igaworks

BIG DATA COMPETITION

권다희 민형규 이해환

고객

개요

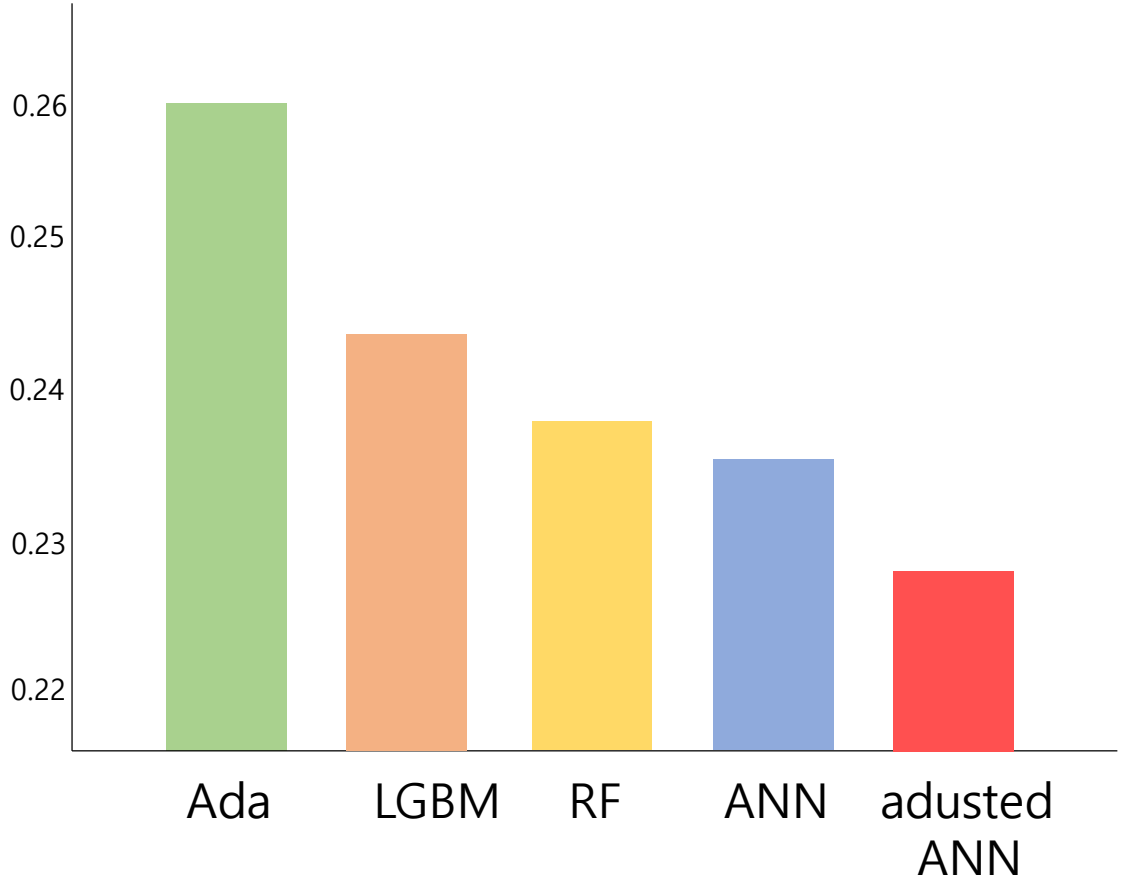
✓ 최종 Log Loss: **0.228**

✓ 사용 알고리즘:

Random Forest

✓ 사용 데이터:

Audience_Profile, Train



사용변수

<Train>

<Audience_Profile>

Age

Marry

Gender

Install_Pack

Cate_Code

Predicted House Price

Asset Index

Train

문제점: 모두 명목형 변수인데 하나의 변수로 들어가 있음 -> one-hot encoding으로 shape 변환
-> 이렇게 되면 변수의 개수가 약 30000개로 증가함.
-> PCA를 통해 변수 개수를 줄이면서도 정보손실을 최소화하도록 함.

-> 총 62개의 변수 생성

Audience_Profile

-Asset, house price는 NaN값이 너무 많아서 (90%) 사용불가
-Install pack, Cate_code shape 변환 + 길쭉한거 캡쳐해서 두개 비교
Install pack는 상위 20%(거의 대부분의 사람이 쓰는 앱들)과 하위 20%(거의 안쓰는 앱들을 뺀 총 95개의 어플에 대해 clustering진행 + 박스플랏 시각자료
=> K-means를 사용할 수 없는 binary data이기 때문에, Kohonen SOM을 통하여 총 9개의 cluste형성. + Kohonen SOM이 어떤 건지 잘 알 수 있게, Unsupervised 설명 & R 그림 첨부
Cate_code는 응답자의 수가 90%가 넘는 13개의 항목들만 추출하여 사용 (문항의 대표성및 결과의 신빙성을 위해)
=추정된 나이, 성별, 결혼 여부는 dummy화하여 사용

Who?

광고가 뜨면 무조건 누르는 I씨 

광고는 무조건 보지 않는 G씨 

관심 있는 광고만 보는 A씨 

[딥러닝 적용모델]

목차

- Aud 전처리
- 모델 스트럭쳐
- 결과

Audience 전처리

1. audience 중에 train, test와 겹치는 것만 분석(81만)

2. 사용할 변수와 그렇지 않을 변수를 선별

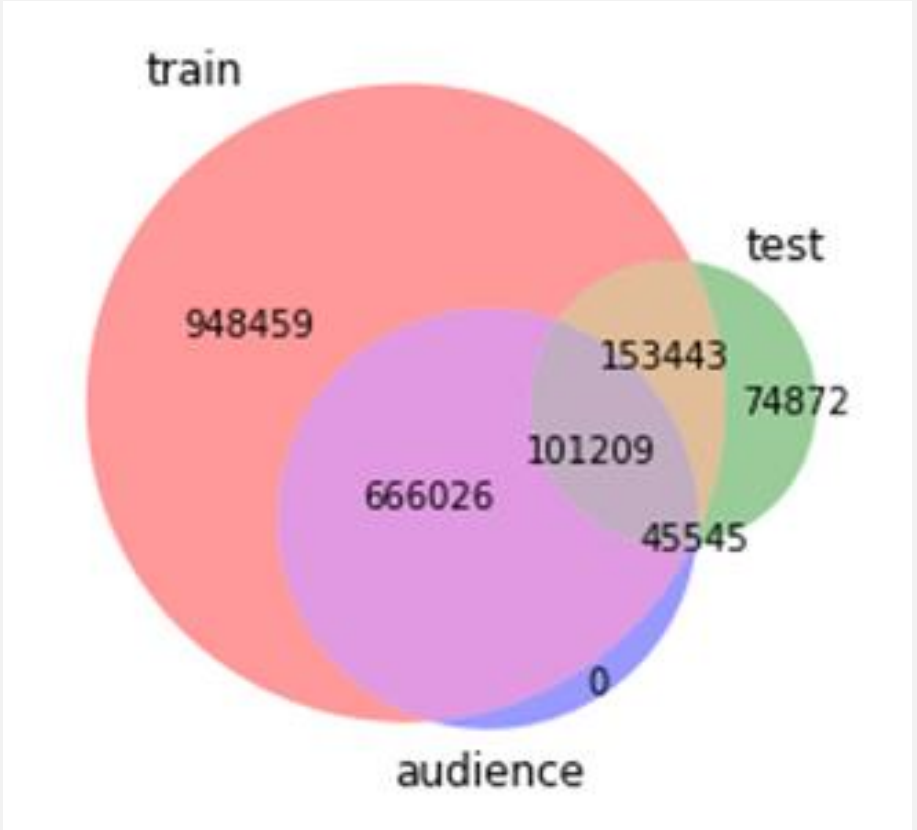
3. 사용하지 않는 변수는 다음과 같음

    1. install :
    2. house : Nan 값이 많음

4. 사용한 변수는 다음과 같이 변형

    1. age, gender, marry : 더미화
    2. cate_code : (1) SOM
                   (2) 더미화

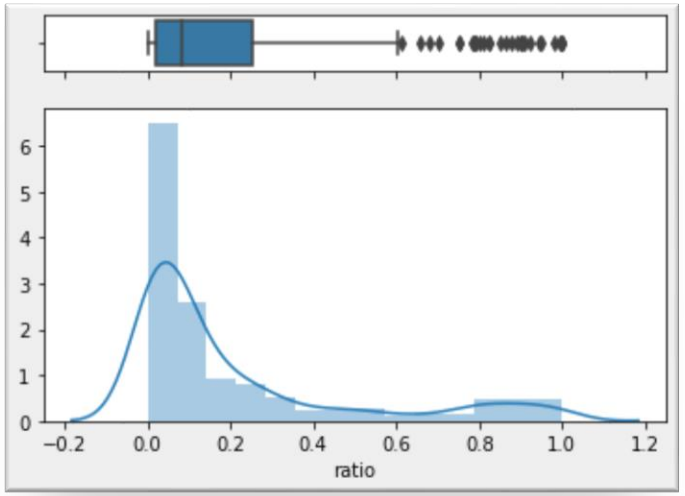
cate_code는 두 가지 중 성능이 좋은 것을 이용
이를 위해서 cate_code EDA를 진행함.

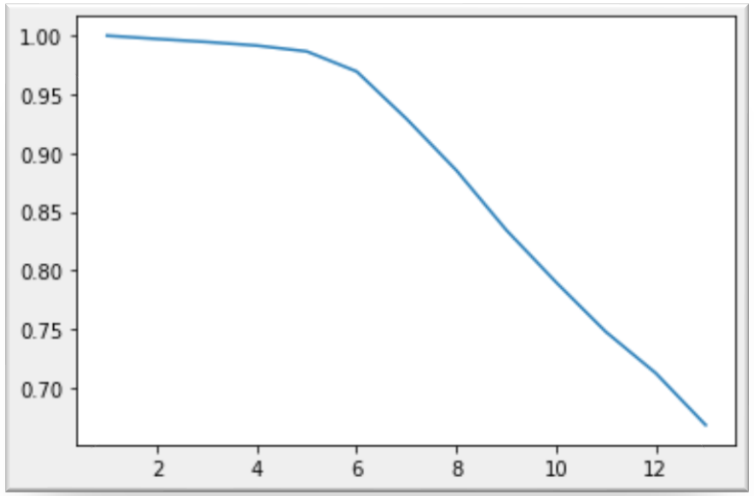


[딥러닝][전처리][audience]

사용 : age, gender, marry, cate_code

미사용 : install_pack, predicted_house





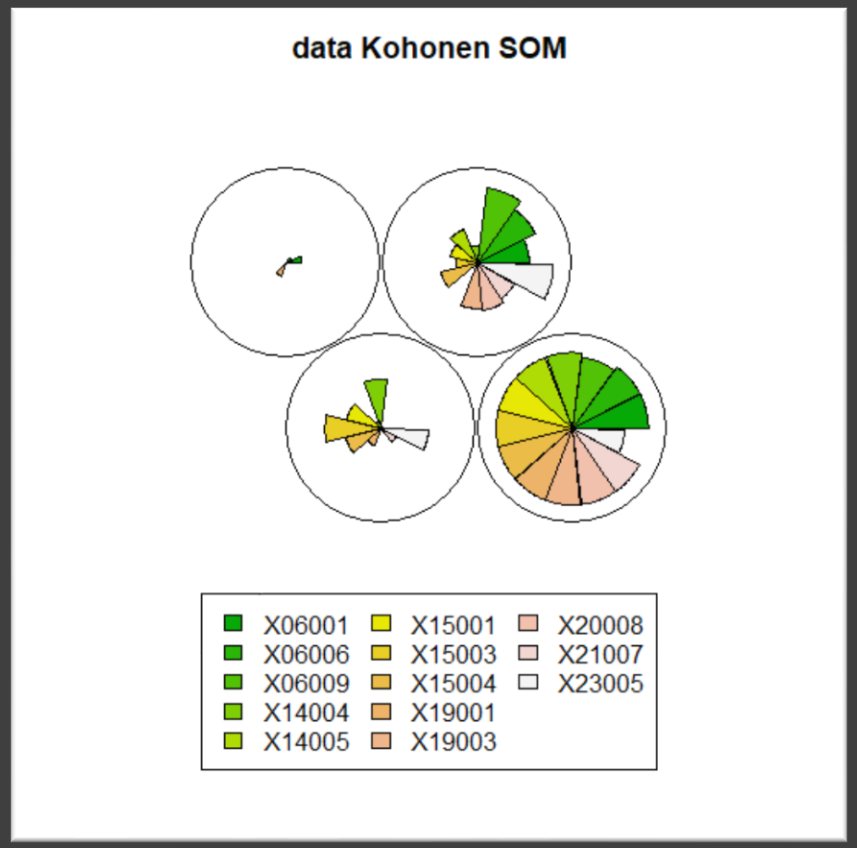
	freq	ratio
14004	812687	0.999886
20008	810422	0.997099
15003	810370	0.997035
19001	810036	0.996624
15004	807806	0.99388
21007	797286	0.980937
6009	769990	0.947354
15001	768317	0.945295
19003	751033	0.92403
6006	738755	0.908924
23005	733217	0.90211
14005	732129	0.900771
6001	731895	0.900484
21001	724646	0.891565

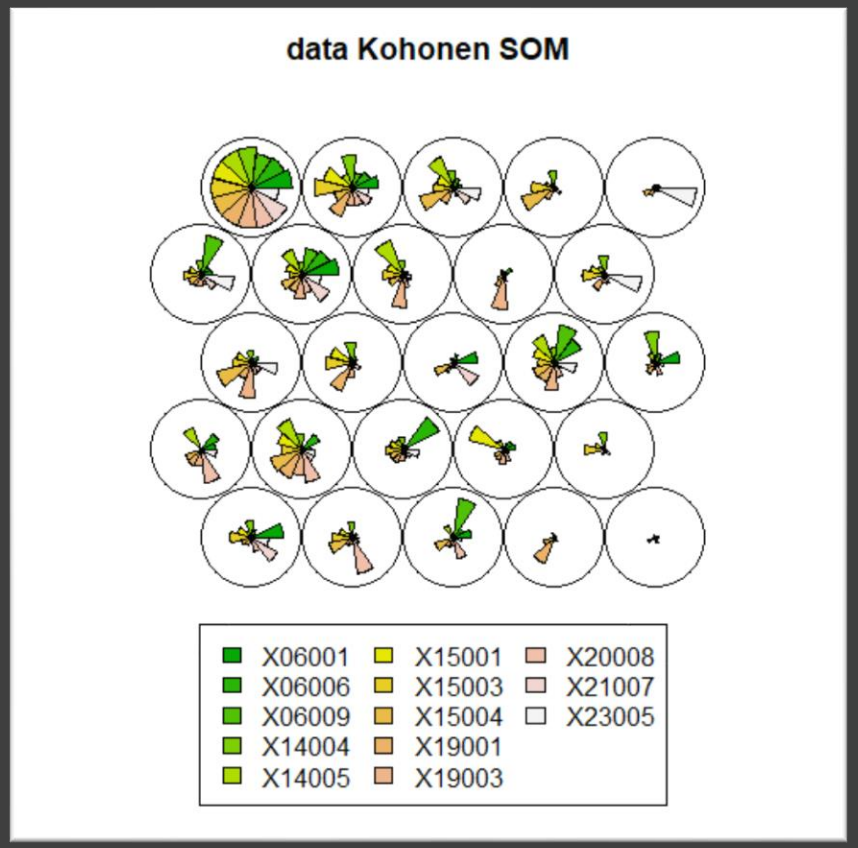
[그림1] 전체 241개의 질문을 받았으며, 그 중 13개는 90% 이상 사용되었다.

[그림2] 많이 등장한 13개의 질문을 상위 순서대로 묶어서 모두 가지고 있는 비율을 살펴보았다.

13개를 피쳐로 쓰면 missing value가 없는 데이터가 불과 65%에 불과하게 된다. (54만)

이 경우에는 기타 27만의 데이터를 새로운 피쳐로 쓰거나 아예 버려야한다.





Kohonen SOM

전체 13개의 피쳐를 사용했기 때문에 54만의 데이터를 가지고 비지도학습을 실행한 결과이다.

[그림1] 4가지 그룹으로 구별한 경우에 해당한다.

[그림2] 25가지 그룹으로 구별한 경우에 해당한다.

Kohonen SOM을 쓰지 않은 이유

1. **Hyper Parameter가 너무 많음.**
    1. "몇 개의 피쳐"를 쓸지 의외에도
    2. "몇 개의 그룹"으로 나눌지까지 생각해야함

2. **Missing value가 있으면, 전혀 다른 카데고리로 분류할 수 밖에 없음.**
    1. 예컨데, 13개 피쳐를 쓸 때 12개 피쳐가 매우 유사한 형태가 있어도 하나만 값이 없어도 전혀 다른 그룹으로 분류가 된다.

3. **Distance 활용 불가**
    1. 각 그룹의 중심으로부터의 거리가 SOM의 특징이지만 이를 활용할 방법이 없음.

	06001	06006	06009	14004	14005	15001	15003	15004	19001	19003	20008	21007	23005
1648210	2	5	2	2	3	4	3	1	2	5	5	3	1
1885479	3	2	3	2	1	1	2	3	2	2	2	4	NaN
1289369	5	2	3	3	2	NaN	4	3	4	1	4	3	5
1415766	2	1	1	3	4	1	4	3	2	2	3	1	5
1928794	4	3	4	1	1	2	1	2	1	5	3	3	5

test	train	test
13	0.24058636286549925	0.24258105381618497
12	0.24129212869389438	0.24161603005799437
11	**0.23800112232660822**	**0.2373734540745832**
10	0.24128660627206772	0.2403213846601593
9	0.24275661014081618	0.24322728316745645
8	0.2435498267927994	0.2440606937400578
7	0.24148697476776687	0.2416347923562318
6	0.2414870190261766	0.2422810056710189
5	0.24510329956492216	0.2472448158064689
4	0.24167016636497124	0.24451112102147018
3	0.2428716311710037	0.24435146753371553
2	0.2406817530859671	0.24245889241564506
1	0.24213330259006352	0.24200724095360607

클러스터링 없이 딥러닝 결과

SOM을 활용하지 않고, 13개의 그룹화된 피쳐를 그대로 집어넣은 경우 결과이다.

[그림1] 13개 피쳐에 대해서 전체 데이터셋이 응답한 값을 dataframe으로 만든 것
[그림2] missing value를 제외한 각각의 피쳐그룹을 딥러닝 트레이닝을 시킨 결과

모델 구조

1. **Audience와 겹치는 데이터 :**
    1. 11개의 피쳐를 사용하여 train + audience 데이터를 모델피팅
    2. 모델은 4개의 layer, 각각 150개의 node를 사용

2. **Audience와 겹치지 않는 데이터 :**
    1. 11개의 피쳐를 모두 가지고 있지 않은 데이터는 사용하지 않음.
    2. 모델 구조는 위와 동일하나, node 개수를 160개로 늘림.

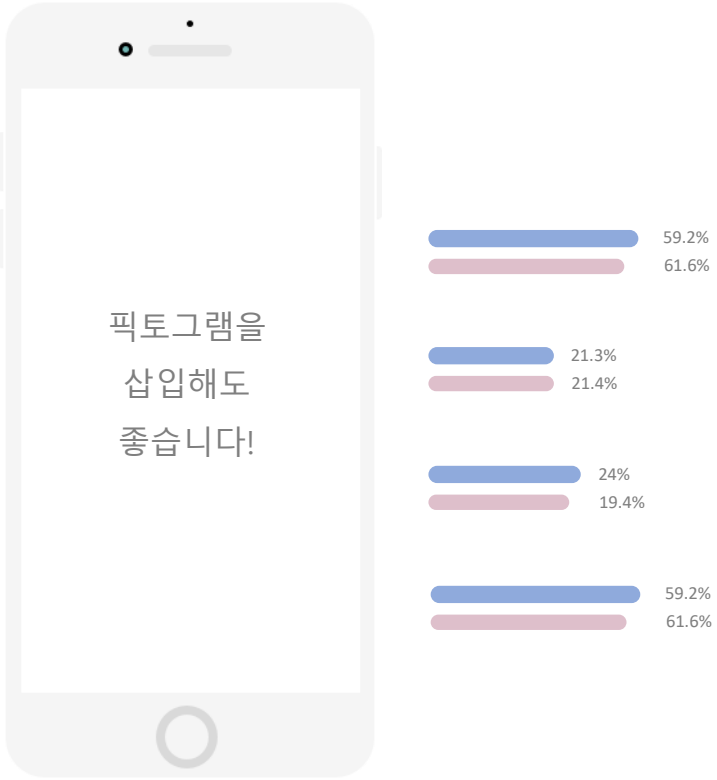
최종 결과

1. Audience 데이터를 사용하면 모델의 성능이 더 좋아짐.
2. 그러나 둘 다 딥러닝을 사용할 경우보다 뒤에 나오는 앙상블이 더 좋음.
3. 따라서 앙상블 모델에다가 audience 데이터만 갖다 붙이는 것을 추천.
4. **혹은 임베딩 레이어를 사용했어야 했다.**

01

대제목을 입력해 주세요

대제목을 입력해 주세요



강조할 내용을 입력해 주세요

위의 표 혹은 그래프를 설명할 수 있는 내용을 입력해 주세요.
이때 주의할 점은, 줄글 형식 보다는 두괄식이 좋다는 점과
하이라이트를 해 두면 더욱 눈에 띈다는 사실을 기억해 두는 것 입니다!

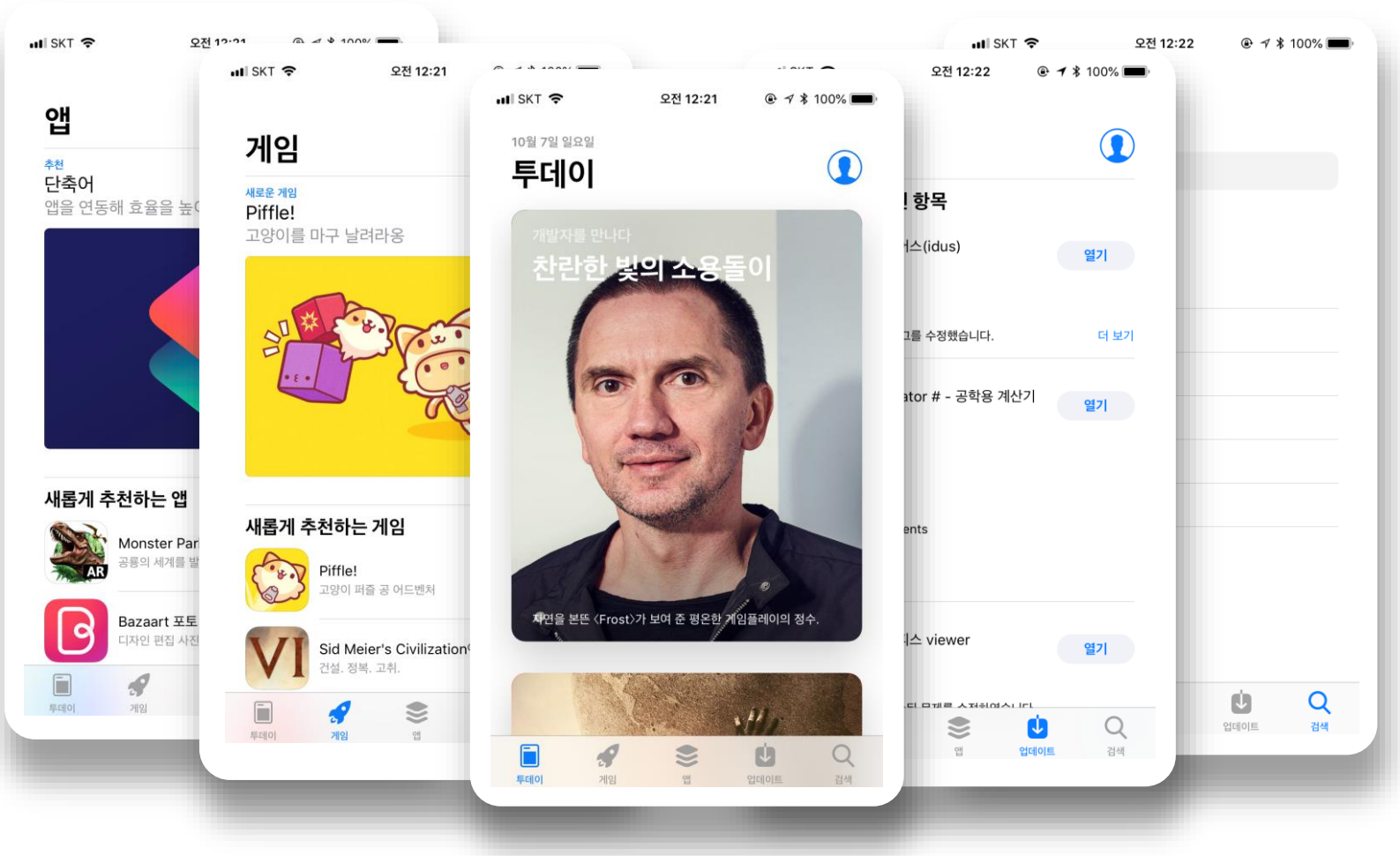
강조할 내용을 입력해 주세요

위의 표 혹은 그래프를 설명할 수 있는 내용을 입력해 주세요.
이때 주의할 점은, 줄글 형식 보다는 두괄식이 좋다는 점과
하이라이트를 해 두면 더욱 눈에 띈다는 사실을 기억해 두는 것 입니다!

03

대제목을 입력해 주세요

대제목을 입력해 주세요



강조할 내용을 입력해 주세요

앱을 소개하기에 적절한 페이지

앱을 소개하기에 적절한 페이지

앱을 소개하기에 적절한 페이지

04

대제목을 입력해 주세요

대제목을 입력해 주세요

결론 1 입력하기

최종 마무리 하는 결론이나 관련 내용에 대해서 입력해 주세요.

최종 마무리 하는 결론이나 관련 내용에 대해서 입력해 주세요.

결론 2 입력하기

최종 마무리 하는 결론이나 관련 내용에 대해서 입력해 주세요.

최종 마무리 하는 결론이나 관련 내용에 대해서 입력해 주세요.

결론 3 입력하기

최종 마무리 하는 결론이나 관련 내용에 대해서 입력해 주세요.

최종 마무리 하는 결론이나 관련 내용에 대해서 입력해 주세요.

강조할 내용을 입력해 주세요

감사합니다.