TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
TRUNG TÂM TIN HỌC

ĐỒ ÁN TỐT NGHIỆP DATA SCIENCE

PROJECT RECOMMENDATION SYSTEM

Nguyễn Nhật Tố Trân vs Nguyễn Vũ Mai Phương

DL07_K302_Apr2025

TABLE OF CONTENTS

01

BUSINESS UNDERSTANDING

BUSINESS UNDERSTANDING

Problem

- Shopee Việt Nam là nền tảng mua sắm trực tuyến với nhiều sản phẩm đa dạng và khuyến mãi hấp dẫn. Giả sử hiện tại Shopee vẫn chưa có Recommendation System.
- Shopee muốn tung ra các chiến dịch quảng cáo, truyền thông, và cũng để bán hàng cho từng đối tượng khách hàng khác nhau một cách hiệu quả.

Solution

- Xây dựng hệ thống gợi ý sản phẩm nhằm giới thiệu đúng sản phẩm đến đúng khách hàng.
- Cá nhân hóa trải nghiệm mua sắm của khách hàng, cũng như thiết kế chương trình ưu đãi hoặc khuyến mãi mới.
- Giới thiệu sản phẩm mới đến đúng đối tượng tiềm năng.

PROCESS

1

EDA

Tiền xử lý dữ liệu và khám phá insights

2

BUILD MODEL

CONTENT-BASED FILTERING

- Gensim
- Cosine Similarity

COLLABORATIVE FILTERING

- ALS (PySpark)
- Surprise (KNNBasic, SVD, KNNBaseline, KNNWithMeans, KNNWithZScore, CoClustering, NMF, BaselineOnly)

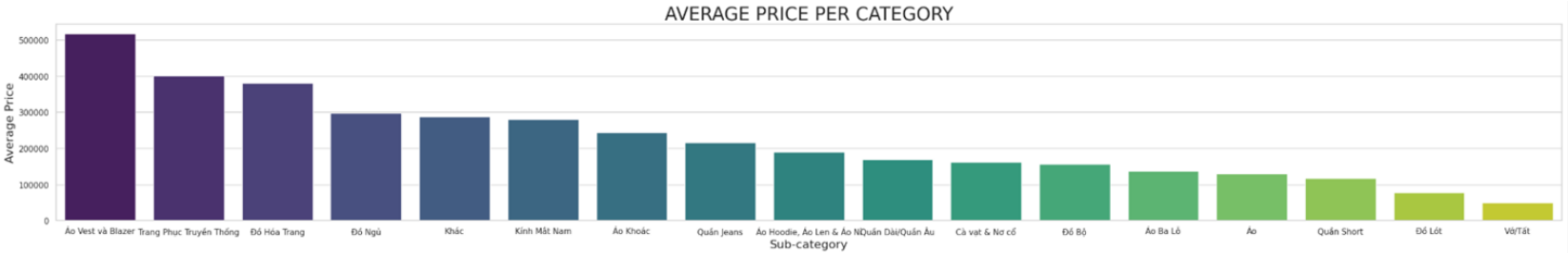
3

MODEL EVALUATION

02

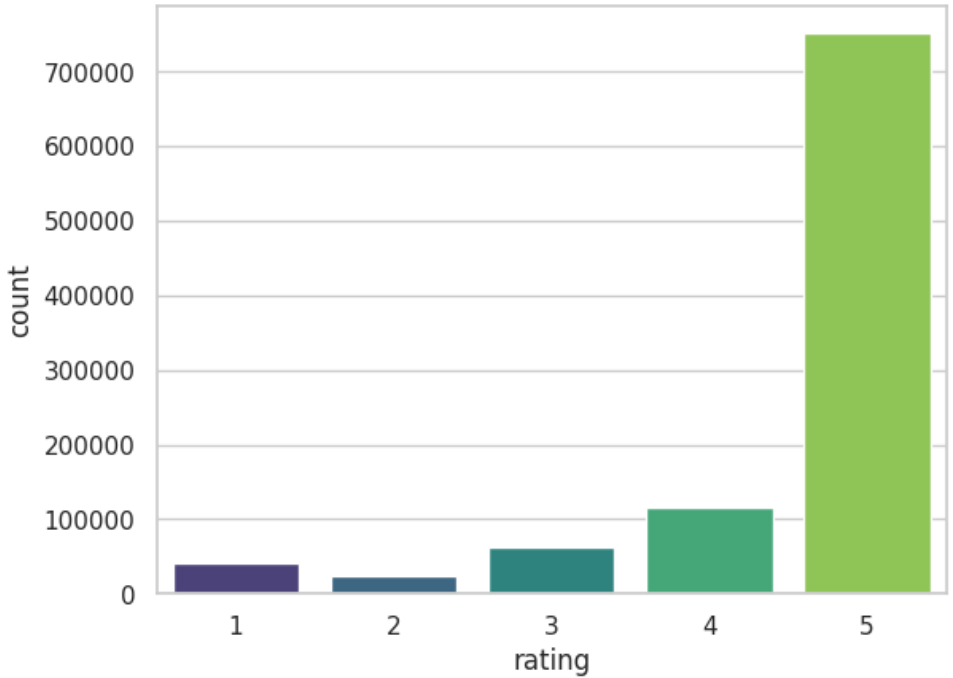
EDA

EDA



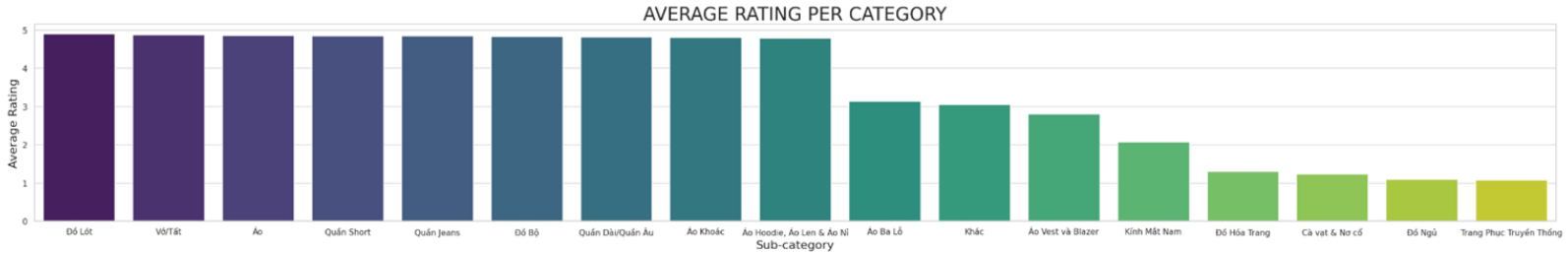
sub_category	avg price	avg_rating
Áo Vest và Blazer	517,251	2.803592
Trang Phục Truyền Thống	400,608	1.080431
Đồ Hóa Trang	379,890	1.297859
Đồ Ngủ	296,850	1.095000
Khác	286,988	3.053328
Kính Mắt Nam	280,364	2.075777
Áo Khoác	242,988	4.811914
Quần Jeans	216,218	4.841467
Áo Hoodie, Áo Len & Áo Nỉ	190,409	4.790267
Quần Dài/Quần Âu	168,677	4.818667
Cà vạt & Nơ cổ	161,959	1.237253
Đồ Bộ	155,651	4.830333
Áo Ba Lỗ	136,848	3.134267
Áo	129,296	4.865436
Quần Short	117,270	4.852333
Đồ Lót	78,140	4.901133
Vớ/Tất	49,269	4.870087

EDA



Hầu hết users đều cho đánh giá 5*

EDA



sub_category	avg price	avg_rating
Áo Vest và Blazer	517,251	2.803592
Trang Phục Truyền Thống	400,608	1.080431
Đồ Hóa Trang	379,890	1.297859
Đồ Ngủ	296,850	1.095000
Khác	286,988	3.053328
Kính Mắt Nam	280,364	2.075777
Áo Khoác	242,988	4.811914
Quần Jeans	216,218	4.841467
Áo Hoodie, Áo Len & Áo Nỉ	190,409	4.790267
Quần Dài/Quần Âu	168,677	4.818667
Cà vạt & Nơ cổ	161,959	1.237253
Đồ Bộ	155,651	4.830333
Áo Ba Lỗ	136,848	3.134267
Áo	129,296	4.865436
Quần Short	117,270	4.852333
Đồ Lót	78,140	4.901133
Vớ/Tất	49,269	4.870087

EDA

	product_id	user_id	user	rating
0	190	1	karmakyun2nd	5
1	190	2	tranquangvinh_vv	5
2	190	3	nguyenquoctoan2005	5
3	190	4	nguyenthuyhavi	5
4	190	5	luonganh5595	5

```
Number of unique values:
product_id : 31267
user_id : 650636
user : 650636
rating : 5
```

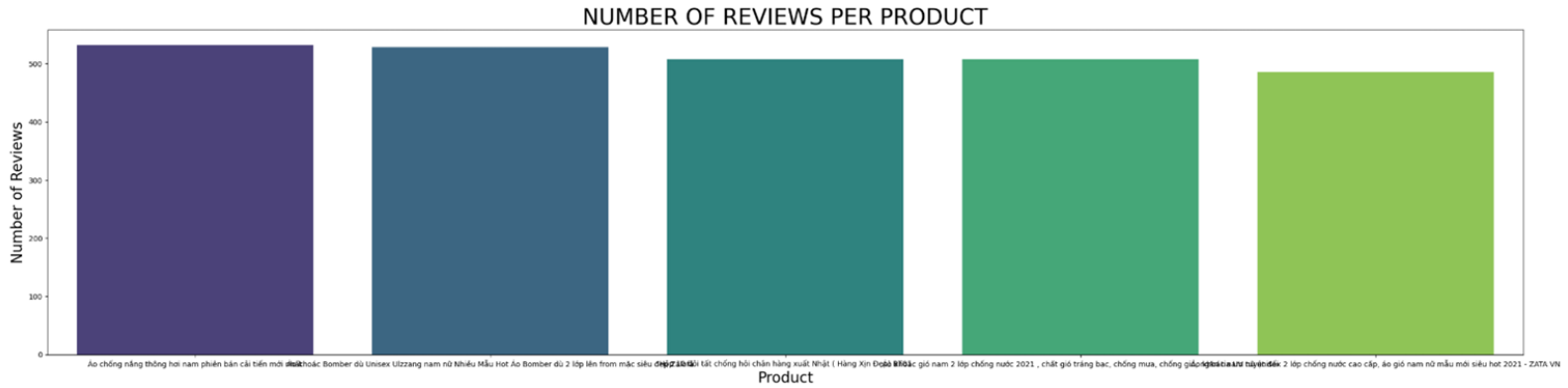
Có 650,636 users, thực hiện review cho tổng cộng 31,267 sản phẩm

EDA

	user	product_name	sub_category
count	994751	994751	994751
unique	647647	30365	17
top	Người dùng Shopee	Áo chống nắng thông hơi nam phiên bản cải tiến...	Đồ Bộ
freq	12951	532	196135

- Có 647647 users, với tổng cộng 994751 lượt mua
- User có tên 'Người dùng Shopee' đứng top 1 mua hàng với tổng lượt mua là 12,951
- Sản phẩm 'Áo chống nắng thông hơ nam phiên bản cải tiến' được reviews nhiều nhất với 532 lượt
- Phân loại hàng 'Đồ bộ' là sản phẩm được mua nhiều nhất với tổng lượt mua là 196,135 lượt mua

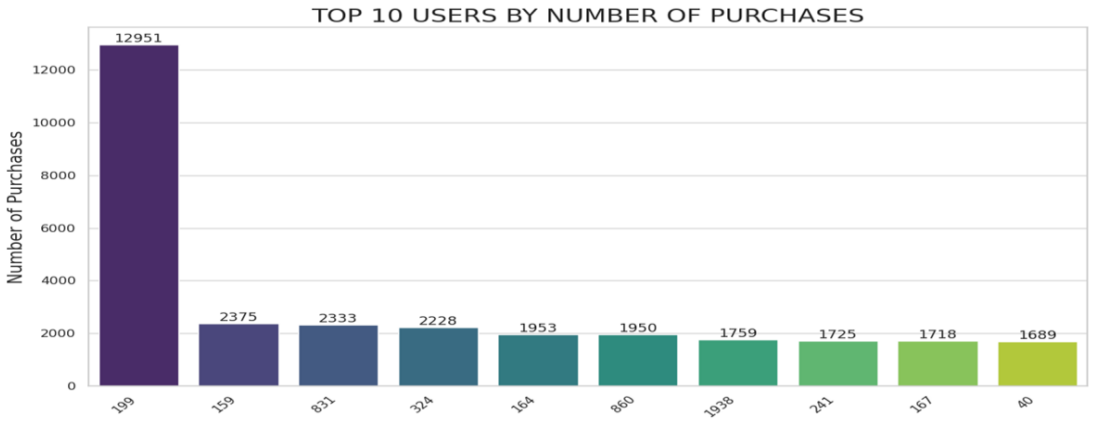
EDA



product_name	count_no_reviews
Áo chống nắng thông hơi nam phiên bản cải tiến mới nhất	532
Áo khoác Bomber dù Unisex Ulzzang nam nữ Nhiều Mẫu Hot Áo Bomber dù 2 lớp lên from mặc siêu đẹp Zalofa	528
Hộp 10 đôi tất chống hôi chân hàng xuất Nhật (Hàng Xịn Đẹp) BT01	508
Áo khoác gió nam 2 lớp chống nước 2021 , chất gió tráng bạc, chống mưa, chống gió, ngăn tia UV tuyệt đối	508
Áo khoác nam nữ unisex 2 lớp chống nước cao cấp, áo gió nam nữ mẫu mới siêu hot 2021 - ZATA VN	486

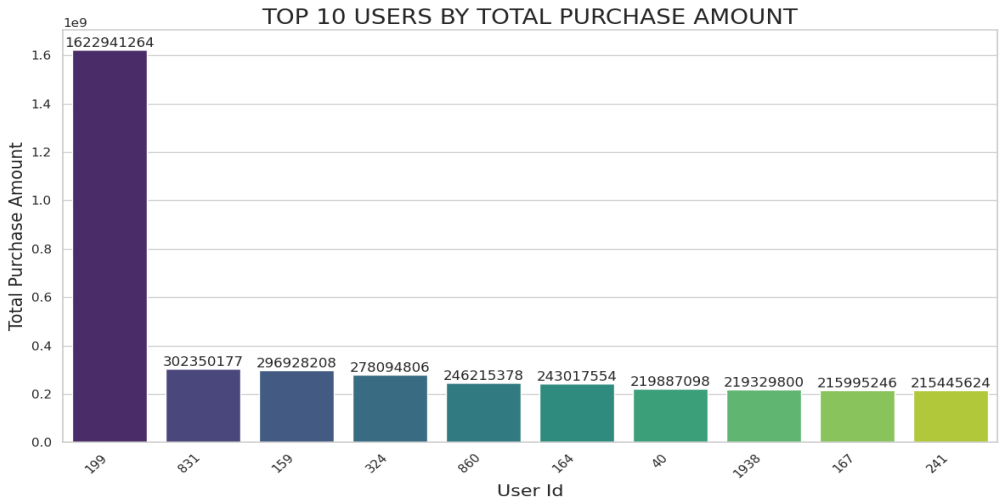
"Áo chống nắng thông hơi nam phiên bản cải tiến mới nhất" là sản phẩm có nhiều lượt đánh giá từ KH đã mua nhất

EDA



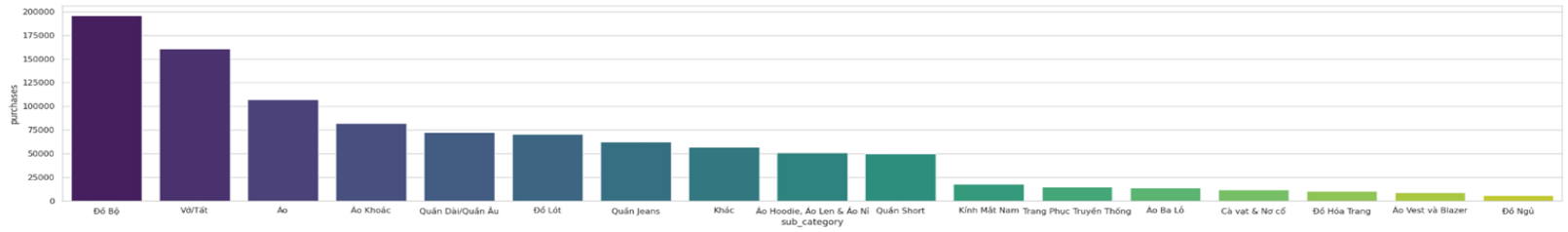
user_id	purchases	purchase amount
199	12951	1,622,941,264
159	2375	296,928,208
831	2333	302,350,177
324	2228	278,094,806
164	1953	243,017,554
860	1950	246,215,378
1938	1759	219,329,800
241	1725	215,445,624
167	1718	215,995,246
40	1689	219,887,098

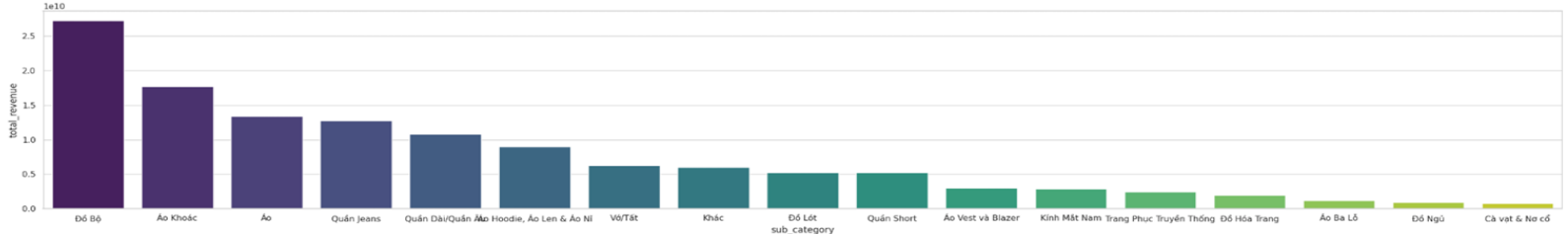
EDA



user_id	purchases	purchase amount
199	12951	1,622,941,264
831	2333	302,350,177
159	2375	296,928,208
324	2228	278,094,806
860	1950	246,215,378
164	1953	243,017,554
40	1689	219,887,098
1938	1759	219,329,800
167	1718	215,995,246
241	1725	215,445,624

EDA

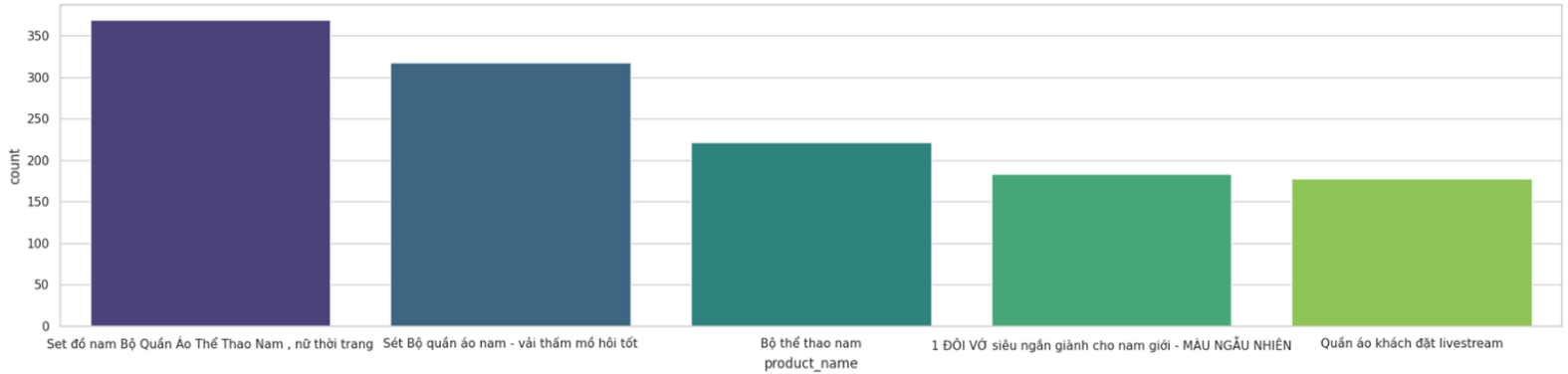




sub_category	most_reviewed_product	purchases	total revenue	percent_amount
Đồ Bộ	Set đồ nam Bộ Quần Áo Thể Thao Nam , nữ thời t...	196135	27,280,362,712	21.539333
Áo Khoác	Áo chống nắng thông hơi nam phiên bản cải tiến...	81928	17,697,549,394	13.973180
Áo	Áo sơ mi nam nữ dài tay Unisex Basic màu trắng...	107339	13,361,829,098	10.549892
Quần Jeans	Quần jean nam baggy kiểu ống rộng dáng suông c...	62617	12,785,761,923	10.095056
Quần Dài/Quần Âu	Quần Thể Thao Nam 3 Sọc Quần Nam Thu Đông Co G...	72755	10,770,930,866	8.504237

Mặt hàng 'Đồ Bộ' có tổng số lượt mua là 196,135 lần, chiếm doanh thu cao nhất khoảng 27,3 tỷ đồng. Đóng góp 21,5% vào tổng doanh thu.

EDA



product_name	count
Set đồ nam Bộ Quần Áo Thể Thao Nam , nữ thời trang	369
Sét Bộ quần áo nam - vải thấm mồ hôi tốt	318
Bộ thể thao nam	222
1 ĐÔI VỚ siêu ngắn giành cho nam giới - MÀU NGẪU NHIÊN	184
Quần áo khách đặt livestream	178

03

CONTENT-BASED FILTERING

COSINE SIMILARITY

Cosine Similarity

```
# Gọi function
recommendations = get_recommendations_1(1964, cosine_sim, 3) # product_id = 1964 là áo thun ba lỗ
print(recommendations)
```

```
99      [MS038] Áo Thun Sát Nách Thời Trang Mùa Hè Cho...
201               áo ba lỗ tập gym cao cấp hàng gym shark
1228                        Combo 3 Áo ba lỗ nam thể thao
Name: product_name, dtype: object
```

Thời gian: 136 giây

COSINE SIMILARITY USING CHUNK_SIZE

```
# Gọi function
recommendations = get_recommendations_2(1964, cosine_sim, 3)
print(recommendations)
```

```
99      [MS038] Áo Thun Sát Nách Thời Trang Mùa Hè Cho...
201               áo ba lỗ tập gym cao cấp hàng gym shark
1228                        Combo 3 Áo ba lỗ nam thể thao
Name: product_name, dtype: object
```

Thời gian: 123 giây

GENSIM

GENSIM

04

COLLABORATIVE FILTERING

SVD

Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.8884	0.8871	0.8909	0.8874	0.8862	0.8880	0.0016
MAE (testset)	0.5678	0.5670	0.5697	0.5678	0.5664	0.5677	0.0011
Fit time	30.84	30.46	28.59	29.08	29.18	29.63	0.87
Test time	2.30	2.74	2.26	2.00	1.39	2.14	0.44

Thời gian: 130s

BASELINEONLY

Evaluating RMSE, MAE of algorithm BaselineOnly on 5 split(s).

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.8766	0.8809	0.8833	0.8837	0.8809	0.8811	0.0025
MAE (testset)	0.5754	0.5771	0.5789	0.5794	0.5776	0.5777	0.0014
Fit time	9.88	10.02	10.48	12.68	10.30	10.67	1.03
Test time	1.71	1.00	1.58	1.65	0.99	1.39	0.32

TUNNING PARAMETERS

RMSE score: 0.8751469829756996
MAE score: 0.5598821876670089
Parameters: {'bsl_options': {'method': 'als', 'reg_u': 5, 'reg_i': 5}}

Thời gian: 60s

NFM

Evaluating RMSE, MAE of algorithm NMF on 5 split(s).

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	1.0423	1.0396	1.0388	1.0395	1.0374	1.0395	0.0016
MAE (testset)	0.7687	0.7662	0.7677	0.7678	0.7661	0.7673	0.0010
Fit time	86.49	82.96	84.54	84.14	84.63	84.55	1.14
Test time	2.28	2.67	2.18	1.48	2.18	2.16	0.39

Thời gian: 433s

COCLUSTERING

Evaluating RMSE, MAE of algorithm CoClustering on 5 split(s).

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9925	0.9824	1.0212	0.9883	0.9825	0.9934	0.0144
MAE (testset)	0.6597	0.6553	0.6642	0.6616	0.6529	0.6587	0.0041
Fit time	85.77	90.22	86.28	86.13	87.05	87.09	1.62
Test time	2.52	1.77	1.11	1.07	1.11	1.51	0.57

Thời gian: 532s

KNN BASIC

Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	1.0100	1.0174	1.0158	1.0209	1.0195	1.0167	0.0038
MAE (testset)	0.6166	0.6173	0.6177	0.6226	0.6224	0.6193	0.0026
Fit time	0.75	0.76	0.86	1.01	0.76	0.83	0.10
Test time	1.66	1.45	2.17	1.64	1.68	1.72	0.24

Thời gian: 15s

KNN-BASELINE

Evaluating RMSE, MAE of algorithm KNNBaseline on 5 split(s).

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	1.0122	1.0098	1.0105	1.0088	1.0157	1.0114	0.0024
MAE (testset)	0.6261	0.6251	0.6221	0.6227	0.6289	0.6250	0.0025
Fit time	2.37	1.98	2.26	2.36	2.05	2.20	0.16
Test time	2.47	2.01	2.07	2.91	1.80	2.25	0.39

Thời gian: 22s

KNN WITH MEANS

Evaluating RMSE, MAE of algorithm KNNWithMeans on 5 split(s).

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	1.0030	1.0121	1.0128	1.0082	1.0157	1.0104	0.0044
MAE (testset)	0.6245	0.6305	0.6285	0.6275	0.6332	0.6288	0.0029
Fit time	0.83	0.90	1.15	1.05	1.37	1.06	0.19
Test time	1.53	2.15	2.15	1.81	1.63	1.85	0.26

Thời gian: 15s

KNN WITH ZSCORE

Evaluating RMSE, MAE of algorithm KNNWithZScore on 5 split(s).

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	1.0103	1.0172	1.0104	1.0149	1.0125	1.0131	0.0026
MAE (testset)	0.6225	0.6263	0.6225	0.6273	0.6276	0.6252	0.0023
Fit time	1.39	1.40	1.15	1.33	1.58	1.37	0.14
Test time	3.04	1.89	1.71	2.08	2.13	2.17	0.46

Thời gian: 18s

ALS (PYSPARK)

```
als = ALS(maxIter=15,
          regParam=0.1,
          rank = 20,
          userCol="user_id",
          itemCol="product_id",
          ratingCol="rating",
          coldStartStrategy="drop",
          nonnegative=True)

model = als.fit(train)

predictions=model.transform(test)

evaluator=RegressionEvaluator(metricName='rmse', labelCol='rating', predictionCol='prediction')
rmse=evaluator.evaluate(predictions)
print('With regParam =', 0.1, ', rank =', 20, ', maxIter =', 15 , ': RSME =', rmse)
```

```
With regParam = 0.1 , rank = 20 , maxIter = 15 : RSME = 1.1617218779544844
```

05

MODEL EVALUATION

MODEL EVALUATION

Phương pháp	Thời gian	RMSE	Nhận xét
Gensim			
Cosine Similarity	136s		Cho kết quả tốt, thời gian chạy khá lâu
Cosine Similarity using chunk_size	123s		Cho kết quả tương tự Cosine Similarity truyền thống. Nhanh hơn Cosine Similarity truyền thống
ALS (PySpark)	...	1.16	RMSE cao
SVD	130	0.888	Thời gian chạy rất lâu, tuy nhiên RMSE thấp
Baseline Only	60s	0.875	RMSE thấp nhất, thời gian chạy nhanh

MODEL EVALUATION

Phương pháp	Thời gian	RMSE	Nhận xét
NFM	433ss	1.04	Thời gian chạy lâu, RMSE cao
CoClustering	532s	0.99	Thời gian chạy lâu, RMSE trung bình
KNN with Mean	15s	1.01	Thời gian chạy nhanh nhưng RMSE cao
KNN with ZScore	18s	1.01	Thời gian chạy nhanh nhưng RMSE cao
KNN Basic	15s	1.02	Thời gian chạy nhanh nhưng RMSE cao
KNN Baseline	22s	1.01	Thời gian chạy nhanh nhưng RMSE cao

THANK YOU!

FOR WATCHING