Thompson Sampling을 통한
맞춤형 책 추천 시스템

마음의 양식 with YES24.COM
김소라, 이영송, 주원진, 황경서

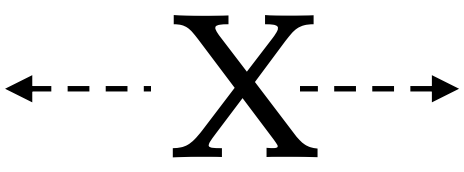
목차

YES24.COM

I. 프로젝트 소개

A. 프로젝트 설명

[특정 User i]

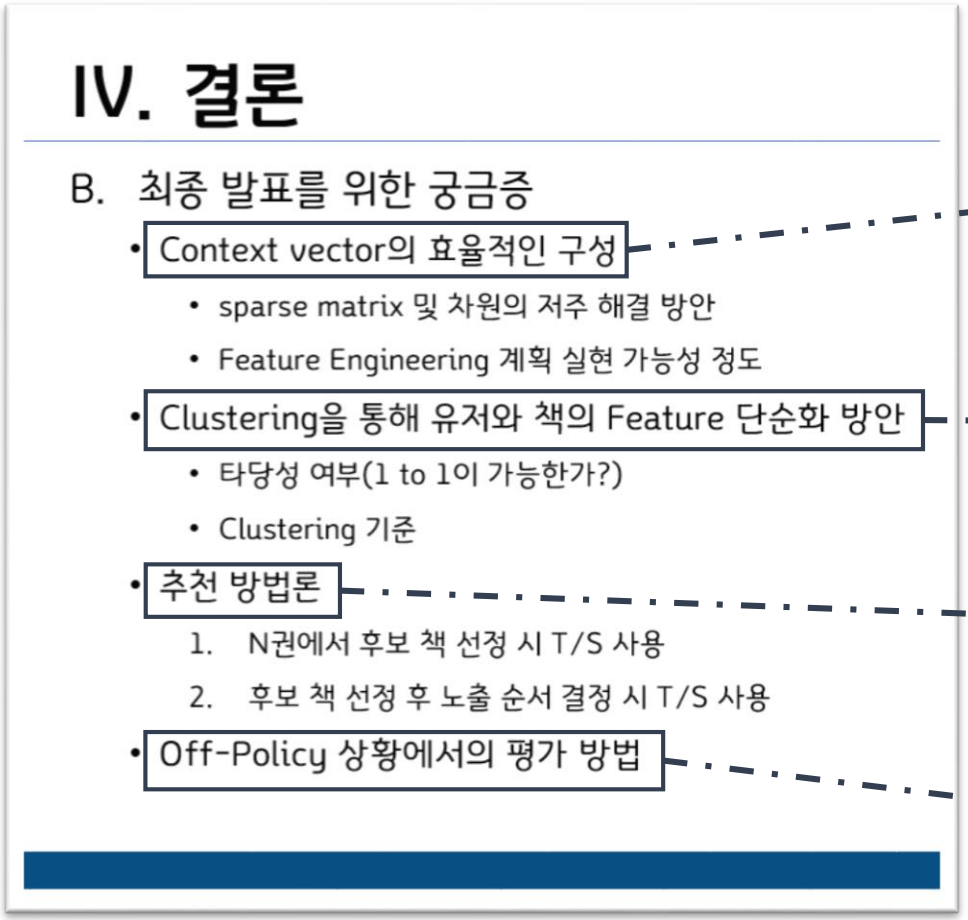


[특정 Book j]

특정 유저와 특정 책이 만나 얻는 행복감 y_{ij}(구매) 를
극대화 할 수 있는 추천 행복 알고리즘 구축

I. 프로젝트 소개

B. 프로젝트 개선 방향



- Context vector 구성
  - 책과 사용자에 대한 세심한 Feature engineering
  - 정규화를 통해 sparse matrix 문제 해소
- Clustering
  - Elbow 방식을 이용한 K-means
- 추천 방법론
  - 사용자에게 가장 적합한 책 집단 후보 선정
- Off-Policy 상황에서 평가 방법
  - Online 전환 시도 -> 시간 제약

II. 데이터 설명

A. Data frame

- User dummy data (1221182 rows x 39 columns)

Account_id	Gender 0	Gender 1	Age 0	···	Address 0	Address 1
102600	1	0	0	···	0	1
103417	1	0	0	···	1	0
105247	0	1	0	···	0	1
···	···	···	···	···	···	···
16832102	0	1	0	···	0	1
16834484	1	0	0	···	0	1
16836022	1	0	0	···	1	0

Train 기간(2020.3~2020.4) 내에 활동한 User에 대한 User Feature dummy 정규화 진행

II. 데이터 설명

A. Data frame

- Book dummy data(270631 rows x 35 columns)

Product_id	Cat 1	Cat 2	Cat 3	…	Price 2	Price 3
51205670	0	0	0	…	0	0
51205708	0	0	1	…	1	0
51205718	0	0	1	…	1	0
…	…	…	…	…	…	…
85352665	1	0	0	…	0	1
85354327	0	0	0	…	0	0
85354780	1	0	0	…	0	1

Train 기간(2020.3~2020.4) 내에 등장한 Book에 대한 Book Feature dummy 정규화 진행

II. 데이터 설명

B. 변수 설명

- User dummy data

Feature	Column 명	변수 설명	비고
User account	Account_id	고객 ID	
Gender	Gender 0	남성	
	Gender 1	여성	
Age	Age 0	12세 이하	
	Age 1	12~19세	
	…	…	
	Age 5	51~64세	
	Age 6	65세 이상	

II. 데이터 설명

B. 변수 설명

- User dummy data _2

Feature	Column 명	변수 설명	비고
책 카테고리 선호도	Most pref 1	수험서 자격증	각 user 별 기간 내의 전체 Click 수 대비 책 카테고리 Click 비율
	Most pref 2	IT 모바일	
	…	…	
	Most pref 33	소설/시/희곡	
	Most pref 34	에세이	
신규 책 선호	New pref 0	올해 출판 된 책	
	New pref 1	2020년 이전에 출판된 책	
주소	Address 0	수도권(서울, 경기, 인천)	
	Address 1	비수도권	

II. 데이터 설명

B. 변수설명

- Book dummy data

Feature	Column 명	변수 설명	비고
Book product	Product_id	책 정보 ID	
책 카테고리	Cat 1	수험서 자격증	
	Cat 2	IT 모바일	
	Cat 3	중고등 학습서	
	…	…	
	Cat 33	소설/시/희곡	
	Cat 34	에세이	

II. 데이터 설명

B. 변수설명

- Book dummy data

Feature	Column 명	변수 설명	비고
출판일	Published 0	3개월 미만	2020년 5월 1일 기준
	Published 1	3개월 이후 6개월 미만	
	Published 2	6개월 이후 1년 미만	
	Published 3	1년 이후 3년 미만	
	Published 4	3년 이후	
책 가격	Price 0	10000원 미만	전체 책 가격의 4분위 수
	Price 1	10000원 이상 13800원 미만	
	Price 2	13800원 이상 19000원 미만	
	Price 3	19000원 이상	

Ⅱ. 데이터 설명

C. Clustered data _User와 Book data에서 도출

- Data frame(24024943 rows x 5 columns)

Account_id	Product_id	Purchase	Book_cluster	User_cluster
100001	88440267	1	3	3
100001	85951536	1	3	3
100001	86895523	0	3	3
…	…	…	…	…
16837207	84678826	1	3	2
16837207	22791986	0	0	2
16837207	84639236	0	3	2

Train 기간(2020.3~2020.4)에서 Click event가 발생한 순간의 Clustered data

Ⅱ. 데이터 설명

C. Clustered data

- 변수 설명

	Column 명	변수 설명	비고
User account	Account_id	고객 ID	Click event 발생 시 생기는 data
Book product	Product_id	책 정보 ID	
구매 유/무	Purchase	구매 1, 비구매 0	
책 그룹	Book Cluster	5 그룹으로 묶음	K-means 사용
유저 그룹	User Cluster	6 그룹으로 묶음	

II. 데이터 설명

D. Clustered dummy data _Clustered data에 User와 Book data 대입

- Data frame(24024943 rows x 5 columns)

Account_id	Product_id	Purchase	Book_cluster	User_cluster
User dummy data 정규화	Book dummy data 정규화	1	3	3
		1	3	3
		0	3	3
		…	…	…
		1	3	2
		0	0	2
		0	3	2

Spars matrix 해소

III. 모델링 결과

A. 모델링 설명

Algorithm Linear Contextual Thompson Sampling

1: 5그룹의 책 cluster별로 각 context의 feature 별 prior distribution을 가정

2: 새로운 Context 발생 시 현재까지의 distribution에서 sampling

3: Context와 sampling 결과 값을 계산하여 가장 높은 scalar 값을 나타내는 책 Cluster 선택

4: 선택된 책 Cluster의 reward 값 관측 후 결과에 따라 해당 책 Cluster의 distribution 업데이트

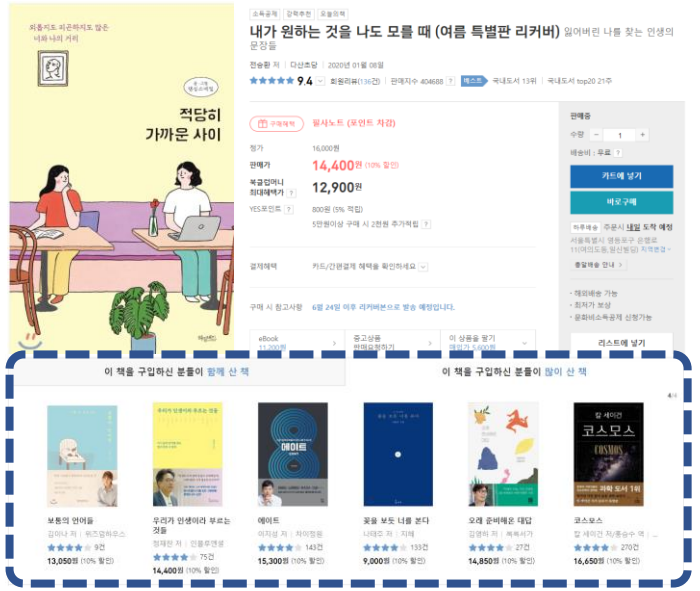
III. 모델링 결과

B. 평가 비교(Off-Policy)_2020년 5월 간 data

1. Yes24 추천 알고리즘

임의의 책 한 권을
Test 기간 동안 Click한
User

Random User 100



24권의 책이 추천된 상태

III. 모델링 결과

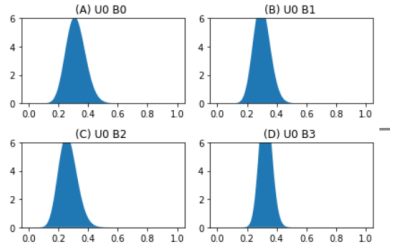
B. 평가 비교(Off-Policy)_2020년 5월 간 data

2. T/S 방식 추천 알고리즘_Model 1

임의의 책 한 권을
Test 기간 동안 Click한
User

Random User 100

User Clustered



β분포

book Clustered

Yes24와 동일한 방식으로 추출

III. 모델링 결과

B. 평가 비교(Off-Policy)_2020년 5월 간 data

2. T/S 방식 추천 알고리즘_Model 1

Algorithm Linear Contextual Thompson Sampling Model 1

3~4월 click data : train data

5월 click data : test data

1: 유저 Cluster에 대한 책 Cluster의 (성공, 실패) 베타 분포 학습

2: Click Event 발생 시 유저 Cluster-책 Cluster의 확률 분포에서 임의의 값을 sampling

3: 해당 유저 Cluster에 대해서 가장 큰 sampling 값을 갖는 책 Cluster를 선택

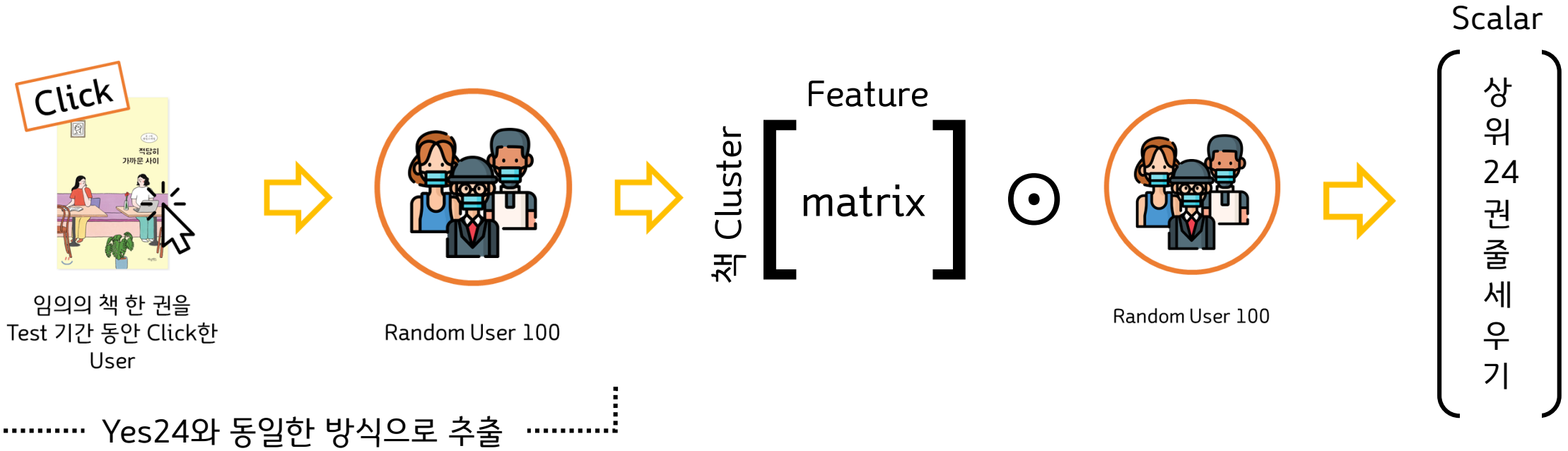
4: 선택된 책 Cluster의 책들에 대해 학습한 Context의 Coefficient를 내적하여 상위 24개를 추출

5: 실제 사용자가 구매한 책 중 추천된 책이 포함된 값을 계산

III. 모델링 결과

B. 평가 비교(Off-Policy)_2020년 5월 간 data

2. T/S 방식 추천 알고리즘_Model 2



III. 모델링 결과

B. 평가 비교(Off-Policy)_2020년 5월 간 data

2. T/S 방식 추천 알고리즘_Model 2

Algorithm Linear Contextual Thompson Sampling Model 2

동일한 random user pool 100명의 3~4월 click data : train data

동일한 random user pool 100명의 5월 click data : test data

1: Train data로 책 Cluster의 feature별 분포 학습

2: Test data 내 Click Event 발생 시 책 Cluster의 feature별 확률 분포에서 임의의 값을 sampling

3: 해당 유저의 context feature와 sampling된 값을 내적하여 가장 큰 값을 갖는 상위 24권을 추출

4: 실제 사용자가 구매한 책 중 추천된 책이 포함된 값을 계산

III. 모델링 결과

B. 평가 비교(Off-Policy)_2020년 5월 간 data

2. T/S 방식 추천 알고리즘_Model 2

Algorithm Linear Contextual Thompson Sampling Model 2

2: Test data 내 Click Event 발생 시 **책 Cluster의 feature별 확률 분포**에서 임의의 값을 sampling

beta_df

5 Cluster

	gender_0.0	gender_1.0	g_age_0.0	g_age_1.0	g_age_2.0	g_age_3.0	g_age_4.0	g_age_5.0	g_age_6.0	most_pref_1.0	...	pub0.0	pub1.0	pub2.0
0	-0.008041	0.008041	0.0	0.000871	-0.045461	0.015704	0.018834	-0.003297	0.0	0.002749	...	-0.017351	-0.010937	-0.014325
1	-0.020075	0.020075	0.0	-0.004530	-0.003938	0.001125	0.010368	-0.010119	0.0	-0.005313	...	-0.044744	-0.015037	-0.003578
2	-0.003361	0.003361	0.0	0.002852	0.016485	-0.004534	-0.004503	-0.003877	0.0	0.000079	...	-0.040404	-0.018523	0.017996
3	-0.008422	0.008422	0.0	0.009094	0.022423	-0.037532	0.022905	0.001765	0.0	0.010594	...	0.092001	0.055470	0.016955
4	-0.006379	0.006379	0.0	0.006108	0.001111	-0.013813	0.015907	-0.002477	0.0	-0.026003	...	-0.032416	-0.028814	-0.023092

III. 모델링 결과

B. 평가 비교(Off-Policy)_2020년 5월 간 data

2. T/S 방식 추천 알고리즘_Model 2

Algorithm Linear Contextual Thompson Sampling Model 2

3: 해당 유저의 context feature와 sampling된 값을 **내적하여 가장 큰 값을 갖는 상위 24권을 추출**

```
df_fin_fin_fin.sort_values(['account_id','val_max'],ascending=False)[0:50]
```

	account_id	product_id	book_cluster	0	1	2	3	4	val_max	val_max_idx
4120	16790454.0	37300128.0	1.0	-0.086427	0.499775	0.050252	-0.174359	-0.039236	0.499775	1.0
4119	16790454.0	79297023.0	1.0	-0.024507	0.482987	0.001415	-0.151331	-0.142064	0.482987	1.0
4117	16790454.0	37300128.0	1.0	-0.133613	0.455308	-0.102648	-0.241132	0.043122	0.455308	1.0
4124	16790454.0	37300128.0	1.0	-0.041165	0.447951	-0.060858	-0.287709	0.044051	0.447951	1.0
4108	16790454.0	37300128.0	1.0	-0.132805	0.393513	-0.038143	-0.239686	-0.025957	0.393513	1.0
4103	16790454.0	37300128.0	1.0	-0.094839	0.323402	-0.048594	-0.128818	-0.032212	0.323402	1.0
					⋮					
4096	16790454.0	37300128.0	1.0	-0.004763	0.218254	-0.200370	-0.272850	0.012328	0.218254	1.0
4084	16790454.0	65282018.0	2.0	0.058522	0.110866	0.191640	-0.078400	0.114816	0.191640	2.0
4069	16790454.0	89309569.0	3.0	-0.063426	-0.035225	0.146071	0.186341	-0.042155	0.186341	3.0
4079	16790454.0	63688657.0	4.0	0.015910	-0.168208	0.083975	-0.042007	0.174044	0.174044	4.0
4071	16790454.0	90061659.0	3.0	0.059242	-0.054407	0.003760	0.133078	0.061722	0.133078	3.0

상위 24권 추출

III. 모델링 결과

C. 평가 비교(Off-Policy)_수익 비교

	Yes24	T/S_1	T/S_2
평가 방법	Click event 이후 추천된 24권 구매 유무	Click event 이후 **개인별** 추천된 24권 구매 유무	
한계점	책의 다양성 X	그룹화된 조합으로 정교한 개인 맞춤형 책 추천은 어려움	Train data와 Test data의 User pool이 동일하여 Overfitting 가능성이 농후함
구매 건 수	46건	81건	391건
매출액	₩625,680	₩940,140	₩5,365,690

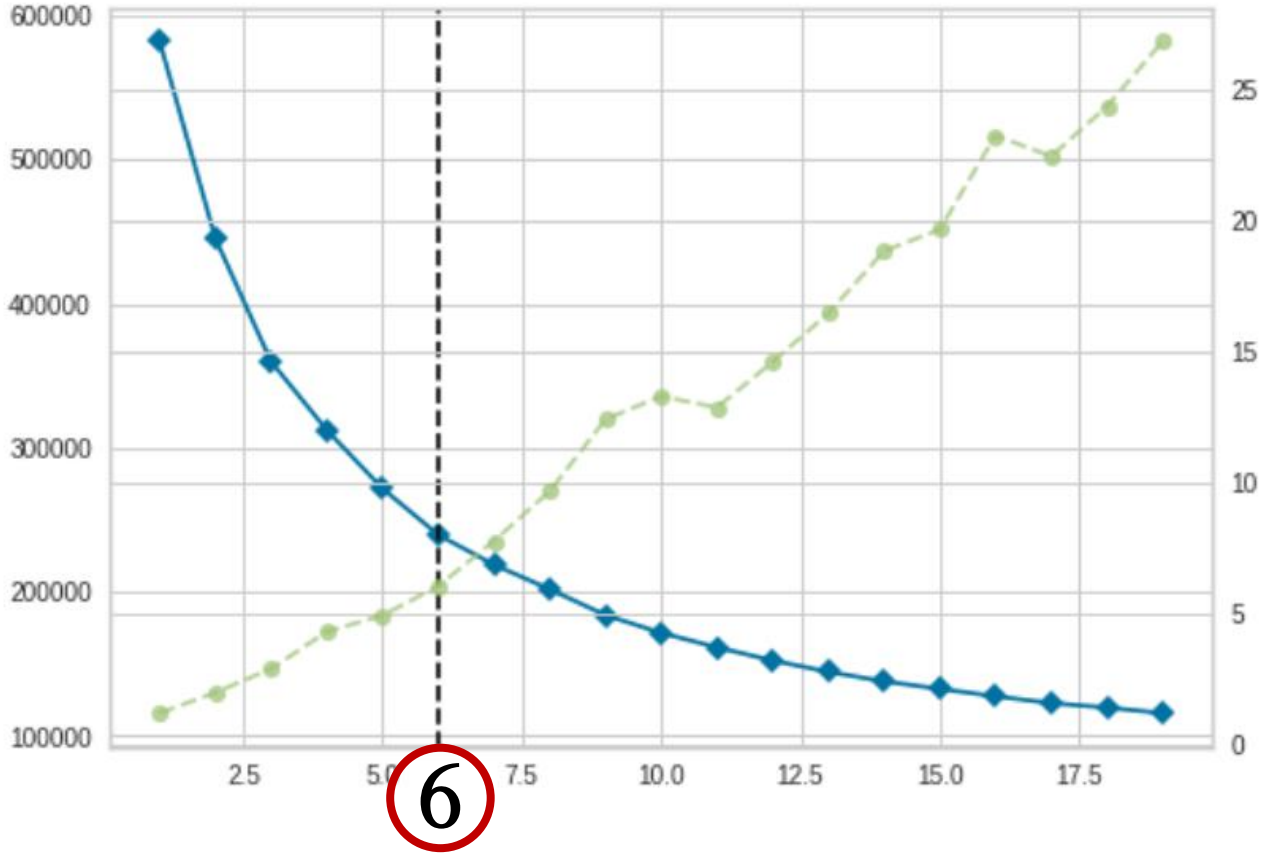
IV. 결론

- 추천 시스템 개선을 통한 이점
  - 다양성(Diversity)과 콜드 스타트(Cold Start) 문제를 해소 가능
  - 개인화 맞춤 추천을 통해 수익 극대화

    - 기존의 추천 방식에 비해 1.5~8배 차이 발생
- 앞으로의 방향성
  - Gibbs sampling을 통해 더욱 정교한 모델 구현 가능
  - 온라인 테스트를 통해 실시간 추천 가능
  - 이미 구매한 책은 추천 list에서 제외

V. 부록

- K-means 군집 수 결정
  - User Cluster

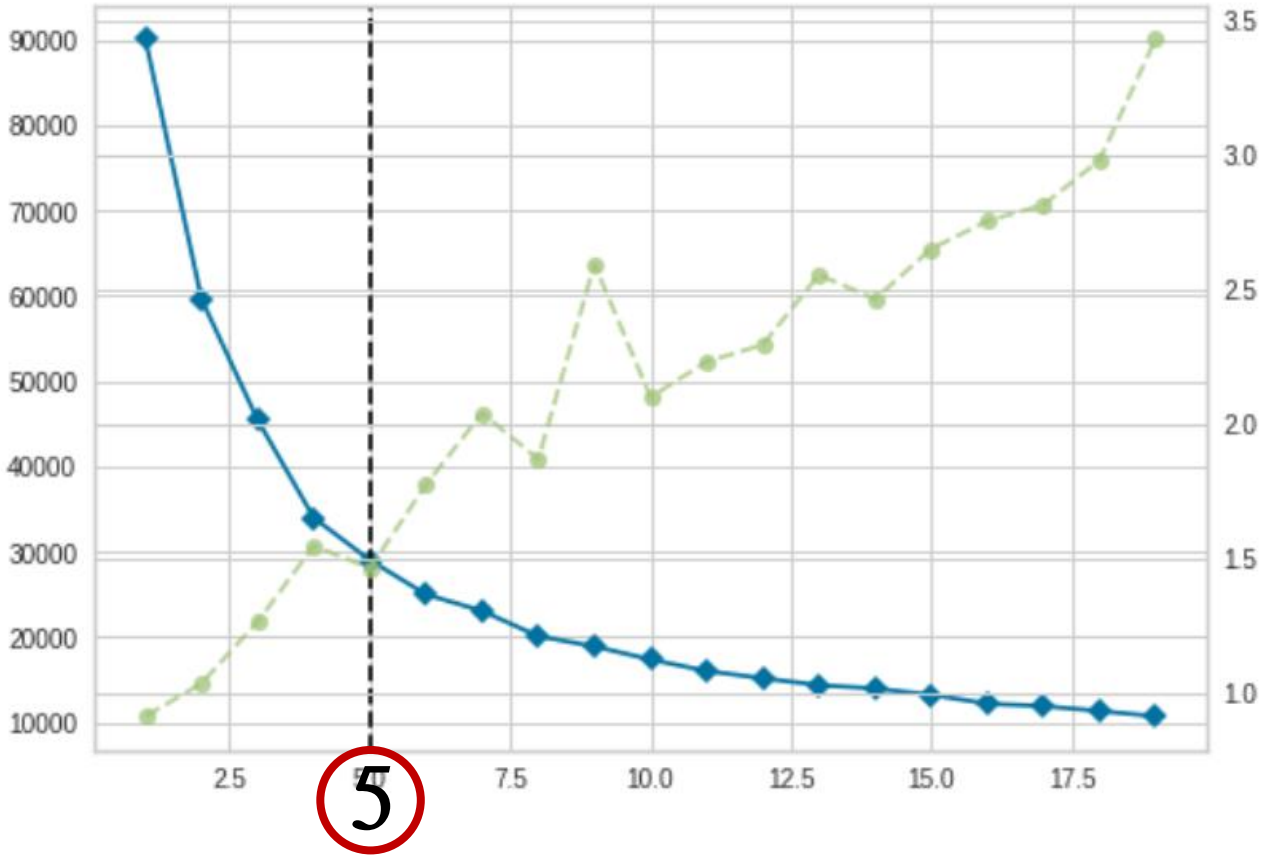


군집 그룹의 변동성

군집 학습 할 때 소요 시간

V. 부록

- K-means 군집 수 결정
  - Book Cluster



V. 부록

- Yes 24 추천 책 목록_1
  - 임의의 책 제목 : “내가 원하는 것을 나도 모를 때”
  - 위 책 기준 추천 책 제목
    - [예스리커버] 1일 1페이지, 세상에서 가장 짧은 교양 수업 365
    - 더 해빙 The Having
    - 1cm 다이빙 (썸머 캣 에디션)
    - 지적 대화를 위한 넓고 얕은 지식 제로
    - 하버드 상위 1퍼센트의 비밀
    - 날씨가 좋으면 찾아가겠어요
    - 애쓰지 않고 편안하게
    - 지쳤거나 좋아하는 게 없거나
    - 스스로 행복하라
    - 당신이 옳다
    - 팩트폴니스

V. 부록

- Yes 24 추천 책 목록_2
  - 나는 나로 살기로 했다
  - 지금 이대로 좋다
  - 이 한마디가 나를 살렸다
  - 사서함 110호의 우편물
  - 타인의 해석
  - 여행의 이유
  - 지리의 힘
  - 보통의 언어들
  - 우리가 인생이라 부르는 것들
  - 에이트
  - 꽃을 보듯 너를 본다
  - 오래 준비해온 대답
  - 코스모스