Deep Click Through Rate Prediction System for Bunjang Ads

Kelly · Emily

Data Team, BUNJANG Corp.

Abstract

CTR(Click Through Rate) 예측은 현재 사회에서 중요하게 생각되며 여러방면에서 활용되고 있다. 클릭 예측을 통해 판매자와 소비자간의 행동 패턴을 분석 / 예측하고, 이를 활용하여 서로의 이익과 실용성을 위해 이용될 수 있다. 번개장터의 광고 효율과 이익을 창출하기 위한 CTR 예측 모델을 총 4가지로 구현했으며, 번개장터 데이터와 크리테오사(Criteo Co.)에서 제시한 데이터를 통해 비교 분석한다.

1. Introduction

E-Commerce 기반의 온라인 디스플레이 광고를 서비스함에 있어서 개별 상품의 CTR(Click Through Rate)을 예측하는 것은 매우 중요하다. 이는 사용자에게는 매력적인 상품의 노출을 기대해볼 수 있으며, 플랫폼을 서비스하는 회사 입장에서는 광고 수익 증대로 나타날 수 있다. 추천시스템과 같은 고도화 기술이 발전하면서 CTR을 예측하는 기술 또한 비슷한 기법들로 발전하게 되었다. 중고 물품 플랫폼을 서비스하는 번개장터에서도 추천 효율 향상 및 광고 노출 지면 확대 등 CTR 효율의 극대화를 위한 많은 노력을 해왔다. 앞으로 번개장터가 더욱 진보화된 광고 최적화 연구 및 기술 향상시킬 수 있도록 기존 연구들을 바탕으로 번개장터에 알맞게 재구현하였으며, 본 글은 해당 기술을 설명하기 위한 기술보고서이다.

2. Methodology

2.1. Deep Neural Networks

희소행렬의 값을 예측할 수 있는 MF(Matrix Factorization)과 같은 기법들은 CTR의 예측을 비롯한 추천기법에 사용되어 왔다. 주로 사용자(User)와 물품(Item)간 행동(Rating)을 행렬로 구성하여 학습을 진행하는 형태였다. 하지만 유저와 물품이라는 비교적 단순한 입력

요소들만을 사용하기 때문에 성능상 한계점이 있었으며, 입력 요소들을 추가하기 어려운 부분이 있었다. 이러한 문제점들을 해결하고 더 정교한 학습을 위해 최근에는 딥러닝의 발전과 함께 뉴럴넷(신경망)을 이용한 고도화된 방법론이 연구되어 왔으며, 그중 하나가 DNN을 이용한 CTR 예측 기술이다. 번개장터에도 DNN 기술을 적용하기 위해 다양한 사례 및 연구들을 조사하였다. 사례 조사를 위해 연구 구글에서 게재한 *Deep Neural Networks for YouTube Recommendations*라는 연구논문을 선택하였다[1]. 아래 그림 같이 유저 정보 및 상세한 컨텍스트 정보 등을 임베딩하여 학습시킨다. 결국 MF와 비슷한 성능을 보여주게 되며, 내적공간(Dot Product Space)에 펼쳐진 잠재요소들 중 가장 가까운 물품(비디오)을 찾는 과정을 보여준다. 특이사항으로는 여러 물품의 임베딩 벡터들을 평균을 내어 사용하였으며, 완전 연결된 계층(Fully-Connected layer) 형태의 아키텍처 구성을 나타내고 있다.

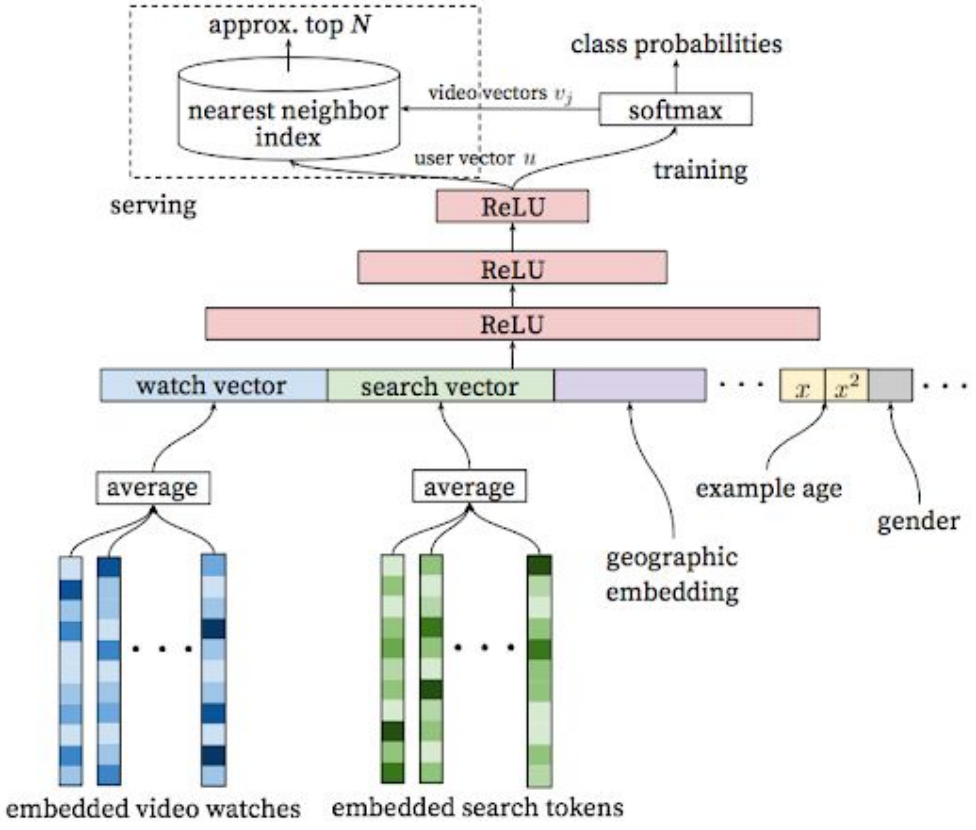

Figure 3: Deep candidate generation model architecture showing embedded sparse features concatenated with dense features. Embeddings are averaged before concatenation to transform variable sized bags of sparse IDs into fixed-width vectors suitable for input to the hidden layers. All hidden layers are fully connected. In training, a cross-entropy loss is minimized with gradient descent on the output of the sampled softmax. At serving, an approximate nearest neighbor lookup is performed to generate hundreds of candidate video recommendations.

Figure1. Youtube Recommendation System Architecture

2.2. Factorization Machine

FM(Factorization Machine)은 2010년에 *Randle, S.*에 의해 발표되었으며, DNN과 마찬가지로 대부분의 추천문제가 User, Item, Rating이라는 튜플셋으로 사용되기 때문에 메타데이터(Tag, Category)를 첨가하기 어렵다는 문제에 기반해 설계된 알고리즘이다[2]. 실제로 서비스에서 사용될 수 있는 추가 Feature들을 손쉽게 추가할 수 있다는 것이 특징이다. 아래는 모델의 목적함수이다.

$$\phi_{\mathrm{FM}}(\boldsymbol{w}, \boldsymbol{x}) = \sum_{j_1=1}^{n} \sum_{j_2=j_1+1}^{n} (\boldsymbol{w}_{j_1} \cdot \boldsymbol{w}_{j_2}) x_{j_1} x_{j_2}.$$

위의 식을 토대로 학습이 진행된다면 $O(k*n^2)$의 시간복잡도를 가지며, 저자는 수학적으로 이를 다시 정리되어 아래처럼 구성하였다.

$$\phi_{\mathrm{FM}}(\boldsymbol{w}, \boldsymbol{x}) = \frac{1}{2} \sum_{j=1}^{n} (\boldsymbol{s} - \boldsymbol{w}_j x_j) \cdot \boldsymbol{w}_j x_j,$$

정리를 통해 FM의 시간복잡도를 선형적($O(k*n)$)으로 바꿔놓았으며, 이는 획기적으로 학습시간을 단축시키는 역할을 하였다. 해당 식의 증명은 아래와 같다[3].

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i \, x_j \\
=&\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i \, x_j - \frac{1}{2} \sum_{i=1}^{n} \langle \mathbf{v}_i, \mathbf{v}_i \rangle \, x_i \, x_i \\
=&\frac{1}{2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{f=1}^{k} v_{i,f} \, v_{j,f} \, x_i \, x_j - \sum_{i=1}^{n} \sum_{f=1}^{k} v_{i,f} \, v_{i,f} \, x_i \, x_i \right) \\
=&\frac{1}{2} \sum_{f=1}^{k} \left(\left(\sum_{i=1}^{n} v_{i,f} \, x_i \right) \left(\sum_{j=1}^{n} v_{j,f} \, x_j \right) - \sum_{i=1}^{n} v_{i,f}^2 \, x_i^2 \right) \\
=&\frac{1}{2} \sum_{f=1}^{k} \left(\left(\sum_{i=1}^{n} v_{i,f} \, x_i \right)^2 - \sum_{i=1}^{n} v_{i,f}^2 \, x_i^2 \right)
\end{aligned}$$

2.3 Deep Factorization Machine (Deep FM)

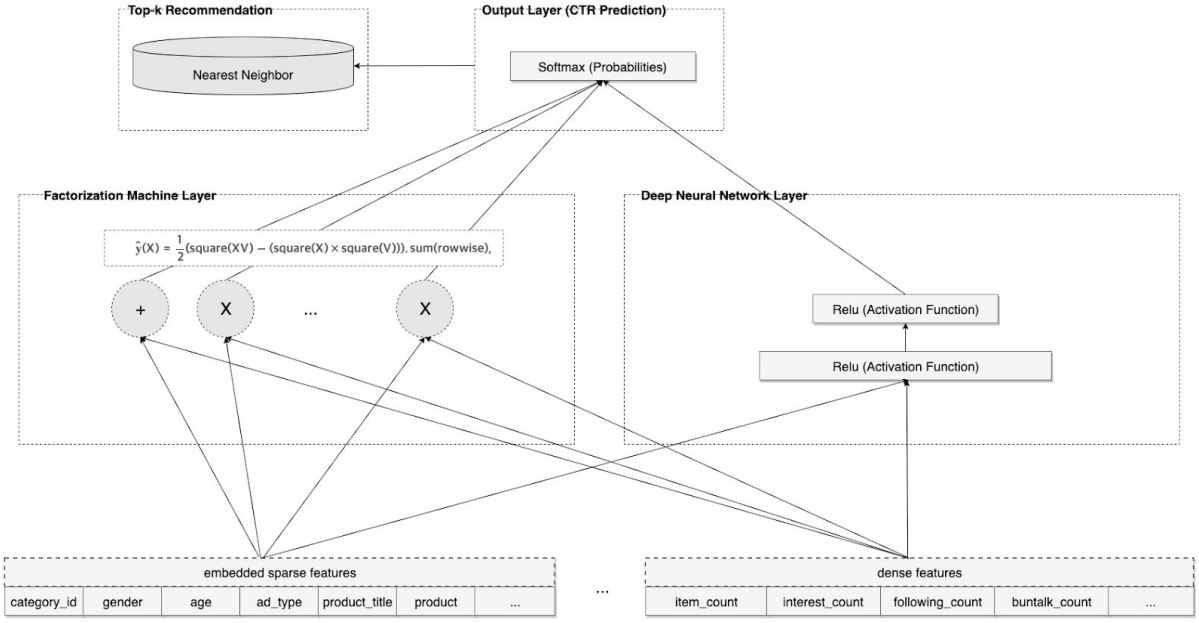

Figure3. Bunjang CTR Prediction System Architecture

DeepFM은 앞서 설명한 DNN과 FM을 합성(Concat)한 뉴럴넷 모델이다[4]. Kaggle 경진대회의 다양한 프로젝트에서 CTR 예측을 위해 사용되었으며 최상위권 성적으로 이어지고 있다. DeepFM과 관련된 논문 및 소스코드를 바탕으로 번개장터에 알맞게 재구성된 모델 구조는 위의 그림과 같다. FM모델을 정의한 레이어와 DNN을 정의한 레이어의 합성을 통해 결과를 추출한다. 결과는 0에서 1까지의 클릭확률로 제공되며, 이를위해 Softmax함수를 사용한다.

2.4. Data analysis

번개장터 데이터 . 번개장터 데이터는 2019년 11월 19일 12시부터 18시까지 6시간동안 노출된 상품의 정보와 해당 상품을 판매하는 상점의 정보를 내포하고 있다. *user, user_extra_info, product_info, searched_pids_v2*의 데이터베이스 테이블에서 데이터 추출하였으며, Table 1은 번개장터 데이터의 feature들과 출처를 알려준다. Feature들은 단순히 상품과 상점의 정보뿐만이 아니라 성별, 나이, 결혼 유무와 같은 개인적인 정보도 포함한다. 타겟(Label)은 binary 숫자로 이루어져 있는 click 행이다. 0은 상품이 노출만 되고 클릭이 안되었다는 것을 뜻하고, 반대로 1은 노출된 상품이 클릭되었다는 것을 뜻한다. Figure 1-3은 전체적으로 노출수에 비해 클릭이 얼마나 되었는지 보여준다. 노출 로그 추출의 어려움으로 인해 해당 데이터는 광고(파워업, 슈퍼업, 상점업) 노출로만 구성되어 있다.

Category	Feature	Description	Feature Type 1	Feature Type 2	Table
Product	p_price	price of the product	numerical(continuous)	integer	product_info
	p_qty	quantity of the product	numerical(continuous)	integer	product_info
	p_image_count	the number of images used to describe the product	numerical(continuous)	integer	product_info
	p_emergency_cnt	the number of the product has been reported	numerical(continuous)	integer	product_ext
	p_comment_cnt	the number of comments of the product	numerical(continuous)	integer	product_ext
	p_interest	the number of view of the product	numerical(continuous)	integer	product_ext
	p_pfavcnt	how many dibs are called on the product (the number of zzim)	numerical(continuous)	integer	product_ext
	p_taekpo	whether the shipping fee is included (0,2: not included, 1: included)	categorical(nominal)	integer	product_info
	p_exchg	whether the trade is available (0,2: unavailable, 1: available)	categorical(nominal)	integer	product_info
	p_category_id	the category id of the product	categorical(nominal)	integer	product_info
Seller	u_favorite_count	the number of followers of the shop	numerical(continuous)	integer	user
	u_comment_count	the number of comments of the shop	numerical(continuous)	integer	user
	u_review_count	the number of reviews of the shop	numerical(continuous)	integer	user
	u_grade	the level of the shop (grade / review count → max. 10)	numerical(discrete)	integer	user
	u_item_count	the number of items the shop sells	numerical(continuous)	integer	user
	u_interest	the number of visits of the shop	numerical(continuous)	integer	user
	u_following_cnt	the number of followers of the shop	numerical(continuous)	integer	user
	u_parcel_post_count	the number of parcel service transactions the shop has provided	numerical(continuous)	integer	user_extra_info
	u_bunpay_count	the number of the shop's usage of bunpay	numerical(continuous)	integer	user_extra_info
	u_transfer_count	the number of the shop's remittance transaction	numerical(continuous)	integer	user_extra_info

	u_bunp_account_count	the number of bunp account transactions the shop has	numerical(continuous)	integer	user_extra_info
	u_bunp_meet_count	the number of direct dealing the shop has done	numerical(continuous)	integer	user_extra_info
	u_bizlicense	whether the shop is official or not (0: normal shop, 1: official shop)	binary(categorical)	integer	user
	u_sex	the owner of the shop's sex (0: none, 1: female, 2: male)	categorical(nominal)	integer	user_extra_info
	u_age	the owner of the shop's age	numerical(continuous)	integer	user_extra_info
	u_married	whether the owner of the shop is married (0: single, 1: married)	binary(categorical)	integer	user_extra_info
Search	ad_type	the type of ad used on the product	categorical(nominal)	string	searched_pids_v2

Table 1. Statistics of features selected in the bunjang's dataset for further bunjang click through rate prediction

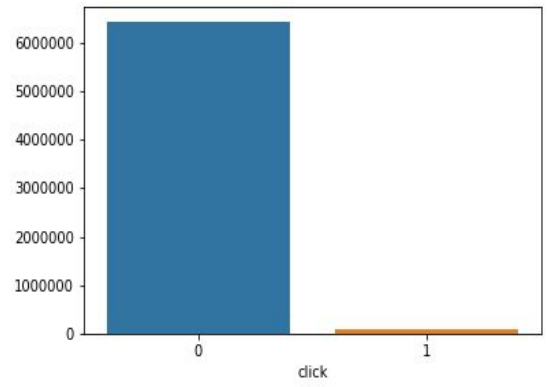

Figure 4. The total number of unclick and that of click of the exposed and advertised items for given 6 hours

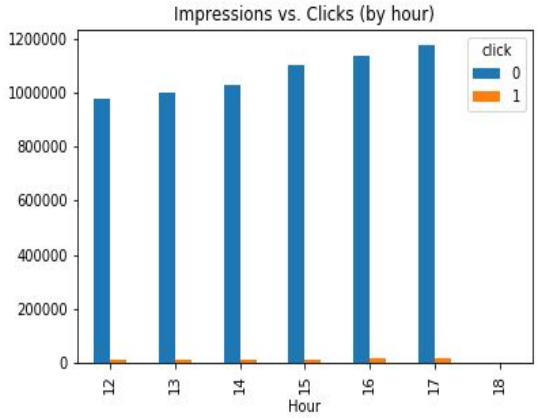

Figure 5. The number of unclick and that of click by hour(for 6 hours)

번개장터에서의 광고로 인한 노출대비 클릭 비율은 1.25%이다. 시간대별 클릭 비율은 Figure 3에서 보여준다. 앞서 말한것과 같이 노출로그의 광고의 종류는 파워업(power up), 슈퍼업(super up), 상점업(shop up)으로 구성되어 있다. 광고로 인해 노출된 상품들 중 가장 높은 노출 수치를 나타내는 것은 Figure 4에서 보여지는 것과 같이 슈퍼업이다. 광고 별 노출대비 클릭 수치도 확인할 수 있다.

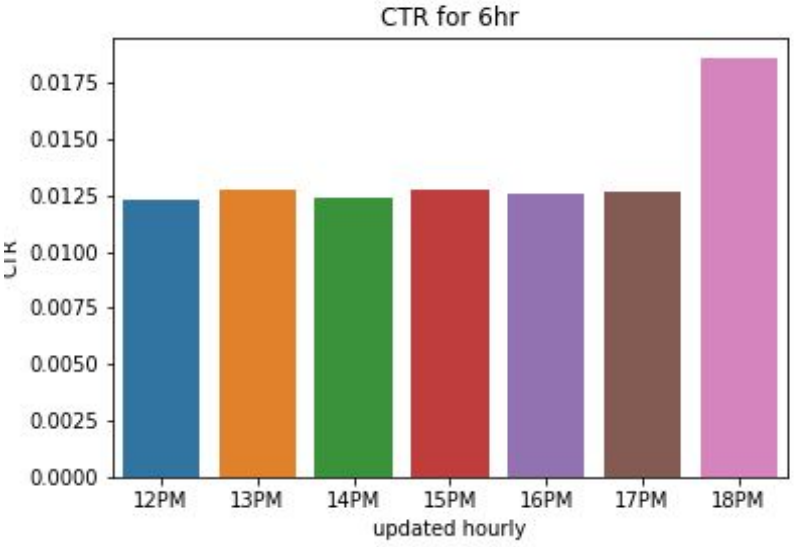

Figure 6. CTR(Click Through Rate) by hour in Bunjang dataset

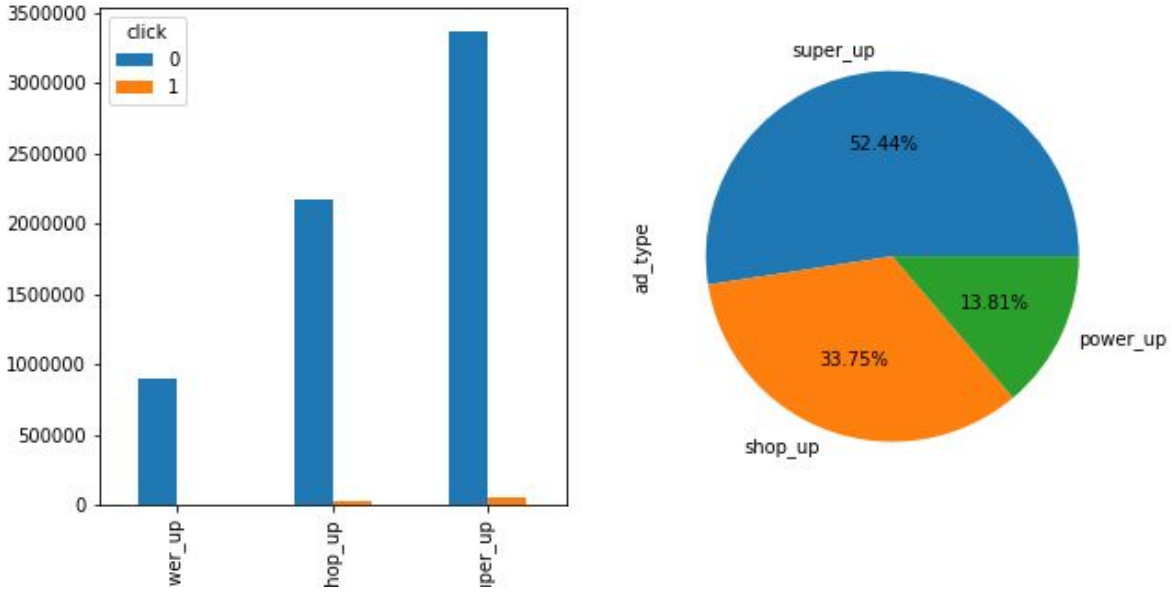

Figure 7. The number of unclick and click for each ad type is shown on the left,
and the proportion of the advertisement type used is shown on the right

Criteo 데이터 . 크리테오사에서 Kaggle의 클릭 예측 모델 대회를 위해 업로드한 7일동안의 공개 로그 데이터이다[5]. 모든 Feature들에 대한 설명은 공개되지 않았고, 이중 범주형데이터는 암호화되어 해쉬 데이터의 형태를 띄고 있다. 해당 데이터 또한 정답(Label)셋은 클릭 여부이며 이진(Binary) 데이터의 형태로 이루어져 있다. 전체적인 정답(Label=1) 비율은 25.62%를 차지하고있다. 번개장터의 데이터가 완벽히 구축되기까지 설계된 클릭예측 모델의 실험을 위해 사용될 것이다.

Feature	Description
I1 - I13	Integer features (anonymous)
C1 - C26	Categorical features (anonymous)

Table 2: Features used in Criteo's click through rate prediction model

3. Results

앞서 구성된 CTR 예측모델과 데이터를 통해 학습과 검증을 진행하였다. 설계된 모델들을 테스트하기 위해 크리테오 데이터를 1/25만큼 샘플링하여 사용하였다. (샘플 데이터 크기: 1,833,625행) 샘플링 된 데이터의 클릭 비율은 25.60%로 기존의 데이터와 비슷하다.

데이터를 모델에 적용하기 전, 크리테오 샘플 데이터에서 수치형(Numerical) 데이터는는 0과 1사이의 Dense Feature로 정규화(스케일링)시키고, 범주형(Categorical) 데이터는 Sparse Feature로 인코딩 및 임베딩하였다. 예측 모델에 설정되어있는 손실 함수는 MSE(Mean Squared Error)를 사용하였다.

$$MSE = \frac{1}{n}\Sigma(y_i - \hat{y}_i)^2 ,$$

n = the size of the test set, y_i *= real value,* $\hat{y}_i$ *= predicted value*

최적화를 위해서는 Adagrad(Adaptive Gradient) 알고리즘(Optimizer)을 사용하였다. 해당 알고리즘에서는 Momentum Term이 없는대신, 경사(Gradient) 제곱의 합을 이용한 식을 나타낸다. 이는 학습률을 올바르게 적용할 수 있는 장점이 있다. 학습률(Learning rate)는 0.0001로 기본값을 두었다.

$$g_0 = 0$$

$$g_{t+1} \leftarrow g_t + \nabla_\Theta L(\Theta)^2$$

$$\Theta_j \leftarrow \Theta_j - \varepsilon \frac{\nabla_\Theta L}{\sqrt{g_{t+1}+1e^{-5}}} ,$$

g *= the sum of squared gradients,*

Θ = *weight parameter*

DNN의 경우에는 layer의 구조를 총 4가지로 나누어 테스트 하였다(Table 3). Hidden layer의 갯수를 다르게 구성하였으며, 활성화 함수도 ReLU(Rectified Linear Unit), Sigmoid 로 나누어 비교하였다(Figure 5).

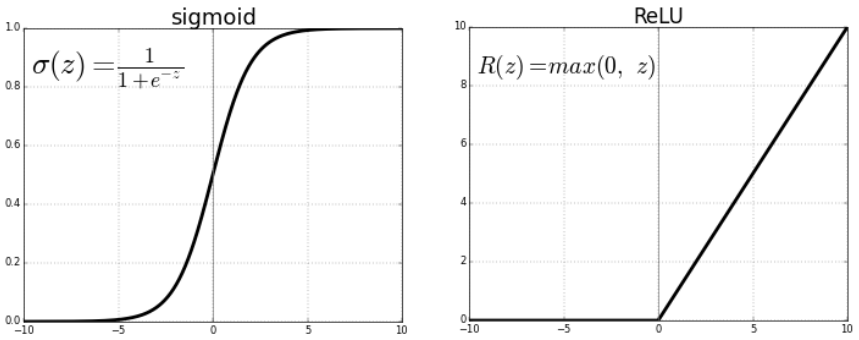

Figure 8. Illustrates two activation functions used in the model.
The left is Sigmoid and the right is ReLU(Rectified Linear Unit)

	Number of hidden layer	Activation Function
I	3	ReLU
II	4	ReLU
III	3	Sigmoid
IV	4	Sigmoid

Table 3: List of constructions of DNN model used to compare performances

학습 데이터와 테스트 데이터의 비율은 8:2 로 나누었고, 테스트 데이터를 통해 결과 값들을 검증하기 위해서 AUC(Area Under Curve)와 Log Loss를 Metric으로 사용하였다. CTR 예측 모델은 결국 이진 분별 문제(Binary Classification Problem)이기 때문에 모델이 얼마나 클래스를 구분 잘하는지 나타나기 위해서 AUC를 사용하였다. AUC는 0~1사이의 값으로 표기되며 높을수록 좋은 성능을 나타낸다고 볼 수 있다. 손실율을 측정하는 수식은 다음과 같다.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i * log(p(y_i)) + (1-y_i) * log(1-p(y_i)),$$

$$y = label(binary\ -0\ or\ 1)$$

$$p(y) = predicted\ probability\ of\ the\ point\ being\ 1(class)\ for\ all\ N\ points$$

선형(Linear), DNN, FM, DeepFM, 총 4가지의 모델을 테스트한 결과는 Table 4와 같다. 활성화 함수로 Sigmoid를 사용한 DNN모델이 가장 높은 AUC를 나타났고, 그중에서도 4개의 은닉층(Hidden Layer)를 사용한 모델이 손실율 더 낮게 나왔다.

Model	Criteo	
	AUC	Log Loss
Linear	0.7127	0.8384
DNN with 3 hidden layer using ReLU	0.7225	1.0276
DNN with 4 hidden layer using ReLU	0.7084	1.1698
DNN with 3 hidden layer using Sigmoid	0.7324	0.7824
DNN with 4 hidden layer using Sigmoid	0.7324	0.6695
FM	0.7148	0.8435
DeepFM	0.7198	1.2149

Table 4: Model comparison on Criteo sample data. DNN(Deep Neural Network) with 3/4 hidden layer using activation function of Sigmoid model results in the highest AUC(Area Under Curve), and DNN with 4 hidden layer using Sigmoid model got the lowest log loss

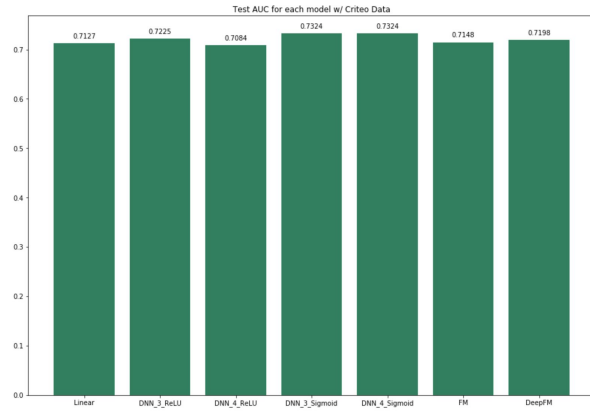

Figure 9. Test AUC(Area Under Curve) for each model with criteo sample data for demonstration. The higher the auc is, the better the model performs on the given data

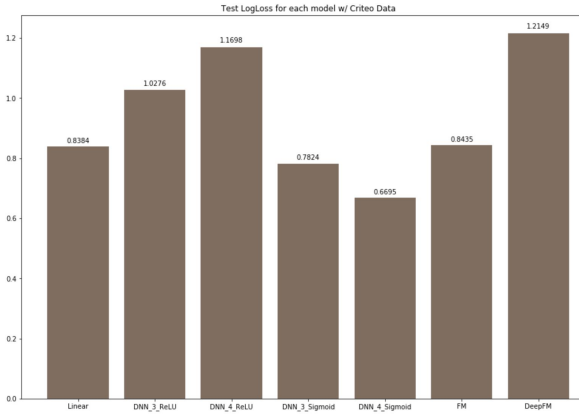

Figure 10. Test log loss for each model with criteo sample data

4. Conclusions

번개장터의 광고효율 극대화를 위해 뉴럴넷 기반의 CTR 예측모델을 구현하였다. 현재 번개장터의 로그 데이터를 수집하는데에 어려움이 있기 때문에 크리테오사의 샘플링 데이터를 통해 테스트 하였다. 테스트 결과 4개의 은닉층과 Sigmoid 활성화 함수로 구성된 DNN이 AUC와 손실율이 가장 이상적으로 나왔다. 하지만, 추후 번개장터 데이터를 사용했을 때에는 다른 모델들의 성능향상을 기대할 여지가 있다. 앞으로 번개장터 데이터 기반의 해당 모델의 튜닝은 지속적으로 이루어질 것이며, CTR 예측 모델로 인해 번개장터의 소비자와 판매자에게 모두 이득이 되는 방향을 기대해 볼 수 있을 것이다.

References

[1] Covington, P., Adams, J., & Sargin, E. (2016, September). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems* (pp. 191-198). ACM.

[2] Rendle, S. (2010, December). Factorization machines. In *2010 IEEE International Conference on Data Mining* (pp. 995-1000). IEEE.

[3] https://yamalab.tistory.com/107

[4] Guo, H., Tang, R., Ye, Y., Li, Z., & He, X. (2017). DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247*.

[5] https://www.kaggle.com/c/criteo-display-ad-challenge