Finding Users Who Act Alike : Transfer Learning for Expanding Advertiser Audiences 리뷰

생성일	@2025년 4월 21일 오후 2:13

📖 [논문 리뷰] Finding Users Who Act Alike: Transfer Learning for Expanding Advertiser Audiences (KDD 2019)"

- **원본 논문**: https://www.pinterestlabs.com/media/phkg2uau/transferlearning-kdd2019.pdf
- **저자**: Pinterest 연구팀
- **키워드**: Audience Expansion, Embedding, Transfer Learning, Look-alike Audience

---

🚩 Abstract | 논문 한눈에 보기

온라인 광고에서 광고주는 기존 고객과 유사한 신규 고객을 찾는 **Audience Expansion** 기술을 활용합니다. 본 논문에서는 Pinterest가 **실제 서비스에 적용한 임베딩 기반 Audience Expansion 모델**을 소개합니다.

핵심 아이디어는 다음과 같습니다.

- Pinterest의 모든 사용자 데이터를 활용해 **전역 사용자 임베딩 모델**을 학습
- 광고주가 제공한 소규모 고객 리스트(Seed)를 전역 임베딩 공간에서 효율적으로 표현하고, 이를 통해 신규 고객의 유사도를 측정
- 기존 광고주별 분류기(Classifier) 모델과 **앙상블(Ensemble)** 하여 성능을 극대화

실험 결과, 제안된 모델은 특히 **소규모 Seed 리스트**에서 기존 분류기 모델의 한계를 크게 극복했으며, 실제 서비스에서 높은 성과를 나타냈습니다.

🔍 1. Introduction

온라인 광고가 보편화되면서, 광고주들은 보다 정확하게 자신들의 타겟 고객층을 설정할 수 있게 되었습니다. 기존의 광고 플랫폼들은 광고주가 직접 인구통계학적 요소(연령, 성별 등), 키워드 검색 내역, 특정 주제에 대한 관심 여부 등을 기준으로 타겟 고객을 지정해야 했기에 광고주에게 높은 통제권을 제공하지만, 광고 플랫폼이 보유한 사용자에 대한 풍부한 데이터를 충분히 활용하지 못하는 한계를 가집니다.

그래서 등장한 기법이 바로 **Audience Expansion(잠재 고객 확장)** 또는 **Look-alike Audience(유사 잠재 고객)**입니다. 이는 광고주가 특정한 타겟 기준을 일일이 설정하는 대신, 자신들의 기존 고객(Seed Users) 리스트만 광고 플랫폼에 제공합니다. 그러면 플랫폼이 이 리스트를 바탕으로 유사한 관심사를 가진 사용자들을 더 넓게 찾아서 잠재 고객으로 제시합니다. 이 방식은 광고주가 관리해야 하는 변수의 수를 줄이면서도, 광고 플랫폼이 사용자에 대해 가진 방대한 데이터와 노하우를 최대한 활용할 수 있다는 큰 장점을 갖습니다.

그러나 이 시스템을 실제 서비스 규모로 확대하려면 여러 가지 기술적 도전 과제가 존재합니다.

- **Seed 크기의 다양성**: 광고주가 제공하는 고객 리스트는 크기가 천차만별, 따라서 어떤 크기에도 안정적으로 작동하는 모델이 필요
- **확장성(Scalability)**: 시스템은 수억 명의 사용자와 수많은 광고주를 처리할 수 있을 만큼 효율적이어야 함
- **빠른 응답성(Responsiveness)**: 새로운 광고주나 신규 사용자가 시스템에 들어올 때 최대한 빠르게 처리
- **실제 성능(Real-World Performance)**: 사용자가 광고를 실제 클릭하거나 보는 등, 현실적이고 구체적인 사용자 행동으로 성능을 평가
- **광고주 데이터 격리(Advertiser Data Isolation)**: 광고주 간의 데이터 누출이 없어야 함. 즉, 특정 광고주가 제공한 데이터가 다른 광고주를 위해 사용되어서는 안됨

✅ 1.1 Current work

Pinterest는 Audience Expansion 문제를 해결하기 위해 "**Act-alikes**"라는 모델을 개발했습니다. Act-alikes는 광고주가 제공한 고객 리스트와 유사한 사용자를 찾는 것을 목표로 하며, 기존에는 광고주별로 분류기(Classifier)를 별도로 학습하는 방식을 사용했습니다. 그러나 이 방법은 **Seed 리스트가 작을 때 성능이 낮아지는 문제**가 있었습니다.

Pinterest는 이 문제를 해결하기 위해 **오가닉 콘텐츠(organic content, 광고가 아닌 일반 콘텐츠)**를 활용한 **전이 학습(Transfer Learning)** 기반 접근법을 도입했습니다. 오가닉 콘텐츠와 광고 콘텐츠가 매우 유사하다는 점을 이용해, 오가닉 콘텐츠와의 사용자 상호작용 데

이터를 활용하여 사용자 간 유사성을 측정하는 **범용(Universal) 사용자 모델**을 구축했습니다.

이렇게 구축된 사용자 유사성 모델을 모든 광고주가 공유하여 개별 광고주 모델을 경량화하고, 소규모 Seed 리스트에도 효과적으로 대응할 수 있도록 했습니다. 또한, 임베딩 방식과 기존 분류기 모델을 결합한 **앙상블 모델**을 통해 성능을 한층 더 개선했습니다.

✅ 1.2 Unique Contributions to Look-alike Advertising

본 논문은 기존 방식의 한계를 해결하기 위해 다음과 같은 주요 기여를 했습니다.

- **사용자 임베딩 모델**을 활용하여 사용자 간 유사도를 효율적으로 계산하는 방법을 제안
- Pinterest의 방대한 **사용자 행동 데이터를 활용한 전이 학습**을 통해 소규모 Seed 리스트의 한계를 극복
- 새로운 광고주가 빠르게 적용 가능한 **경량화된 시드 표현 방법**을 개발
- 기존 분류기와 임베딩 모델을 결합한 **앙상블 방식**을 제안하여 모든 Seed 크기에서 성능을 극대화

🔍 2. Related Work

기존 연구들은 주로 **분류기(Classifier)-기반** 모델로 Audience Expansion을 해결해왔습니다. 즉, Seed 사용자를 Positive, 무작위 사용자를 Negative로 라벨링해 광고주별 분류기를 학습하는 방법입니다. 하지만 이 방식은 Seed가 충분히 클 때 뛰어난 성능을 보이지만, **Seed가 작으면 과소적합** 문제가 심각해집니다. 또한 **광고주별 모델을 따로 학습해야 하는 문제점**도 있었습니다.

임베딩 기술(Word2Vec, GloVe, StarSpace 등)은 NLP 분야에서 성공적으로 사용되었지만 광고 타겟팅 분야에서는 그동안 거의 활용되지 않았습니다. Pinterest는 이러한 임베딩을 전역 사용자 모델에 적용하고, 전이 학습을 통해 각 광고주에게 맞춘 경량 표현을 만드는 새로운 길을 열었습니다.

🔍 3. Ensemble Act-alike Model

✅ 3.1 Classifier-based Approach

첫 번째로 **광고주별 개별적인 로지스틱 회귀 분류기를 학습하는 방식**입니다.

광고주가 제공한 Seed 리스트를 활용하여, 이 사용자들을 긍정(positive) 데이터로 설정합니다. 그리고 동일한 수의 무작위 사용자를 부정(negative) 데이터로 설정해 균형 잡힌 학습

데이터를 만듭니다. 각 사용자는 인구통계학적 정보, 관심사 등 다양한 특성을 **연속형 및 이산형 피처** 형태로 나타내며, 이를 기반으로 로지스틱 회귀 모델을 학습합니다.

모델이 학습된 후, 모든 사용자에 대해 점수를 계산하고, 높은 점수를 받은 사용자들을 광고주의 확장된 잠재 고객으로 선정합니다.

그러나 이 접근법에는 다음과 같은 문제가 있습니다.

- 광고주별로 별도의 모델을 학습해야 하므로 **계산 비용이 큼**
- Seed 리스트의 크기에 따라 성능 차이가 크며, 특히 **작은 Seed 리스트에서는 성능이 저하**

이를 해결하기 위한 새로운 접근법이 바로 다음 절에서 소개하는 **전역 사용자 임베딩 모델(Global User Embedding Model)**입니다.

✅ 3.2 Global User Embedding Model

사용자 간 유사성을 효율적으로 포착하기 위해, 모든 사용자의 행동 로그 데이터를 기반으로 **범용(Universal) 사용자 임베딩 모델**을 구축했습니다. 이 모델은 **StarSpace**라는 임베딩 학습 기술에서 영감을 받아, 사용자를 저차원 벡터 공간으로 표현합니다. 이는 사용자가 과거에 관심을 보였던 콘텐츠 주제와 사용자 자신을 같은 임베딩 공간에 투영하여, 관심 있는 콘텐츠와 사용자가 서로 가까운 위치에 있도록 학습하는 것입니다.

이 모델을 선택한 이유는 아래와 같습니다.

- **데이터 보안**: 광고주들이 제공한 Seed 리스트는 공유할 수 없으므로 범용 데이터로 학습한 모델이 필요
- **방대한 데이터 활용 가능**: Pinterest에는 매월 2억 5천만 명 이상의 사용자와 1,750억 개 이상의 핀(Pin)에 대한 데이터가 있으며, 이를 활용하면 사용자 관심을 잘 포착할 수 있음
- **유사성 정의**: 사용자가 콘텐츠와 상호작용할 가능성을 명확히 표현할 수 있어 광고 클릭 가능성과 직관적으로 연결

🔷 3.2.1 사용자 표현(User Representation)

사용자는 인구통계학적 특성(성별, 국가 등)이나 관심 있는 주제와 같은 **이산형 특성(discrete features)**과 클릭률(CTR), 저장 횟수 등의 **연속형 특성(continuous features)**으로 표현됩니다.

- 이산형 특성은 정해진 사전(dictionary)의 값들 중 하나로 표현하고, 이를 임베딩 벡터로 변환
- 연속형 특성은 정규화한 뒤 직접 사용

이산형 특성은 여러 값이 있을 수 있어, **풀링(Pooling)** 과정을 거쳐 하나의 벡터로 집약한 뒤, 모든 특성을 하나의 긴 벡터로 결합(concatenate)하고, 선형(Dense) 레이어를 통과시켜 최종 사용자 임베딩(**user embedding, u**)을 얻습니다.

🔷 3.2.2 주제 표현(Topic Representation)

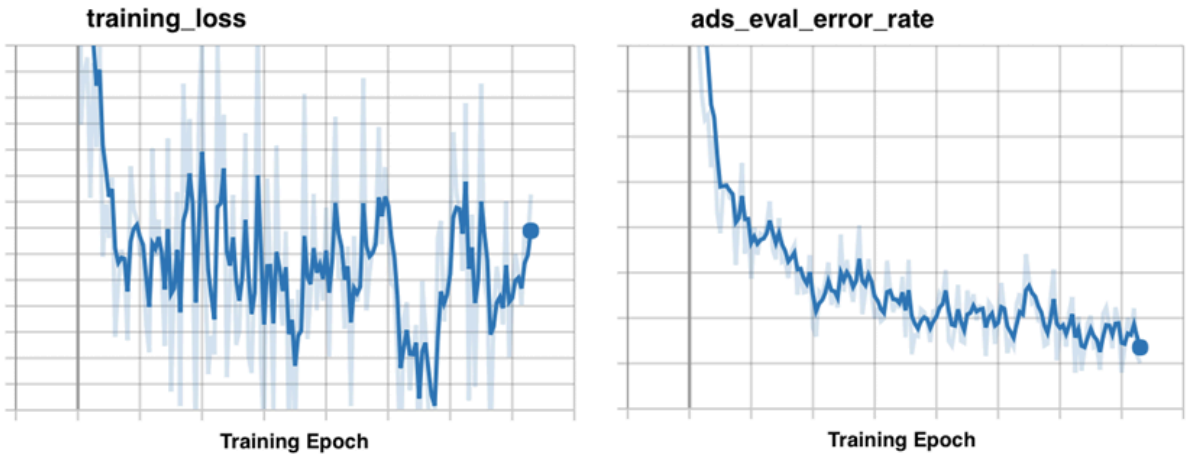
Pinterest의 콘텐츠(핀)는 주제(topic)로 태깅됩니다. 각 주제 역시 사용자와 동일한 차원의 임베딩 벡터(**topic embedding, t**)로 표현됩니다. 사용자가 핀과 상호작용하면, 그 핀의 주제와 사용자가 서로 가깝게 위치하도록 학습이 이루어집니다.

🔷 3.2.3 학습 데이터 구성 및 모델 학습(Training Data)

학습 데이터는 사용자가 상호작용한 핀의 주제(**긍정 사례, (u, t+)**)와 무작위로 선택된 상호작용하지 않은 주제(**부정 사례, (u, t-)**)로 구성됩니다. 이를 **마진 기반의 랭킹 손실(margin ranking loss)**로 학습하여, 사용자가 실제 상호작용한 주제가 랜덤으로 선택한 부정 사례보다 유사도가 높아지도록 합니다. 이 과정을 통해 유사한 관심사를 가진 사용자는 서로 가까이 위치하게 됩니다.

🔷 3.2.4 학습 중단 기준(Stopping criteria)

모델 학습 중 주기적으로 **전이 학습(transfer task)** 평가를 수행하여 평가 에러율이 더 이상 낮아지지 않으면 학습을 종료합니다. 전이 학습 평가는 별도의 광고주 데이터를 사용하지 않고도 사용자 유사성을 간접적으로 평가할 수 있게 해줍니다.



사용자와 콘텐츠 간 학습 loss(오른쪽)은 고려하지 않고, 사용자와 광고주 간 학습(왼쪽)을 고려하고 있음.

🔷 3.2.5 모델 재학습(Model Retraining)

사용자 특성이 서서히 변하므로, 사용자 임베딩을 주기적으로 업데이트합니다. 그러나 모델 전체를 매일 재학습할 필요는 없고, 장기간 주기로 재학습을 수행해 사용자의 관심사 변화를 효과적으로 포착합니다.

✅ 3.3 Embedding-Based User-Advertiser Scoring Model

이 모델의 목표는 광고주의 Seed 사용자들이 밀집한 임베딩 공간 영역을 찾고, 새로운 잠재 고객을 확장하는 것입니다.

🔷 3.3.1 지역 민감 해싱(Locality Sensitive Hashing, LSH)

임베딩 공간을 임의의 초평면(hyperplane)으로 분할하여 사용자들을 2^n개와 같이 여러 영역(region)에 나눕니다. 같은 영역에 위치한 사용자들은 서로 유사할 확률이 높아지며, 이를 통해 계산 효율성을 높입니다.

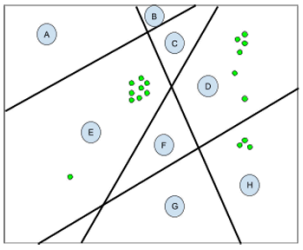
🔷 3.3.2 투표 점수(Voting Score)

각 영역별로 시드 사용자의 수를 집계하여 각 사용자에게 점수를 부여합니다. 시드 사용자가 많은 영역에 속한 사용자는 높은 점수를 받아 확장될 가능성이 큽니다. 하지만 이 방식은 시드가 많다는 이유만으로 특정 영역을 높게 평가하게 되어, 실제 Pinterest 내에서 인기가 많은 행동이나 관심사가 가진 배경 분포(background distribution)를 고려하지 않는 한계가 있습니다.

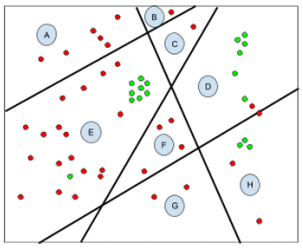
🔷 3.3.3 밀도 점수(Density Score)

하지만 투표 점수는 비시드 사용자 수를 고려하지 않기 때문에, 영역 내 전체 사용자 수를 고려한 밀도 점수로 보완합니다. 이를 통해 보다 균형 있고 정확하게 사용자의 관심 영역을 평가할 수 있게 합니다.

최종적으로 여러 번의 LSH 분할과 점수 계산을 반복해 변동성을 줄이고, 최종 점수를 사용자와 광고주 간 친화성(Affinity)으로 활용합니다.



(a) Example for simple voting



(b) Example for density voting

Region	Voting Score	Density Score ($\alpha, \beta = 0$)	Density Score ($\alpha = 1, \beta = 4$)
A	0	0	0.25
B	0	0	0.25
C	0	0	0.25
D	6	0.86	0.63
E	9	0.33	0.32
F	0	0	0.25
G	0	0	0.25
H	3	0.5	0.4

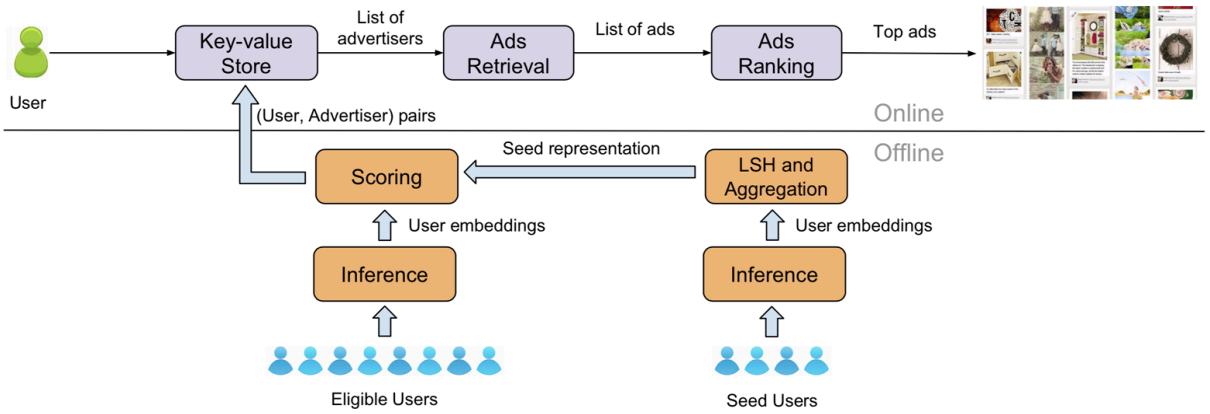
투표 점수와 밀도 점수에서 차이가 있음

이렇게 구축한 임베딩 기반 모델은 광고주별 별도 모델 학습 없이도 효과적이며, 특히 작은 Seed 리스트를 가진 광고주에게 유리합니다. 최종적으로 Pinterest는 이 임베딩 기반 모델과 분류기 기반 모델을 결합하여 앙상블 모델을 구축하고, 두 접근법의 장점을 모두 활용합니다.

🔍 4. End-to-End System

Pinterest는 실제 서비스 운영을 위해 다음과 같은 전체 시스템을 구축했습니다.

모델 결과를 점수화(scoring)하고 서빙(serving)하는 이 시스템은 상단에 온라인 광고 서빙과 하단에 오프라인 데이터 파이프라인으로 구성됩니다. 오프라인 파이프라인은 정기적으로 실행되어 광고주별 최상위 사용자 목록을 생성하고, 이를 온라인에 서빙합니다.



✅ 4.1 User Embedding and Seed Representation - 오프라인

정기적으로 실행되는 파이프라인은 사용자의 최신 특성(feature)을 수집한 뒤, 이전에 학습한 사용자 임베딩 모델을 이용해 새로운 임베딩을 추론하는 것입니다. 이미 필요한 특성으로 거의 완전하게 표현되기 때문에, 임베딩 모델 자체를 자주 재학습할 필요는 없습니다.

광고주가 제공한 시드 사용자들의 임베딩을 모아 Seed 표현을 계산합니다. 이는 (영역 ID, 점수) 쌍으로 이루어진 딕셔너리 형태로 저장하며, 시드 사용자 수나 배경 사용자 수가 너무 적어 점수를 신뢰할 수 없는 영역의 항목은 모두 생략합니다.

✅ 4.2 Audience Scoring - 오프라인

사용자 임베딩과 광고주의 Seed 표현이 준비되었다면, 이제 광고주와 사용자 간의 친화도(Affinity Score)를 계산합니다. 친화도 점수는 각 사용자와 각 광고주의 Seed 표현 사이의 관계를 나타내는 점수로, 논문에서 제안한 수식을 활용해 계산됩니다.

이후 후보 사용자 선정 및 효율적인 계산을 위해 다양한 방법을 활용합니다. 우선 모든 Pinterest 사용자가 매일 서비스를 방문하는 것은 아니기 때문에, 계산의 효율성을 높이기 위해 최근 Pinterest를 방문한 사용자만 계산 대상으로 선정합니다.

또한 계산 작업은 매우 많기 때문에 **MapReduce 프레임워크**를 사용하여 효율적으로 분산 처리합니다. 구체적으로 친화도 점수 계산 작업은 '사용자 x 광고주' 모든 조합을 대상으로 이루어지므로, 계산 대상 수가 매우 많아집니다. 이를 효과적으로 나누어 여러 컴퓨터에서 동시에 처리하기 위해서 논문은 **Fragment-and-Replicate Join**이라는 방법을 활용합니다. 이 방식은 데이터를 작은 조각(fragment)으로 나누고, 필요한 만큼 데이터를 복제(replicate)하여 병렬 처리 속도를 극대화하는 기법입니다.

🔷 사용자 점수 정렬 방법 : Bucket Sort

많은 사용자의 점수를 계산한 후에는 광고주가 원하는 상위 p%의 사용자를 추출해야 합니다. 단순한 정렬 방법은 시간이 오래 걸리는(O(NlogN)) 단점이 있습니다. 또한, 점수를 샘플링하여 정렬하는 방식도 정확성 측면에서 한계가 있죠.

Pinterest는 이를 해결하기 위해 빠르고 정확한 **버킷 소팅(Bucket Sort)** 방법을 사용합니다. 버킷 소팅은 점수를 일정한 구간으로 나누고, 각 구간(버킷)에 점수를 할당하여 매우 빠르게(O(N)) 원하는 퍼센트 내 사용자들을 추출할 수 있게 합니다.

🔷 임베딩과 분류기 모델의 조합 : Blending

실험 결과 임베딩 기반 모델은 작은 Seed 리스트에서 우수했지만, 큰 Seed 리스트에서는 기존의 분류기 모델(Classifier)이 오히려 더 좋은 성능을 보였습니다. Pinterest는 이 두 가지 모델의 강점을 모두 활용하기 위해 두 모델의 결과를 혼합(blending)하는 전략을 선택했습니다. 혼합 과정은 다음과 같습니다.

- 두 모델이 각각 추출한 사용자 리스트의 **교집합**을 우선 선택합니다.
- 그 이후 나머지 사용자를 **라운드 로빈(round-robin)** 방식으로 교차 선택하여, 광고주가 원하는 사용자 수만큼 채웁니다.
- 이때 Seed 리스트 크기에 따라 임베딩 기반 모델 비율과 분류기 모델 비율을 다르게 하여 최적의 성능을 얻도록 합니다.

실험 결과 이러한 혼합 전략이 단일 모델을 사용하는 것보다 훨씬 좋은 성능을 보였으며, 실제 서비스에서도 성공적으로 적용되었습니다.

✅ 4.3 Ads Serving - 온라인

실제 Pinterest를 방문한 사용자에게 광고를 제공하는 과정입니다. 지금까지의 과정을 통해 **(광고주, 사용자 리스트)** 형태로 얻어진 데이터를 실제 광고를 제공하는 형태인 **(사용자, 광고주 리스트)**로 변환하여 키-값 데이터베이스에 저장합니다.

사용자가 Pinterest를 방문하면, 저장된 광고주 리스트를 조회하여 적합한 광고를 추출합니다. 다른 추천 방식으로 얻은 광고 후보와 함께 모아, 가장 관련성 높은 광고를 선정한 후, 이를 광고 경매(second-price auction)에 올립니다. 이 과정에서 가장 적합한 광고가 최종적으로 사용자의 피드에 노출됩니다. 광고는 유기적인(organic) 콘텐츠와 함께 섞여 사용자 경험을 방해하지 않으면서도 효율적인 광고 집행을 가능하게 합니다.

🔍 5. Results

✅ 5.1 Offline Evaluation

- **오프라인 평가**에서 제안 모델은 기존 방식 대비 작은 Seed 리스트에서 정밀도 11.2%, 재현율 33.1% 향상

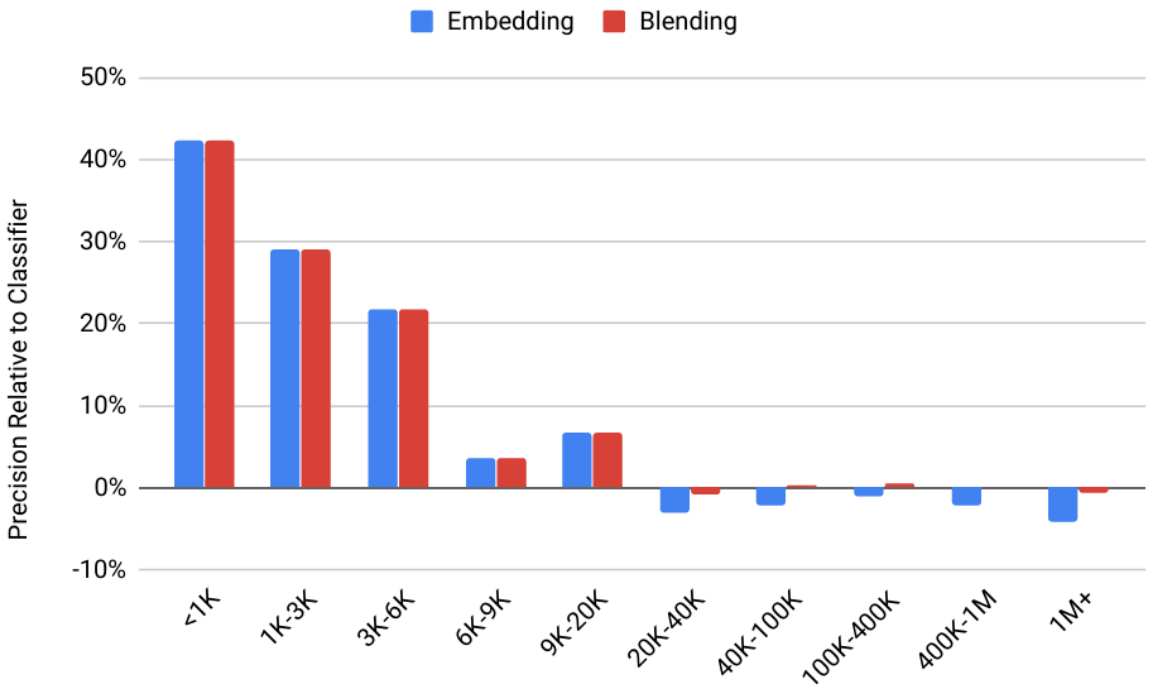
모델이 광고주의 시드 리스트를 얼마나 잘 재현하는지를 평가하기 위해 정밀도(precision)와 재현율(recall)을 사용했습니다.

광고주의 시드 사용자 중 10%를 보류 세트(Hs)로 떼어내고, 비시드 사용자 집합에서 동수의 보류 사용자(Hn)를 샘플링합니다. 이는 양성(positive)과 음성(negative) 예시의 수를 균형 있게 유지하기 위함입니다. 남은 90%의 시드 사용자 데이터를 사용해 사용자-광고주 스코어링 모델을 학습한 뒤, 이 모델로 모든 유효 사용자에게 점수를 부여하고 친화도 점수 순으로 정렬합니다. 모델은 광고주가 원하는 잠재 고객 크기 범위 전반에 걸쳐 우수한 성능을 내야 하므로, 여러 임계값(threshold)에서 정밀도와 재현율을 측정하고, 이때 보류된 시드 사용자 중 임계값을 넘는 사용자를 성공으로 간주하여 집계합니다.

Table 2: Precision (P) and Recall (R) Relative to Classifier Model

Model	ΔP@100K	ΔP@1M	ΔP@5M
Embedding	+9.67%	+9.55%	+7.31%
Blending	+11.16%	+10.55%	+8.70%
Model	**ΔR@100K**	**ΔR@1M**	**ΔR@5M**
Embedding	+32.71%	+24.69%	+18.41%
Blending	+33.09%	+31.99%	+26.55%

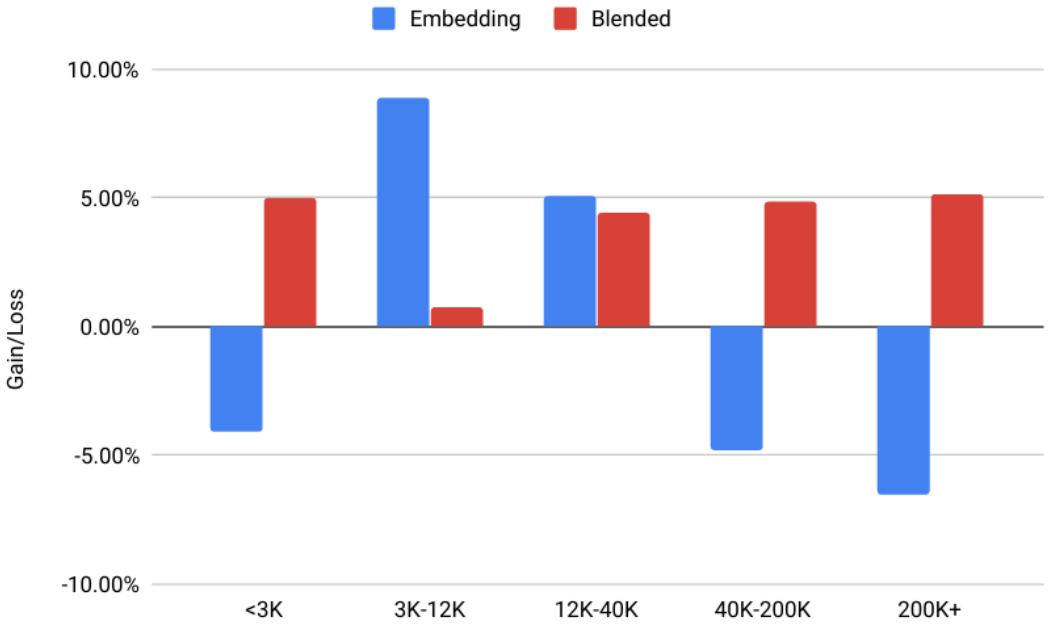
임베딩 모델은 분류모델 기준 대비 정밀도와 재현율 모두에서 크게 개선되었으며, 혼합 모델은 더욱 향상시켰음을 확인할 수 있습니다.



그래프는 Seed 리스트 크기에 따라 정밀도 값을 추가로 세분화했습니다. Seed 리스트가 작은 경우 임베딩 모델 및 혼합 모델이 분류기 모델보다 훨씬 우수한 성능을 보였고, Seed 리스트가 큰 경우에는, 분류기 모델과 비교할 때 임베딩 모델의 성능이 떨어지는 반면, 혼합 모델은 거의 동등한 성능을 발휘했습니다.

✅ 5.2 Online Evaluation

온라인 A/B 실험은 모든 실험군에 대해 사용자 ID별로 트래픽을 할당했으며, 분류기 모델, 임베딩 기반 모델과 혼합 모델의 CPC(클릭당 비용) 광고에서의 성능을 평가했는데, 평가 지표로는 일반적으로 사용되는 클릭률(CTR)을 사용했습니다.



희소한 클릭 데이터를 희석시키지 않기 위해 다섯 개 버킷만 사용했으며, 광고주가 캠페인 설정을 변경함에 따라 사용자가 노출 가능한 광고가 계속 바뀌기 때문에 본질적으로 노이즈가 많습니다. 임베딩 모델은 다섯 개 버킷 중 두 개에서 분류기 모델 대비 CTR을 개선했으며, 혼합 모델은 모든 다섯 개 버킷에서 CTR을 개선했습니다.

Treatments	Δ CTR
Embedding Only	-4.1%
Blended	+2.1%

또한 모든 시드 크기를 통합한 결과, 전반적으로 임베딩 전용 모델은 분류기 모델과 비교해 CTR이 감소했으나, 혼합 모델은 CTR을 개선되었습니다. 때문에 혼합된 분류기·임베딩 기반 모델은 Pinterest 프로덕션에 배포되었습니다.

🔍 6. Conclusions & Future Work

본 논문에서는 범용 사용자 임베딩 모델과 광고주 시드 리스트 표현을 중심으로 구축된 End to End 광고 look-alike 시스템을 소개했습니다. 전이 학습(transfer learning) 접근법을 통해 방대한 과거 사용자 로그를 효과적으로 활용했으며, 시드 리스트의 각 사용자로부터 얻을 수 있는 정보를 최대한 활용하여 소규모 시드 리스트에 대해서도 분류기 모델보다 우수하게 모델링할 수 있었습니다. 또한, 분류기 모델과 임베딩 기반 모델을 혼합(blending)하여 모든 시드 리스트 크기 구간에서 오프라인·온라인 성능을 모두 개선하는 방법을 제시하였습니다.

이후로 사용자 임베딩 모델 개선을 위해 멀티태스크 학습 적용 및 사용자 임베딩 위에 분류기 모델을 학습하는 하이브리드 접근 방식 등을 연구할 것이라고 합니다. 또한 사용자-광고주 유사도 결정 시 내적(dot product) 함수 대신 광고주별 특화된 유사도 함수를 학습하는 방법도 모색할 것이라고 합니다.

또 다른 향후 방향은 현재의 End to End 시스템을 실시간(candidate) 생성기로 전환하여 사용자가 사이트를 방문할 때마다 온더플라이(on-the-fly)로 사용자 임베딩을 계산하고, LSH를 사용하여 가장 유망한 시드 리스트 후보를 찾은 뒤, 해당 사용자에 대한 점수를 그 시드 리스트의 점수 임계값과 비교할 수 있도록 하는 것이라고 합니다. 이렇게 하면 사용자의 최신 행동을 실시간으로 반영할 수 있으며, 오프라인에서 방문하지 않는 사용자에 대한 불필요한 계산을 방지할 수 있기 때문입니다.