번개장터 CTR을 높이기 위한 광고 추천 알고리즘 개발

SSAC 6조
문영주, 백승재, 손희서, 조지민

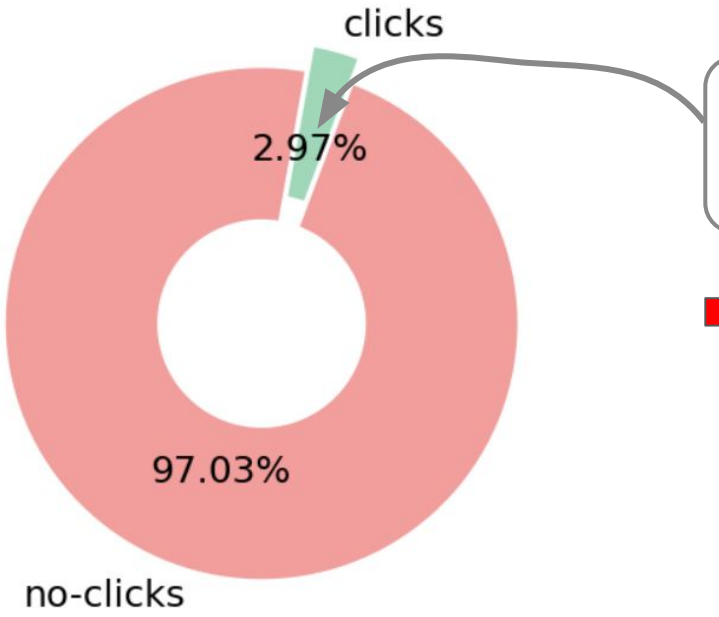
목차 Table of Contents

1 문제정의

CTR(Click-through rate, 클릭률)

- 온라인 광고의 노출횟수 대비 클릭 수를 의미
- 광고가 노출된 횟수(Impression) 중, 실제 클릭을 통해 이동한 경우의 비율
- CTR = (클릭 수 / 노출된 횟수) × 100
- 온라인 광고 효과를 측정하는 데 있어 CTR은 중요 지표
- 온라인 광고시장 평균 CTR 약 0.3%

출처 : [네이버 지식백과] 클릭률 [Click-through rate] (ICT 시사상식 2015, 2014.12.31)

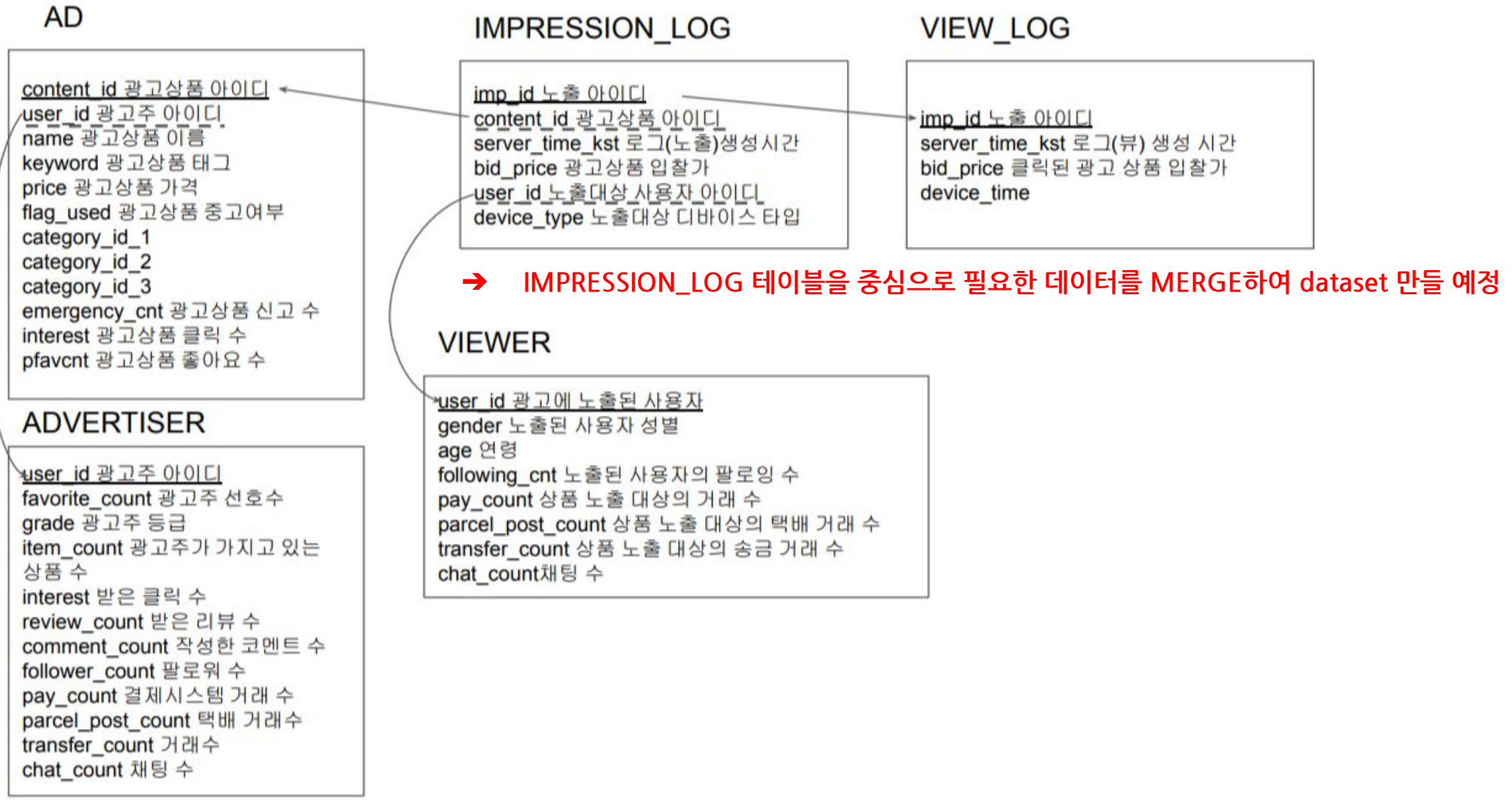


2021년 8월 31일 impression_log의
imp_id 중복 값을 제거한 후
impression(노출) 대비 view(광고클릭) 비율

**CTR을 예측하고,
CTR 비율을 높이는데 도움 되도록
알맞는 viewer에게 알맞은 content를
추천하는 알고리즘 모델링 개발.**

2 데이터정의

데이터 관계도



➔ IMPRESSION_LOG 테이블을 중심으로 필요한 데이터를 MERGE하여 dataset 만들 예정

2 데이터정의

INPUT DATA

{ 개인고객 정보 }

	col_name	설명
1	'user_id'	광고에 노출된 대상의 아이디
2	'gender'	상품 노출 대상의 성별
3	'age'	상품 노출 대상의 나이
4	'following_cnt'	상품 노출 대상의 팔로잉 수
5	'pay_count'	상품 노출 대상의 거래수
6	'parcel_post_count'	상품 노출 대상의 택배 거래수
7	'transfer_count'	상품 노출 대상의 송금 거래수
8	'chat_count'	상품 노출 대상의 채팅수

{ 광고상품 정보 }

	col_name	설명
1	'content_id'	광고 상품 아이디
2	'user_id'	광고주 아이디
3	'name'	광고 상품 이름
4	'keyword'	광고 상품 태그
5	'price'	광고 상품 가격
6	'flag_used'	광고 상품 중고 여부
7	'category_id_1'	광고 상품 1차 카테고리
8	'category_id_2'	광고 상품 2차 카테고리
9	'category_id_3'	광고 상품 3차 카테고리
10	'emergency_cnt'	광고 상품 신고수
11	'comment_cnt'	광고 상품에 달린 코멘트수
12	'interest'	광고 상품 클릭수
13	'pfavcnt'	광고 상품 좋아요 수

{ 광고주 정보 }

	col_name	설명
1	'user_id'	광고주 아이디
2	'favorite_count'	광고주 선호수
3	'grade'	광고주 등급
4	'item_count'	광고주가 가지고 있는 상품수
5	'interest'	광고주가 받은 클릭수
6	'review_count'	광고주가 받은 리뷰수
7	comment_count	광고주가 작성한 코멘트 수
8	'follower_count	광고주의 팔로워수
9	'pay_count'	광고주의 결제거래시스템 거래수
10	'parcel_post_count''	광고주의 택배 거래수
11	'transfer_count'	광고주의 거래수
12	'chat_count'	광고주의 채팅수

2 데이터정의

INPUT DATA 클릭 정보 데이터 : IMPRESSION_LOG& VIEW_LOG 테이블을 imp_id를 기준으로 merge

	imp_id	content_id	server_time_kst_x	bid_price_x	user_id	device_type_x	server_time_kst_y	bid_price_y	device_type_y	click_time	click	label
851020	9919612e429724183042	152627901	2021-08-31 23:54:24.948000+09:00	70	9140358	a	2021-08-31 23:54:31.027000+09:00	70.0	a	0 days 00:00:06.079000	1	
851021	97f9612e4295183f6e39	160836652	2021-08-31 23:54:20.398000+09:00	165	1668136	a	NaT	NaN	NaN	NaT	0	
851022	998c612e426b027d43c3	140116266	2021-08-31 23:53:43.939000+09:00	85	6309266	a	NaT	NaN	NaN	NaT	0	

- 851,023 row
- 노출아이디, 상품 아이디, 노출시간, PPC 등
- 추가 변수
  - click_label : click 여부 (1 = 클릭, 0= 클릭x)
  - click_time : click까지 걸린 시간 (impression_log['server_time_kst_x'] - view_log['server_time_kst_y'])

3 모델링

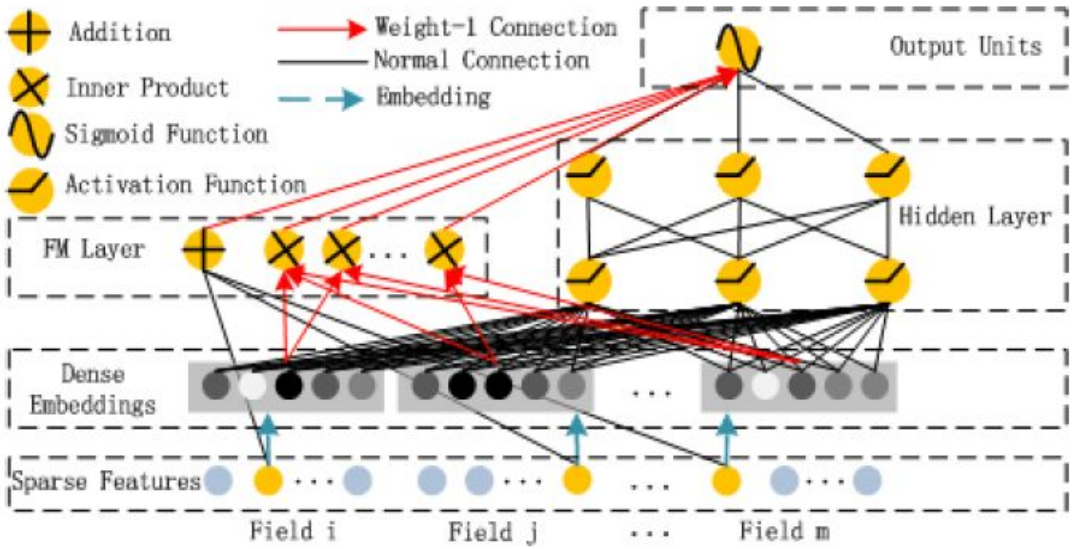
Main model : Deep FM

- Main ML models: **Deep FM**(NN, 인공신경망 계열)
  - pytorch 혹은 tensorflow 라이브러리 사용
  - **인공신경망 (Deep Learning) + Factorization Machine**

    ⇒ Factorization machine Model이 추천영역에서 갖는 강점 + Deep learning이 feature학습에서 갖는 강점
  - pre-training이 필요없고 별도의 feature engineering 이 필요없음.
  - AUC와 logloss면에서 여타 최신 모델들보다 뛰어난 성능을 보였고 효율성 측면에서도 우수함.
  - low, high order feature interaction을 모두 표현.

Deep FM = Factorization Machine + Deep Learning

low-order feature interaction / high-order feature interaction



참고 및 출처: https://leehyejin91.github.io/post-deepfm/

3 모델링

성능향상을 위한 추가 고려 모델

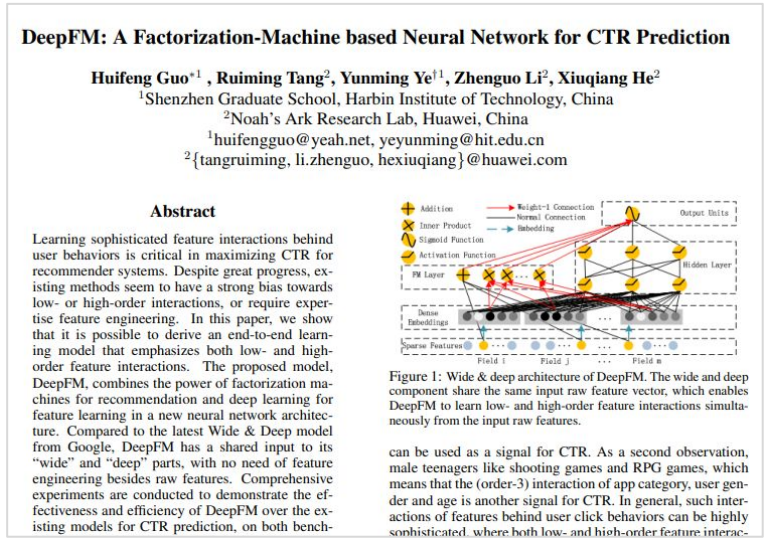
- **Machine Learning**:
  - 분류모델 평가지표: **ROC-AUC, confusion matrix, feature importance**
  - 여러 머신러닝 모델링 후 **평가 성능이 높은 모델**을 주 사용
    1. **Ensemble (Classifier)**:
       - CTR에 따른 상품별 추천 여부 분류 및 user_id 당 CTR 예측
       - LightGBM & XGBoost
    2. **Clustering**:
       - 고객의 상품 선호도 판단 및 고객 집단 분류 및 어떤 상품이 중고거래에 많이 이용되는지를 파악
       - KNN & Fuzzy Clustering각 객체가 어느 군집에 속할지를 가중치(weight)나 확률(probability)로 정도를 나타내주며 데이터와 모든 클러스터링 간의 높은 유사성을 표현해주는 기법
    3. **Naïve Bayes**:
       - CTR의 조건부 확률값 계산

- **Text Analysis/Image Analysis:**
  - **LDA** / image classification (**CNN-CIFAR10**)
  - ad 데이터의 텍스트 컬럼, url 컬럼의 데이터들을 활용하여 텍스트 분석/이미지 분석에 이용
  - 텍스트 분석의 시각화: LDA(형태소 분석 및 토픽모델링) 시각화 & 워드클라우드

3 모델링

References

논문



DeepFM: A Factorization-Machine based Neural Network for CTR Prediction

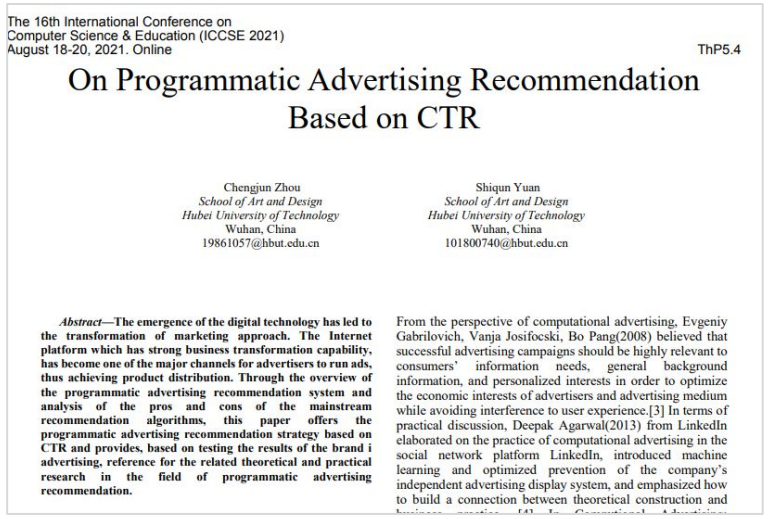
Huifeng Guo*[1] , Ruiming Tang[2], Yunming Ye†[1], Zhenguo Li[2], Xiuqiang He[2]
[1]Shenzhen Graduate School, Harbin Institute of Technology, China
[2]Noah's Ark Research Lab, Huawei, China
[1]huifengguo@yeah.net, yeyunming@hit.edu.cn
[2]{tangruiming, li.zhenguo, hexiuqiang}@huawei.com

Abstract

Learning sophisticated feature interactions behind user behaviors is critical in maximizing CTR for recommender systems. Despite great progress, existing methods seem to have a strong bias towards low- or high-order interactions, or require expertise feature engineering. In this paper, we show that it is possible to derive an end-to-end learning model that emphasizes both low- and high-order feature interactions. The proposed model, DeepFM, combines the power of factorization machines for recommendation and deep learning for feature learning in a new neural network architecture. Compared to the latest Wide & Deep model from Google, DeepFM has a shared input to its "wide" and "deep" parts, with no need of feature engineering besides raw features. Comprehensive experiments are conducted to demonstrate the effectiveness and efficiency of DeepFM over the existing models for CTR prediction, on both bench-

Figure 1: Wide & deep architecture of DeepFM. The wide and deep component share the same input raw feature vector, which enables DeepFM to learn low- and high-order feature interactions simultaneously from the input raw features.

can be used as a signal for CTR. As a second observation, male teenagers like shooting games and RPG games, which means that the (order-3) interaction of app category, user gender and age is another signal for CTR. In general, such interactions of features behind user click behaviors can be highly sophisticated, where both low- and high-order feature interac-



The 16th International Conference on Computer Science & Education (ICCSE 2021)
August 18-20, 2021. Online
ThP5.4

On Programmatic Advertising Recommendation Based on CTR

Chengjun Zhou
School of Art and Design
Hubei University of Technology
Wuhan, China
19861057@hbut.edu.cn

Shiqun Yuan
School of Art and Design
Hubei University of Technology
Wuhan, China
101800740@hbut.edu.cn

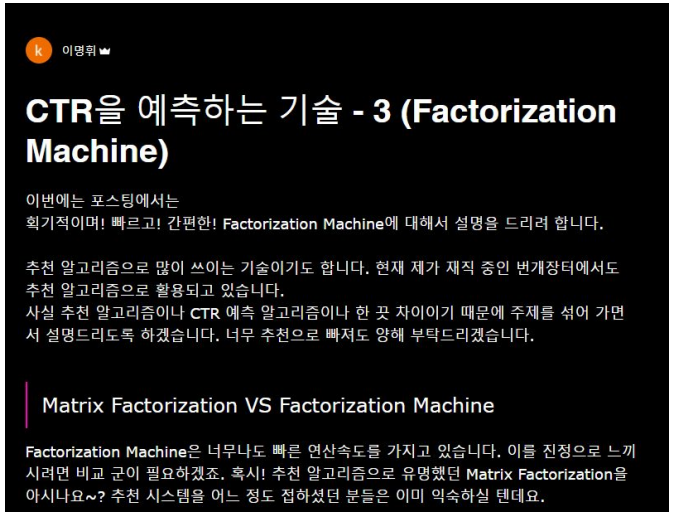
***Abstract*—The emergence of the digital technology has led to the transformation of marketing approach. The Internet platform which has strong business transformation capability, has become one of the major channels for advertisers to run ads, thus achieving product distribution. Through the overview of the programmatic advertising recommendation system and analysis of the pros and cons of the mainstream recommendation algorithms, this paper offers the programmatic advertising recommendation strategy based on CTR and provides, based on testing the results of the brand i advertising, reference for the related theoretical and practical research in the field of programmatic advertising recommendation.**

From the perspective of computational advertising, Evgeniy Gabrilovich, Vanja Josifocski, Bo Pang(2008) believed that successful advertising campaigns should be highly relevant to consumers' information needs, general background information, and personalized interests in order to optimize the economic interests of advertisers and advertising medium while avoiding interference to user experience.[3] In terms of practical discussion, Deepak Agarwal(2013) from LinkedIn elaborated on the practice of computational advertising in the social network platform LinkedIn, introduced machine learning and optimized prevention of the company's independent advertising display system, and emphasized how to build a connection between theoretical construction and

Kaggle.com

Github.com

Blog



이명휘

CTR을 예측하는 기술 - 3 (Factorization Machine)

이번에는 포스팅에서는
획기적이며! 빠르고! 간편한! Factorization Machine에 대해서 설명을 드리려 합니다.

추천 알고리즘으로 많이 쓰이는 기술이기도 합니다. 현재 제가 재직 중인 번개장터에서도 추천 알고리즘으로 활용되고 있습니다.
사실 추천 알고리즘이나 CTR 예측 알고리즘이나 한 끗 차이이기 때문에 주제를 섞어 가면서 설명드리도록 하겠습니다. 너무 추천으로 빠져도 양해 부탁드리겠습니다.

Matrix Factorization VS Factorization Machine

Factorization Machine은 너무나도 빠른 연산속도를 가지고 있습니다. 이를 진정으로 느끼시려면 비교 군이 필요하겠죠. 혹시! 추천 알고리즘으로 유명했던 Matrix Factorization을 아시나요~? 추천 시스템을 어느 정도 접하셨던 분들은 이미 익숙하실 텐데요.

DeepFM분야의 핵심 논문을 읽고 학습한 후
추가적으로 필요한 자료는 kaggle, github, blog를 참고한다.

BLOG출처 : https://www.kgsbdata.com/post /ctr%EC%9D%84-%EC%98%88%EC%B8%A1%ED%95%98%EB%8A%94-%EA%B8%B0%EC%88%A0-3-factorization-machine

3 모델링

Modeling OUTPUT

OUTPUT DATA

CTR을 예측하는데 있어서 user, ad item, 기타 광고 플랫폼 정보 등을 복합적으로 활용할 때
전통적인 추천시스템에 비해 더 효과적으로 CTR을 예측할 수 있다.

개인고객 정보 field				광고상품 정보 field			광고주 정보 field			OUTPUT
USER_ID	GENDER	AGE	...	CONTENT_ID	PRICE	...	USER_ID	GRADE	...	CLICK_LABEL
513808	2	48	...	157826057	250000	...	2487730	7673	...	1
10764680	1	25	...	162431268	439000	...	1462217	5520	...	0

$$\hat{y} = sigmoid(y_{FM} + y_{DNN})$$

$$x = [x_{field_1}, x_{field_2}, \ldots, x_{field_j}, \ldots, x_{field_m}]_{(1\times d)}$$
$$y \in \{0, 1\}$$

최종 OUTPUT : user당 광고를 클릭할 확률(0~1)

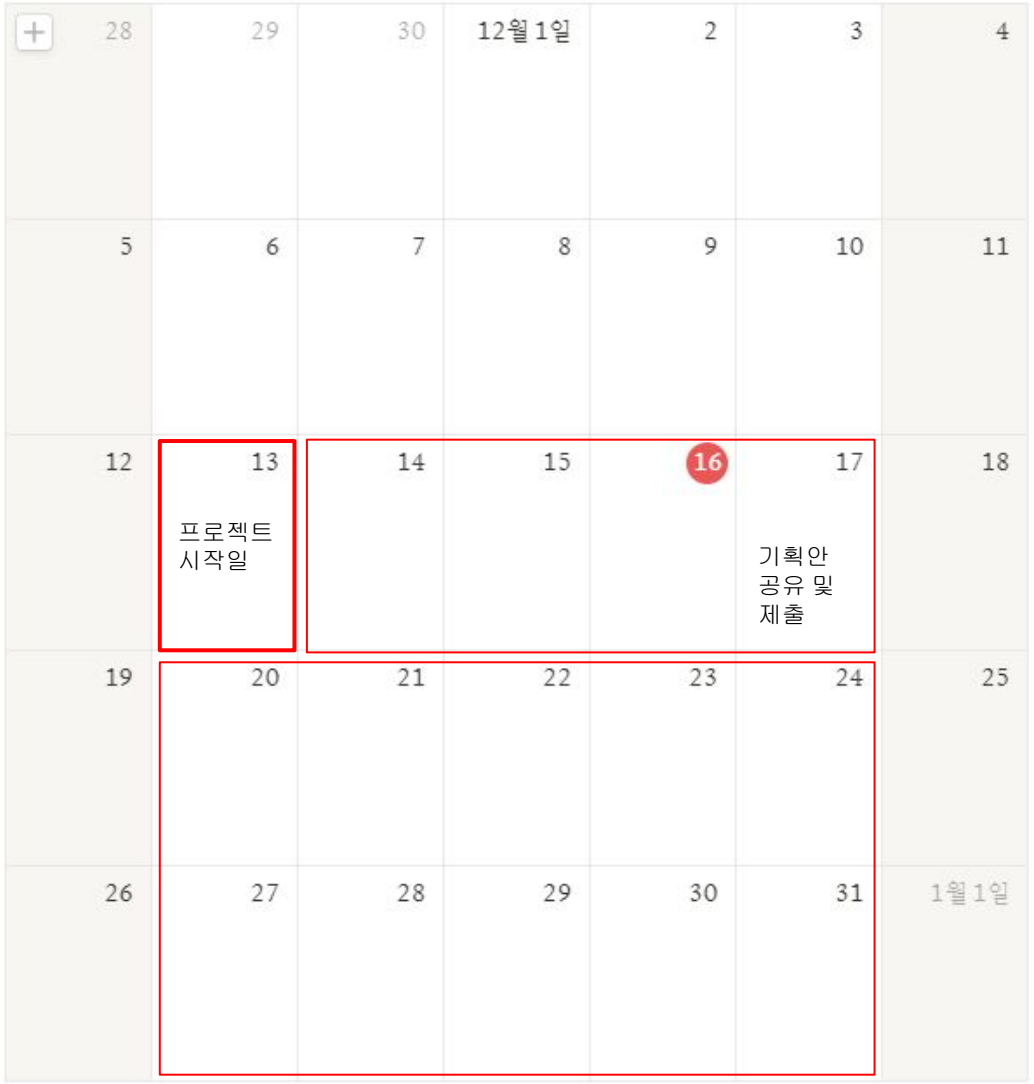
*programmatic advertising recommendation algorithm based on CTR 논문 참조 *그림 참조: https://leehyejin91.github.io/post-deepfm/

4 프로젝트 수행 환경

- 사용할 협업 툴
  - Slack : 멤버 & 강사님과 의사소통
  - GitHub : 프로젝트 코드 공유
  - Notion : 보고서 및 업무 분담 등 기록

- 수행 환경
  - Google Colab / Jupyter Notebook : 기본 환경
  - SQL, pytorch, tensorflow : 필요하다면 추가적으로 고려
  - Tableau : 데이터 시각화가 필요하다면 태블로를 활용

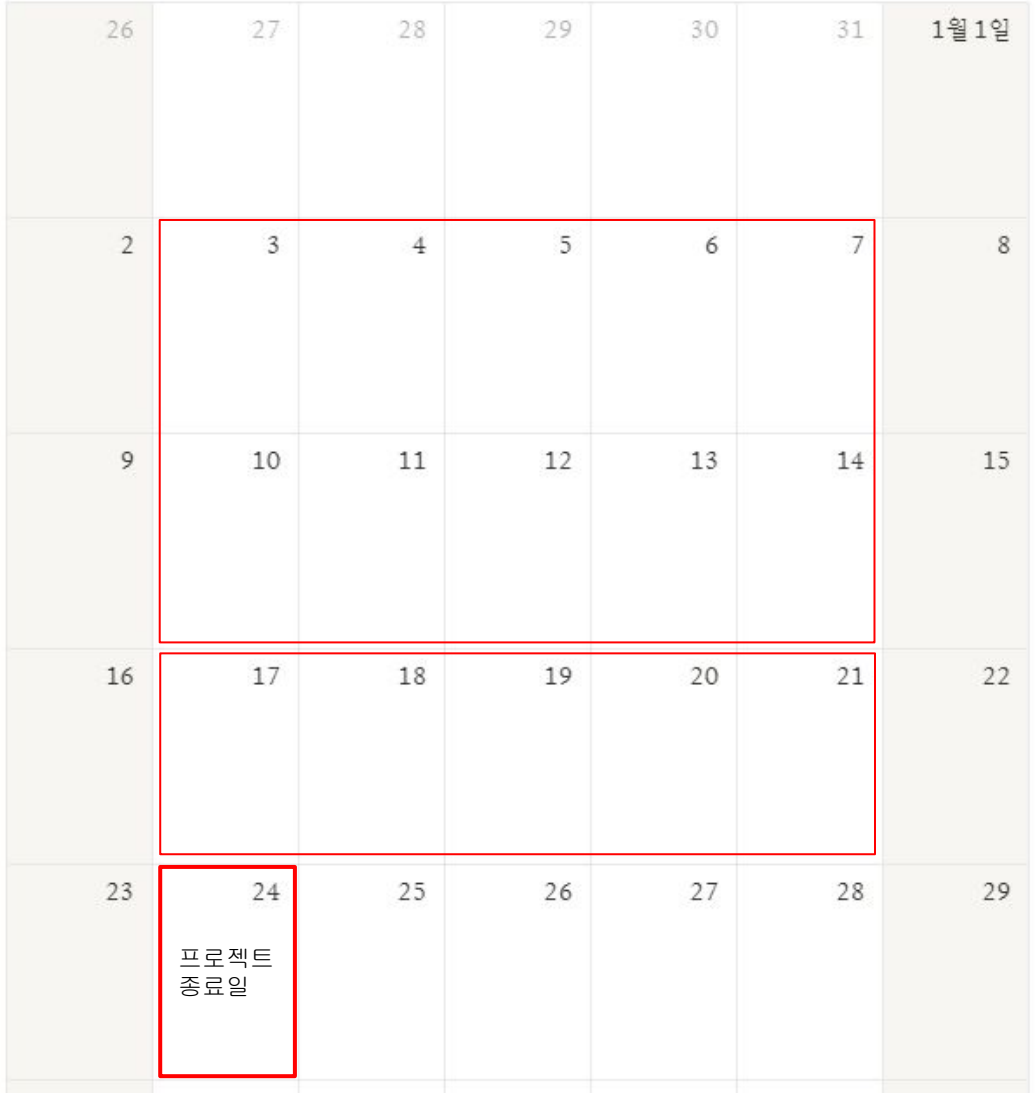
5 프로젝트 계획



[12/13 - 12/17] 데이터 탐색 및 기획서 작성

[12/20 - 12/31] DeepFM 학습 및 필요 데이터 전처리

5 프로젝트 계획



26	27	28	29	30	31	1월 1일
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24 프로젝트 종료일	25	26	27	28	29

[1/3 - 1/14] 기본 모델링(DeepFM) 진행
및 추가 성능향상 위한 모델링 고려

[1/17 - 1/21] 결과 정리 및 보고서 작성

- 프로젝트를 진행하는 과정에서 필요하다면 마지막주까지 모델링을 진행할 수 있음


Thank You
감사합니다