

Amazon Product Recommendation System

Rank-based, Collaborative Filtering, Clustering-based, Matrix-Factorization, Latent Features, Surprise

'Femi Bolarinwa
f3bolarinwa@yahoo.com

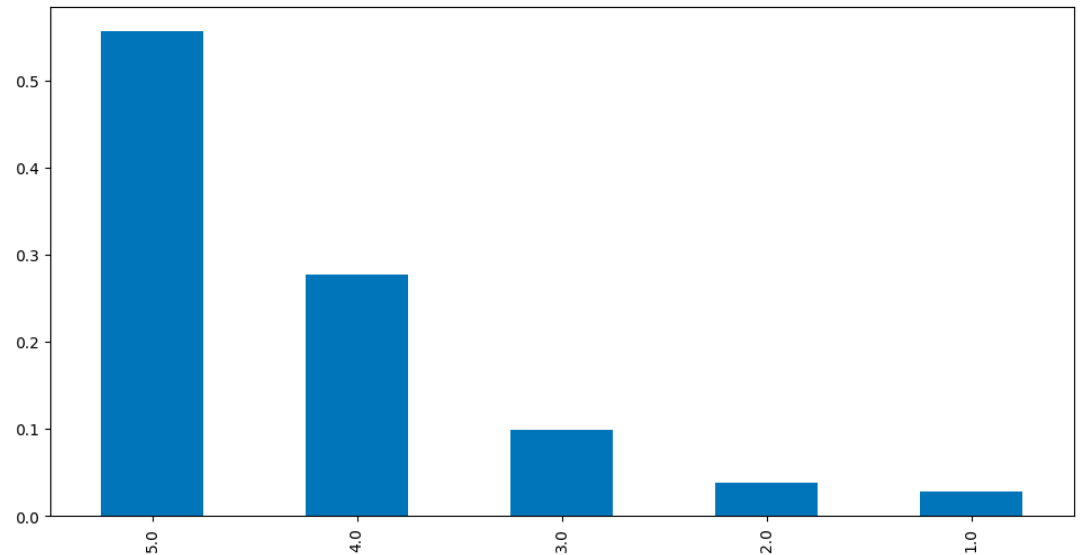
Data Info

Amazon Electronic Products

- 7 million+ datapoints available
- 4 **features** per datapoint: user id, product id, user rating, timestamp
- **Sieved** data down to **users** and **items** with 50 ratings and 5 ratings or more respectively
- 65,290 datapoints left afterwards with **1540** and **5689 unique** users and items respectively
- $(1540 \times 5689 =)$ **8,761,060** possible **user-item interactions**
- Goal is to build recommendation systems/engines for:
 - $(8,761,060 - 65,290 =)$ **8,695,770** unavailable user-item interactions
- Data is split into training set and test set (except for ranked based model)

Distribution of Ratings

- More than 80% of ratings are 5 or 4
- Appears items are generally liked



Model 1

Rank/Popularity Based Recommendation system

- Items are ranked based on there **average rating**
- Ideal for **cold start** for new users who do not have any history

	avg_rating	rating_count
prod_id		
B00LGQ6HL8	5.0	5
B003DZJQQI	5.0	14
B005FDXF2C	5.0	7
B00I6CVPVC	5.0	7
B00B9KOCYA	5.0	8
B003CK10DG	5.0	5
B003B41XYO	5.0	5
B0007WK8KS	5.0	6
B0084FM5JC	5.0	5
B005GI2VMG	5.0	10

Model 1

Rank/Popularity Based Recommendation system

Adjustment of the Ratings and Ranking

- **Rating count** matters when ranking product based on average rating
- Example: a product with 5.0 average rating and 100 rating count is ***not necessarily better*** liked than a product with 4.5 average rating and 1000 rating count
- Empirically, an item's true rating is directly proportional to the inverse of the ***square root*** of the rating count of the item.

	avg_rating	rating_count	corrected_rating
prod_id			
B0052SCU8U	4.979592	49	4.836735
B003ES5ZUU	4.864130	184	4.790409
B000FQ2JLW	5.000000	19	4.770584
B001TH7T2U	4.925000	40	4.766886
B00ISFNSBW	5.000000	18	4.764298
B00BQ4F9ZA	4.911111	45	4.762040
B001TH7GUU	4.871795	78	4.758567
B001TH7GUA	5.000000	17	4.757464
B00HZWJGS8	5.000000	17	4.757464
B000IJY8DS	5.000000	17	4.757464

Model 2

User-user similarity based recommendation system

- Model is based on ***similarity measure*** between every possible pair of users
- Users are then ***clustered*** based on their similarity measure
- ***Average*** rating is then calculated for each cluster
- Top 10 recommendations for user "**A3LDPF5FMB782Z**" is shown

	prod_id	predicted_ratings
0	B000067RT6	5.000000
1	B000BQ7GW8	5.000000
2	B001TH7GUU	5.000000
3	B005ES0YYA	5.000000
4	B00834SJSK	5.000000
5	B00E3W15P0	5.000000
6	B00006RVPW	5.000000
7	B004RORMF6	5.000000
8	B002WE6D44	4.937476
9	B000N99BBC	4.874904

Model 2

User-user similarity based recommendation system

Adjustment of the Ratings and Ranking

- **Rating count** matters when ranking product based on average rating

prod_id	rating_count	predicted_ratings	corrected_ratings
B00834SJSK	101	5.000000	4.900496
B001TH7GUU	78	5.000000	4.886772
B000BQ7GW8	54	5.000000	4.863917
B00E3W15P0	48	5.000000	4.855662
B000067RT6	41	5.000000	4.843826
B002WE6D44	100	4.937476	4.837476
B004RORMF6	34	5.000000	4.828501
B00006RVPW	31	5.000000	4.820395
B005ES0YYA	26	5.000000	4.803884
B000N99BBC	167	4.874904	4.797522

Model 3

Item-item similarity based recommendation system

- Model is based on ***similarity measure*** between every possible pair of item
- Items are then ***clustered*** based on their similarity measure
- ***Average*** rating is then calculated for each cluster
- Top 10 recommendations for user “**A1A5KUIIIHFF4U**” is shown

	prod_id	predicted_ratings
0	1400532655	4.292024
1	1400599997	4.292024
2	9983891212	4.292024
3	B00000DM9W	4.292024
4	B00000J1V5	4.292024
5	B00000JDF5	4.292024
6	B00000JDF6	4.292024
7	B00000K135	4.292024
8	B00000K4KH	4.292024
9	B00001P4XA	4.292024

Model 3

Item-item similarity based recommendation system

Adjustment of the Ratings and Ranking

- **Rating count** matters when ranking product based on average rating

	prod_id	rating_count	predicted_ratings	corrected_ratings
0	B00001P4XA	12	4.292024	4.003349
1	9983891212	8	4.292024	3.938471
2	B00000J1V5	7	4.292024	3.914060
3	B00000JDF5	7	4.292024	3.914060
4	B00000JDF6	7	4.292024	3.914060
5	1400532655	6	4.292024	3.883776
6	1400599997	5	4.292024	3.844810
7	B00000DM9W	5	4.292024	3.844810
8	B00000K135	5	4.292024	3.844810
9	B00000K4KH	5	4.292024	3.844810

Model 5

Matrix Factorization recommendation system

- Model is based on *latent features* of the user-item interaction *matrix*
- *Singular value decomposition* helps to extract the latent features
- Latent features could be product manufacturer, country of origin, etc
- Top 10 recommendations for user “**A3LDPF5FMB782Z**” is shown

	prod_id	predicted_ratings
0	B00006HSYC	4.775916
1	B0010Y414Q	4.765719
2	B00AR95ESC	4.748074
3	B00IVPU6AA	4.734200
4	B002EEP3MK	4.733098
5	B000FQ2JLW	4.732380
6	B0033PRWSW	4.725349
7	B000JV9LUK	4.725264
8	B000F7QRTG	4.718166
9	B004ING996	4.716102

Model 5

Matrix Factorization recommendation system

Adjustment of the Ratings and Ranking

- **Rating count** matters when raking product

prod_id	rating_count	predicted_ratings	corrected_ratings
B000FQ2JLW	19	4.732380	4.502964
B000F7QRTG	20	4.718166	4.494559
B0033PRWSW	15	4.725349	4.467150
B00IVPU6AA	14	4.734200	4.466939
B0010Y414Q	11	4.765719	4.464207
B00006HSYC	8	4.775916	4.422362
B00AR95ESC	9	4.748074	4.414741
B000JV9LUK	10	4.725264	4.409036
B002EEP3MK	9	4.733098	4.399765
B004ING996	6	4.716102	4.307854

Model Evaluation

Performance Metrics on Unseen test data

- **Recall:** fraction of actually relevant items that are recommended to the user,
- **Precision:** fraction of recommended items that are actually relevant
- **f1-score:** harmonic mean of recall and precision
- **RMSE/MAE:** how far the overall predicted ratings are from the actual ratings.

	RMSE	MAE	Precision	Recall	f1-score
User-User similarity-based Model (default)	1.001000	0.709000	0.855000	0.858000	0.856000
User-User similarity-based Model (tuned)	0.953000	0.728000	0.847000	0.894000	0.870000
Item-Item similarity-based Model (default)	0.995000	0.704000	0.838000	0.845000	0.841000
Item-Item similarity-based Model (tuned)	0.958000	0.721000	0.839000	0.880000	0.859000
Bi-clustering Model (default)	0.944000	0.665000	0.853000	0.821000	0.837000
Bi-clustering Model (tuned)	0.942000	0.664000	0.854000	0.823000	0.838000
Matrix-Factorization Model	0.888000	0.653000	0.853000	0.880000	0.866000
Matrix-Factorization Model (tuned)	0.882000	0.656000	0.854000	0.884000	0.869000

Recommendation

- The ultimate goal of a recommendation engine is to make predictions as close as possible to users' true taste. That means minimizing **false positive** and maximizing **true positive** as much as possible and therefore predicting ratings as close as possible to the true value.
- In terms of performance metrics, this means highest **f1-score** and lowest rating **MAE/RMSE** possible on unseen test data.
- For this project work, **User-user** similarity based model and **Matrix Factorization** Model (with tuned parameters) give the best performance. They would be my recommendations for implementation