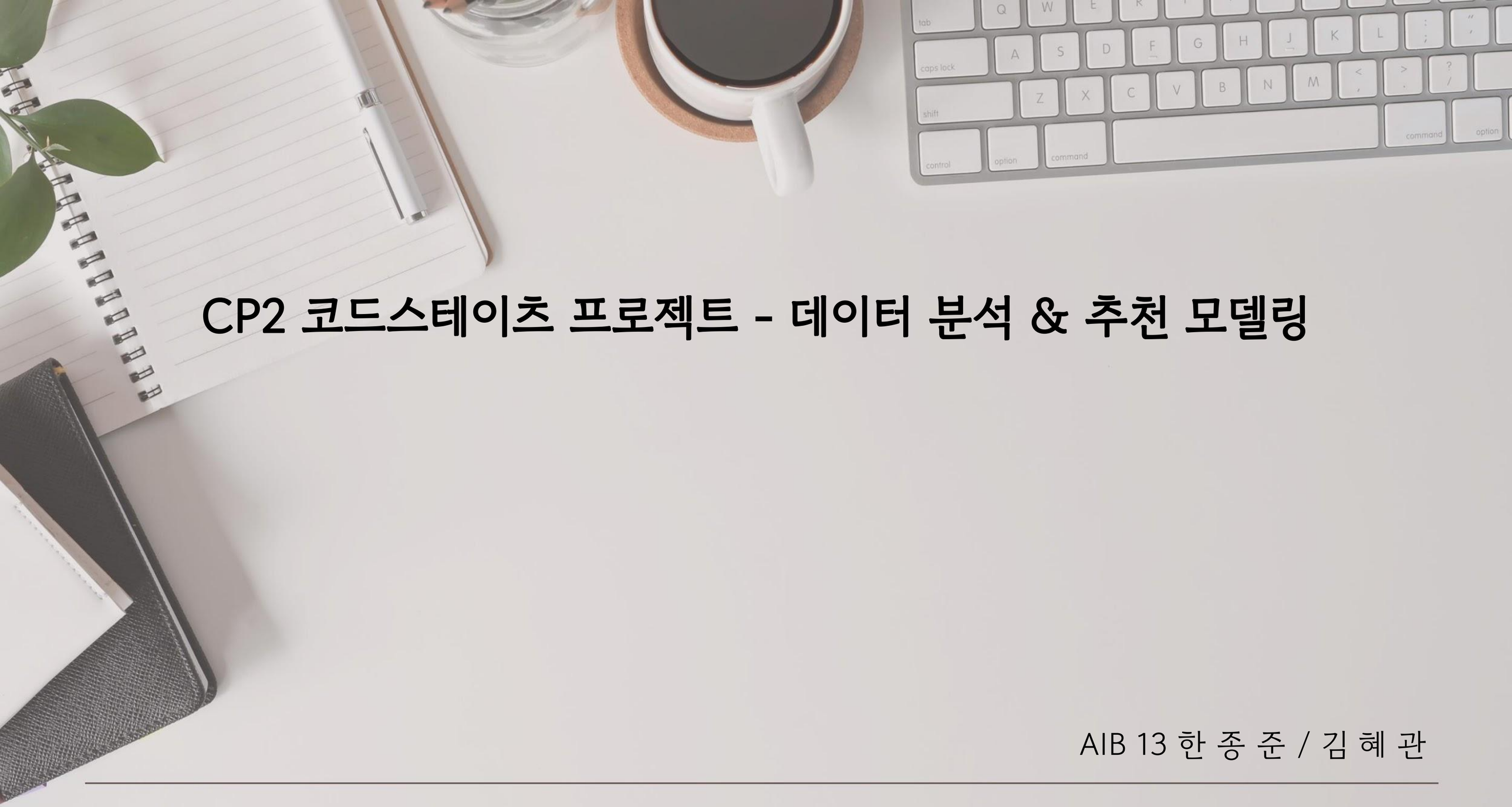


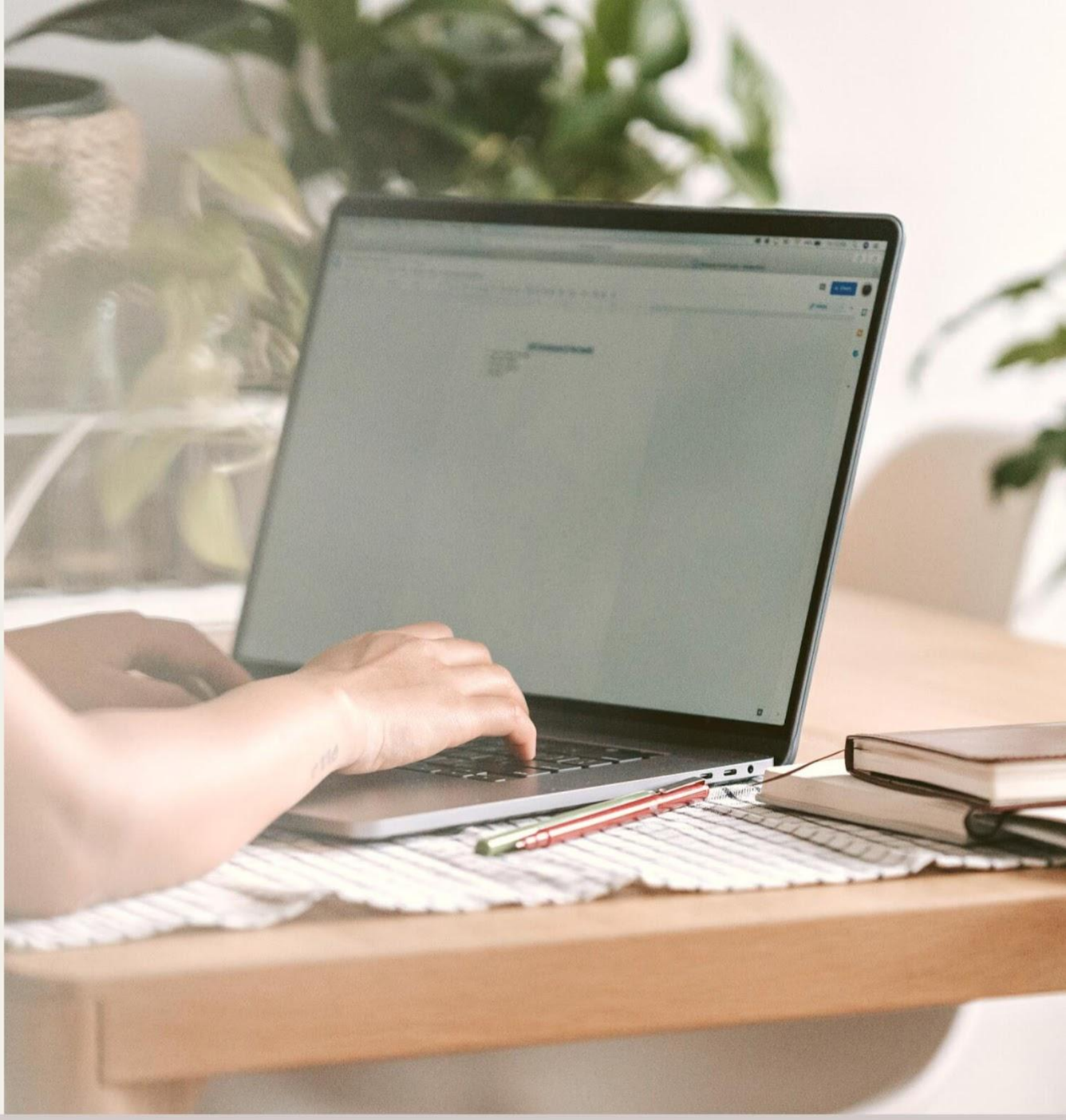
CP2 코드스테이츠 프로젝트 - 데이터 분석 & 추천 모델링

AIB 13 한 종 준 / 김 혜 관

CONTENTS

1. 프로젝트 개요

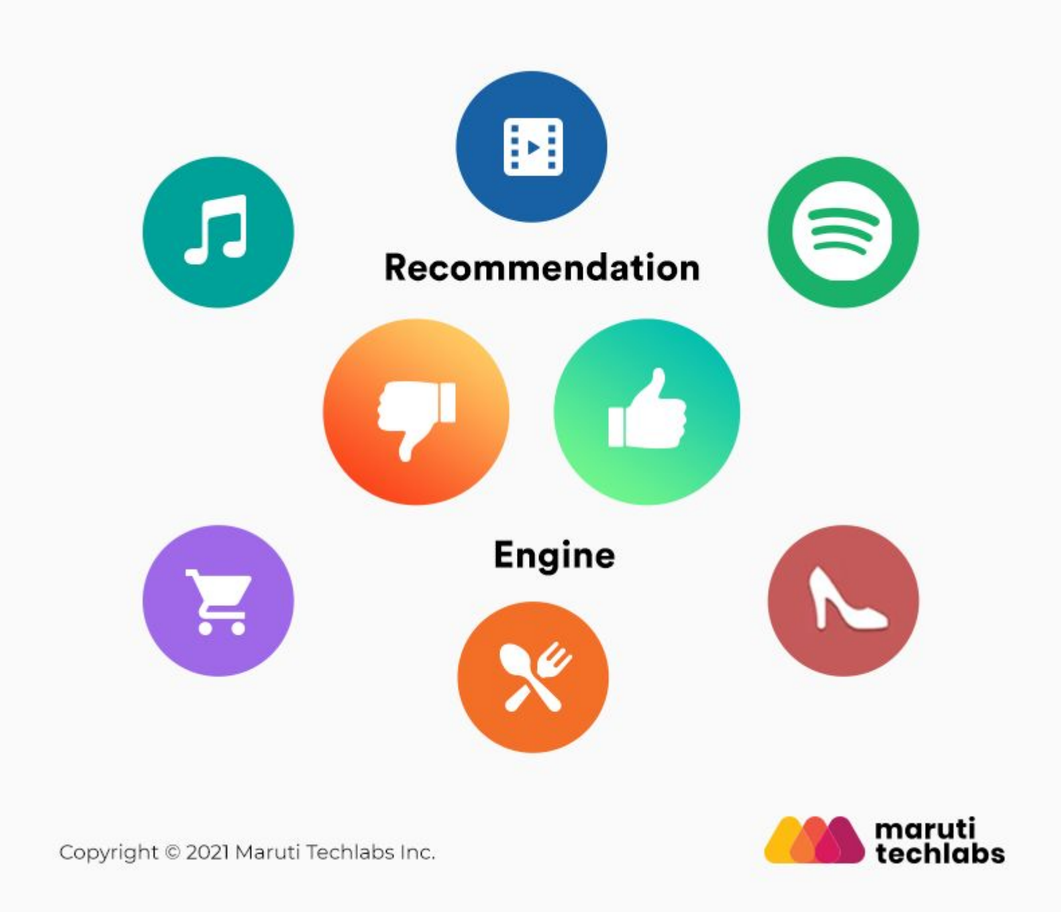
1. 이커머스란
2. 추천 시스템이란
3. 프로젝트 목표와 방향성



1. 이커머스란

- 전자상거래라 하며 일반적으로는 온라인 구매에 중점
- 마케팅, 판매, 주문 등과 고객과 기업간의 소통에 대한 외부 프로세스가 포함
- 정보 기술의 발전으로 현 시대에 그 중요도는 꾸준히 올라가고 있다

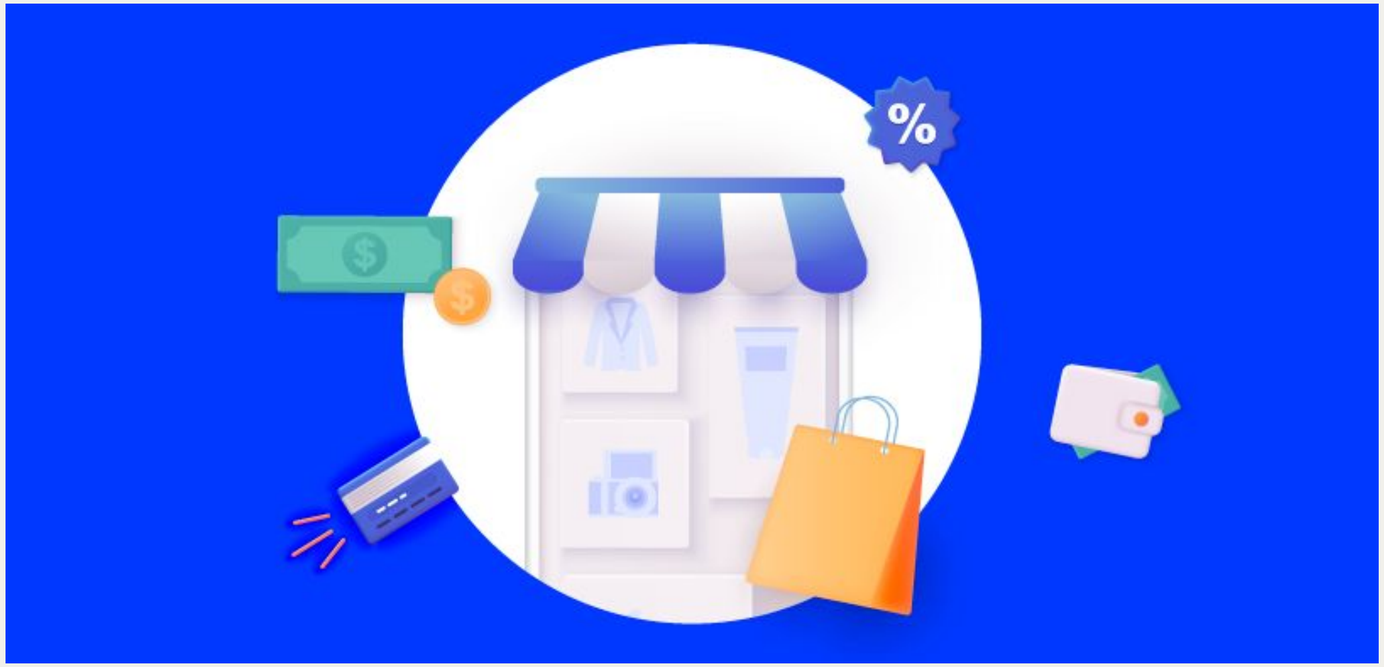
2. 추천시스템이란



- 특수 알고리즘과 기계 학습 솔루션을 사용
  - 특정 사용자에 대한 필터링과 그에 맞춰 매출의 증대 도모

- 협업 필터링 / 콘텐츠 기반 필터링 / 하이브리드 추천 시스템

3. 프로젝트 목표와 방향성

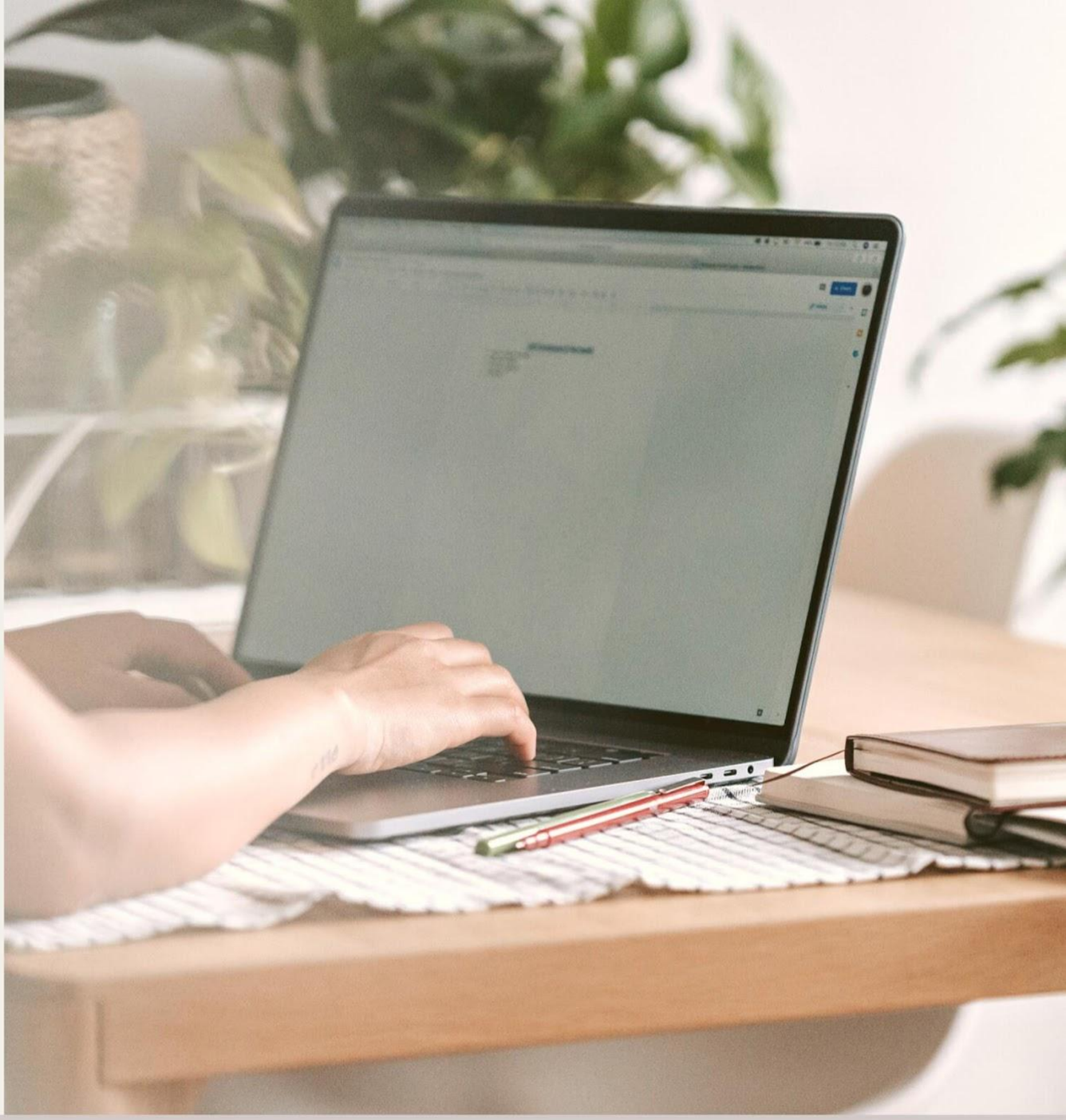


- 목표
  - 이커머스 도메인과 추천시스템에 학습
  - 히스토리 데이터의 분석을 통한 액션아이템 도출

- 방향성
  - 전환률과 유저 방문 수 등을 통한 매출의 증대 도모

2. 프로젝트 타임라인

1. 팀 구성 및 역할
2. 프로젝트 수행 절차



1. 팀 구성 및 역할

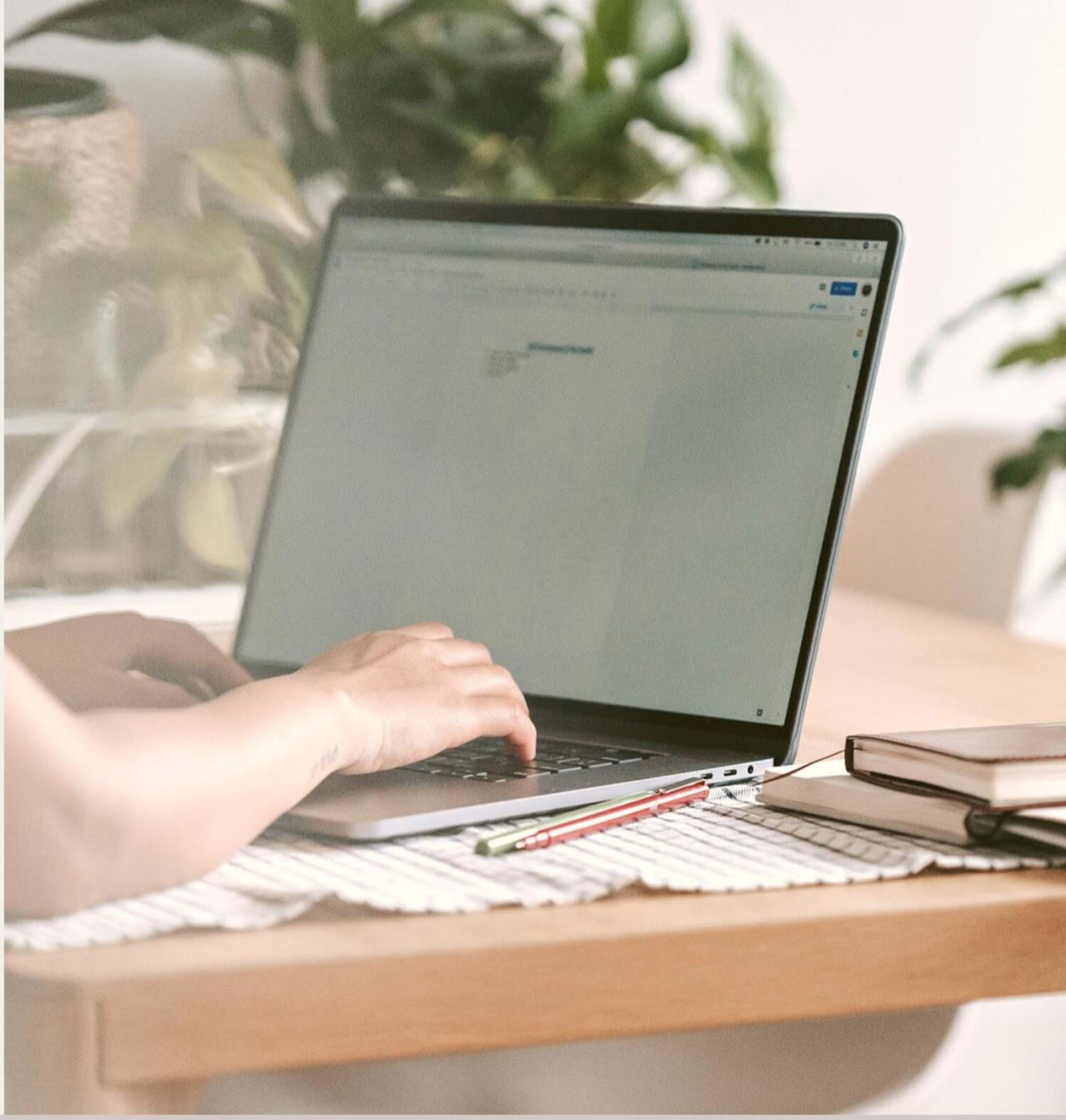
팀원	역할
한종준	• 데이터 분석 • 데이터 정제 • 모델링 • 모델을 통한 추천 알고리즘 제작
김혜관	• 데이터 분석 • 데이터 정제 • 모델링 • 모델을 통한 추천 알고리즘 제작

2. 프로젝트 수행 절차

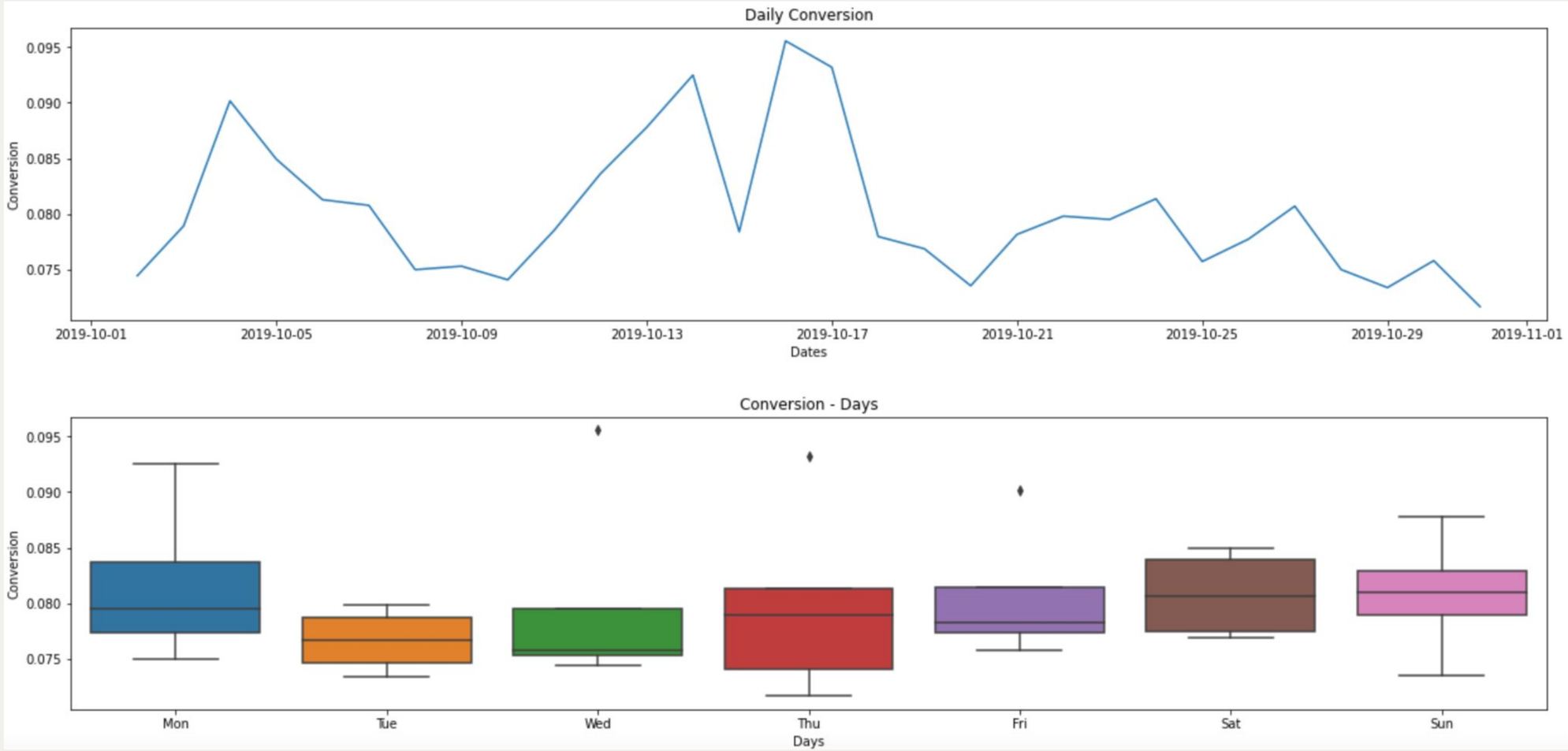
구분	기간	활동	비고
도메인 지식 학습	➜ 9/ 16 ~ 9/ 18	➜ 도메인 지식에 대한 학습과 분석 방향성 선정	이커머스 도메인에 관한 학습
데이터 분석	➜ 9/ 19 ~ 9/ 25	➜ 분석을 통한 중요 사항 파악 ➜ 액션 아이템 도출	주간보고
추천 시스템 학습	➜ 9/ 26 ~ 9/ 28	➜ CB / CF / Hybrid	오피스아워
데이터 전처리	➜ 9/ 29 ~ 10/ 3	➜ sparse matrix 데이터 제작을 위한 전처리	주간보고
추천 시스템 모델링	➜ 10/ 4 ~ 10/ 8	➜ ALS / LightFM 모델 제작	오피스아워
추천 알고리즘 제작	➜10/ 9 ~ 10/ 12	➜ 모델을 통한 추천 알고리즘 제작	최종보고

3. 프로젝트 세부 내용

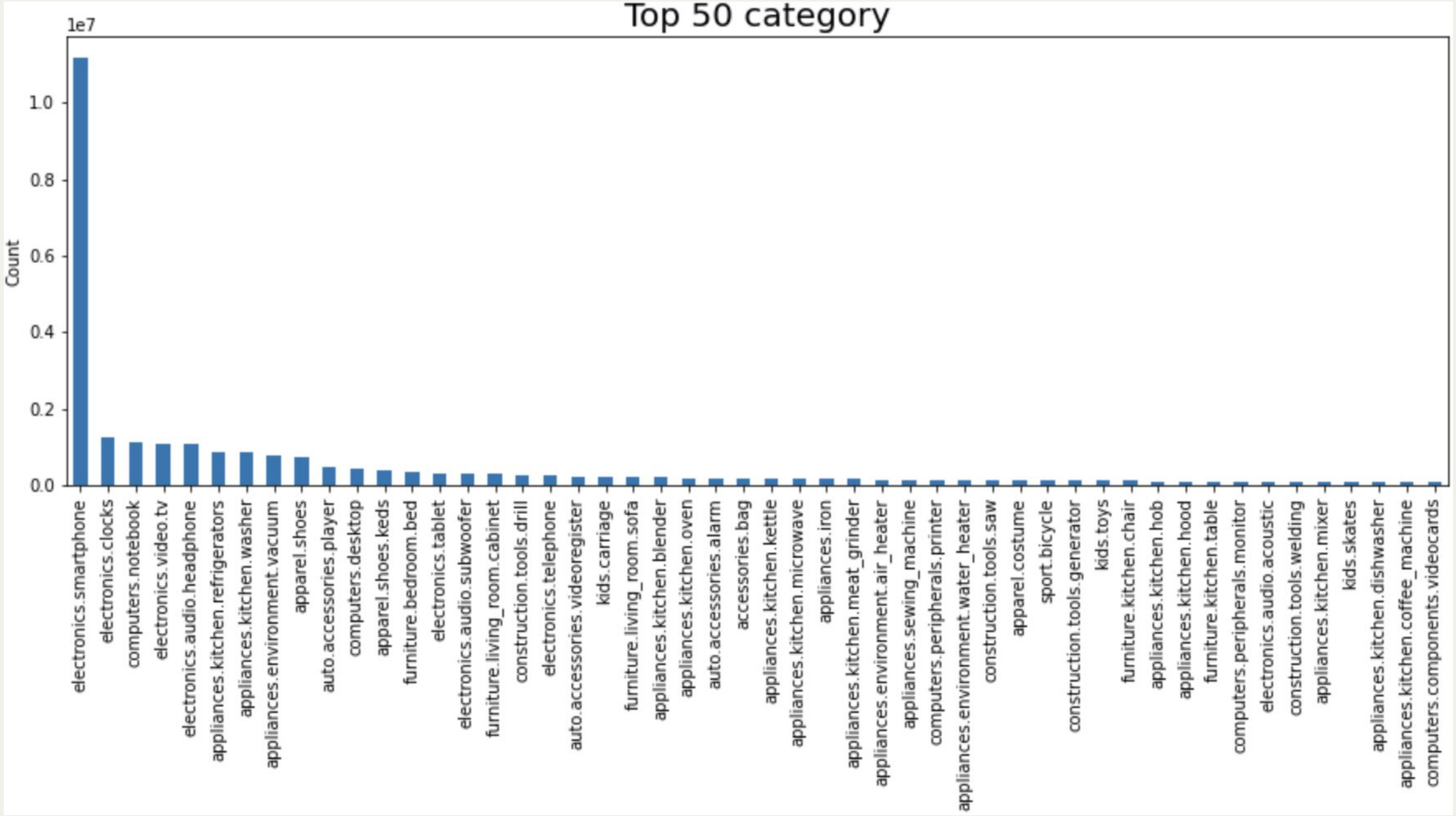
1. 데이터 분석
2. 데이터 전처리 및 모델 선정 이유
3. 모델링 내용 및 추천 알고리즘 결과



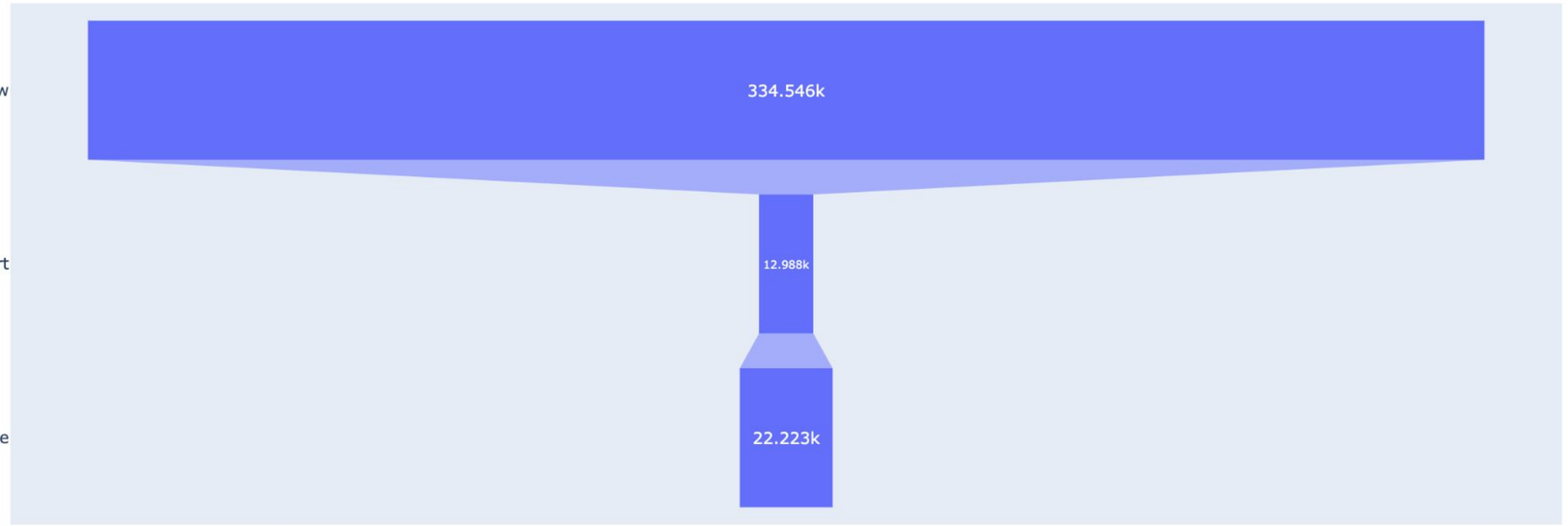
1. 데이터 분석

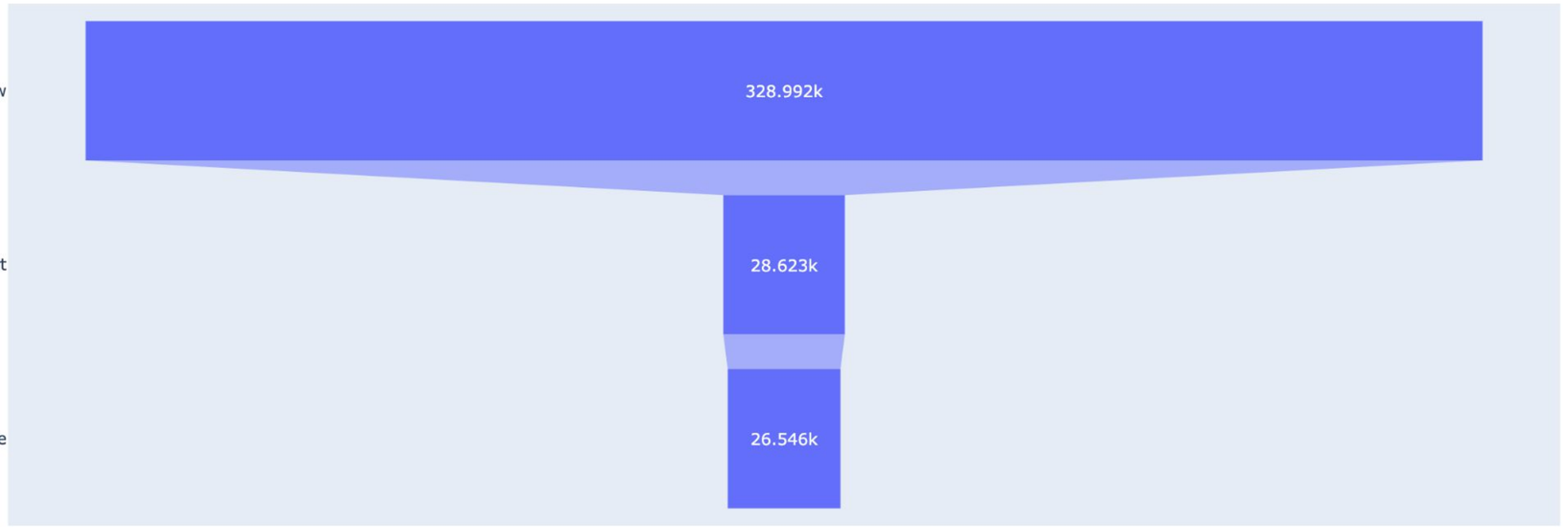


전환률에 대한 일별 분석


Top 50 category
1e7
Count
1.0
0.8
0.6
0.4
0.2
0.0
electronics.smartphone
electronics.clocks
computers.notebook
electronics.video.tv
electronics.audio.headphone
appliances.kitchen.refrigerators
appliances.kitchen.washer
appliances.environment.vacuum
apparel.shoes
auto.accessories.player
computers.desktop
apparel.shoes.keds
furniture.bedroom.bed
electronics.tablet
electronics.audio.subwoofer
furniture.living_room.cabinet
construction.tools.drill
electronics.telephone
auto.accessories.videoregister
kids.carriage
furniture.living_room.sofa
appliances.kitchen.blender
appliances.kitchen.oven
auto.accessories.alarm
accessories.bag
appliances.kitchen.kettle
appliances.kitchen.microwave
appliances.iron
appliances.kitchen.meat_grinder
appliances.environment.air_heater
appliances.sewing_machine
computers.peripherals.printer
appliances.environment.water_heater
construction.tools.saw
apparel.costume
sport.bicycle
construction.tools.generator
kids.toys
furniture.kitchen.chair
appliances.kitchen.hob
appliances.kitchen.hood
furniture.kitchen.table
computers.peripherals.monitor
electronics.audio.acoustic
construction.tools.welding
appliances.kitchen.mixer
kids.skates
appliances.kitchen.dishwasher
appliances.kitchen.coffee_machine
computers.components.videocards

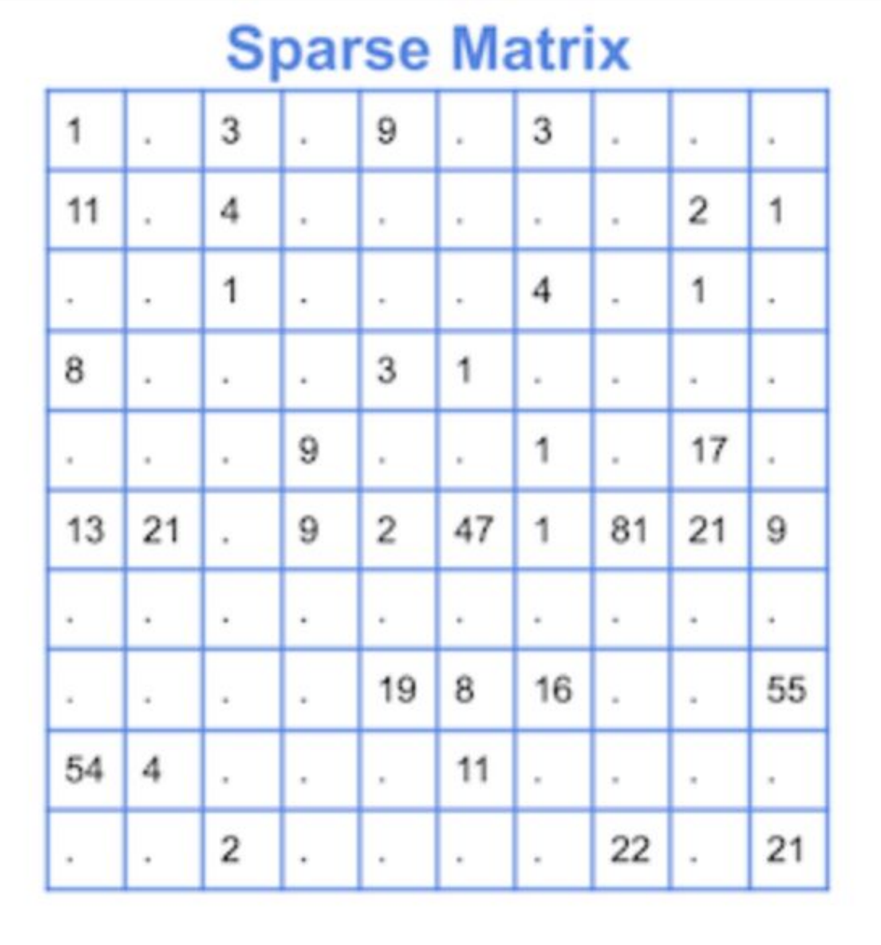
카테고리 선호도 조사





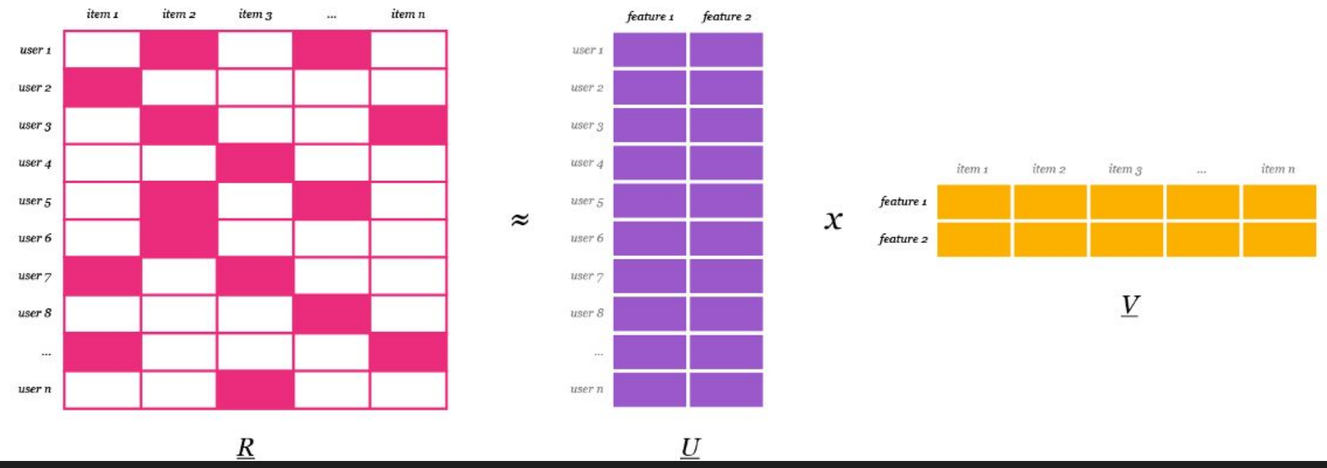
- 전환률 최대, 최소 날짜에 대한 Funnel 분석
- 유의미한 결과는 없다고 판단
- 결과적으로 카테고리 특성이 가장 의미있다 판단
- 카테고리에 대한 세션 로그 노출 빈도를 올리는 방향으로 액션아이템 도출

2. 데이터 전처리 및 모델 선정 이유



- 데이터 분석을 통한 event type 에 따른 가중치 조절
- 가격에 대한 라벨링 진행과 각 라벨 별 전환률에 비례한 가중치 조절
- 이러한 가중치들을 통한 Rating 생성
- Train, Test 셋으로 나눔으로써 Sparse Matrix에 Mask를 씌워 노이즈 생성
- Sparse Matrix 를 통한 분석을 위해 코사인 유사도를 통한 ALS 모델과 ALS 모델보다 복잡하고 정교한 선호도 개념의 알고리즘이 추가 된 LightFM 모델 선정
- 이 후에도 가중치에 대한 세세한 조절을 통한 모델 최적화

3. 모델링 내용 및 추천 알고리즘 결과



	category_code	brand	price
0	electronics.smartphone	samsung	197.43
1	electronics.smartphone	apple	735.05
2	electronics.video.tv	samsung	368.04
3	electronics.smartphone	apple	360.08
4	electronics.smartphone	samsung	92.64

```
evaluation.mean_average_precision_at_k(als_model, csr_train, csr_test, K = 3, show_progress = False, num_threads = 0)
```

```
0.010880683174413633
```

```
CPU times: user 4.5 ms, sys: 6.99 ms, total: 11.5 ms
Wall time: 5 ms
```

- 사용 모델 : ALS (베이스라인)
  - ALS 모델은 CF 기반의 모델
  - Implicit 한 Sparse Matrix 데이터를 사용
  - 알고리즘의 경우 비슷한 제품군을 잘 나타냄
  - 모델 속도는 빠른 편
  - 성능 자체는 0.011 정도의 낮은 정도
    - 이는 로그 자체가 하나뿐인 사용자들이 많기 때문

product rated by user

user id

$$\begin{pmatrix} 1.0 & 0 & 5.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.0 & 0 & 0 & 0 & 0 & 11.0 & 0 \\ 0 & 0 & 0 & 0 & 9.0 & 0 & 0 & 0 \\ 0 & 0 & 6.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7.0 & 0 & 0 & 0 & 0 \\ 2.0 & 0 & 0 & 0 & 0 & 10.0 & 0 & 0 \\ 0 & 0 & 0 & 8.0 & 0 & 0 & 0 & 0 \\ 0 & 4.0 & 0 & 0 & 0 & 0 & 0 & 12.0 \end{pmatrix}$$

	category_code	brand	price
0	electronics.smartphone	apple	975.57
1	electronics.smartphone	samsung	130.76
2	electronics.smartphone	samsung	254.82
3	electronics.smartphone	apple	1415.48
4	electronics.smartphone	apple	464.13

```
print('Test precision at k={}:\t{:.4f}'.format(k, precision_at_k(light_model, csr_test1, k=k).mean()))
```

```
Test precision at k=3:  0.0498
```

```
CPU times: user 16.5 ms, sys: 2.75 ms, total: 19.3 ms
Wall time: 22.5 ms
```

- 사용 모델 : LightFM
- 추천 알고리즘의 경우 같은 카테고리의 비슷한 형태로 잘 나타남
- 모델 속도의 경우 빠른편
- 성능 평가 결과 0.05에 가까운 성능
  - 이는 ALS 모델과 대비해 더 좋은 상태
  - 이 역시 수치가 낮은 편이라고 볼 수는 없음

회고

김혜관

- 가설을 수립에 필요한 이커머스 도메인 지식의 중요성을 배움
- 현재 데이터의 고객 정보에 대한 아쉬움
- 실제로는 추천 시스템을 적용한 후의 데이터를 수집하고, 해당 결과를 바탕으로 추천 알고리즘에 사용되는 가중치를 조절하는 과정을 거쳐 추천 시스템 성능을 개선할 수 있을 것으로 보임

한종준

- 현재 데이터 셋의 경우 기간이 한 달로 짧다고 느껴진다
- 기간이 긴 히스토리 데이터를 통해 분석을 할 경우 더 많은 방향성이 제시될 것이라 예상
- 추가적으로 딥러닝 모델을 학습하고 모델링하고자 했으나 실행하지 못해 아쉽다