0.14 | 0.0
Brand	.05|.03	.06| .03
Product Bought	.02|.01	.02| .007

Best AUC and precision value for Hybrid model with Brand as an Item feature trained on Warp Loss function.Although looking at table 3 it feels like that the precision value is really low for such high AUC attained in test case but in reality we are computing top 10 precision against >27000 products so the results are still reasonable. I believe that we could have gotten much better result by using combination of Brand and the related metric as an item feature and by tuning the hyperparameters but i wasn't able to do it because of RAM and CPU constraints on my laptop

In Conclusion with the result that we have got i believe that if we want to use a Pure Collaborative model than we should use the one trained on SVD algorithm as it had the lowest RMSE value, on the other hand Hybrid model have performed better and if we want to use that we should use Brands as the feature for it to be trained on as it had really good AUC value and decent precision.

PLEASE NOTE: If you want to run the hybrid model on notebook i have provided two .csv files [brand.csv and short2.csv]. I have also provided the code for generation of both the csv files at the end of the code in the Additional Code subsecttion.